KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI

DOKTORA TEZİ

DOKÜMANLARIN ANLAMSAL BENZERLİKLERİNE DAYALI ÖZGÜN BİR KONU MODELLEME YÖNTEMİ

EKİN EKİNCİ

KOCAELİ 2019

KOCAELİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ

ANABİLİM DALI

DOKTORA TEZİ

DOKÜMANLARIN ANLAMSAL BENZERLİKLERİNE DAYALI ÖZGÜN BİR KONU MODELLEME YÖNTEMİ

EKİN EKİNCİ

Doç. Dr. Sevinç İLHAN OMURCA
Danışman, **Kocaeli Üniversitesi** .....

Prof. Dr. Yaşar BECERİKLİ
Jüri Üyesi, **Kocaeli Üniversitesi** .....

Prof. Dr. Banu DİRİ
Jüri Üyesi, **Yıldız Teknik Üniversitesi** .....

Prof. Dr. Nevcihan DURU
Jüri Üyesi, **Kocaeli Üniversitesi** .....

Dr. Öğr. Üyesi Adem TUNCER
Jüri Üyesi, **Yalova Üniversitesi** .....

Tezin Savunulduğu Tarih: 15.02.2019

ÖNSÖZ VE TEŞEKKÜR

Bu tez çalışması, kullanıcı yorumlarından ürün özelliklerinin çıkartılması amacıyla anlamsal konu modelleme yöntemlerini geliştirmek için gerçekleştirilmiştir.

Doktora eğitimim süresince benden desteğini esirgemeyen, tezimin her aşamasında bilgi ve tecrübesini benimle paylaşarak çalışmalarıma katkıda bulunan ve yoğun akademik çalışma hayatında değerli zamanından bana ayıran saygıdeğer hocam, tez danışmanım Doç. Dr. Sevinç İLHAN OMURCA'ya tüm içtenliğimle teşekkür ederim.

Tez çalışmama bilgi ve tavsiyeleri ile katkıda bulunan saygıdeğer tez ilerleme jürim Prof. Dr. Yaşar BECERİKLİ'ye ve Prof. Dr. Banu DİRİ'ye,

Akademik çalışmalarım sırasında, birçok aşamada bana destek olan değerli çalışma arkadaşlarım Arş. Gör. Dr. Fidan KAYA GÜLAĞIZ'a, Arş. Gör. Dr. Süleyman EKEN'e ve Arş. Gör. Abdurrahman GÜN'e,

Maddi ve manevi desteklerini tüm hayatı boyunca benden esirgemeyen başta merhum babam Murat EKİNCİ olmak üzere, annem Nergiz EKİNCİ'ye ve kız kardeşim Başak EKİNCİ'ye teşekkürü bir borç bilirim.

Ocak – 2019 Ekin EKİNCİ

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

A	:	Modellerin çalışma sürelerini içeren dizi
α	:	Dirichlet hiperparametresi
a_w	:	Veri kümelerinden uzmanlar tarafından çıkartılan ürün özelliklerinin sayısı
b'	:	Değişken
β	:	Dirichlet hiperparametresi
C_g	:	g. dokümandaki toplam kelime sayısı
$C_{g,k}$	:	g. yorumda k. konuya atanan kelime sayısı
$c_{w,k}$	:	w. kelimesinin k. konuya tüm koleksiyonda kaç kere atandığının sayısı
D	:	Doküman koleksiyonu
D'	:	D koleksiyonunun kavram ve adlandırılmış varlıklar ile temsil edilmesi ile oluşan yeni koleksiyon
d_M	:	Doküman koleksiyonu D'de yer alan m. doküman
d'_M	:	D' kümesindeki M. doküman
$D(v_l^{(k)})$	:	v_l kelimesinin kaç adet dokümanda bulunduğunun sayısı
$D(v_n^{(k)}, v_l^{(k)})$:		v_n ve v_l kelimelerinin birlikte geçtiği doküman sayısı
e	:	e sayısı
E	:	Kavram ve adlandırılmış varlıklar ile genişletilen dokümandaki toplam kelime sayısı
φ	:	Kelimelerin konulardaki dağılımı
φ_k	:	Kelimelerin k. konudaki dağılımı
$\varphi_{w,k}$	:	w. kelimesinin k. konuya atanma olasılığı
G	:	MD kümesindeki toplam doküman sayısı
Γ	:	Gama fonksiyonu
k	:	k. konu
K	:	Gizli konu sayısı
m	:	m. konu
M	:	Koleksiyonda yer alan toplam doküman sayısı
MD	:	D kümesindeki dokümanların birleşiminden oluşan yeni koleksiyon
md_G	:	MD kümesindeki G. doküman
μ	:	Rastgele değişken
$n_{i,k}$	:	i. yorumda k. konuya atanan kelime sayısı
$n_{w,k}$	:	w. kelimesinin k. konuya tüm koleksiyonda kaç kere atandığının sayısı
N_m	:	m. dokümandaki toplam kelime sayısı
p	:	Kesinlik
θ	:	Konuların dokümanda bulunma olasılığını
$\theta_{i,k}$	:	k. konunun i. dokümanda bulunma olasılığı
θ_m	:	Konuların m. dokümanda bulunma olasılığı
r	:	Duyarlılık
t	:	Değişken
T	:	Simpleksin boyutu
t_a	:	a_w ile t_w'nin kesişim kümesi
t_w	:	Konu modelleri tarafından çıkartılan kelimeler

u	:	Bilinen temel ölçüt
u_g	:	g. doküman için bilinen temel ölçüt
V	:	Doküman koleksiyonundan elde edilen sabit sözlük
$V^{(k)}$	:	k. konudaki en olası S kelime
$v_S{}^k$	:	k. konudaki S. Kelime
v	:	Modelin saniye cinsinden çalışma süresi
v'	:	Modelin saniye cinsinden çalışma süresinin normalize edilmiş hali
$w_{m,n}$	:	m. dokümanda n. konumda bulunan kelime
x	:	Değişken
Y	:	Değişken
$z_{m,n}$	:	m. dokümanda n. konumda bulunan kelimenin konusu
z_w	:	w. kelimesinin konusu

Kısaltmalar

ADM-LDA:		Aspect Detection Model is based on Latent Dirichlet Allocation
AEP-LDA	:	Appraisal Expression Patterns LDA
CGS	:	Collapsed Gibbs Sampling (Collapsed Gibbs Örnekleme)
CL	:	Cannot-link
Corr-LDA	:	Correspondence Latent Dirichlet Allocation
CTM	:	Correlated Topic Models
DAG	:	Directed Acyclic Graph (Yönlü Döngüsüz Graf)
DDİ	:	Doğal Dil İşleme
DTAS	:	Dependency Topic Affects Sentiment LDA
DTM	:	Dynamic Topic Models (Dinamik Konu Modelleri)
ELDA	:	Enriched LDA
EM	:	Expectation Maximization (Beklenti Maksimizasyonu)
JMTS	:	Joint Mult-grain Topic
JST	:	Joint Sentiment/Topic Model
LDA	:	Latent Dirichlet Allocation (Gizli Dirichlet Ayırımı)
L-LDA	:	Labeled Latent Dirichlet Allocation (Etiketli Gizli Dirichlet Ayırımı)
LSA	:	Latent Semantic Analysis (Gizli Anlamsal Analiz)
LTM	:	Lifelong Topic Model
MCMC	:	Markov Chain Monte Carlo
MedLDA	:	Maximum Entropy Discrimination Latent Dirichlet Allocation (Maksimum Entropi Ayrımı Gizli Dirichlet Ayırımı)
MG-LDA	:	Multi Grain LDA
ML	:	Must-link
PAM	:	Pachinko Allocation Model (Pachinko Ayırımı Modeli)
pLSA	:	Probabilistic Latent Semantic Analysis (Olasılıksal Gizli Anlamsal Analiz)
SLDA	:	Supervised Latent Dirichlet Allocation (Denetimli Gizli Dirichlet Ayırımı)
SVD	:	Singular Value Decomposition (Tekil Değer Ayrışımı)
SVM	:	Support Vector Machines (Destek Vektör Makineleri)
WSA	:	Word Sense Ambiguation (Kelime Anlamı Belirsizliği)

DOKÜMANLARIN ANLAMSAL BENZERLİKLERİNE DAYALI ÖZGÜN BİR KONU MODELLEME YÖNTEMİ

ÖZET

Yapısal ve yapısal olmayan milyarlarca içeriği biz kullanıcılarına sunan Web, günümüzün önemli veri kaynaklarından birisi haline gelmiştir. Sunulan içerik her geçen gün büyümekte, bu içerikten istenilen bilginin otomatik bir şekilde çıkartılması ve çıkartılan bilginin organize edilme, analiz edilme ve anlaşılması adımında ise daha yeni ve daha etkili yöntemlerin geliştirilmesi gerekmektedir. Konu modelleri ise bahsedilen bu görevleri gerçekleştirme aşamasında güçlü ve başarılı bir yöntem olarak karşımıza çıkmaktadır. İlk olarak 1990 yılında ortaya çıkan konu modelleri içerisinde ise en yeni ve başarılı olanı Gizli Dirichlet Ayırımıdır (LDA).

Doküman gibi ayrık verileri modellemek ve dokümanı meydana getiren konuları ortaya çıkarmak için kullanılan üretici grafiksel bir yöntem olan LDA, sadece kelimelerin doküman koleksiyonunda birlikte geçme durumlarını dikkate almaktadır. Buna karşın içerdikleri anlamsal bilgiyi ise dikkate almamaktadır. Bu durum önemli bir dezavantaj oluşturmaktadır.

Bu tez çalışmasında kavram ve adlandırılmış varlıklar şeklindeki anlamsal bilgiyi LDA'ya dahil ederek anlamsal olarak ilişkili, uyumlu, detayları yakalayabilen ve daha anlamlı konuları elde etmek amacıyla iki konu modeli önerilmiştir. Concept-LDA olarak adlandırılan birinci yöntemde, LDA'nın temel varsayımı olan kelime torbası yaklaşımı, {kelime+kavram+adlandırılmış varlık} torbası olacak şekilde genişletilerek anlamsal bir zenginleştirme yöntemi hedeflenmiştir. Geliştirilen Concept-LDA alandan bağımsız bir yöntemdir. NET-LDA olarak adlandırılan ikinci yöntemde ise, anlamsal olarak benzer dokümanlar birleştirilmiş ve birleştirme adımında elde edilen anlamsal benzerlik bilgisi yeni bir adaptif parametre olarak modele dahil edilmiştir. NET-LDA hem alandan hem de dilden bağımsız olup her iki yöntem ile başarılı konuların çıkartılması sağlanmıştır. Anlamsal bilginin elde edilmesi adımında ise graf tabanlı bir yaklaşım olan Babelfy kullanılmıştır.

Geliştirilen yöntemlerin performansları hem niceliksel hem de niteliksel olarak değerlendirilmiştir. Concept-LDA'nın değerlendirilmesi adımında on iki farklı ürüne ait İngilizce kullanıcı yorumları kullanılmıştır; NET-LDA'nın değerlendirilmesinde ise biri Türkçe diğer on iki tanesi İngilizce olmak üzere on üç farklı ürüne ait kullanıcı yorumları kullanılmıştır. Ayrıca, geliştirilen yöntemler hem niceliksel hem de niteliksel olarak üç temel yöntemden elde edilen sonuçlar ile karşılaştırılmıştır. Yapılan deneyler sonucunda anlamsal bilginin modele dahil edilmesi ile anlamsal olarak ilişkili, uyumlu, detayları yakalayabilen ve daha anlamlı konuların elde edildiği görülmüştür. Geliştirilen yöntemlerin temel yöntemlere kıyasla da oldukça başarılı oldukları yapılan deneylerde ispatlanmıştır.

Anahtar Kelimeler: Anlamsal Bilgi, Babelfy, Gizli Dirichlet Ayırımı, Kavram ve Adlandırılmış Varlıklar, Konu Modelleri, Özellik Çıkarımı.

AN ORIGINAL TOPIC MODEL METHOD BASED ON SEMANTIC SIMILARITY OF DOCUMENTS

ABSTRACT

The Web, which provides billions of structural and non-structural content to its users, has become one of today's important data sources. The content provided is growing day by day, newer and more effective methods need to be developed in the process of automatically extracting desired information from this content and organizing, analyzing and understanding this extracted information. Topic models come across as a powerful and successful method for performing these tasks. Among the topic models themselves, which first appeared in 1990, Latent Dirichlet Allocation (LDA) is the most recent and successful topic model.

LDA, which is a generative graphical method used to model discrete data such as documents and reveal the topics that compose the documents, considers only word co-occurrence distribution in the document. On the other hand, LDA does not considers the semantic information documents contain. This poses a significant drawback.

In this thesis, two topic models have been devised by incorporating semantic knowledge in the form of concepts and named entities into the LDA in order to obtain semantically related, coherent, detailed and more meaningful topics. In the first method called Concept-LDA, bag-of-words which is the basic assumption of LDA is expanded to be a bag of {words+concepts+named entities} as a semantic enrichment method is aimed. The proposed Concept-LDA is independent of domain. In the second method called NET-LDA, semantically similar documents are merged and semantic similarity obtaining in the merging step is injected into the model as a new adaptive parameter. NET-LDA is independent both of domain and language. In the step of obtaining semantic knowledge a graph based approach Babelfy is used.

The performances of the proposed methods are evaluated both quantitatively and qualitatively. In the evaluation of Concept-LDA, user reviews of twelve different domains are used; in the evaluation of NET-LDA, user reviews of thirteen different domains one in Turkish and the other twelve in English are used. Besides, the proposed methods are compared both quantitatively and qualitatively with the results obtained from three baselines. As a result of the experiments conducted, it is seen that the incorporating semantic knowledge into the model semantically related, coherent, detailed and more meaningful topics are obtained. It has been proved with the experiments that the proposed methods are also fairly successful compared to the baselines.

Key Words: Semantic Knowledge, Babelfy, Latent Dirichlet Allocation, Concepts and Named Entities, Topic Models, Aspect Extraction.

GİRİŞ

Web, 1990'lardan itibaren aşamalı olarak gelişme gösterirken; sınırlı kaynak sunmasından ötürü kullanıcılar bu yeni tanıştıkları dünya ile etkileşim konusunda ilk başlarda pasif kalmışlardır (Jiménez-Zafra ve diğ., 2016). Ancak, zaman içerisinde Web teknolojilerinde yaşanan gelişmeler özellikle Web 3.0 ile internetin günlük hayatın önemli bir parçası haline gelmesi ve çevrimiçi aktiviteler (sosyal medya kullanımı, elektronik alışveriş, blog yazarlığı, internet bankacılığı, gazete okuma, çevrimiçi yorum yapma, vb.) ile birlikte büyük kolaylıklar gelmiştir. Sonuç olarak yapısal ve yapısal olmayan büyük miktardaki verilerin depolandığı bir ortam biz kullanıcılara özellikle de veriler üzerinden çalışmalar yapan araştırmacılara sağlanmıştır. Bununla birlikte her geçen gün artan veri miktarı içerisinden bilginin otomatik çıkarımı ise zor hale gelmiştir. Dolayısıyla, bu büyük miktardaki veriyi organize etmeye, analiz etmeye ve anlamaya yardımcı olma adımında konu modelleri önemli bir araç olarak karşımıza çıkmıştır.

İlk olarak Deerwester ve diğ. (1990) tarafından önerilen Gizli Anlamsal Analiz (LSA) yöntemi ile ortaya çıkan konu modelleri son yıllarda makine öğrenmesi ve metin madenciliği uygulamalarında aktif bir araştırma alanı haline gelmiştir. Konu modellerindeki konu; dokümanlardaki gizli tematik bilgidir, yani dokümanın temasıdır ve konu modellerindeki birincil amaç yapısal olmayan doküman koleksiyonlarındaki bu gizli tematik bilgiyi küçük boyutlu uzaya çevirerek keşfetmektir (Blei ve diğ., 2003; Steyvers ve Griffiths, 2007; Boyd-Graber ve Blei, (2009, 2011); Lu ve diğ., 2011). Konu modellerinin bu gizli tematik bilgiyi keşfederken dayandığı temel fikir ise; kelimeler üzerinden olasılık dağılımına sahip olan konuların rastgele bir araya gelerek dokümanları oluşturması şeklinde açıklanmaktadır (Hofmann, (1999, 2001); Griffiths ve Steyvers, (2002a, 2002b, 2004); Blei ve diğ., 2003; Steyvers ve Griffiths, 2007).

Bilinen ilk konu modeli olan LSA ile doküman koleksiyonundaki gizli anlamsal ilişkileri keşfederek düşük boyutlu anlamsal bir uzay elde etmek için doküman terim

matrisi üzerinden tekil değer ayrışımı (SVD) uygulanmıştır (Deerwester ve diğ., 1990).

LSA'nın istatistiksel bir görünümü olarak Hofmann (1999) tarafından geliştirilen olasılıksal gizli anlamsal analiz (pLSA) LSA'ya kıyasla daha karmaşık bir yaklaşımdır. Üretici ve grafiksel bir yöntem olan pLSA, dokümanları gizli konuların bir karışımı olarak modellemeye yönelik olasılıksal bir yaklaşım sağlayan ilk yöntemlerden birisidir. Ancak model sadece kelime seviyesinde bir olasılık modeli sunmaktadır. Dolayısıyla bu da pLSA'nın tam bir üretici model olarak çalışmasını engellemektedir. Aşırı öğrenmeye meyilli olması ve daha önce görmediği dokümanlar üzerinde genelleme yapamaması ise modelin önemli bir dezavantajıdır (Popescul ve diğ., 2001).

Blei ve diğ. tarafından 2003 yılında geliştirilen Gizli Dirichlet Ayırımı (LDA), pLSA'nın bazı dezavantajlarını ortadan kaldıran tam bir üretici modeldir. Tam bir üretici model olmasını ise konuların dokümanlardaki dağılımını temsil eden parametrelerini bir Dirichlet dağılımından gelen değişkenler olarak ele alıp pLSA modelini genişleterek sağlamaktadır. Ayrıca tamamen denetimsiz bir yöntem olan LDA kelime torbası yaklaşımına dayalı çalışmaktadır yani kelimelerin doküman içerisindeki yerleşimi göz ardı edilmektedir.

LDA, doküman gibi ayrık verileri modellemek ve dokümanı meydana getiren konuları ortaya çıkarmak için kullanılan üretici olmasının yanında grafiksel bir modeldir de (Blei ve diğ., 2003). Ancak bir doküman koleksiyonunda mümkün konu yapısı oldukça fazla olduğu için konuların elde edilmesinde örnekleme adımına ihtiyaç duyulmaktadır. LDA'da kullanılan örnekleme yöntemlerinin başında ise Beklenti Maksimizasyonu (EM) ve Gibbs örnekleme gelmektedir.

EM; verilen bir dizi gözleme dayalı olarak modelin olasılığını maksimum yapmaya çalışan denetimsiz bir öğrenme yöntemidir (Banko, 2018). Markov Chain Monte Carlo (MCMC)'nun özel bir türü olan Gibbs örnekleme ise Bayesian çıkarımındaki sonsal dağılım için kullanılmaktadır.

LDA'nın başarılı bir konu modeli olması ile birlikte araştırmacılar daha sonraki yıllarda yeni konu modelleri tasarlamak yerine LDA tabanlı modeller geliştirmeye ve

pek çok farklı alana uygulamaya başlamışlardır (Jelodar ve diğ., 2018). LDA tabanlı geliştirilen ilk ve başlıca yaklaşımlar; correspondence LDA (Corr-LDA), Konu-Yazar Modeli, Correlated Konu Modelleri (CTM), Dinamik Konu Modelleri (DTM), Pachinko Ayırımı Modeli (PAM), Denetimli Gizli Dirichlet Ayırımı (SLDA), Etiketli Gizli Dirichlet Ayırımı (L-LDA) ve Maksimum Entropi Ayrımı Gizli Dirichlet Ayırımı (MedLDA) şeklinde sıralanmaktadır.

Corr-LDA; görüntü etiketleme, otomatik bölge etiketleme ve metin-tabanlı görüntü erişimi görevlerini yerine getirmek üzere geliştirilmiş, parametre tahmini ve kestiriminde varyasyonel EM kullanan bir konu modelidir (Blei ve Jordan, 2003). Amaçlanan model, görüntüdeki bölgelerin ve bu bölgeleri etiketlemede kullanılacak kelimelerin gizli değişkenler ile temsilleri arasındaki koşullu ilişkiyi bulmayı hedeflemektedir. Deneysel çalışma 7000 adet etiketli görüntüyü içeren Corel veritabanı ve etiketlemede kullanılan 168 kelimeden oluşan sözlük üzerinden gerçeklenmiştir. Her görüntü 6-10 arası bölgeye ayrılmış olup, 2-4 arası etiket ile ilişkilendirilmiştir. Veri kümesinin %25'i test kümesi olarak kullanılmış ve %80-%90 arası bir başarı elde edilmiştir.

Yazar-Konu Modeli her dokümanı konular üzerinden dağılım ile ilişkilendirmek yerine her yazarı konular üzerinden dağılım ile ilişkilendirmektedir (Steyvers ve diğ., 2004; Rosen-Zvi ve diğ., 2004). Parametre kestiriminde Gibbs Örnekleme kullanan bu modeldeki temel fikir, birden fazla yazar tarafından oluşturulan dokümanlar, birden fazla yazarın ve bu yazarların üzerinde durduğu birden fazla konunun birleşiminden oluşmaktadır. Bunun için modelde, yazar bilgisi ile doküman içeriği birleştirilir ve dokümanların içeriği ve yazarların ilgi alanları eş zamanlı olarak modellenmiş olur. Model NIPS ve CiteSeer'den elde edilen akademik makalelerin özetleri üzerine uygulanmıştır.

LDA'nın önemli bir kısıtlaması elde edilen konular arasındaki korelasyonu modellemiyor olmasıdır. Bunun nedeni konu oranları arasındaki değişkenlik durumunu Dirichlet dağılımı ile modelliyor olmasıdır. Blei ve Lafferty (2006a) tarafından geliştirilen CTM konular arasındaki korelasyonu yakalamak amacıyla lojistik normal dağılımı kullanmıştır. CTM'deki temel fikir; gizli bir konunun dokümanda bulunması başka bir konuyla ilişki olabilir şeklindedir. CTM'nin LDA ile

Science dergisinde yer alan OCRed makaleleri üzerinden karşılaştırıldığında daha başarılı olduğu gözlemlenmiştir. Yine bu modelde parametre tahmini ve kestirimi için EM modeli kullanılmıştır.

Blei ve Lafferty (2006b) tarafından geliştirilen DTM sıralı bir şekilde organize edilmiş doküman koleksiyonlarındaki konuların yıllara göre gelişimini analiz etmek ve hangi yılda hangi konunun popüler olduğunu belirlemektedir ve olasılıksal zaman serisi modelleri ailesinden gelmektedir. DTM, parametre tahmini ve kestirimi adımında varyasyonel Kalman filtresi ve varyasyonel Dalgacık (Wavelet) Regresyonunu kullanmaktadır. Konular arasındaki ilişkiyi dikkate alan bir diğer konu modeli ise PAM'dır (Li ve McCallum, 2006). Model konular arasındaki keyfi, iç içe ve muhtemelen seyrek ilişkileri yönlü döngüsüz graf (DAG) kullanarak tespit etmektedir. Bu yöntemde parametre tahmini ve kestirimi adımında Gibbs Örnekleme kullanılmıştır.

LDA denetimsiz bir yöntem olmakla birlikte LDA tabanlı denetimli konu modelleri de geliştirilmiştir. SLDA kullanıcı yorumlarından film derecelendirme puanı tahmini ve tanımlardan web sayfalarının popülaritesinin tahmini problemleri üzerine uygulanmış denetimli bir konu modelidir (Blei ve McAuliffe, 2007). Parametre tahmini ve kestiriminde EM kullanılmıştır.

Kredi atama problemini ele alan, çok etiketli dokümanlar için önerilen L-LDA da SLDA gibi denetimli bir konu modelidir (Ramage ve diğ., 2009). Bu yöntem ile dokümandaki her kelime en uygun etiket ile ya da her etiket en uygun kelime ile eşleştirilir. Modelin LDA'dan farkı; konu modelinin, sadece gözlemlenen dokümanın etiket kümesi ile ilişkili konuların kullanılması ile kısıtlanmasıdır, yani denetimin modele dahil edilmesidir. Model destek vektör makineleri (SVM) ile karşılaştırılmıştır ve SVM'ye göre oldukça başarılı olduğu yapılan deneyler sonucunda tespit edilmiştir. Bu modelde de parametre tahmini ve kestirimi adımında Gibbs Örnekleme kullanılmıştır.

Denetimli konu modellerinden bir diğeri olan Med-LDA regresyon ve sınıflandırma problemleri için 2009 yılında geliştirilmiştir (Zhu ve diğ., 2009). Geliştirilen yöntem, denetimli konu modellerini eğitme adımında max-margin prensibini kullanmaktadır. Regresyon ve sınıflandırma problemleri için daha başarılı konu temsilleri elde

edebilmek adına, tek amaçlı fonksiyonunun beklenen pay kısıtları ile optimize edilmesiyle max-margin prensibi gizli konuları keşfetme sürecine dahil edilmektedir. Bu modelde de parametre tahmini ve kestirimi adımında EM kullanılmıştır.

Diğer bir taraftan, LDA literatürde kaynak kod analizinden (Linstead ve diğ., 2007; Lukins ve diğ., 2008; Lukins ve diğ. 2010; Savage ve diğ., 2010; Mahmoud ve Niu, 2015) etiket önerisine (Bundschusve diğ., 2009; Krestel ve Fankhauser, 2009; Krestel ve diğ., 2009; Si ve Sun, 2009; Lu ve Lee, 2015; Zhao ve diğ., 2016), görüntü sınıflandırma ve etiketlemeden (Blei ve Jordan, 2003; Barnard ve diğ. 2003; Bissacco ve diğ., 2006; Rasiwasia ve Vasconcelos, 2013; Bahmanyar ve diğ., 2018) olay tespitine (Ritter ve diğ., 2012; Hu ve diğ., 2012; Rule ve diğ., 2018), duygu sınıflandırmadan (Bao ve diğ., 2009; Bao ve diğ. 2012; Liang ve diğ. 2018) kullanıcı yorumlarından özellik çıkarmaya (Titov ve McDonald, 2008; Atıcı ve diğ. 2017; Ekinci ve İlhan Omurca 2017a; Wang ve diğ. 2018) kadar pek çok alana uygulanmaktadır.

Gizli uzaydaki konuların anlamsal olarak uyumlu olduğu söylense bile, LDA sadece kelimelerin doküman koleksiyonunda birlikte geçme durumlarını dikkate almaktadır, içerdikleri anlamsal bilgiyi ise dikkate almamaktadır (Chang ve diğ., 2009). Bu durum LDA için bir dezavantaj oluşturmaktadır. Bu dezavantajın üstesinden gelebilmek için bu tez çalışmasında anlamsal bilgiyi modele dahil eden Concept-LDA ve NET-LDA olmak üzere iki farklı konu modeli geliştirilmiştir. Burada bahsedilen anlamsal bilgi kavramlar ve adlandırılmış varlıklardır. Kavram ve adlandırılmış varlıkları çıkartmak amacıyla ise Babelfy kullanılmıştır. Babelfy, varlık bağlama (entity linking) ve kelime anlamı belirginleştirme yöntemlerine dayalı graf tabanlı bir yaklaşım olup, aday anlamları alt graf şeklinde verip yüksek tutarlılık gösteren anlamları sezgisel olarak seçmektedir (Moro ve diğ., 2014b). Concept-LDA İngilizce kullanıcı yorumlarından ürün özelliklerini çıkartmak amacıyla kelime torbası yaklaşımı yerine {kelime+kavram+adlandırılmış varlık} torbası yaklaşımını kullanmaktadır. Model on iki farklı veri kümesine uygulanmıştır. NET-LDA ise dokümanlar arasındaki anlamsal benzerliği kavram ve adlandırılmış varlıklar üzerinden hesaplayarak benzer dokümanları birleştirip; birleştirilen doküman sayısını doküman-konu dağılımına etki ettirerek LDA'nın temel varsayımı olan kelimelerin birlikte geçme durumlarını anlamsal olarak güçlendirmektedir. LDA tabanlı çalışmaların büyük bir kısmı simetrik

önseller kullanılarak gerçeklenmekte iken NET-LDA doküman-konu dağılımına etki eden bilinen temel ölçütün kullanılması ile asimetrik önseller ile gerçeklenmiştir. Ayrıca dilden bağımsız olarak geliştirilen NET-LDA hem Türkçe hem de İngilizce dokümanlar üzerinden konuları çıkarmaktadır ve model biri Türkçe olmak üzere on üç farklı veri kümesine uygulanmıştır. Tez kapsamında önerilen her iki konu modeli ile, kullanılan kavram ve adlandırılmış varlıklar sayesinde anlamsal olarak ilişkili, uyumlu, detayları yakalayabilen ve daha anlamlı konuların alandan bağımsız bir şekilde elde edilmesi hedeflenmiştir. Yapılan deneyler niceliksel ve niteliksel değerlendirildiğinde ise geliştirilen bu iki yöntemin her anlamda başarılı olduğu temel yöntemler ile yapılan karşılaştırmalar sonucunda gözlemlenmiştir. Ayrıca geliştirilen yöntemler çalışma süreleri açısından da karşılaştırılmıştır. Bu açıdan ise NET-LDA diğer yöntemlere üstünlük sağlamıştır.

Tez çalışmasının birinci bölümünde kullanıcı yorumlarından ürün özelliklerini çıkarma adımının temeli olan LDA ayrıntılı bir şekilde anlatılacaktır. İkinci bölümde Babelfy'dan, üçüncü bölümde tasarlanan mimariden; önişleme adımı, isim öbeklerinin dokümanlardan Babelfy ile çıkartılması, kavram ve adlandırılmış varlıkların elde edilmesi ve Concept-LDA ile NET-LDA ile konuların elde edilmesi olacak şekilde bahsedilecektir. Dördüncü bölümde; yapılan deneysel çalışmalar, önerilen yöntemlerin temel yöntemler ile karşılaştırılması ve elde edilen sonuçların niceliksel ve niteliksel değerlendirilmesi yapılacaktır. Sonuçlar ve öneriler bölümünde, elde edilen sonuçlar yorumlanacak, çalışmanın bilime ve günümüz teknolojisine sağlayabileceği katkıları tartışılacaktır. Ayrıca gelecekte yapılacak çalışmalar için önerilerde bulunulacaktır.

Literatürde yer alan konu modelleme ile ilgili yapılan ulusal ve uluslararası çalışmalar incelendiğinde {kelime+kavram+adlandırılmış varlık} torbasına ve doküman benzerliğine dayalı konu modellerine rastlanmamıştır. Ayrıca Babelfy ile edilen kavram ve adlandırılmış varlıklar da konu modellerine ilk kez bu tez çalışması ile dahil edilmiştir. Tüm bunlar göz önünde bulundurulduğunda, özgün konu modelleme yaklaşımları önerilmekte dolayısıyla günümüz araştırmacılarına ve ileride yapılacak çalışmalara önemli katkılar sağlayacağı düşünülmektedir.

1. KONU MODELLERİ

Dijitalleşen dünya ile birlikte İnternet, kullanıcılarına haber siteleri, bloglar, forumlar, sosyal ağlar, kütüphaneler vb. ortamları sunmaya başlamıştır. Bu ortamların biz kullanıcılara sağladığı ve her geçen gün artan büyük miktardaki veriye erişim ve bu veri içerisinden aradığımız bilgiyi ortaya çıkarmak ise normal bir insan için zor bir görevdir. Dolayısıyla da bu büyük miktardaki veriyi organize etmeye, analiz etmeye ve anlamaya yardımcı olma adımında yeni ve otomatik yöntemlere ihtiyaç duyulmaktadır.

Tamamen denetimsiz olan konu modelleri bu çok büyük miktardaki veriyi otomatik olarak organize etme, analiz etme, anlama, özetleme ve bu veri içerisinde arama yapmamızı sağlayan yöntemleri biz kullanıcılarına sunmaktadır (Blei, 2013). Böylece; doküman içerisindeki gizli tematik bilgi yani dokümanın konusu keşfedilmiş olur, dokümanlar bu konulara göre etiketlenebilir ve bu etiketler doküman koleksiyonun organize edilmesinde, özetlenmesinde, koleksiyon üzerinde arama yapılmasında kullanılabilir. Ayrıca, konuların birbiri ile olan ilişkisi, zaman içerisinde gösterdikleri değişimleri ("Makine öğrenmesi yöntemlerinin 2000-2018 yılları arasında uygulandığı alanlar ve değişimler nelerdir?") keşfetmeye de yardımcı olmaktadır. Şekil 1.1'de Jo ve Oh'un (2011) çalışmalarında kullandıkları restoran yorumlarından elde edilen konular gösterilmiştir.

Konu 32	**Konu 67**	**Konu 78**	**Konu 98**
Kahvaltı	Salata	İçki	Pazar kahvaltısı
Kahve	Pancar	Kokteyl	Yumurta
Patates	Keçi peyniri	Akşam yemeği	Beklemek
Yumurta	Salata sosu	Liste	Mimoza
Meyve	Taraf	Martini	Omlet
Sosis	Salatalık	Alkol	Yumurtalı ekmek
Fransız tostu	Marul	Tur	Krep
Gözleme	Göğüs biftek	New York	Çılbır
İrmik	Kızarmış ekmek	Karışım	Öğleyin
Fransız ekmeği	Öneri	Likör	Somon füme

Şekil 1.1. Restoran yorumlarından elde edilen dört konu

Konu modelleme için önerilen algoritmalar istatistiksel yöntemler olup, dokümanı oluşturan kelimeleri analiz ederek bir sonuca varmayı amaçlar. Konu modelleme yöntemleri üzerine literatürde pek çok başarılı çalışma olmakla birlikte, hala daha teorikte anlaşılması güç bir konu olarak karşımıza çıkmaktadır.

Konu modelleri ilk olarak Deerwester ve diğ. (1990) tarafından önerilen Gizli Anlamsal Analiz (LSA) yöntemi ile ortaya çıkmıştır. LSA ile doküman terim matrisi üzerinden tekil değer ayrışımı (SVD) uygulanarak doküman koleksiyonundaki gizli anlamsal ilişkiler keşfedilip düşük boyutlu anlamsal bir uzay elde edilmiştir (Deerwester ve diğ., 1990).

LSA'nın istatistiksel bir görünümü olarak geliştirilen ve olasılıksal gizli anlamsal analiz (pLSA) olarak adlandırılan daha karmaşık bir yaklaşım ise Hofmann (1999) tarafından geliştirilmiştir. Üretici ve grafiksel bir yöntem olan pLSA, dokümanları gizli konuların bir karışımı olarak modellemeye yönelik olasılıksal bir yaklaşım sağlayan ilk yöntemlerden birisidir. Modelin sadece kelime seviyesinde bir olasılık modeli sunması tam bir üretici model olmasını engellemektedir. Aşırı öğrenmeye meyilli olması ve daha önce görmediği dokümanlar üzerinde genelleme yapamaması ise modelin önemli bir dezavantajıdır (Popescul ve diğ., 2001).

Blei ve diğ. tarafından 2003 yılında geliştirilen Gizli Dirichlet Ayırımı (LDA), konuların dokümanlardaki dağılımını temsil eden parametrelerini bir Dirichlet dağılımından gelen değişkenler olarak ele alarak pLSA modelini genişletir ve böylece pLSA'nın bazı dezavantajlarını ortadan kaldıran tam bir üretici modeli tanımlar.

Bu tez çalışması kapsamında ise kullanıcı yorumlarında geçen ürün özellikleri dokümanın konusu olarak ele alınmış ve bu özelliklerin çıkartılması amacıyla en yaygın ve başarılı konu modelleme yöntemlerinden birisi olan LDA'nın kullanılmasına karar verilmiştir.

1.1. Gizli Dirichlet Ayırımı

LDA, doküman gibi ayrık verileri modellemek ve dokümanı meydana getiren konuları ortaya çıkarmak için kullanılan üretici bir konu modelidir (Blei ve diğ., 2003). LDA tamamen denetimsiz bir yöntemdir dolayısıyla herhangi bir önbilgiye ihtiyaç duymaz

ve kelime torbası yaklaşımına dayalı çalışmaktadır. Kelimelerin doküman içerisindeki yerleşimi göz ardı edilirken, kelimelerin birlikte bulunması göz önünde bulundurulur. Gizli Dirichlet Ayrımındaki "gizli" ile ifade edilmek istenilen gizli konuların keşfedilmesiyle dokümanın temasının bulunmasıdır (Jadhav, 2018). Dirichlet; çokterimli değişkenler için eşlenik önsel dağılımdır (Bishop, 2006). Üreticilik ise LDA'nın dayandığı temel fikirdir. Üretici ile kastedilen ise kelimeler üzerinden olasılık dağılımına sahip olan konuların rastgele bir araya gelerek dokümanları oluşturması şeklinde açıklanmaktadır (Hofmann, (1999, 2001); Griffiths ve Steyvers, (2002a, 2002b, 2004); Blei ve diğ., 2003; Steyvers ve Griffiths, 2007). LDA'ya ait üretici model Şekil 1.2'de verilmiştir.

1. Her konu k için $k \in [1, K]$
    a. Kelimelerin konular içerisindeki dağılımını belirle: $\varphi_k \sim Dirichlet(\beta)$
2. Her doküman m için $m \in [1, M]$
    a. Konuların doküman içerisindeki dağılımını belirle: $\theta_m \sim Dirichlet(\alpha)$
    b. Doküman m'deki her kelime w_n için
        i. Rastgele bir konu belirle: $z_{m,n} \sim Mult(\theta_m)$
        ii. Rastgele bir kelime seç: $w_{m,n} \sim Mult(\varphi_k, z_{m,n})$

Şekil 1.2. LDA için üretici model

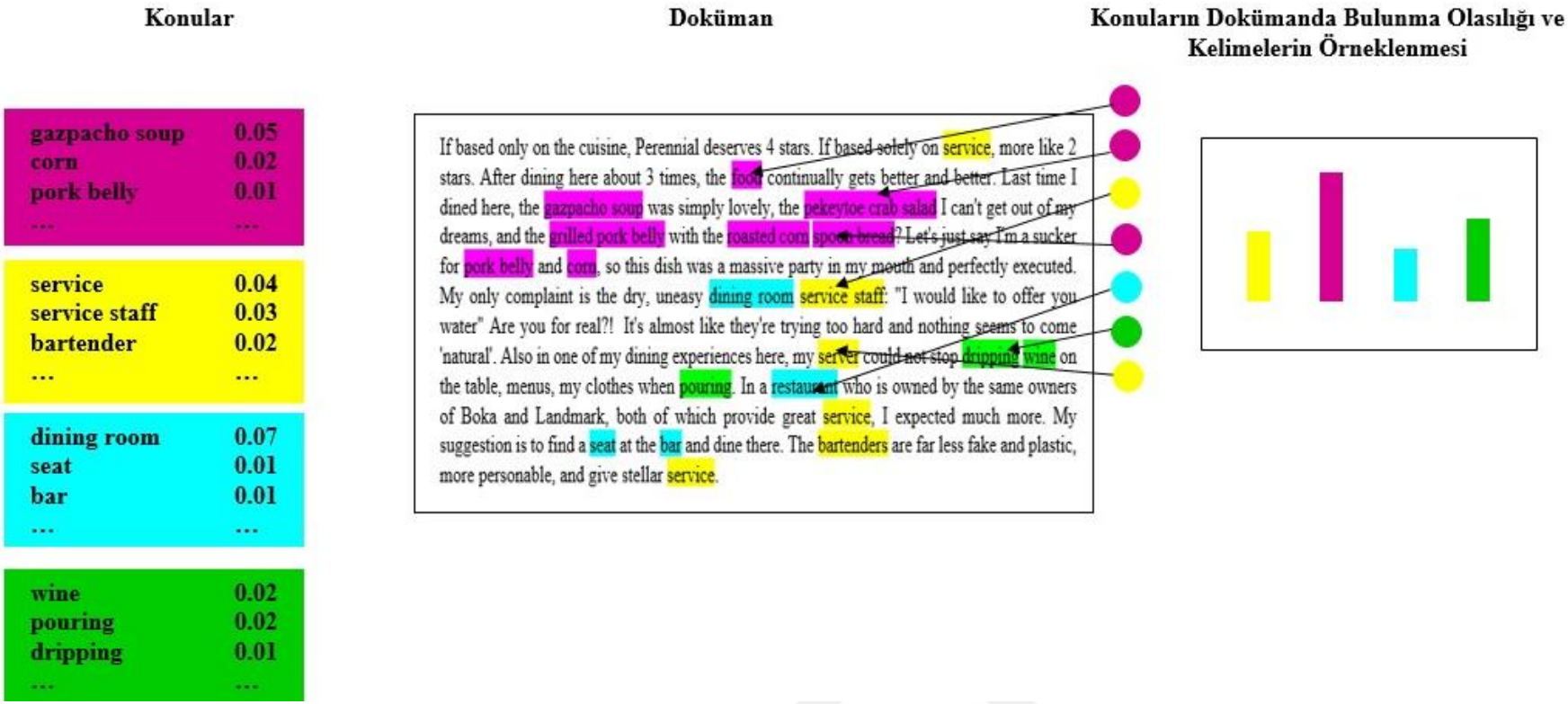
Bu istatistiksel model dokümanların birden fazla konunun karışımından oluştuğu varsayımına dayanmaktadır. Üretici modelin birinci adımında konular, sabit bir sözlük olan V'de yer alan kelimeler üzerinden Dirichlet dağılımına göre olasılık dağılımı göstermektedir. Bu sabit sözlük doküman koleksiyonundaki kelimelerden oluşmaktadır. İkinci adımda her doküman için her konunun ilgili dokümanda bulunma olasılığı yine Dirichlet dağılımına göre belirlenir. Dokümanda yer alan her kelime için konular çok terimli dağılıma göre örneklenmektedir. Son olarak da ilgili konu için kelime çok terimli dağılıma göre örneklenmektedir. LDA için üretici model örnek bir restoran yorumu üzerinden Şekil 1.3'ten itibaren anlatılmaktadır.

If based only on the cuisine, Perennial deserves 4 stars. If based solely on service, more like 2 stars. After dining here about 3 times, the food continually gets better and better. Last time I dined here, the gazpacho soup was simply lovely, the pekeytoe crab salad I can't get out of my dreams, and the grilled pork belly with the roasted corn spoon bread? Let's just say I'm a sucker for pork belly and corn, so this dish was a massive party in my mouth and perfectly executed. My only complaint is the dry, uneasy dining room service staff: "I would like to offer you water" Are you for real?! It's almost like they're trying too hard and nothing seems to come 'natural'. Also in one of my dining experiences here, my server could not stop dripping wine on the table, menus, my clothes when pouring. In a restaurant who is owned by the same owners of Boka and Landmark, both of which provide great service, I expected much more. My suggestion is to find a seat at the bar and dine there. The bartenders are far less fake and plastic, more personable, and give stellar service.

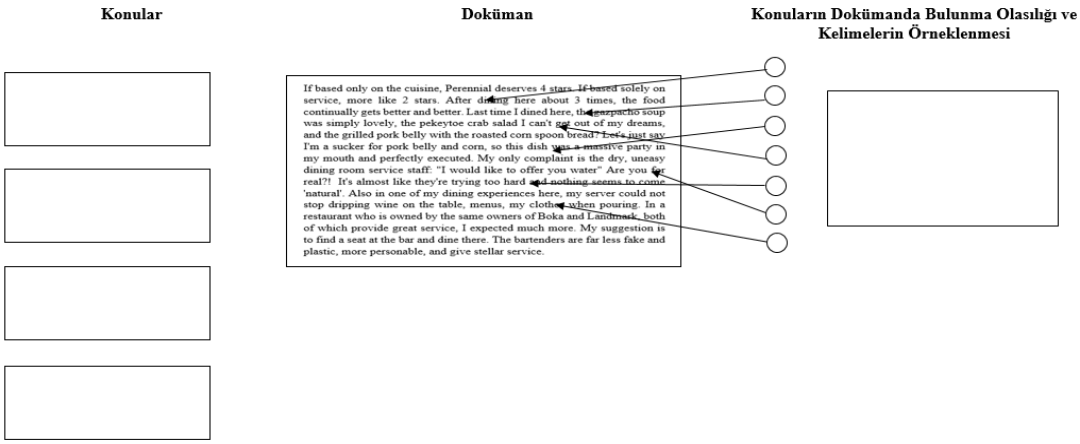
Şekil 1.3. Bir dokümanın birden fazla konunun karışımı olması

Yukarıdaki örnek yorumda doküman; servis, yiyecek, ortam ve içecek konuları üzerinden olasılık dağılımı göstermekteyken, dokümandaki kelimeler bu konulardan birisi altında olasılık dağılımı göstermektedir. Bu durum Şekil 1.4'te gösterilmiştir.

Şekil 1.4'ün sol tarafında gösterildiği üzere tüm doküman koleksiyonu için sabit sözlük üzerinden olasılık dağılımı gösteren belli sayıdaki konuların olduğu varsayılır. Dokümanın üretilmesi ise şu şekilde gerçeklenir: i) Histogramda gösterildiği gibi konuların doküman üzerinden olasılık dağılımı belirlenir, ii) Her bir kelime için konuların örneklenmesi gerçeklenir, renkli daireler konuların örneklenmelerini temsil etmek için kullanılmıştır, iii) Son adımda örneklenen konu için sol kısımda yer alan kelimelerden bir tanesi seçilir. Ancak burada bir hayal dünyasından bahsedilmektedir. Öğrenilmek istenen gizli değişkenler gözlemleniyor gibi davranılmaktadır. Gerçek dünyada ise sadece Şekil 1.5'teki gibi doküman yani dokümanı oluşturan kelimeler gözlemlenebilmektedir.

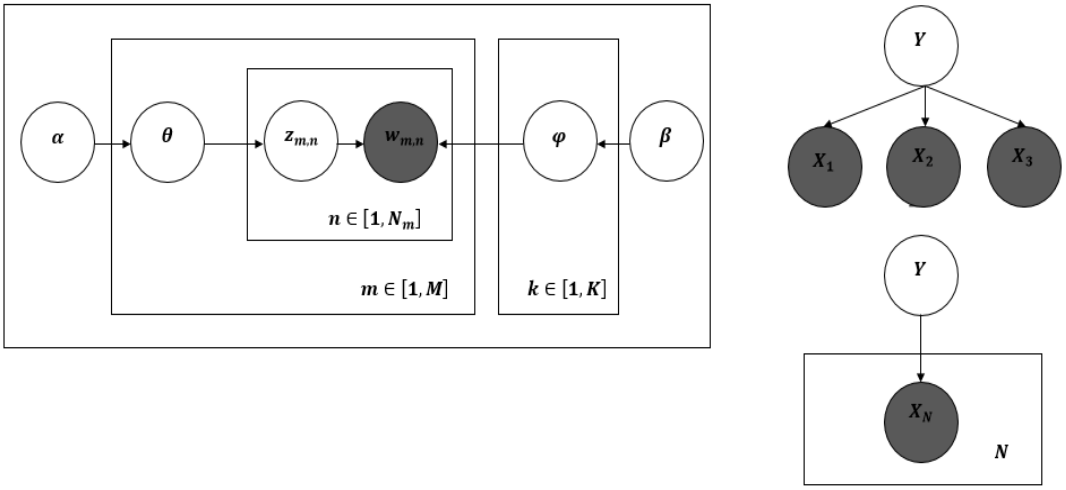


Şekil 1.4. LDA'nın altında yatan temel fikir



Şekil 1.5. Gerçek dünya görüntüsü

LDA üretici bir model olmasının yanında ayrıca grafiksel bir modeldir ve grafiksel temsilinde plate notasyonu kullanılmaktadır. Plate notasyonu, aynı tipteki birden fazla nesnenin tekrarlama durumunu ifade etmektedir (Ekinci ve İlhan Omurca, 2017a). Plate notasyonu LDA için gözlemlenen verinin yani dokümanı oluşturan kelimelerin rastgele değişkenler yani gözlemlenemeyen veriler (konular, konuların dokümanda bulunma olasılığı ve kelimelerin konulara atanma olasılığı) ve bu değişkenlerin yönlü kenarlar üzerinden nasıl üretildiğini anlatmaktadır. LDA için plate notasyonu Şekil 1.6'da verilmiştir. Elimizde sadece dokümanlar gözlenebilir durumda olup; konular, konuların dokümandaki ve kelimelerin konulardaki dağılımları gizlidir. Bu nedenle grafiksel modelde gözlemlenen değişkenler gri renkle temsil edilirken gözlenemeyenler beyaz renk ile temsil edilmiştir.



Şekil 1.6. LDA'nın grafiksel temsili (Blei, 2018)

Şekil 1.6'da verilen grafiksel modelde M koleksiyonda yer alan toplam doküman sayısını, N_m ise m. dokümandaki toplam kelime sayısını temsil etmektedir. $w_{m,n}$ m. dokümanda n. konumda bulunan kelimeyi, $z_{m,n}$ ise m. dokümanda n. konumda bulunan kelimenin konusunu temsil etmektedir. K toplam konu sayısıdır. θ konuların dokümanda bulunma olasılığını, φ ise kelimelerin konulardaki dağılımını göstermektedir. α ve β Dirichlet parametreleridir. Verilen grafiksel modele göre tüm gizli ve gözlemlenen rastgele değişkenlerin birleşik dağılımı $p(\varphi_{1:K},\theta_{1:M},z_{1:M},w_{1:M})$ Eşitlik (1.1)'de verilmiştir.

$$\left(\prod_{k=1}^{K} p(\varphi_k \mid \beta)\right)\left(\prod_{m=1}^{M} p(\theta_m \mid \alpha)\right)\left(\prod_{n=1}^{N} p(z_{m,n} \mid \theta_m)(w_{m,n} \mid z_{m,n}, \varphi_k)\right) \tag{1.1}$$

LDA ile asıl hedeflenen gizli değişkenlerin yani model parametrelerinin elde edilmesidir. Bu amaçla Eşitlik (1.2)'deki sonsal dağılım kullanılmaktadır.

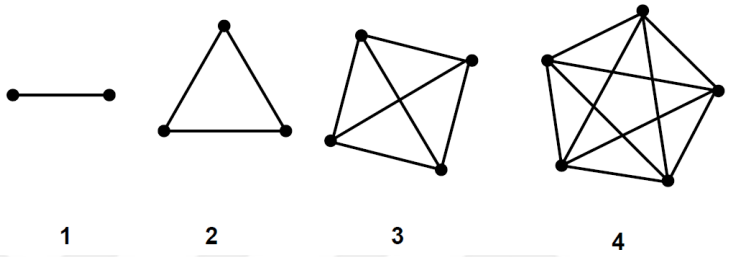
$$p(\varphi_{1:K}, \theta_{1:M}, z_{1:M} \mid w_{1:M}) = \frac{p(\varphi_{1:K}, \theta_{1:M}, z_{1:M}, w_{1:M})}{p(w_{1:M})} \tag{1.2}$$

Eşitlik (1.2) incelendiğinde pay kısmının tüm rastgele değişkenlerin ortak dağılımı olduğu görülmektedir ve pay kolayca hesaplanabilir. Ancak paydaya bakıldığında gözlemlerin marjinal olasılığı olduğu görülmektedir yani bu marjinal olasılık doküman kümesinin herhangi bir konu modeli altındaki olasılığına karşılık gelmektedir. Hesaplanabilmesi için gizli konu yapısının tüm örnekleri üzerinden ortak dağılımı toplamak gerekmektedir. Yalnız mümkün konu yapısı oldukça fazla olduğu için bu toplamın hesaplanması mümkün değildir. Bu nedenle sonsal dağılıma yakınsamak gerekmektedir ve örnekleme algoritmalarından yararlanılmaktadır. Bu örnekleme algoritmalarından en yaygın kullanılan Gibbs Örneklemenin standart bir gerçekleştirimi olan Collapsed Gibbs Örnekleme (CGS) algoritmasıdır. Bu tez çalışmasında sonsal dağılıma yakınsamak için CGS algoritmasından yararlanılmıştır. Ancak öncesinde Dirichlet Dağılımından bahsedilecektir.

1.1.1. Dirichlet dağılımı

Dirichlet dağılımı Beta dağılımının çok değişkenli versiyonu olup, bir simpleks ile sınırlandırılmış rastgele vektörler için başlıca çok değişkenli bir dağılım olarak

tanımlanmaktadır, başka bir deyişle de toplamı bire eşit olan pozitif vektörleri tanımlamaktadır (Ng ve diğ., 2011). Geometride hiper-tetrahedron olarak adlandırılan simpleks tetrahedral bir bölgenin rastgele n boyunun genellemesi olarak tanımlanmaktadır (Li ve diğ., 2015). Bir T simpleks ise T+1 nokta kümesinden oluşmaktadır. Bu noktaların hepsi her yerde sıfır değerine sahip olmayan T boyutlu bir hacim elemanı tanımlamaktadır. Örneğin iki nokta 1 boyutta bir çizgi oluştururken, üç nokta ile 2 boyutta üçgen, dört nokta ile 3 boyutta dörtyüzlü oluşmaktadır. 1 ile 4 boyut arasındaki simplekslerin 2 boyuttaki izdüşümü Şekil 1.7'de verilmiştir.

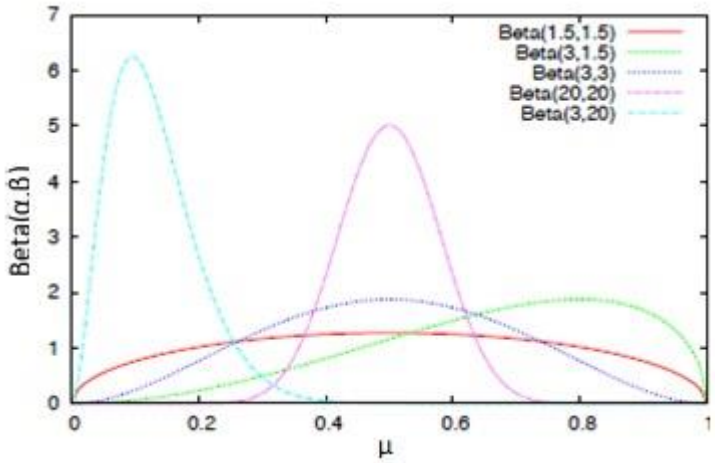


Şekil 1.7. Simplekslerin iki boyuttaki izdüşümü (Hanson, 1994)

Beta dağılımı binom için önsel eşlenik iken, Dirichlet çok terimli dağılımlar için önsel eşleniktir. Dirichlet dağılımı için öncelikle Beta dağılımın verilmesi gerekmektedir. Varsayalım ki rastgele bir değişken olan $\mu(0\leq\mu\leq1)$ $\alpha(\alpha>0)$ ve $\beta(\beta>0)$ hiperparametreleri ile birlikte Beta dağılımına sahip olsun. Bu durumda μ değişkeni sürekli bir dağılıma sahip olup, olasılık yoğunluk fonksiyonu Eşitlik (1.3) ile verilmiştir.

$$p(\mu \mid \alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1} \tag{1.3}$$

Beta dağılımının α ve β'nın çeşitli değerleri için almış olduğu değerler Şekil 1.8'de verilmiştir.



Şekil 1.8. Beta dağılımı (Hockenmaier, 2018)

Çok terimli dağılımlardaki önsel olasılık Eşitlik (1.4)'te verilmiştir.

$$p(\mu \mid \alpha) \propto \prod_{t=1}^{T} \mu_t^{\alpha_t - 1} \tag{1.4}$$

$0 \leq \mu \leq 1$ ve $\sum_{t=1} \mu_t = 1$ dağılımın kısıtları, $\alpha_1, \alpha_{2,\ldots,} \alpha_T$ dağılımın parametreleridir, $\alpha = (\alpha_1, \alpha_{2,\ldots,} \alpha_T)^T$ şeklinde temsil edilmektedir. Dağılımdaki toplama kısıtından ötürü μ_t T-1 boyutlu simpleks ile sınırlandırılmıştır. Bu dağılımın normalize edilmiş hali Eşitlik (1.5) ile verilmiştir.

$$\mathrm{Dir}(\mu \mid \alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)..\Gamma(\alpha_T)} \prod_{t=1} \mu_t^{\alpha_t - 1} \tag{1.5}$$

Burada $\Gamma(\alpha_0)$ Gama fonksiyonunu temsil etmektedir. $\Gamma(x)$ şeklindeki Gama fonksiyonu Eşitlik (1.6)'da verilmiştir.

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt \tag{1.6}$$

α=100 için Dirichlet dağılımını hesaplayan R kodu Şekil 1.9'da, α'nın çeşitli değerleri için elde edilen Dirichlet dağılımının grafiksel temsili ise Şekil 1.10'da verilmiştir.

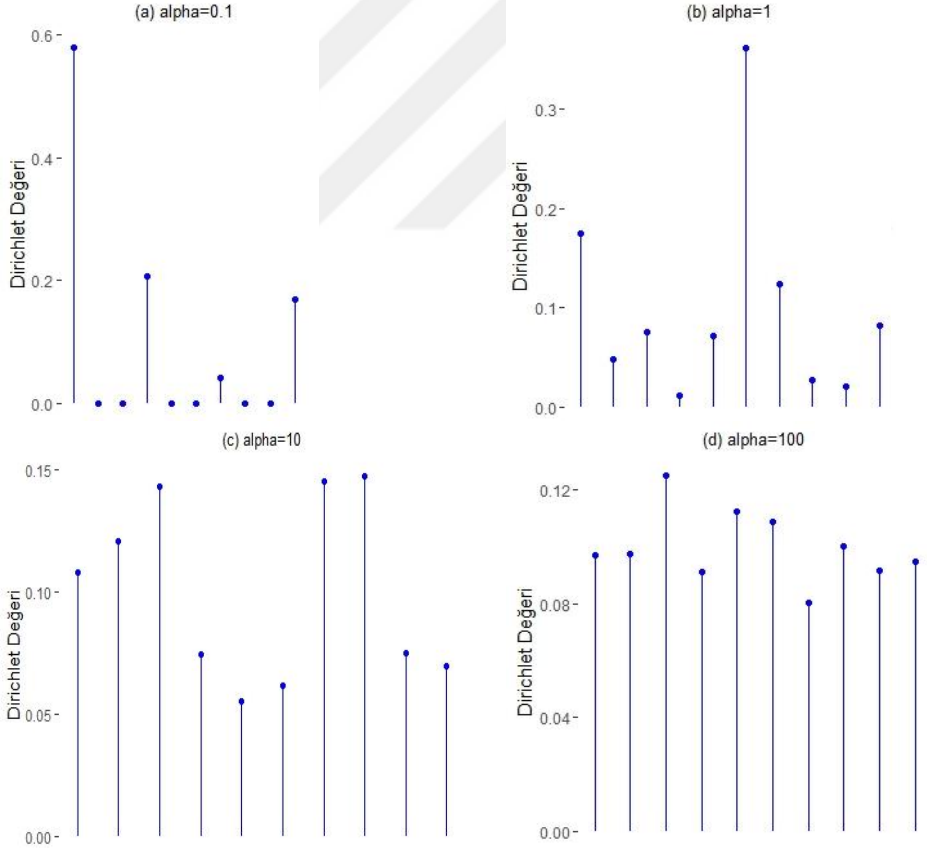
DirichletHesaplama.R

```r
library(ggplot2)
a=c(100,100,100,100,100,100,100,100,100,100)
n<-length(a)
z<-rep(0,n-1)
for(j in 1:(n - 1)) {
  z[j] <- rbeta(1, sum(a[1:j]), a[j + 1])
}
y_axis <- rep(0, n)
y_axis[1] <- prod(z)
for(i in 2:(n - 1)) {
  y_axis[i] <- (1 - z[i - 1]) * prod(z[i:(n - 1)])
}
y_axis[n] <- 1 - sum(y_axis[1:(n - 1)])

data=data.frame(x=1:length(a), y=y_axis)

# Plot
ggplot(data, aes(x=x, y=y)) +
  geom_point(color="blue") +
  geom_segment(aes(x=x, xend=x, y=0, yend=y),color = "blue")+
  labs(x="",y="Dirichlet Değeri")+
  ggtitle("(d) alpha=100") +
  theme(plot.title = element_text(hjust = 0.5,size=10))+
  theme(plot.background = element_rect(fill = 'white', colour = 'white'))+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),panel.background = element_blank())+
  coord_fixed(ratio = 80)+
  theme(axis.ticks.x=element_blank(),axis.text.x=element_blank())
```

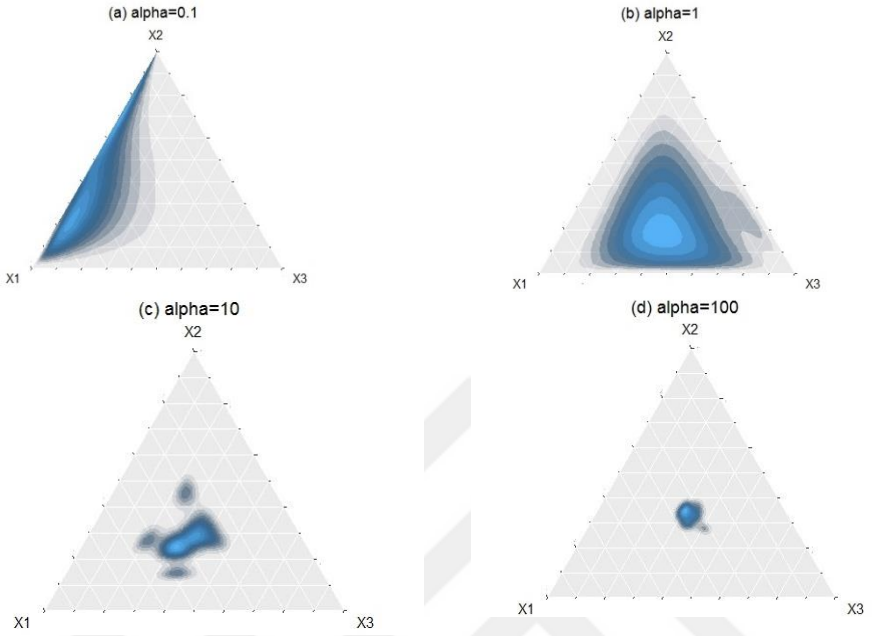
Şekil 1.9. Verilen α değeri için Dirichlet dağılımını veren R kodu



Şekil 1.10. α'nın çeşitli değerleri için elde edilen Dirichlet dağılımının grafiksel temsili

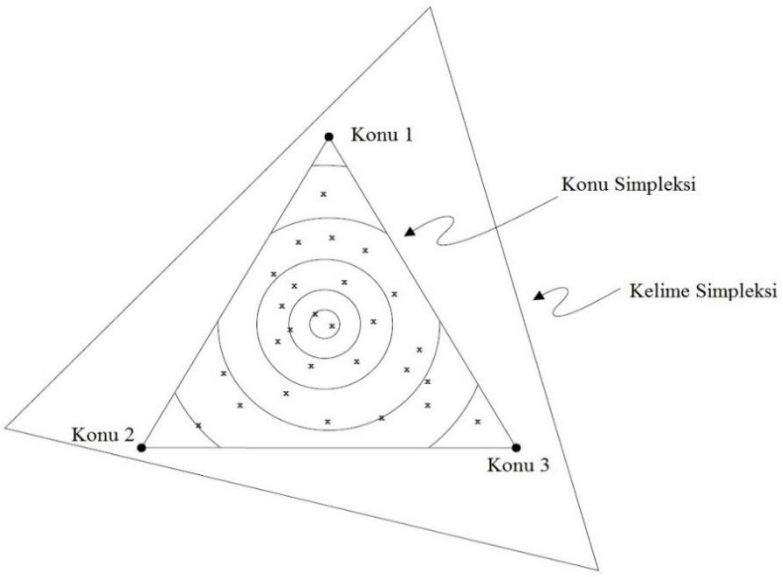
Şekil 1.10 LDA için değerlendirildiğinde α değerinin 1,0'dan küçük olması az sayıda konunun birleşiminden oluşan dokümanların üretildiğini göstermektedir. Simpleks

üzerinde ise yoğunluk daha çok köşelerde olmaktadır. Eğer α değeri 1,0'dan büyükse daha çok konunun birleşiminden oluşan dokümanlar üretilmektedir. Şekil 1.10'daki α ile elde edilen noktaların simpleksteki görünümleri Şekil 1.11'de verilmiştir.



Şekil 1.11. α'nın çeşitli değerleri için elde edilen Dirichlet dağılımının simpleks ile temsili

Simpleksler LDA'nın geometrik temsilinde de Şekil 1.12'de gösterildiği gibi kullanılmaktadır.



Şekil 1.12. LDA'nın simpleks üzerinden geometrik temsili (Blei ve diğ., 2003)

Şekil 1.12, 3 kelimenin temsil edildiği kelime simpleksine 3 konunun temsil edildiği konu simpleksinin yerleştirilmesini temsil etmektedir. Her iki simpleksin köşelerinde olasılık değeri 1'e eşit olmaktadır. Unigramların karışımı modelinde her doküman için konulardan bir tanesi seçilmektedir; yani konu simpleksinin köşelerinden bir tanesi rastgele seçilmektedir. Dokümanı oluşturan tüm kelimeler ise bu konu ile ilişkili dağılımdan gelmektedir. pLSA'da dokümanı oluşturan her kelime için ilgili konu dokümana özgü konu dağılımına göre belirlenmektedir. Bu da konu simpleksi içerisinde yer alan bir noktaya göre (Şekil 1.12'de x ile belirtilmektedir) kelimelerin konu dağılımının belirlenmesi anlamına gelmektedir. LDA'da ise dokümanda yer alan her bir kelime rastgele belirlenen bir parametreye göre dağılımı belirlenen konulardan birinin rastgele seçilmesi ile üretilmektedir. Bu parametre ise konu simpleksindeki dağılımlardan her doküman için bir kere örneklenmektedir. Dokümanlar simplekte kontur ile temsil edilmektedir.

1.1.2. Gibbs örnekleme

Gibbs örnekleme; özellikle Bayesian çıkarımındaki sonsal dağılım için kullanılan, dağılımlar ile ilgili bilgi veren popüler metot olan MCMC örneklemenin özel bir türüdür. Buradaki Monte Carlo, dağılımdan rastgele örnekler alarak dağılımın özelliklerini incelemeyi sağlamaktadır. Mesela bir normal dağılımın otalamasını dağılımın eşitliğinden bulmak yerine Monte Carlo ile rastgele örneklerden büyük bir küme kurulup bu kümenin ortalaması hesaplanır. Oluşan bu yeni kümenin ortalamasını hesaplamak dağılımın formülü üzerinden hesaplamaya göre daha kolaydır. MCMC'nin Markov chain özelliği ise rastgele örneklerin özel sıralı bir süreç ile örneklenmesini ifade etmektedir. Her rastgele örnek bir sonraki rastgele örneği üretmek için kullanılmaktadır. Aşağıdaki örnek MCMC'yi basit bir örnekleme yöntemi olan Metropolis algoritması ile anlatmaktadır.

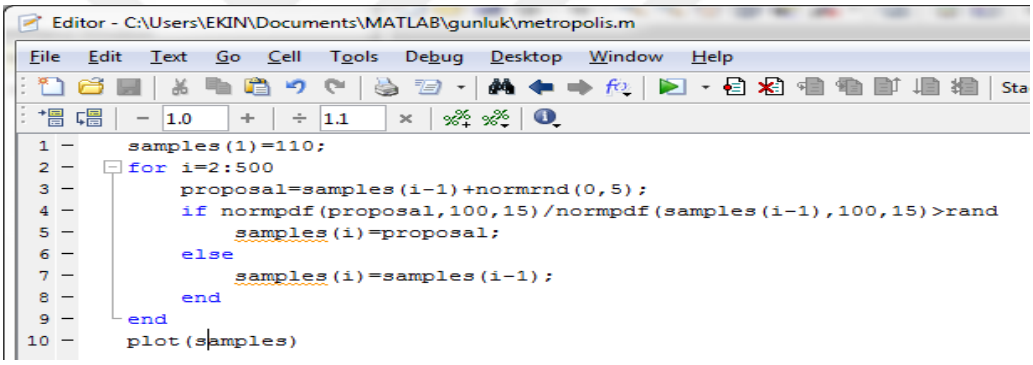
Diyelim ki, öğrencilerin test skorlarının ortalaması öğrenilmek isteniyor, dolayısıyla ortalama bilinmiyor. Skorlar normal dağılım göstermektedir ve standart sapma 15'tir. Bir öğrencinin notu da 100 olarak gözlemlenmiştir. MCMC ile hedef dağılımdan örnekler seçilir. Bu durumda sonsal tek bir gözlem değeri verilmişken (100) popülasyon ortalamasının her bir mümkün değeri için olasılık değerini temsil eder. Metropolis Hastings algoritmasının adımları Şekil 1.13'te verilmiştir.

Algoritma 1: Metropolis Hastings Algoritması

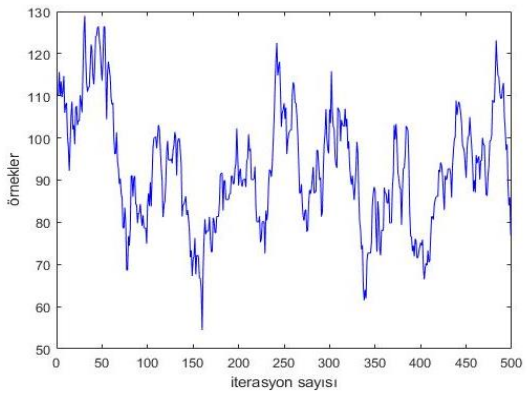
1. Uygun bir başlangıç tahmini ile başla
2. MCMC bu tahminden yeni örnekler zinciri üretir. Bunu yaparken de son örneğe gürültü ekler. Rastgele olan bu gürültü de normal dağılımdan üretilir.
3. Yeni örnek ile bu örneğin üretildiği örnek karşılaştırılır.
4. Eğer yeni örneğin sonsalı üretildiği örneğin sonsalından büyükse yeni örnek kabul edilir.
5. Eğer büyük değilse kabul ya da ret rastgele yapılır.
6. Eğer örnek kabul edilirse bu örnek MCMC zincirindeki bir sonraki örnektir. Eğer kabul edilmezse de bir önceki değer yeni örnek olarak aynen kullanılır.
7. Böylece MCMC'nin bir iterasyonu tamamlanmış olur.
8. Yeterince örnek üretilene kadar bu süreç devam eder.

Şekil 1.13. Metropolis Hastings algoritmasına ait sözde kod (van Ravenzwaaij ve diğ., 2018)

Algoritmanın Matlab kodu ile yazılmış hali Şekil 1.14'te elde edilen grafik ise Şekil 1.15'te verilmiştir.



Şekil 1.14. Metropolis Hastings algoritmasının Matlab'da yazılmış kodu



Şekil 1.15. Metropolis Hastings algoritması ile elde edilen örnekler

Eğer örnekleme yapılacak problemdeki parametreler arasında güçlü bir ilişki varsa Metropolis Hastings yetersiz kalmaktadır. Bu durumda Gibbs örnekleme işin içine

girmektedir. Gibbs örneklemenin algoritması iki değişkenli bir sonsal dağılım üzerinden Şekil 1.16'da verilmiştir.

Algoritma 2: Gibbs Örnekleme Algoritması

1. İki değişkenimiz b' ve Y olsun. Sırasıyla değerleri 1 ve 0,5 olsun. Sonsal dağılım bu iki değişkenin tüm kombinasyonları üzerinden tanımlanmaktadır.
2. b' için Metropolis'te olduğu gibi yeni değer üretilir. Bu değer de 1,2 olsun.
3. Eğer verilen Y değeri için b' nün yeni değeri popülasyonun dağılımı için daha uygunsa bu değer kabul edilir. Bu durumda Y=0,5, b'=1,2 yi kabul etmiş olur.
4. Y için yeni değer üretilir. Bu değerin üretilmesi için de bir dağılıma ihtiyaç vardır. Y'nin yeni değeri 0,6 olarak üretilmiş olsun.
5. Yani Y değerinin kabulü de aynı b' değerinin kabulü gibi yapılır. Belli bir b' değeri için yeni Y değeri daha uygunsa bu değer kabul edilir. Eğer kabul edilmezse de Y aynı kalır.
6. Böylece Gibbs'in bir iterasyonu tamamlanmış olur.
7. Adım 2'ye dönülüp bir sonraki iterasyona geçilir.

Şekil 1.16. Gibbs örnekleme algoritmasına ait kod

1.1.3. Collapsed Gibbs örnekleme

İlk kez 2004 yılında Griffiths ve Steyvers tarafından tanıtılan Collapsed Gibbs örnekleme, Gibbs örneklemenin standart bir gerçekleştirimidir. CGS ile model parametreleri olan θ ve φ dışarlanmakta, kelimelere konu atamada kullanılan parametre z ilk olarak bir dokümandaki her kelime için daha sonra koleksiyonundaki diğer dokümanlarda yer alan her kelime için iteratif olarak yeniden örneklenmektedir. Standart Gibbs örneklemede bir dokümandaki ya da doküman koleksiyonundaki kelimelerin konulara atanmasında diğer kelimeler dikkate alınmazken, CGS'de model parametreleri dışarlandığı için bu kelimeler model parametrelerinin vekili olarak kullanılmaktadır. CGS algoritması aşağıda yer alan örnek üzerinden adım adım anlatılmaktadır.

Birinci adımda rastgele bir yorum seçilerek işe başlanmaktadır. Varsayalım ki, seçilen yorum 5 kelimeden oluşsun ve konu sayısı da 3 olarak belirlenmiş olsun. Örnek yorum Şekil 1.17'de verilmiştir.

biscuit	jam	place	corncake	drinks

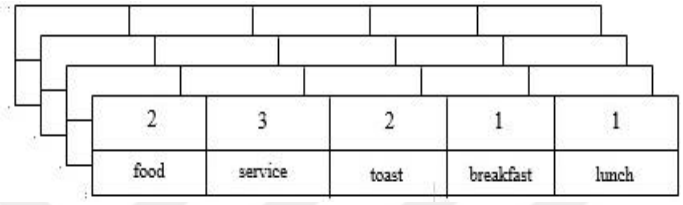
Şekil 1.17. Örnek yorum

İkinci adımda yorumdaki kelimeler konulara rastgele atanmaktadır. Şekil 1.18'de görüldüğü üzere birinci ve üçüncü konuya ikişer kelime, ikinci konuya bir kelime atanmıştır.

3	2	1	3	1
biscuit	jam	place	corncake	drinks

Şekil 1.18. Kelimelerin konulara rastgele atanması

Şekil 1.18'deki işlem koleksiyondaki tüm yorumlara uygulanır ve Şekil 1.19 elde edilir.



Şekil 1.19. Koleksiyondaki tüm kelimelerin konulara rastgele atanması

Koleksiyonda yer alan her yorum için rastgele konu atama işlemi tamamlandıktan sonra yorum bazında istatistikler yani yerel istatistikler çıkartılır. Yerel istatistik; yorumda her konuya kaçar tane kelime atandığını vermektedir. Şekil 1.17'deki yorum için yerel istatistikler Tablo 1.1'de verilmiştir.

Tablo 1.1. Şekil 1.17'deki yoruma ait yerel istatistikler

Konu$_1$	**Konu$_2$**	**Konu$_3$**
2	1	2

Yerel istatistikler koleksiyondaki tüm yorumlar için elde edildikten sonra global istatistikler koleksiyondan çıkartılır. Yani tüm koleksiyon için her kelimenin her konuya kaç kere atandığı hesaplanır. Temsili global istatistikler Tablo 1.2'de verilmiştir.

Tablo 1.2. Koleksiyondan elde edilen temsili global istatistikler

	Konu$_1$	**Konu$_2$**	**Konu$_3$**
biscuit	1	0	35
jam	10	8	2
place	42	1	0
corncake	0	0	20
drinks	50	0	1
…	…	…	…

Tüm istatistiksel bilgileri elde ettikten sonra koleksiyonu oluşturan her yorumdaki her kelime için yeniden konu ataması adımı işletilir. Bu işlem tüm koleksiyon üzerinde iteratif olarak gerçekleştirilir. Şekil 1.20'de "jam" kelimesi için yeni konu ataması yapılması örneklenmiştir. İlk olarak "jam" kelimesi için mevcut atama kaldırılır yani "jam" için hangi konuya atandığı bir soru işaretidir. Dolayısıyla istatistiklerin de güncellenmesi gerekmektedir. Bu durumda ilgili yorumda $Konu_2$'ye atanan kelime sayısı 0'a düşmektedir, yine "jam" kelimesi için global istatistiklerde $Konu_2$'ye atanan kelime sayısı 8'den 7'ye düşmektedir. Yoruma ait güncellenmiş yerel istatistikler Tablo 1.3'te, global istatistikler ise Tablo 1.4'te verilmiştir.

3	?	1	3	1
biscuit	jam	place	corncake	drinks

Şekil 1.20. "jam" kelimesi için yeni konu ataması

Tablo 1.3. Şekil 1.17'deki yoruma ait güncellenmiş yerel istatistikler

$Konu_1$	**$Konu_2$**	**$Konu_3$**
2	**0**	2

Tablo 1.4. Güncellenmiş global istatistikler

	$Konu_1$	**$Konu_2$**	**$Konu_3$**
biscuit	1	0	35
jam	10	**7**	2
place	42	1	0
corncake	0	0	20
drinks	50	0	1
…	…	…	…

"jam" kelimesi için yeni konuya atanma olasılığının hesaplanmasında iki faktör rol almaktadır. Bunlardan ilki; mevcut yorumun konular ile hangi oranlarda ilişkili olduğudur. Bunun için yorumu oluşturan kelimelere (jam kelimesi dışındakilere) bakıp bir konunun ne sıklıkta geçtiği Eşitlik (1.7)'ye göre hesaplanmaktadır.

$$\theta_{i,k} = \frac{n_{i,k} + \alpha}{N_i - 1 + K\alpha} \tag{1.7}$$

Burada $n_{i,k}$ i. yorumda k. konuya atanan kelime sayısını göstermektedir. α Dirichlet parametresidir. N_i i. yorumda yer alan mevcut kelime sayısıdır ve 1 çıkartılmasının

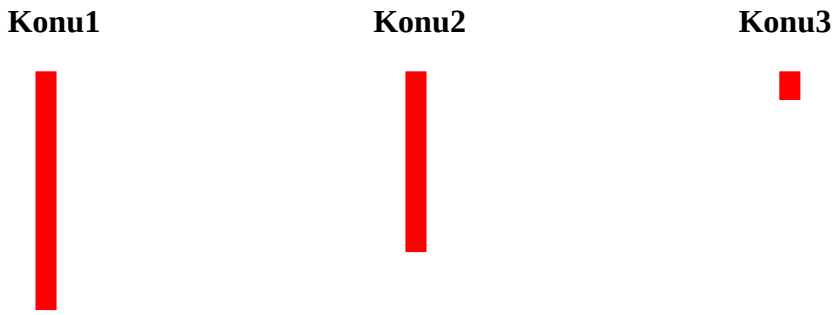
sebebi jam kelimesinin yok sayılmasıdır. K ise konu sayısıdır. Eşitlik (1.6)'ya göre konuların Şekil 1.17'deki yorum ile olan ilişkisi Şekil 1.21'da verilmiştir.

Şekil 1.21. Mevcut yorumun her konu ile olan ilişkisi

"jam" kelimesi için yeni konuya atanma olasılığının hesaplanmasındaki ikinci faktör ise "jam" kelimesinin konular ile ne kadar "ilişkili olduğunun" hesaplanmasıdır. Bu hesaplama adımında ise global istatistiklerden yararlanılmış olup, kelime verilen konu altında ne kadar kullanılmış bilgisi gerekmektedir. Hesaplama adımı Eşitlik (1.8)'de verilmiştir.

$$\varphi_{jam,k} = \frac{n_{jam,k} + \beta}{\sum_{w \in V} n_{w,k} + V\beta} \tag{1.8}$$

$n_{jam,k}$ ile jam kelimesinin k. konuya tüm koleksiyonda kaç kere atandığı bulunmaktadır. β Dirichlet parametresidir. $n_{w,k}$ k. konunun tüm koleksiyonda kaç kere kullanıldığını, V ise sabit sözlükte bulunan toplam kelime sayısını göstermektedir. Eşitlik (1.8)'e göre mevcut kelimenin her konu ile olan ilişkisi Şekil 1.22'de verilmiştir.

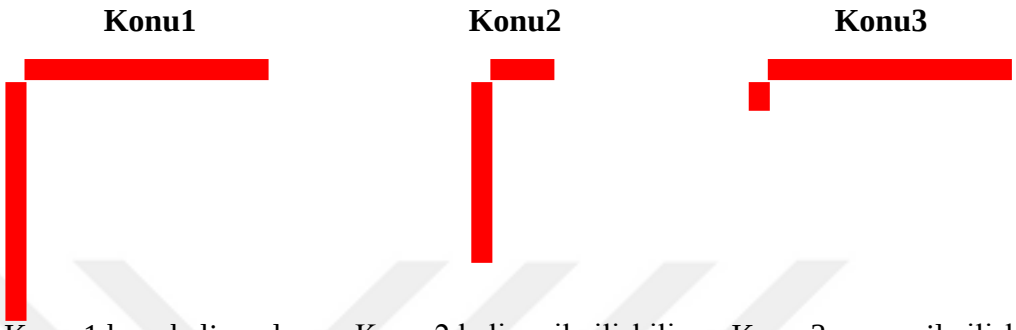


Şekil 1.22. Mevcut kelimenin her konu ile olan ilişkisi

Eşitlik (1.7) ve Eşitlik (1.8)'e dayanarak jam kelimesinin yeni konuya atanmasında Eşitlik (1.9)'dan yararlanılmıştır.

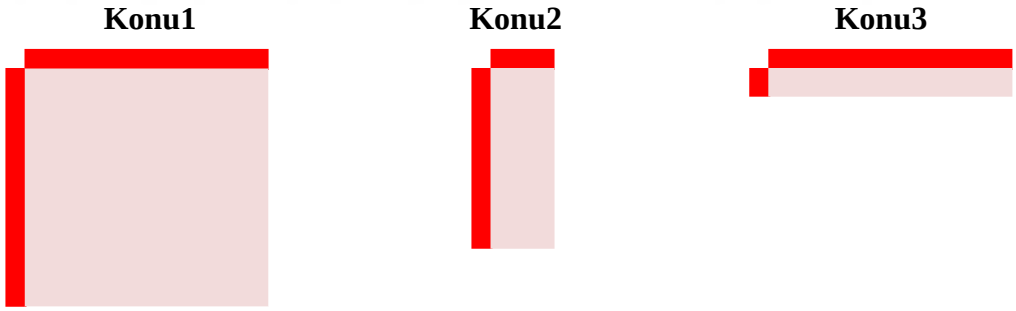
$$p(z_{jam} = k \mid w = jam, i, \alpha, \beta, .) = \frac{n_{i,k} + \alpha}{N_i - 1 + K\alpha} \frac{n_{jam,k} + \beta}{\sum_{w \in V} n_{w,k} + V\beta} \quad (1.9)$$

"jam" kelimesinin yeni konuya atanmasının şekil üzerinden gösterimi ise Şekil 1.21 ve Şekil 1.22'ye dayalı olarak Şekil 1.23'te verilmiştir.



Şekil 1.23. Konuların kelime ve yorum ile olan ilişkisi

Eşitlik (1.8)'in geometrik yorumu Şekil 1.24'de verilmiştir.



Şekil 1.24. "jam" kelimesi için yeni konu belirlemenin geometrik yorumu

Eşitlik (1.9)'dan elde edilen Şekil 1.24'teki taralı alanlar incelendiğinde jam kelimesinin ilgili yorum için yeni konusunun Konu 1 olduğu görülmektedir. Bu durum Şekil 1.25 ile gösterilmiştir.

3	1	1	3	1
biscuit	jam	place	corncake	drinks

Şekil 1.25. "jam" kelimesinin CGS'ye göre yeni konuya atanması

“jam” kelimesinin yeni konu atamasın yapıldıktan sonra yerel ve global istatistiklerde Tablo 1.5 ve Tablo 1.6’daki gibi güncellenmiştir.

Tablo 1.5. Şekil 1.17’deki yoruma ait CGS sonrası güncellenmiş yerel istatistikler

Konu$_1$	**Konu$_2$**	**Konu$_3$**
3	0	2

Tablo 1.6. CGS sonrası güncellenmiş global istatistikler

	Konu$_1$	**Konu$_2$**	**Konu$_3$**
biscuit	1	0	35
jam	**11**	7	2
place	42	1	0
corncake	0	0	20
drinks	50	0	1
…	…	…	…

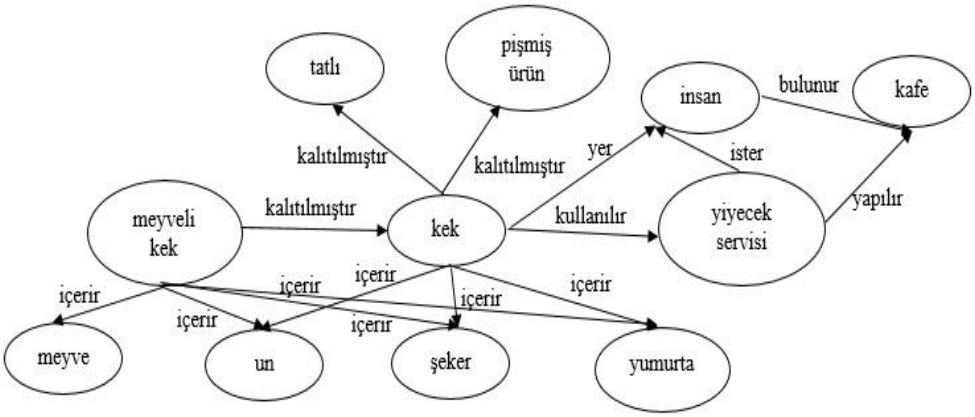
CGS koleksiyondaki tüm yorumlarda yer alan bütün kelimelere uygulandıktan sonra birinci iterasyonunu tamamlamış olur ve belirlenen iterasyon sayısı boyunca tekrar tekrar uygulanır.

2. ANLAMSAL AĞLAR

Doğal dil işlemede (DDİ), kelime anlamı belirsizliği düşük performansa ve etiketleme adımında modellerin başarımını etkileyen problemlere neden olmaktadır (Sanderson, 1994). Bir örnek üzerinde bu durumu inceleyecek olursak, "Türk mutfağı çok lezzetli." cümlesindeki "mutfak" "yiyecek kültürünün tamamı" mı yoksa "dergi ismi" midir? Bu soruya cevap vermek için anlamsal ipuçlarına ihtiyaç duyulmaktadır. Bu cümledeki anlamsal ipuçları "Türk" ve "lezzetli" kelimeleridir. Dolayısıyla bu cümledeki "mutfak" "yiyecek kültürünün tamamı"dır.

Aynı şekilde bir dokümandaki adlandırılmış varlığı tanımlayan ifadeleri bilgi tabanındaki ilgili varlık ile eşleme de son yıllarda üzerine çalışılan ve varlık bağlama olarak tanımlanan bir DDİ problemidir (Rao ve diğ., 2013). Bir metin içerisinde geçen "THK" ile "Türk Hava Kurumu" ifadelerinin aynı varlığı işaret ettiğini tespit etmek oldukça önemlidir. DDİ'de sıklıkla karşılaşılan bu tür problemlerin çözümünde ise güçlü bir yaklaşıma ihtiyaç duyulmaktadır. Anlamsal bilgiyi temsil etmede kullanılan anlamsal ağlar bu amaçla kullanılan önemli bir bilgi kaynağıdır.

İlk olarak Collins ve Quillian'ın klasik ağ teorisi ile ortaya çıkan anlamsal ağlar, doğal dildeki her türlü bilgiyi, etiketli ve yönlü kenarlar ile birbirine bağlı düğümler ile anlamsal bir graf üzerinden temsil etmeyi amaçlamaktadır (Steyvers ve Tenenbaum, 2005; Lehman, 1992). Bu ağlar, bilginin temsili ile sözcüklerin ardındaki gizli anlamların daha iyi anlaşılmasını, bu gizli anlamlar üzerinden çıkarım yapılmasını dolayısıyla da doğal dil uygulamalarının performansının arttırılmasını sağlamaktadır. Şekil 2.1'de restoran yorumlarında sıklıkla geçen "kek" kelimesi için örnek bir anlamsal ağ verilmiştir.



Şekil 2.1. "kek" kelimesi için örnek bir anlamsal ağ

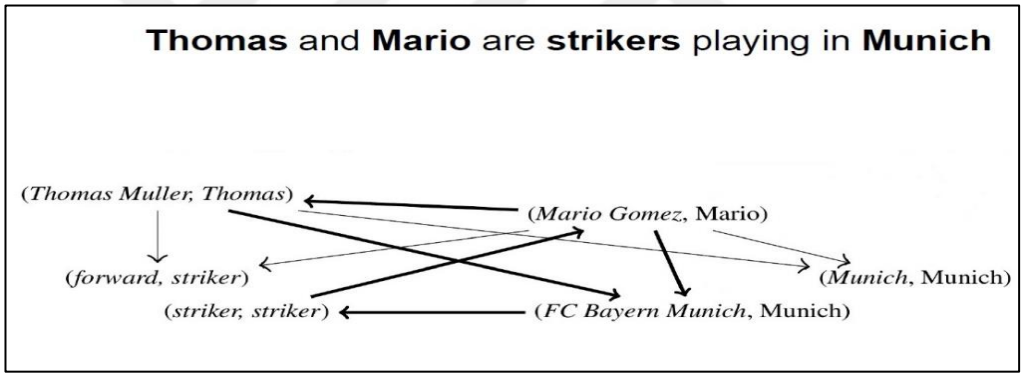
Şekil 2.1'de görüldüğü üzere nesneleri temsil eden düğümler elipsler ile gösterilmekte iken yönlü kenarlar bu nesneler arasındaki ikili ilişkiler ile etiketlenmiştir. Nesneler arasındaki ilişkiler bilgiyi organize etmek için temel oluşturmaktadır. Anlamsal ağlarda en sık kullanılan ilişki "kalıtılmış" ilişkisi olup, nesneler arasındaki ilişkiler her zaman Şekil 2.1'deki gibi somut olmak zorunda değildir.

Anlamsal ağlar literatürde de oldukça geniş yer kaplayan bir konu olarak karşımıza çıkmaktadır. Pek çok anlamsal ağın da temelini oluşturan WordNet 1998 yılında İngilizce için geliştirilmiş büyük kapsamlı bir anlamsal ağdır (Miller, 1995). WordNet'te isimler, fiiller, sıfatlar ve zarflar, her biri ayrı bir kavram ifade eden kavramsal eşanlamlılar kümeleri halinde gruplandırılır. Her kavramsal eşanlamlılar kümesi ise birbirlerine kavramsal-anlamsal ve sözcüksel ilişkiler ile bağlıdırlar. Bu ilişkilerden bazıları; alt sınıf, üst sınıf ve kalıtılmış olma şeklindedir. EuroWordNet, çeşitli Avrupa dillerindeki anlamsal ağları kullanarak geliştirilmiş çok dilli bir anlamsal ağdır (Vossen, 1998). EuroWordNet'in içerdiği diller; Almanca, İspanyolca, İtalyanca, İngilizce, Flemenkçe, Fransızca, Çekçe ve Estçe'dir. Bir diğer çok dilli anlamsal ağ ise ConcepNet'tir. ConceptNet; WordNet gibi anlamsal ağlar, Wiktionary gibi sözlükler, belli bir amaç için geliştirilen oyunlar gibi çeşitli kaynaklardan elde ettiği veriyi kullanarak anlamsal ağı oluşturmaktadır (Speer ve diğ., 2017). Çok dilli anlamsal ağ BabelNet 284 farklı dili desteklemekle birlikte yaklaşık 16 milyon giriş içermektedir. BabelNet; terimler, tanımlar, görseller, çeviriler, istatistikler, bibliyografi ve anlamsal ağı görüntülemek için kullanıcılarına web arayüzü sunmaktadır (URL-1, 2018). Bu tez kapsamında ise BabelNet temelli bir anlamsal ağ olan ve cümleler üzerinde çalışan Babelfy kullanılmıştır. Babelfy'ın kelime anlamı

belirsizliği giderme, varlık bağlama görevleri dışında öbekleri de eldeki yorumlardan çıkartabilmesi tercih nedenlerinden biri olmasında önemli rol oynamıştır.

2.1. Babelfy

Sango ve Tagalogca gibi birkaç milyon insan tarafından konuşulan diller ile İngilizce ve Türkçe gibi yüz milyonlarca insan tarafından konuşulan diller de dahil olmak üzere 284 farklı dili kapsayan Babelfy, varlık bağlama ve kelime anlamı belirginleştirme problemlerine çözüm öneren birleşik, çok dilli ve graf tabanlı anlamsal bir yaklaşımdır (Moro ve diğ., 2014b; URL-2, 2018). Babelfy bu amaçla geliştirilmiş ilk yaklaşım olup, kavramlar ve adlandırılmış varlıklar arasındaki anlamsal ilişkiyi BabelNet altyapısını kullanarak çıkarmaktadır (Navigli ve Ponzetto, 2010; URL-1, 2018). Babelfy'ın kavram ve adlandırılmış varlıklar arasında çıkarmış olduğu ilişki Şekil 2.2 ile verilmiştir.



Şekil 2.2. Kavram ve adlandırılmış varlıklar arasındaki ilişki (Navigli, 2018)

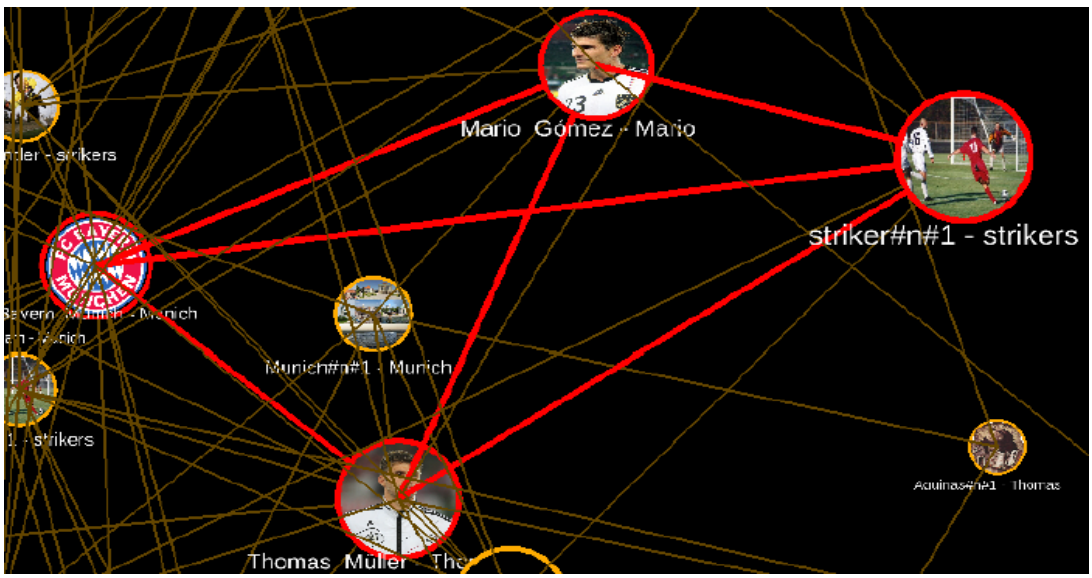
Kavramlar, her biri benzersiz bir anlam taşıyan bilgi birimleri olarak tanımlanmaktadır (Moro ve diğ., 2014a). Kavramı bir örnek kelime üzerinden açıklayacak olursak "Restoran" kelimesini ele alabiliriz. "Restoran" kelimesi için kavramlar "müşteri" ve "servis" olarak elde edilmektedir. Adlandırılmış varlıklar ise kişi, organizasyon veya konum gibi gerçek dünyadaki nesnelerin isimlerine karşılık gelmektedir. Örneğin; Kocaeli Üniversitesi, Türkiye Büyük Millet Meclisi adlandırılmış varlıklara örnek olarak verilebilir.

BabelNet ise on üç farklı veri kaynağından (WordNet, Wikipedia, OmegaWiki, Wiktionary, Wikidata, Wikiquote, VerbNet, Microsoft Terminology, GeoNames, ImageNet, FrameNet, WN-Map, Open Multilingual WordNet) sözlüksel ve

ansiklopedik bilgi içeren bir ansiklopedik sözlük ve kavram ve adlandırılmış varlıkları anlamsal ilişkiye göre ilişkilendiren bir anlamsal ağ olarak tanımlanmaktadır (Navigli ve Ponzetto, 2010; Ehrmann ve diğ., 2014).

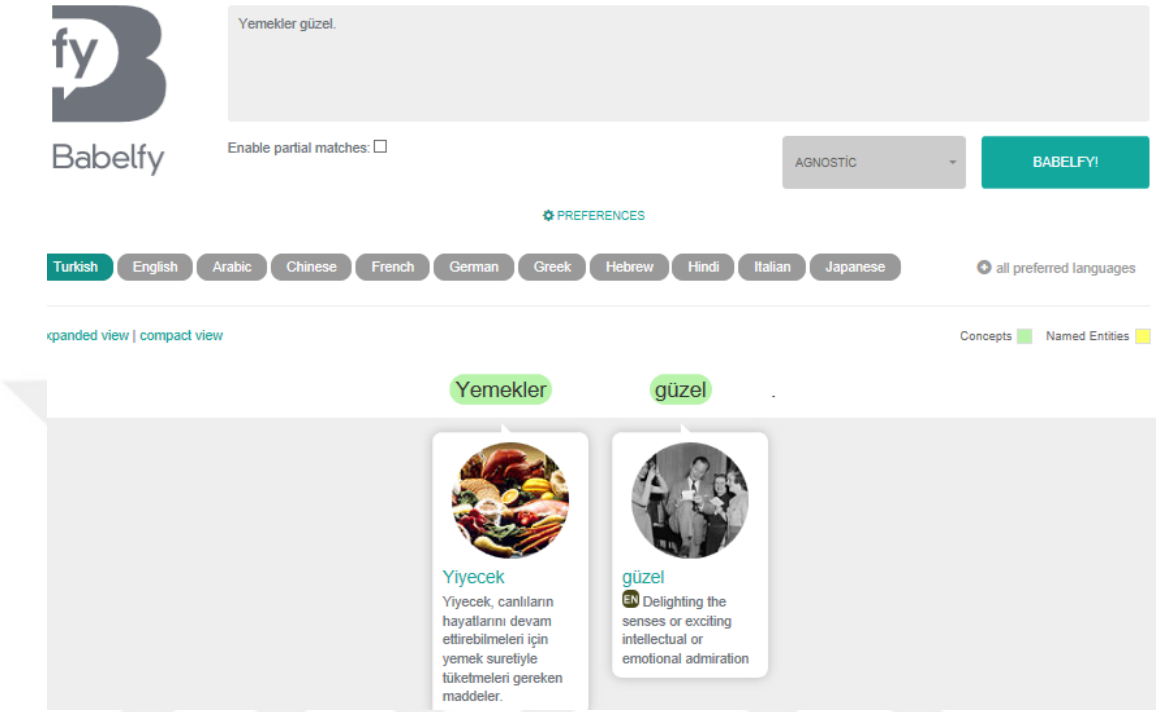
Babelfy, varlık bağlama ve kelime anlamı belirginleştirme görevlerini üç adımda gerçekleştirmektedir (Moro ve diğ., 2014);

- BabelNet kullanılarak ilgili dokümandaki her kavram ve adlandırılmış varlık ilişkili kavram ve adlandırılmış varlıkların kümesi ile ilişkilendirilir. Yani anlamsal imzalar elde edilmiş olur. Bu adım ilgili dokümandan bağımsız olarak bir kere gerçekleştirilir. Mesela Şekil 2.2'de yer alan cümlede "Mario" için elde edilen adlandırılmış varlıklar "Mario Gomez" ve "Mario Basler" şeklindedir.
- Verilen doküman için bilgi tabanında yer alan tüm kelimeler belirlenir, bunlar için aday anlamlar BabelNet kullanılarak çıkartılır. Bu adımda "Mario Gomez" için (forward, striker), (Munich, Munich), (FC Bayern Munich, Munich) ve (striker, striker) aday anlamları, "Mario Basler" için (Munich, Munich) aday anlamı çıkartılır.
- Birinci adımda elde edilen anlamsal imzalar kullanılarak ikinci adımda elde edilen aday anlamlar arasında bağlantı kurulur. Son olarak bu bağlantılar arasından en uygun anlamlılar seçilerek Şekil 2.2'deki bir yoğun alt graf elde edilir. Yoğun grafa ait ağ yapısı Şekil 2.3'te verilmiştir.



Şekil 2.3. Yoğun grafa ait ağ yapısı (Navigli, 2018)

Babelfy ayrıca kendisine verilen metnin dili bilinmiyorsa bile “Agnostic” özelliği ile metnin dilini belirleyip varlık bağlama ve kelime anlamı belirginleştirme görevlerini yerine getirir. Babelfy’ın web arayüzü Şekil 2.4’te verilmiştir.



Şekil 2.4. Babelfy’ın web arayüzü

3. GELİŞTİRİLEN YÖNTEMLER

LDA sadece kelimelerin doküman koleksiyonunda birlikte geçme durumlarını dikkate alırken, dokümanı oluşturan kelimeler arasındaki anlamsal ilişkiyi veya dokümanlar arasındaki anlamsal bilgiyi dikkate almamaktadır (Chang ve diğ., 2009). Bu özellik LDA için bir dezavantajdır; buradan yola çıkılarak anlamsal bilginin LDA ile birlikte kullanımı ile başarılı yöntemlerin geliştirilmesi hedeflenmiştir.

Tez çalışması kapsamında kullanıcı yorumlarından ürün özelliklerinin çıkartılması amacıyla iki farklı konu modeli geliştirilmiştir. Literatürde de kullanıcı yorumlarından ürün özelliklerini çıkarmak amacıyla geliştirilen pek çok farklı konu modeli çalışması bulunmaktadır.

Titov ve McDonald (2008) geliştirdikleri Multi Grain LDA (MG-LDA) ile otel yorumlarından ürün özelliklerini çıkartmak için yerel ve global olmak üzere iki farklı konu tipini ele almışlardır. Yaptıkları varsayıma göre bir dokümandaki sözcük ya yerel konuların karışımından ya da global konuların karışımından örneklenebilmektedir. Lokal konular değerlendirmeye alınan ürün özelliklerinden oluşmakta iken, global konular ürünün genel özelliklerinden oluşmaktadır. Deneysel çalışma için TripAdvisor.com'dan elde ettikleri 27.564 adet yorumu kullanmışlardır. Perceptron tabanlı online öğrenme metodu olan PRanking algoritmasını ürün özelliklerinin değerlendirilmesinde kullanmışlardır. MG-LDA için Gibbs örnekleme algoritmasını 800 iterasyon çalıştırmışlar ve yöntemin, karşılaştırıldığı diğer yöntemleri geride bıraktığı görülmüştür.

Lin ve He (2009) geliştirdikleri tamamen denetimsiz bir konu modeli olan Joint Sentiment/Topic Model (JST) ile sinema yorumları üzerinden ürün özelliklerini ve duygu ifadelerini eş zamanlı bir şekilde çıkarmışlardır. Ayrıca yaptıkları çalışma ile kullanıcı yorumlarını polaritelerine göre de sınıflandırmışlardır. Brody and Elhadad (2010) ürün özelliklerini çıkartmak amacıyla denetimsiz Local LDA yöntemini geliştirmişlerdir. Ortak bilginin kullanımı ile ürün özellikleri için bu özellikleri temsil eden kelimeleri elde etmişlerdir. "value" özelliği için "portions, quality, worth, size,

cheap,…" şeklindeki kelimeler "value" özelliğini temsil etmektedir. Yarı-denetimli bir konu modeli olan Co-LDA ile ürün özellikleri ve duygu ifadeleri eş zamanlı modellenmiş yani sentiment LDA ve topic LDA olmak üzere iki farklı dil modeli oluşturulmuştur. Model CNET'te yer alan kullanıcı yorumları üzerine uygulanmıştır (Wang, 2010). Model LDA ile karşılaştırılmış ve LDA'ya göre üstün olduğu tespit edilmiştir.

Sentence LDA'nın altında yatan temel fikir; aynı cümle içerisinde yer alan kelimelerin aynı konuya ait olduklarıdır (Jo ve Oh, 2011). Sentence LDA ile kullanıcı yorumlarındaki detayları yakalama açısından LDA'ya göre daha başarılı sonuçlar elde edilmiştir. Kullanıcı yorumlarındaki ürün özelliği, duygu ifadesi çiftlerini yakalayabilmek için ise Sentence LDA geliştirilerek Aspect Sentiment Unification Model isimli konu modeli Jo ve Oh (2011) tarafından önerilmiştir. Bu model bir duygu ifadesi ile yakından ilişkili önemli ürün özelliklerini yakalama açısından temel yöntemlere göre doğruluk değerlendirmesi açısından oldukça üstünlük sağlamıştır. Kısıtlamalı LDA ilk olarak Zhai ve diğ. (2011) tarafından önerilmiştir. Bu model ile must-link (ML) ve cannot-link (CL) kısıtları kullanılarak ürün özelliklerinin konular altında kümelenmesi aşamasında LDA'nın performansının yükseltilmesi amaçlanmıştır. ML, aynı konu altında bulunması gereken ürün özellikleri çiftlerini tanımlarken; CL aynı konu altında bulunmaması gereken ürün özellikleri çiftlerini tanımlamaktadır.

Xianghua ve diğ. (2013) Çince yorumların bulunduğu MSA-COSRs veri kümesinden LDA ile global konuları, kayan pencereler ile lokal konuları ve ilgili duygu ifadelerini çıkartmışlardır. Bagheri ve diğ. (2014) LDA'nın temel yaklaşımı olan kelime torbası yerine bir ürün için o ürünün özelliklerinin cümlede Markov zinciri oluşturduğu varsayımını dikkate alan Aspect Detection Model is based on Latent Dirichlet Allocation (ADM-LDA)'ı önermişlerdir. ADM-LDA ile ayrıca eşdizim şeklindeki ürün özellikleri de çıkartılabilmektedir. Wang ve diğ (2014) Fine-grained Label LDA (FL-LDA) ve Unified Fine-grained Label LDA (UFL-LDA) olmak üzere iki tane yeni yarı denetimli konu modeli geliştirmiştir. FL-LDA'da ürün özelliklerini içeren çekirdek sözlüğün kullanımı ile ürün özellikleri kullanıcı yorumlarından çıkarılmakta iken, UFL-LDA ile etiketsiz dokümanlardan yüksek frekanslı ürün özellikleri çıkartılmaktadır. Aynı cümle içerisinde yer alan kelimelerin aynı konu altında yer

alacağı fikrini savunan diğer bir model Appraisal Expression Patterns LDA (AEP-LDA)'dır (Zheng ve diğ., 2014). Yin ve diğ. (2014)'te geliştirdikleri Dependency Topic Affects Sentiment LDA (DTAS)'da kelime torbası yaklaşımı yerine Markov zincirini kullanmışlardır. Lifelong Topic Model (LTM) yalnızca ML; topic modeling with Automatically generated Must-links and Cannot-links (AMC) hem ML hem de CL kısıtını önbilgi olarak kullanan LDA tabanlı yöntemlerdir (Chen ve Liu, (2014a, 2014b)).

Chen ve diğ. (2015) LDA tabanlı etkileşimli bir duygu görselleştirme sistemi ile otel yorumlarını özetleyen bir sistem tasarlamışlardır. Tasarladıkları sistem duygu-ürün özelliği çiftlerini verirken aynı zamanda bu çift için polarite hesabını da gerçeklemektedir. LDA bu çalışmada kullanıcı yorumlarını konular altında kümeleyerek anlaşılırlık sağlanması amacıyla kullanılmış olup, Tayland'daki bir otele ait 10,000 kullanıcı yorumundan 36 adet konu çıkartılmıştır.

Lee ve diğ. (2016) tüketici yorumlarından değerli bilgileri elde etmek amacıyla algısal bir harita ve radar diyagramı oluşturarak farklı firmalara ait ürünlerin karşılaştırılmasını amaçlamışlar ve bu amaçla Mining Conceptual Map isimli 4 aşamalı metodu tasarlamışlardır. Sanal dokümanlar oluşturarak oluşturulan dokümanlar üzerinden ağırlıklı LDA ile ürün özelliklerinin çıkartılması amaçlanmıştır. Poria ve diğ. (2016) geliştirdikleri Sentic-LDA'da, LDA'yı anlamsal benzerlik ve kurallar kullanarak iyileştirmişlerdir. Model kelimeler arasındaki anlamsal ilişkiyi skorlayarak bu değeri LDA'ya dahil etmektedir. Böylece anlamsal olarak benzer kelimeler aynı konu altında kümelenmeye zorlanmaktadır. Kelimelerin anlamsal ilişkisi model için bir denetim mekanizması oluşturmakta ve temel yöntemlerle karşılaştırıldığında Sentic-LDA başarılı konuların elde edildiği görülmüştür. Alam ve diğ. (2016) alandan bağımsız olarak duygu yönelimli ürün özelliklerini bulmak amacıyla Joint Multi-grain Topic (JMTS) tasarlamışlardır. JMTS, MG-LDA'nın gelişmiş bir versiyonudur. MG-LDA konular ile ilişkili kelimeler, içerikle ilişkili konular, içerikler ile ilişkili pencereler ve pencereler ile ilişkili dokümanlar olmak üzere dört hiyerarşik katmandan oluşmakta iken, JMTS pencere ve konu arasına anlamsal bir katman ekleyerek duygu ifadeleri ile bu ifadeler ile ilişkili ürün özelliklerinin çıkartılmasını sağlamaktadır. Deneyler sonucunda yöntemin mevcut yöntemlere göre daha başarılı olduğu görülmüştür.

Yang ve diğ. (2017) LDA tabanlı iki-katmanlı kategorik bilgiyi kullanan CAT-LDA'yı Amazon.com'dan elde ettikleri beş farklı kategorideki kullanıcı yorumlarına uygulamışlardır. Kelimelerin birlikte geçme durumlarının önsel bilgi olarak kullanıldığı Enriched LDA (ELDA) İngilizce ve Persçe kullanıcı yorumlarından oluşan veri kümelerine uygulanarak makul bir doğruluk elde edilmiştir (Shams ve Baraani-Dastjerdi, 2017). Ekinci ve İlhan Omurca (2017b) İngilizce restoran yorumlarına LDA'yı uygulayarak restoran ile ilgili üzerine yorum yapılan özellikleri elde etmişlerdir. Fu ve diğ. (2018) weakly supervised topic sentiment joint model with word embeddings (WS-TSWE) isimli LDA tabanlı modelde konuların ve duygu ifadelerinin çıkartılmasındaki başarıyı modele word-embedding ve HowNet sözlüğünü dahil ederek arttırmışlardır. WS-TSWE, İngilizce ve Çince'deki kitap, otel, bilgisayar ve sinema ile ilgili gerçek kullanıcı yorumları üzerine uygulanmıştır. Heng ve diğ. (2018), çevrimiçi yiyecek ve market alışverişi yorumlarında altı çizilen konuları tespit etmek için LDA'yı uygulamışlardır.

Türkçe için ise yapılmış olan çalışma sayısı sınırlıdır. "www.otelPuan.com" üzerinden elde edilmiş olan kullanıcı yorumlarına LDA uygulanarak hem tek kelimeden oluşan özellikler hem de eşdizim şeklindeki ürün özellikleri elde edilmiştir (Ekinci ve İlhan Omurca, 2017a). "www.sikayetvar.com"dan elde edilen altı farklı firmaya ait kullanıcı şikayetlerine LDA uygulayarak kullanıcıların en çok şikayette bulundukları özellikler belirlenmiştir (Atıcı ve diğ., 2017). Ekinci ve İlhan Omurca (2018a) yaptıkları bir diğer çalışmada otel yorumları üzerinden ürün özelliklerini ve duygu ifadelerini eş zamanlı modelleyerek ürün özellikleri ve ilişkili duygu ifadelerini çıkartmışlardır.

Bu tez çalışmasında kullanıcı yorumlarından ürün özelliklerinin çıkartılması amacıyla dokümanı oluşturan kelimeler arasındaki anlamsal ilişkiyi dikkate alan Concept-LDA ve dokümanlar arasındaki anlamsal bilgiyi dikkate alan NET-LDA isimli iki yöntem geliştirilmiştir. Bu bölümde geliştirilen iki yöntem ayrıntılı bir şekilde anlatılacaktır.

3.1. Concept-LDA

Metinlerin doğru analizinde kelime vektörlerinin anlamsal olarak başarılı bir şekilde temsil edilmesi oldukça önemlidir. LDA'nın temel varsayımı kelimelerin birlikte geçme durumlarının göz önünde bulundurulmasıdır; dolayısıyla LDA anlamsal bilgiyi

dikkate almamaktadır. Bu durum LDA sonucu elde edilen konular altında anlamsal olarak ilişkili kelimelerin yer almasını engellemektedir.

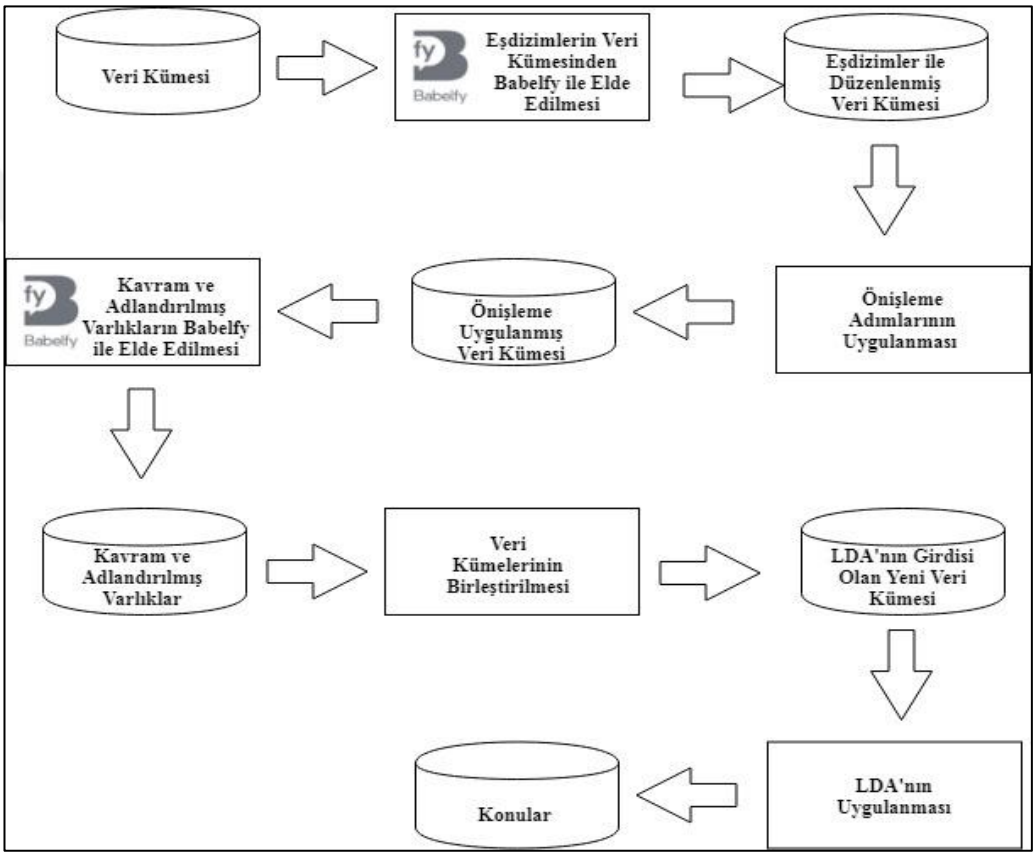
Tez kapsamında önerilen Concept-LDA modelinde LDA'nın bu dezavantajı göz önünde bulundurulmuştur. Kullanıcı yorumları, Babelfy'dan elde edilen kavramlar ve adlandırılmış varlıklar ile genişletilerek anlamsal bir zenginleştirme yöntemi hedeflenmiştir. Bir örnek üzerinden Concept-LDA'nın katkısını açıklayacak olursak; LDA ile doküman koleksiyonu kelime-torbası şeklinde temsil edilmektedir. Dolayısıyla, anlamsal olarak ilişkili "waiter (A person whose occupation is to serve at table (as in a restaurant)" ve "waitress (a woman waiter)" aynı konu içerisinde yer alamamaktadır. Ancak, "waiter" ve "waitress" için yorumlar bu kelimelerin kavramları ile genişletilirse – ki bu kavramlar "waiter" için "person" ve "restaurant", "waitress" için "woman" ve "waiter"- "waiter" ve "waitress" kelimelerini birlikte içeren daha doğru konular elde edilebilmektedir. Yani kelime torbası yaklaşımı yerine, {kelime+kavram+adlandırılmış varlıklar} torbası yaklaşımı ile anlamsal olarak ilişkili konular elde edilebilmektedir.

Concept-LDA ile ana varsayımımız; eğer kelimeler ortak kavramlara veya adlandırılmış varlıklara sahipse veya bir kelime diğeri için kavram veya adlandırılmış varlık ise bu kelimeler benzer anlamlara sahiptirler dolayısıyla da aynı konu içerisinde yer almaları gerekmektedir. Bu adımda WSA probleminin göz önünde bulundurulması gerekmektedir. Kelimelerin cümle içerisindeki anlamlarının doğru bir şekilde belirlenmesi ve bu anlama göre kavram ve adlandırılmış varlıkların elde edilebilmesi için, literatürde WSA problemi çözümünde başvurulan bir araç olan Babelfy kullanılmıştır.

Bu bağlamda, çıkarılan kavramlar ve adlandırılmış varlıklar LDA modelinde kullanılarak daha derin anlambilimsel bilginin dahil edilmesine dayanan yeni bir yöntem önerilmiştir. Model alan bağımsız olup, bunu ispatlamak için ise on iki farklı İngilizce kullanıcı yorumu modele girdi olarak verilmiş ve çıkış olarak daha doğru konular elde edilmiştir.

Concept-LDA Şekil 3.1'de gösterildiği üzere şu beş adımdan oluşmaktadır: (I) Eşdizimler Babelfy kullanılarak orijinal veri kümesinden elde edilir, (II) Birinci adımda elde edilen veri kümesi üzerine noktalama işaretleri, sayı ve özel karakterlerin

eldeki metinlerden çıkartılması, gövdeleme, küçük harfe dönüştürme, durak kelimelerinin elenmesi, yazım hatalarının düzeltilmesi adımları uygulanır, (III) Önişleme adımlarının uygulandığı veri kümesinden Babelfy ile kavram ve adlandırılmış varlıklar çıkartılır, (IV) Önişleme adımlarının uygulandığı veri kümesindeki her bir eleman bir önceki adımda elde edilen kavram ve adlandırılmış varlıklar ile genişletilip yeni bir veri kümesi elde edilir, (V) Son olarak bu veri kümesine LDA uygulanarak anlamsal konular elde edilir.



Şekil 3.1. Concept-LDA akış diyagramı

Şekil 3.1'deki adımlar alt başlıklarda detaylı bir şekilde anlatılmıştır.

3.1.1. Eşdizimlerin veri kümesinden elde edilmesi

Eşdizimler, herhangi bir konuyu, durumu, varlığı olağan bir şekilde ifade etmek için değişik şekillerde bir araya gelen iki veya daha fazla kelime şeklinde tanımlanmaktadır (Manning ve Schütze, 1999). Eşdizimler genelde dört tip olarak el alınmaktadır (Pecina, 2009; Petrovic ve diğ., 2010):

- tamlamaların oluşturduğu kısıtlandırılmış ifadeler: bir cümle içerisinde tek bir varlık olarak ele alınan sürekli kelime dizilerinin oluşturduğu eşdizimlerdir. Tur operatörü, krem peynir, acı biber, vb. eşdizimler bu gruba girmektedir.
- adlandırılmış varlıklar: kişi, organizasyon veya konum gibi gerçek dünyadaki nesnelerin isimlerine karşılık gelen iki veya daha fazla kelimeden oluşan, Vivica Fox, Washington Bulvarı gibi eşdizimlerdir.
- teknik terimler: ilgili alan ile ilgili teknik bilgi taşıyan eşdizimlerdir. "Bilgisayar" alanı ile ilgili kullanıcı yorumları incelendiğinde pek çok teknik terim içerdiği görülmektedir. Bunlardan bazıları; cpu hızı, dokunmatik ekran, ölü piksel şeklinde verilebilir.
- anlamca kaynaşmış bileşik fiiller: eşdizimi oluşturan kelimelerden birinin anlam kaybına uğraması ile oluşmaktadırlar. Karar vermek, hizmet etmek, yerine getirmek bileşik fiillere örnek olarak verilmektedir.

Literatürde farklı diller için eşdizimlerin elde edilmesi problemi de farklı diller için sıklıkla çalışılmıştır. Xu ve diğ. (2003) Çince haberlerden oluşan veri kümesinden eşdizimleri çıkarmak için istatistiksel ve sözdizimsel özellikleri birlikte çıkartmışlardır. Lu ver diğ. (2003) Xtract'ı yeniden düzenleyerek yine Çince için eşdizimleri elde etmişlerdir. Xtract'ın gelişmiş bir versiyonu olan CXtract3 farklı ayırt edici özellikler ile çok aşamalı stratejiyi kullanarak eşdizimleri çıkartmaktadır (Xu ve Lu, 2005a). Xu ve Lu (2005b) CXtract3'ü kabul edilen ve reddedilen eşdizimlere göre geliştirerek %15'lik bir başarı artışı elde etmişlerdir. Bir başka çalışmada İngilizce, Fransızca, İtalyanca ve İspanyolca veri kümlerinden eşdizimlerin elde edilmesi için çok dilli ayrıştırma ve istatistiksel yöntemler kullanılmıştır (Seretan ve Wehrli, 2006). Live diğ. (2006) TCract ile istatistiksel model ile kural tabanlı dilbilimsel bilgiyi birleştirerek isim öbeği şeklindeki eşdizimleri Çince veri kümesinden elde etmişlerdir. Li ve diğ. 2007 yılında yaptıkları çalışmada ise Çince veri kümesinden isim ve fiil öbekleri şeklindeki eşdizimleri çıkartmak için sözdizimsel eşdizim örüntülerinden faydalanmışlardır. Lin ve diğ. (2008) İngilizce için eşdizimleri çıkartmak için lojistik doğrusal regresyon modeliyle birleştirdikleri çoklu sözcüksel ilişki ölçülerini kullanmışlardır. Heid ve Weller (2008) Almanca veri kümesinden isim+fiil şeklindeki eşdizimlerin çıkartılmasında doküman koleksiyonu sorgu diline bağlı çıkarım örüntüleri kullanmışlardır. Lin ve diğ. (2009) birlikte geçme sıklığı, ortak bilgi ve t-

testini kullanarak web tabanlı bir otomatik eşdizim çıkarma sistemi geliştirmişlerdir. Todirascu ve diğ. (2009) İngilizce ve Romanca için fiil+isim şeklindeki eşdizimleri çıkarmak amacıyla istatistiksel yöntemleri ve dile özgü dilbilimsel filtreleri kullanan hibrid bir sistem geliştirmişlerdir. Seretan ve Wehrli (2009) yaptıkları bir diğer çalışmada yine çok dilli sözdizimsel ayrıştırıcı tabanlı bir eşdizim çıkarım sistemi geliştirmiştir. İngilizce ve Hırvatça için geliştirilen TermeX sözcüksel ilişki ölçülerini kullanarak eşdizimleri çıkarmaktadır (Delac ve diğ., 2009). Türkçe eşdizimler için Kumova Metin ve Karaoğlan (2009) geçme sıklıklarını, ortak bilgiyi ve hipotez testlerini kullanmışlardır. Yaptıkları deneyler sonucunda ki-kare testinin ve ortak bilginin Türkçe için en başarılı yöntemler olduklarını tespit etmişlerdir. Haber veri kümesinden göreli uzunluktaki eşdizimlerin çıkartılması için genişletilmiş sözcüksel ilişki ölçülerinden faydalanılmıştır (Petrovic ve diğ., 2010). Çok dilli bir koleksiyondaki eşdizimleri çıkarmak için Liu ve diğ. (2011) herhangi bir dil bilgisi olmaksızın istatistiksel iki dilli sözcük eşleme yöntemlerini kullanmışlardır. Eşdizimlerin dilbilimsel özelliklerinden faydalanarak üç katmanlı bir sistem yine eşdizim çıkarma için geliştirilmiştir (Li ve diğ., 2015; Cao ve diğ., 2015). Birinci katmanda belli bir frekansın üzerindeki eşdizimler, ikinci katmanda hem belli frekansın üzerinde olup, hem de doğrudan sözdizimsel ilişkisi olan eşdizimler, son katmanda ise hem birinci ve ikinci katmandaki özellikleri taşıyan hem de eşdizimdeki kelimelerin eş anlamları ile değiştirilemez olduğu eşdizimler çıkartılmaktadır. Suarez ve diğ. (2015) tarafından İspanyolca bir veri kümesinden eşdizim elde etmek için göreli frekans, ortak bilgi, z-skor, t-skor ve Dunning'in testi uygulanmıştır. Türkçe için yine Kumova Metin ve Karaoğlan (2015) tarafından yapılan çalışmada geçme frekansı, noktasal karşılıklı bilgi katsayısı ve hipotez testleri uygulanarak eşdizimlerin çıkartılması hedeflenmiştir. Al-Thubaity and Baazeem (2017) Arapça koleksiyondan 5-gramlara kadar olan eşdizimleri çıkartmak için Musaheb isimli bir araç geliştirmiştir. Ekinci ve diğ. (2017a, 2017b) Türkçe otel veri kümesinden eşdizim şeklindeki ürün özelliklerini çıkarmak için iki farklı yöntem geliştirmiştir. Birinci yöntemde n-gramlar ile dile özgü sezgisel kurallar kullanılırken, ikinci yöntemde ise Apriori algoritması ile dile özgü sözdizimsel kurallar kullanılmıştır. Ekinci ve İlhan Omurca (2018b) yine Türkçe otel veri kümesini kullanarak tüm eşdizimleri çıkartmak için Babelfy'ı tercih etmişlerdir.

Concept-LDA kapsamında ise İngilizce yorumlardan eşdizimlerin çıkartılması amacıyla Babelfy'ın Java kütüphanesi olan Babelfy-online-API-1.0 kullanılmıştır. Eşdizimlerin elde edilmesinde geliştirilen kod parçası Şekil 3.2'de verilmiştir.

```
BabelfyConstraints constraints = new BabelfyConstraints();
SemanticAnnotation a = new SemanticAnnotation(new
TokenOffsetFragment(0, 0), "bn:03083790n",
        "http://dbpedia.org/resource/BabelNet", Source.OTHER);
constraints.addAnnotatedFragments(a);
BabelfyParameters bp = new BabelfyParameters();
bp.setAnnotationResource(SemanticAnnotationResource.BN);
//bp.setMCS(Source.OTHER);
bp.setScoredCandidates(ScoredCandidates.ALL);
Babelfy bfy = new Babelfy(bp);
List<SemanticAnnotation> bfyAnnotations =
        bfy.babelfy(inputText, Language.EN, constraints);
//bfyAnnotations is the result of Babelfy.babelfy() call
HashSet<String> hs=new HashSet<>();
ArrayList<String> al=new ArrayList<>();
for (SemanticAnnotation annotation : bfyAnnotations)
{
    //splitting the input text using the CharOffsetFragment start a
    String frag = inputText.substring
(annotation.getCharOffsetFragment().getStart(),
            annotation.getCharOffsetFragment().getEnd() + 1);
    al.add(frag);
    hs.add(frag);
}
if(al.size()!=0){
ArrayList<String> arl=new ArrayList<>();
int k=1;
arl.add(0, al.get(0));

for(int i=1;i<al.size();i++){
    if(!al.get(i).equals(al.get(i-1))){
        arl.add(k, al.get(i));
        k++;
    }
}

int t=0;
ArrayList<String> arlNew=new ArrayList<>();
arlNew.add(0, arl.get(0));
for(int i=1;i<arl.size()-1;i++){
    if(arl.get(i).contains(arl.get(i-1))&&
            arl.get(i).contains(arl.get(i+1))&&
            arl.get(i).contains(arlNew.get(arlNew.size()-1))){
        arlNew.set(arlNew.size()-1, arl.get(i));
        i+=2;
    }
    else if(!arl.get(i-1).contains(arl.get(i))){
        arlNew.add(arl.get(i));

    }
}
if(arl.size()>=2&&
        !arl.get(arl.size()-2).contains(arl.get(arl.size()-1)))
    arlNew.add(arl.get(arl.size()-1));
String str="";
for(int i=0;i<arlNew.size();i++){
    String kk="";
    if(arlNew.get(i).contains(" "))
        kk=arlNew.get(i).replace(" ", "_");
        else
        kk=arlNew.get(i);
    str+=kk;
    str+=" ";
}

str=str.trim();
reviews.add(str);
}
```

Şekil 3.2. Babelfy ile eşdizimlerin çıkartılmasında geliştirilen kod parçacığı

3.1.2. Önişleme adımlarının uygulanması

Metinler üzerinde yapılacak önişleme çalışılacak amaca göre farklılıklar göstermekle birlikte temel önişleme adımları;

- noktalama işaretleri, sayı ve özel karakterlerin eldeki metinlerden çıkartılması,

- büyük küçük harf duyarlı olmamasından ötürü büyük harflerin küçük harflere dönüştürülmesi,
- metni meydana getiren ve çok sık tekrarlanan ancak doküman için önemli olmayan durak kelimelerinin eldeki metinlerden ayıklanması,
- yazım hatalarının düzeltilmesi,
- gövdelemenin gerçeklenmesi şeklinde sıralanmaktadır.

Bu önişleme adımları doğal dil işleme sürecini oluşturmaktadır. İlk kez 1950'lerde uygulanmaya başlanan doğal dil işleme, doğal dildeki metinleri yapısal ve anlamsal olarak çözümleme şeklinde tanımlanmaktadır (Nadkarni ve diğ., 2011).

Mevcut veri kümesi üzerinden konuların çıkartılması için öncelikli olarak önişleme adımlarının gerçekleştirilmesi gerekmektedir. Bu tez çalışması kapsamında ilk olarak eldeki yorumlardaki yazım hatalarının düzeltilmesi adımı gerçeklenmiştir. Bu amaçla hazır bir kütüphane olan açık kaynak kodlu LanguageTool isimli hata düzeltme aracından yararlanılmıştır (URL-3, 2018). LanguageTool kullanılarak yazım hatalarının düzeltilmesine dair kod parçacığı Şekil 3.3'te verilmiştir.

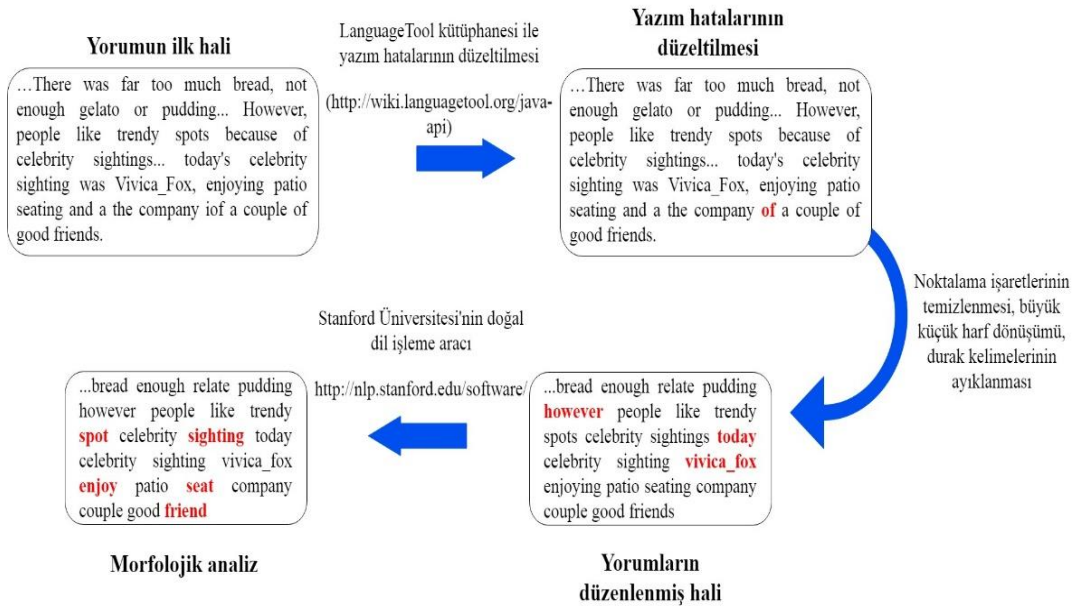
```
Collection c=dictionaryWords;
JLanguageTool langTool = new JLanguageTool(new BritishEnglish());
for(ReviewStructure revStructure:reviewStructure){

    String str=revStructure.getReview();
    String [] tmpStr=str.split(" ");
    String review="";
    for(String tmp:tmpStr){
        if(!c.contains(tmp)){
            List<RuleMatch> matches = langTool.check(tmp);
            int k=0;
            for(int i=0;i<matches.size();i++){
                if(!matches.get(i).getSuggestedReplacements().isEmpty()&&
                        c.contains(matches.get(i).getSuggestedReplacements().get(0))){
                    review=review.concat(matches.get(i).getSuggestedReplacements().get(0));
                    review=review.concat(" ");
                    k=1;
                    break;
                }
            }
            if(k==0){
                review=review.concat(tmp);
                review=review.concat(" ");
            }
        }
        else{
            review=review.concat(tmp);
            review=review.concat(" ");
        }
```

Şekil 3.3. LanguageTool ile yazı hatalarının düzeltilmesi için kod parçacığı

Yazım hataları düzeltildikten sonra yorum içerinde yer alan noktalama işaretleri, sayı ve özel karakterler eldeki metinlerden çıkartılmıştır. Büyük harfler küçük harflere dönüştürülmüş ve eldeki yorumlar durak kelimelerinden temizlenmiştir. Son olarak da morfolojik analiz yapılarak önişleme adımı tamamlanmıştır. Morfolojik analizde gövdeleme işlemi gerçekleştirilerek kelimelerin kök veya gövde durumuna dönüştürülmesi ve kelimeler için standart bir formun oluşturulması amaçlanmıştır. Bu adımda Stanford Üniversitesi'nin doğal dil işleme kütüphanesi kullanılmıştır (URL-4, 2018). Bu kütüphane, belirtkelemeden anafor çözümlemeye kadar olan temel doğal dil işleme adımlarını kullanıcısına sağlayan Java tabanlı bir etiketleme zinciri yapısını sunmaktadır (Manning ve diğ., 2014).

Gerçekleştirilen önişleme adımları örnek bir yorum üzerinden adım adım Şekil 3.4'te verilmiştir.



Şekil 3.4. Örnek bir yorum üzerinde önişleme adımlarının gerçeklenmesi

3.1.3. Kavram ve adlandırılmış varlıkların çıkartılması ile doküman uzayının genişletilmesi

Concept-LDA'nın ardındaki temel yaklaşım; kelime torbası yerine {kelime+kavram+adlandırılmış varlıklar} torbasının kullanılmasıdır. Eldeki dokümanlardan kavram ve adlandırılmış varlık çıkarımı adımında Babelfy'dan yararlanılmıştır. Bu şekilde bir genişletme adımına ihtiyaç duyulmasının nedeni

anlamsal olarak ilişkili olan kelimelerin aynı konu başlığı altında bulunmadığının gözlemlenmesidir. Veri kümesinde yer alan "cocktail pomegranate martini twist halibut lamb price" yorumundan çıkartılan kavram ve adlandırılmış varlıklar "mixed_drink drink Asia Shrub southwestern_Asia fruit cocktail dry_vermouth gin vodka vermouth cocktail_garnish citrus_zest cocktail flesh flatfish Atlantic Pacific money amount_of_money" şeklindedir. Kavram ve adlandırılmış varlıklar orijinal yorum ile birleştirilerek yorum genişletilmektedir ve yorum "cocktail pomegranate martini twist halibut lamb price mixed_drink drink Asia Shrub southwestern_Asia fruit cocktail dry_vermouth gin vodka vermouth cocktail_garnish citrus_zest cocktail flesh flatfish Atlantic Pacific money amount_of_money" şeklinde güncellenmektedir. Yorumda yer alan her bir kelime için elde edilen kavram ve adlandırılmış varlıkların her biri ise anlaşılabilirlik açısından Tablo 3.1'de ayrı ayrı verilmiştir.

Tablo 3.1. Yorumda yer alan kelimelerin her biri için ilgili kavram ve adlandırılmış varlıklar

Kelime	**Kavram veya adlandırılmış varlıklar**
cocktail	mixed drink, drink
pomegranate	Asia, shrub, Southwestern Asia
martini	cocktail, dry vermouth, gin, vodka
twist	cocktail garnish, citrus zest, cocktail
halibut	flesh, flatfish, Atlantic, Pacific
price	money, amount of money

Tablo 3.1 incelendiğinde "twist" kelimesinin Babelfy'da ilk sırada çıkan anlamı "An unforeseen development" şeklindedir. Ancak "twist" kelimesinin içerisinde geçtiği yorum incelendiğinde gerçek anlamının "A twist is a piece of citrus zest used as a cocktail garnish, generally for decoration and to add flavor when added to a mixed drink" anlamında kullanıldığı görülmektedir. Dolayısıyla Babelfy ile yapılan genişletme ile kelimenin gerçek anlamı yakalanmış olup, kendisi ile anlamsal olarak benzerlik gösteren kelimeler ile bir konu içerisine dahil olma olasılığı da yükselmiş olacaktır. Yani "twist" kelimesi doğru konu içerisinde yer alacaktır.

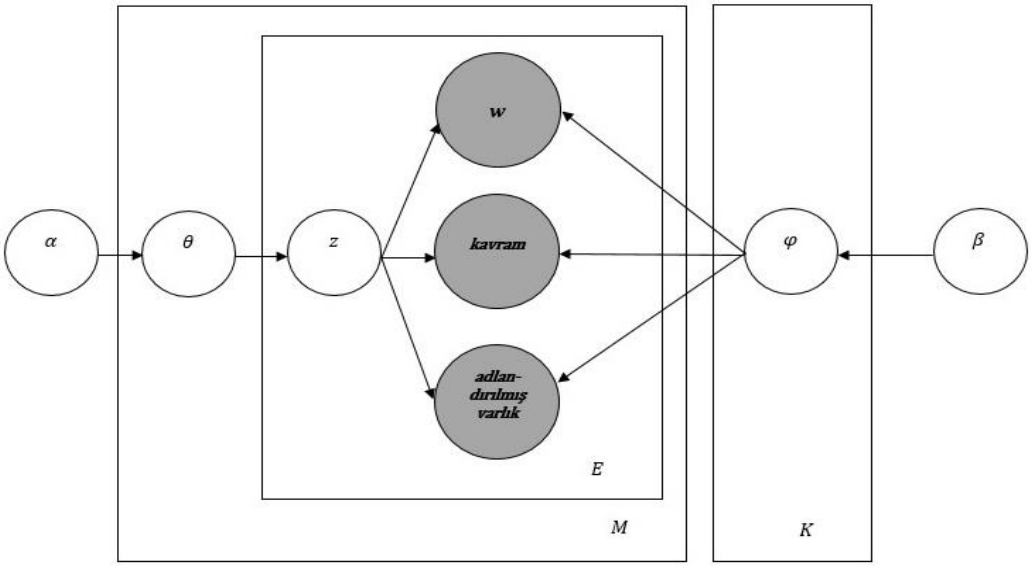
Bir yorumun kavram ve adlandırılmış varlıklar ile Babelfy üzerinden genişletilmesi için yazılan kod parçası Şekil 3.5'te verilmiştir.

```
for(int i=0;i<concepts.size();i++){
    String str="";
    String inputText = concepts.get(i);
    String[] tempArry=inputText.split(" ");
    List<SemanticAnnotation> bfyAnnotations = bfy.babelfy(inputText, Language.EN);
    for(SemanticAnnotation annotation : bfyAnnotations){
        String frag = inputText.substring(annotation.getCharOffsetFragment().getStart(),
                annotation.getCharOffsetFragment().getEnd() + 1);
        tempList.add(frag);
        URL url=new URL("http://babelnet.io/v4/getSynset?id="+annotation.getBabelSynsetID()+
                "&lang=EN&key=...");
        InputStream response = url.openStream();
        try (BufferedReader reader = new BufferedReader(new InputStreamReader(response))) {
            for (String line; (line = reader.readLine())!= null;) {
                //System.out.println(line);
                str+=JSonProcessor.JsonProcessor(line);
                str+=" ";
```

Şekil 3.5. Yorumun kavram ve adlandırılmış varlıklar ile genişletilmesinde kullanılan kod parçası

3.1.4. Konu çıkarımı

Kavram ve adlandırılmış varlıklar ile genişletilen kullanıcı yorumlarından konuların çıkartılması için geliştirilen Concept-LDA grafiksel temsili Şekil 3.6'da verilmiştir.



Şekil 3.6. Concept-LDA'nın grafiksel temsili

Şekil 3.6'da bulunan E ilgili dokümanın kelimeler ve adlandırılmış varlıklar ile genişletilmesinden sonra bu dokümanda bulunan toplam kelime sayısını temsil etmektedir. M toplam doküman sayısını, K toplam konu sayısını ifade etmektedir. θ konuların dokümanda bulunma olasılığını, φ ise kelimelerin konulardaki dağılımını göstermektedir. α ve β Dirichlet parametreleridir. w ilgili yorumdaki kelimeyi, kavram

w'nun genişletilmesiyle elde edilen kavramı ve adlandırılmış varlık w'nun genişletilmesiyle elde edilen adlandırılmış varlığı ifade etmektedir.

3.2. NET-LDA

NET-LDA tasarlanırken Concept-LDA'da olduğu gibi LDA'yı kısıtlayan anlamsal olarak ilişkili, uyumlu, detayları yakalayabilen ve anlamlı konuları elde edememe durumu göz önünde bulundurulmuştur. Literatürde LDA'nın bu dezavantajının üstesinden gelmek amacıyla anlamsal bilginin kullanıldığı çalışmalar mevcuttur (Griffiths ve Steyvers, 2002a; Kim ve Hovy, 2006; Griffiths ve diğ., 2007; Chemudugunta ve diğ., 2008; Godin ve diğ, 2013; Poria ve diğ., 2016; Zhang ve diğ., 2016). Ancak bu çalışmalarda kullanılan anlamsal bilgi dokümanların anlamsal benzerliği olarak ele alınmamıştır. Bu tez çalışmasında dokümanların anlamsal benzerliği ilk kez farklı bir LDA modelini geliştirmek üzere kullanılmıştır.

Bu tez çalışması kapsamında, geliştirilen NET-LDA'nın ardındaki temel fikir; dokümanlar anlamsal olarak birbirine ne kadar benzer ise içerdikleri konular da o derece benzer olacaktır şeklindedir. Birbirine anlamsal olarak benzer dokümanlar birleştirilip NET-LDA'ya girdi olarak verildiğinde; LDA'nın temel olarak kullandığı kelimelerin birlikte geçme sıklıkları da anlamsal olarak güçlendirilmiş olacaktır. Buna dayalı olarak dokümanlar arasındaki anlamsal benzerlik hesaplanmış ve bu benzerlik adaptif bir parametre olarak modele dahil edilmiştir. Bu parametre bilinen temel bir ölçüt olarak NET-LDA'ya dahil edilmekte olup, doküman-konu dağılımına etki etmektedir. Parametrenin belirlenmesi adımında ise doküman benzerlikleri üzerinden benzerlik grafı oluşturulmuştur. Grafın düğümlerini dokümanlar oluşturmakta, birbirine benzer dokümanlar kenarlar ile birbirine bağlanmakta, kenar ağırlıkları ise dokümanların birbirlerine benzerlikleri olacak şekilde tanımlanmaktadır. Dokümanların anlamsal benzerlikleri ise Babelfy ile elde edilen kavram ve adlandırılmış varlıklar üzerinden hesaplanmıştır. Önerilen LDA modeli, anlamsal olarak birbirine bağlanmış dokümanlar üzerinden bir graf yapısı sunması nedeniyle NET-LDA olarak adlandırılmıştır.

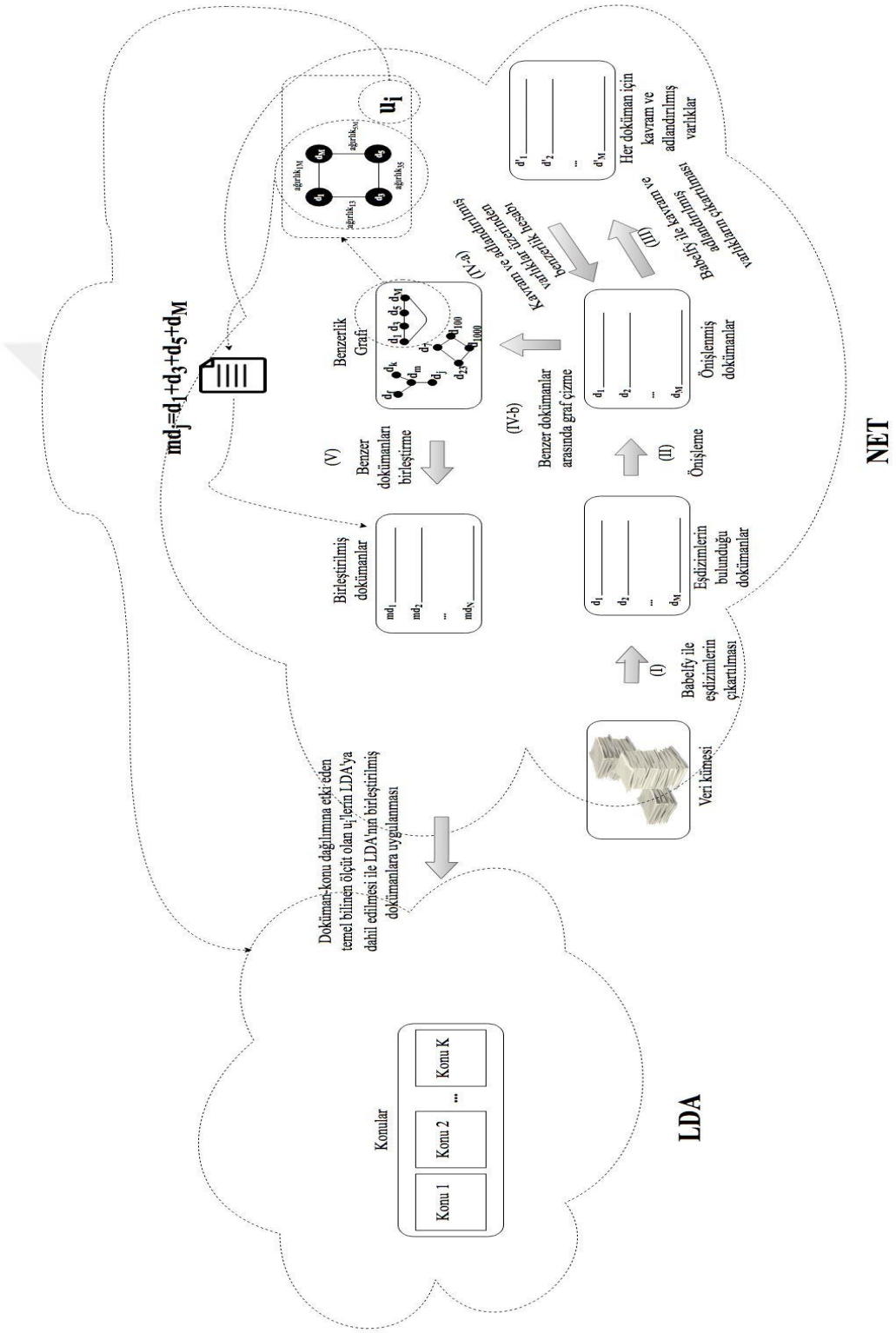
Mevcut LDA literatürü incelendiğinde, anlamsal olarak ilişkili ve tutarlı konuların elde edilmesi amacıyla konu modelleri ve anlamsal benzerliğin avantajlarının birleştirildiği ilk model NET-LDA'dır. Önerilen NET-LDA dilden bağımsızdır; Babelfy tarafından

sağlanan 284 farklı dildeki dokümanlara uygulanabilir. Tez çalışması kapsamında İngilizce ve Türkçe kullanıcı yorumlarına uygulanmıştır. Concept-LDA'da olduğu gibi NET-LDA'da alan bağımsız bir modeldir ve 13 farklı veri kümesine uygulanarak alan bağımsız olduğu ispatlanmıştır.

NET-LDA Şekil 3.7'de gösterildiği üzere 2 farklı modülden oluşmaktadır: ilk modül anlamsal grafın oluşturulması (NET), ikinci modül ise doküman-konu dağılımına etki eden parametrenin dahil edilmesi ile konuların çıkartılması (LDA). Birinci modül olan NET'i oluşturan 5 adım: (I) Babelfy kullanılarak orijinal veri kümesinden eşdizimlerin elde edilmesi, (II) Noktalama işaretleri, sayı ve özel karakterlerin ilk adımdan gelen veri kümesinden çıkartılması, yine bu veri kümesine gövdeleme, küçük harfe dönüştürme, durak kelimelerinin elenmesi, yazım hatalarının düzeltilmesi adımlarının uygulanması, (III) Önişlenmiş veri kümesinden Babelfy ile kavram ve adlandırılmış varlıkların elde edilmesi, (IV) (a) Kavram ve adlandırılmış varlıklar üzerinden doküman benzerliklerinin hesaplanıp (b) önişlenmiş veri kümesindeki benzer dokümanlar arasında grafın çizilmesi, (V) Son olarak da benzer dokümanların birleştirilmesi şeklindedir.

Önerilen yaklaşımda farklı alandaki ve dildeki doküman koleksiyonlarının her biri ayrı ayrı NET-LDA'ya girdi olarak verilmiştir ve her alan için ürün özellikleri kümesi çıktı olarak elde edilmiştir. Burada "alan" ile kastedilen belli bir kişi, ürün, servis vb. ile ilgili dokümanların kümesidir. Doküman koleksiyonu $D=\{d_1, d_2, \ldots, d_M\}$ ile temsil edilsin. Ürün özelliklerinin başarılı bir şekilde temsil edilmesi için eşdizimlerin veri kümesinden çıkartılması oldukça önemlidir. Bu nedenle Adım (I)'de eşdizimler veri kümesinden çıkartılmıştır. Adım (II)'de temel önişleme adımları veri kümesine uygulanmıştır. Adım (III)'te kelimelerin birlikte geçme durumlarını anlamsal olarak güçlendirmek amacı ile D kümesinde yer alan her d_j için kavram ve adlandırılmış varlıklar elde edilmiştir. Doküman koleksiyonundaki her d_j için d_j'nin kavram ve adlandırılmış varlıklar ile temsili olan yeni doküman d'_j oluşturulmuştur. d'_j'lerden oluşan doküman kümesi $D'=\{d'_1, d'_{2,\ldots,} d'_M\}$ şeklinde ifade edilmiştir. Adım (IV)'te D''de yer alan d'_j'ler arasındaki anlamsal benzerlik için cosinüs benzerliği kullanılmıştır. Elde edilen benzerlikler kullanılarak D kümesinde yer alan d_j'ler arasında benzerlik grafı çizilmiştir. Adım (V)'te, Adım (IV)'te oluşturulan benzerlik grafına bağlı olarak D kümesinde yer alan dokümanlar birleştirilerek yeni dokümanlar

dolayısıyla yeni bir koleksiyon MD={md_1, md_2, …, md_G} oluşturulmuştur. Her oluşturulan yeni doküman md_i için bilinen temel ölçüt u_i'de bu adımda belirlenmiştir.

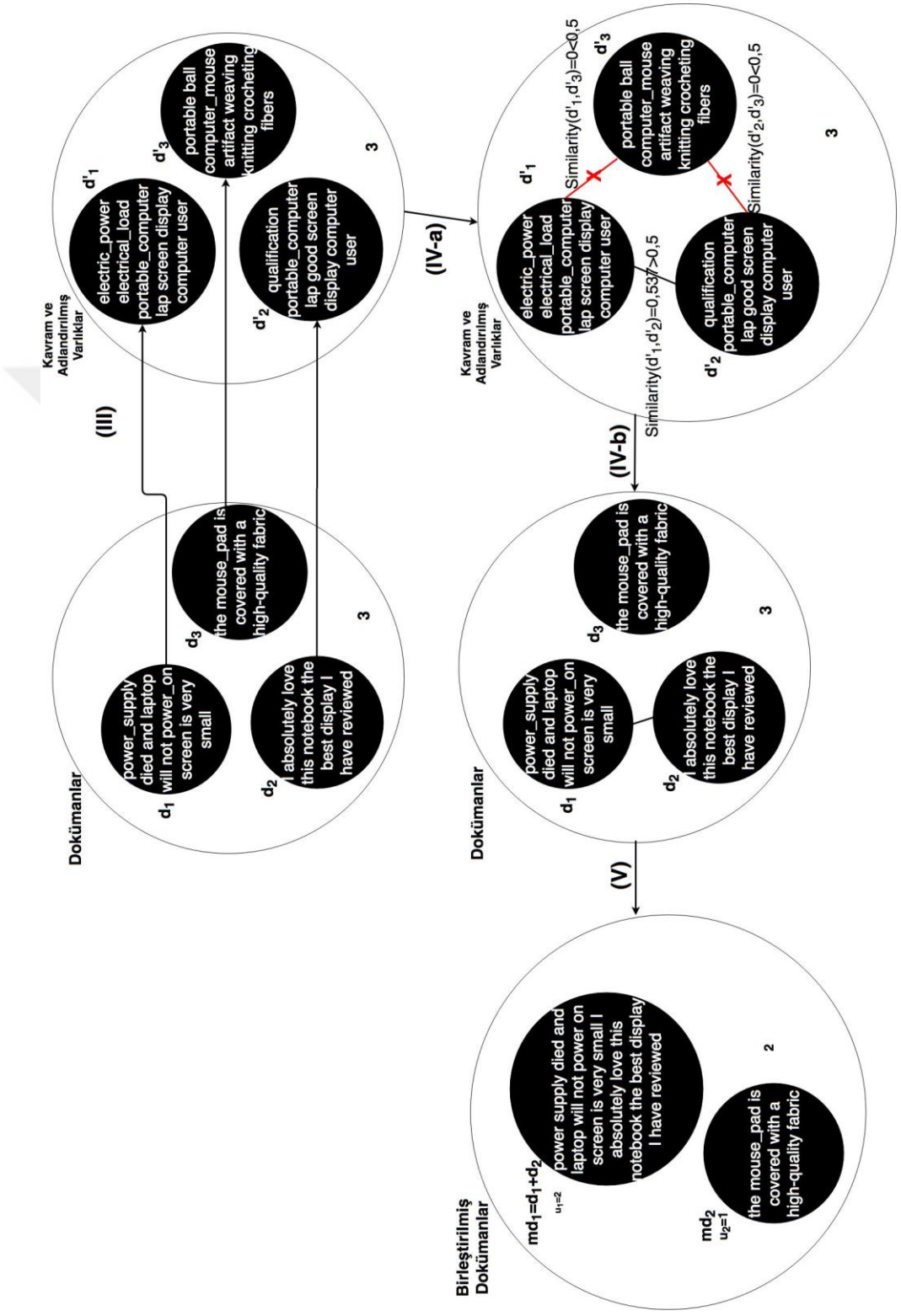


Şekil 3.7. NET-LDA akış diyagramı

Örneğin Şekil 3.7'de gösterildiği üzere d'_1, d'_3, d'_5 ve d'_M benzer olsunlar. Adım (IV)'te d_1, d_3, d_5 ve d_M arasında kurulan benzerlik grafına göre bu dört doküman birleşerek md_i dokümanını oluşturacaklardır. md_i 4 dokümanın birleşiminden oluştuğu için md_i için u_i de 4 olacaktır. Ayrıca M olan orijinal doküman boyutu G'ye düşmüş olacaktır. Böylece modelin NET kısmı tamamlanmış olup, yeni dokümanlar LDA'ya girdi olarak verilirken bu dokümanlar için elde edilen u_i değerleri de LDA'daki doküman-konu dağılımına etki edecek ve çıktı olarak konular elde edilecektir. NET-LDA'nın adımlarından (III), (IV) ve (V) Şekil 3.8'de ayrıntılı bir şekilde örnek yorumlar üzerinden verilmiştir.

Bu durumu veri kümesindeki üç yorum üzerinden anlatalım: d_1= "Power supply died and laptop won't power on. Screen is very small.", d_2= "I absolutely love this notebook. The best display I've reviewed." ve d_3= "The mouse pad is covered with a high-quality fabric.". Bu üç yorumun birbirine olan benzerliği sıfırdır. Ancak d_1 ve d_2 yorumları eş anlamlı kelimeler içermektedir. Yani bu yorumlar anlamsal açıdan birbirlerine benzemektedirler. Bu iki yorumdaki "laptop" ile "notebook" ve "screen" ile "display" kelimeleri kendi içlerinde aynı kavrama karşılık gelmektedir. Dolayısıyla bu kelimelerin aynı konular altında yer alması istenmektedir. Bu kelimeler için Babelfy üzerinden elde edilen kavramlar ise "laptop" ve "notebook" için "portable computer" ve "lap" şeklindeyken "screen" ve "display" için "screen", "display", "computer" ve "user" şeklindedir. Bu duruma dayalı olarak ilk olarak bu yorumlar kavram ve adlandırılmış varlıklar ile temsil edilmiştir. Bu durumda, d'_1= "electric_power electrical_load portable_computer lap screen display computer user", d'_2= "qualification portable_computer lap good screen display computer user" ve d'_3= "portable ball computer_mouse artifact weaving knitting crocheting fibers" halini almışlardır. Elde edilen kavram ve adlandırılmış varlıklar üzerinden benzerlik hesabı yapıldığında d'_1 ve d'_2 yorumlarının birbiri olan benzerliği 0,537 olarak elde edilmişken bu iki yorumun d'_3 yorumu ile olan benzerliği sıfır olarak hesaplanmıştır. d'_1 ve d'_2 arasındaki benzerlik 0,5'ten büyük olduğu için d'_1 ve d'_2'nin orijinal halleri olan d_1 ve d_2 birleştirilmiş ve yeni doküman md_1 elde edilmiştir. md_1 yorumu için adaptif parametre u_1 ise bu yorumu oluşturan dokümanların sayısı 2 olarak belirlenmiştir. d_3 yorumu ise diğer iki yorumla birleştirilemediği için bu yorum için

adaptif parametre u_2 1 değerini almıştır. Başlangıçta 3 olan doküman koleksiyonundaki eleman sayısı ise 2'ye düşmüştür.



Şekil 3.8. NET-LDA'nın alt adımlarının ayrıntılı anlatımı

Alt bölümlerde ise NET-LDA'nın tüm adımları ayrıntılı olarak verilmiştir.

3.2.1. Eşdizimlerin veri kümesinden elde edilmesi

İngilizce veri kümesi için eşdizimler Concept-LDA için elde edilmişti. Türkçe veri kümesi için eşdizimlerin elde edilmesinde adımında da İngilizce'de olduğu gibi Babelfy'dan yararlanılmıştır. Türkçe veri kümesi üzerinden tamlamaların oluşturduğu kısıtlandırılmış ifadeler, adlandırılmış varlıklar ve anlamca kaynaşmış bileşik fiiller olmak üzere 3 farklı eşdizim çeşidi elde edilmiştir. Veri kümesi herhangi bir teknik terim içermediği için bu gruba ait eşdizim elde edilememiştir. Türkçe veri kümesinden elde edilen bazı eşdizimler etiketleri ile birlikte Tablo 3.2'de verilmiştir.

Tablo 3.2. Türkçe veri kümesinden elde edilen eşdizimler ve etiketleri (Ekinci ve İlhan Omurca, 2018b)

Eşdizim	Etiket
basketbol sahası çakıl taşı insan kaynakları kurutma makinesi müşteri hizmetleri	Tamlamaların oluşturduğu kısıtlandırılmış ifadeler
Abdullah bey Halil bey	Adlandırılmış varlıklar
hizmet etmek tebrik ederim	Anlamca kaynaşmış bileşik fiiller

3.2.2. Önişleme adımlarının uygulanması

Concept-LDA'dan farklı olarak NET-LDA'da Türkçe veri kümesi için de önişleme adımları gerçekleştirilmiştir. Önişleme adımları gerçekleştirilirken Türkçe DDİ Kütüphanesi olan Zemberek kullanılmıştır. Zemberek, tüm Türk dilleri için yazım denetimi, gövdeleme ve morfolojik analiz gibi pek çok görevi yerine getiren açık kaynak kodlu bir kütüphanedir (Akın ve Akın, 2007). Türkçe veri kümesi üzerinde ilk olarak yazım denetimi yapılmıştır. Bu amaçla Zemberek'in sağladığı spellChecker.suggestForWord() fonksiyonu kullanılmıştır. Örneğin "peyni" şeklinde yanlış yazılmış bir kelime için bu fonksiyon [peynir, Peynir] şeklinde düzeltme işlemi yapmaktadır. Yazım hatalarından düzeltilmesi adımından sonra, noktalama işaretleri, sayılar, özel karakterler, durak kelimeleri veri kümesinden temizlenmiştir. En son adım olarak gövdeleme işlemi veri kümesine uygulanmıştır. Bu amaçla da yine Zemberek'ten yararlanılmıştır. Bir yorum içerisinde yer alan "odalar" kelimesi

gövdeleme işlemi sonrasında "oda" olarak değiştirilmiştir. Gövdeleme adımında kullanılan kod parçası Şekil 3.9'da verilmiştir.

```
public String analyze(String word){
    System.out.println("Word = " + word);
    System.out.println("Parses: ");
    List<WordAnalysis> results = morphology.analyze(word);
    for(WordAnalysis result : results){
        System.out.println("\tStems = " + result.getStems());
        System.out.println("\tLemmas = " + result.getLemmas());
    }
    ArrayList <String> stems = new ArrayList<>();
    stems = (ArrayList<String>) results.get(0).getStems();
    return stems.get(0);
}
```

Şekil 3.9. Gövdeleme adımında kullanılan kod parçası

3.2.3. Kavram ve adlandırılmış varlıkların çıkartılması

Kavram ve adlandırılmış varlıkların çıkartılması Babelfy ile sağlanmıştır. Yine bu adımda Concept-LDA'dan farklı olarak Türkçe için kavram ve adlandırılmış varlıklar elde edilmiştir. Kavram ve adlandırılmış varlıklar sadece anlamsal benzerlik için kullanıldığı için Türkçe için çıkartılan kavram ve adlandırılmış varlıklar İngilizce'dir. "otel hizmet tur operatörü" şeklindeki bir kullanıcı yorumundan elde edilen kavramlar "travelers pay meal services work person benefits travel package_holiday" şeklindedir. Her bir kelime için Babelfy'ın verdiği kavramlar Tablo 3.3'te verilmiştir.

Tablo 3.3. Yorumda yer alan kelimelerin her biri için ilgili kavramlar

Kelime	Kavram veya adlandırılmış varlıklar
otel	travelers, pay, meal, services
hizmet	work, person, benefits
tur operatörü	travel, package holiday

3.2.4. Benzerlik grafının oluşturulması ve dokümanların birleştirilmesi

Dokümanlar için benzerlik grafının oluşturulması adımında D kümesinden elde edilen kavram ve adlandırılmış varlıklardan oluşan kümesi D' kullanılmıştır. Bu adımda kavram ve adlandırılmış varlıkların kullanılmasının nedeni WSA probleminin çözümü ile ilgili yorumların gerçek anlamlarının elde edilmiş olması bu sayede daha doğru benzerlik değerlerini elde edilmiş olmasıdır. Dokümanlar arası benzerlik grafının oluşturulması amacıyla geliştirilen algoritma Şekil 3.10'da verilmiştir.

Algoritma 3: Benzerlik Grafı Oluşturma Algoritması

Girdi: D={d_1, d_2, …, d_M} kümesinin kavram ve adlandırılmış varlıklarından oluşan küme D'={d'_1, d'_2,…, d'_M}.

Çıktı: DokümanBenzerlikGrafı(D): düğümlerini dokümanların oluşturduğu, benzer dokümanların birer alt gra oluşturduğu birden fazla alt graf

Düğümleri $d_j \in D$ olan bir ağ kurulur

for each $d'_j \in D'$ **do**
    d'_j'yi tf-idf ağırlık matrisi ile temsil et
end for
for each $d'_j \in D'$ **do**
    maksimum(d'_j)= Ø
    benzerlik(d'_j)= -∞
    for each $d'_t \in D'$ **do**
        if j≠t **then**
            d'_j ile d'_t arasındaki cosinüs benzerliği tf-idf ağırlık matrisi üzerinden hesaplanır
            if benzerlik(d'_j,d'_t)>eşik değeri **and** benzerlik(d'_j, d'_t)> benzerlik(d'_j) **then**
                maksimum(d'_j)= d'_t, benzerlik(d'_j)= benzerlik(d'_j, d'_t) olarak ata
            end if
        end if
    end for
end for
for each $d_j \in D$ **do**
    if maksimum(d'_j)≠Ø **then**
        maksimum(d'_j)=d'_t'i bul
        d_j ile d_t arasında grafı çiz
    end if
end for

Şekil 3.10. Benzerlik grafı oluşturma algoritmasına ait sözde kod

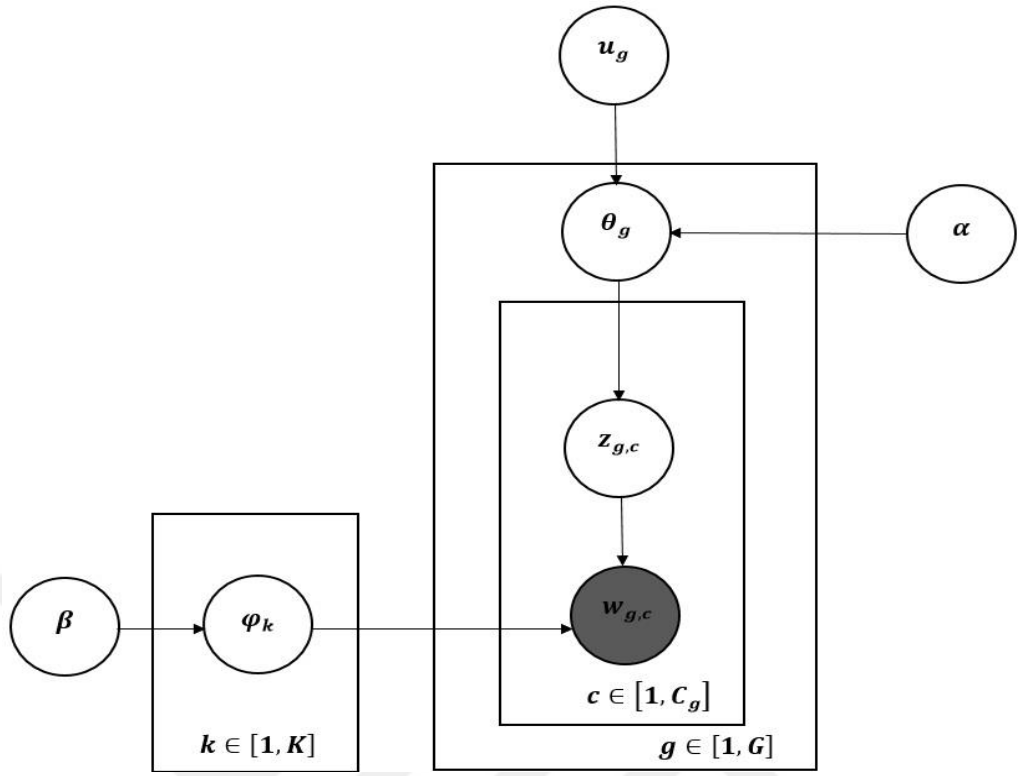
Şekil 3.10'da yer alan algoritma incelendiğinde grafın oluşturulması şu şekilde açıklanmaktadır: ilk olarak D' kümesinin elemanlarının her birinin tek düğümlü graflar olarak yer aldığı bir ağ oluşturulur. Kavram ve adlandırılmış varlıklardan oluşan bu ağ algoritmaya girdi olarak verilir. Ağdaki her bir eleman tf-idf ağırlıklandırma matrisi ile temsil edilip, elemanlar arasındaki benzerlik için cosinüs benzerliği kullanılır. Eğer iki eleman arasındaki benzerlik eşik değerinden büyükse ve d'_j'nin maksimum benzerlik gösterdiği eleman d'_t ise d_j ile d_t arasında graf çizilir ve bu iki düğüm arasındaki benzerlik yeni karşılaştırmalar için eşik değer olarak kullanılır. Sonuç olarak; tek düğümlü grafların, iki düğümlü grafların ve çok düğümlü grafların oluşturduğu bir ağ elde edilmiş olur. Her bir alt graftaki düğümler kendi içlerinde birleştirilerek yeni doküman uzayı MD={md_1, md_2, …, md_G} oluşturulmuştur. Yeni doküman koleksiyonu

MD'deki eleman sayısı alt graf sayısı olacak şekilde M'den G'ye azalmıştır. Ayrıca her alt graftaki elaman sayısını temsil eden bilinen temel ölçüt u'da bu ağ üzerinden elde edilmiştir. Hem yeni doküman kümesi MD hem de MD'deki her eleman md_i için doküman-konu dağılımına etki edecek u_i değerleri NET-LDA'nın ikinci modülü olan LDA'ya girdi olarak gönderilecektir.

3.2.5. Konu Çıkarımı

Dirichlet önsellerinin LDA'nın üzerindeki etkisinin çok az olduğu genel kanısından ötürü literatürdeki çalışmaların pek çoğunda LDA'nın simetrik önseller kullanılarak uygulandığı görülmüştür. Ancak Wallach ve diğ. (2009) yaptıkları çalışma ile doküman-konu dağılımında asimetrik önsellerin kullanılmasının modelin performansında büyük etkisinin olduğunu kanıtlamışlardır. Dirichlet önsellerinin optimizasyonun performans iyileştirmedeki etkisi ve benzer dokümanların benzer konu dağılımı göstermeleri göz önünde bulundurulmuş ve NET-LDA'da asimetrik Dirichlet önselleri kullanılmıştır. Asimetrik önsellerin kullanılmasının nedeni ise asimetrik önseller ile niceliksel ve niteliksel açıdan daha başarılı bir modelin elde edilmesidir.

Tasarlanan NET-LDA yapısının ikinci modülü Şekil 3.10'da verilmiştir. NET-LDA'nın sonsal dağılım üzerinden örneklemesi adımında CGS kullanılmıştır. Örnekleme adımı: (I) Sabit sözlük V'deki kelimelerin konular altında örneklenmesi simetrik Dirichlet önseli β'dan elde edilen φ ile temsil edilmektedir, (II) her doküman için her konunun ilgili dokümanda bulunma olasılığı θ Wallach ve diğ. (2009) göz önünde bulundurularak asimetrik Dirichlet önseline göre örneklenmiştir. Dirichlet önseli α'nın asimetrik olması ise her MD'de bulunan her doküman md_i için bilinen temel ölçüt u_i'nin her konu atamasına kenar üzerinden dahil edilmesi ile sağlanmıştır. (III) Dokümanda yer alan her kelime için konular çok terimli dağılıma göre örneklenmektedir ve (IV) son olarak da ilgili konu için kelime çok terimli dağılıma göre örneklenmektedir.



Şekil 3.11. NET-LDA'nın grafiksel temsili

Şekil 3.11'de yer alan parametrelerin açıklamaları Tablo 3.4'te verilmiştir.

Tablo 3.4. NET-LDA parametreleri

Parametre	Açıklaması
G	MD kümesinde yer alan doküman sayısı
K	Gizli konuların sayısı
C_g	g. dokümandaki kelime sayısı
α, β	Dirichlet hiperparametreleri
u_g	g. doküman için bilinen temel ölçüt
θ_g, φ_k	Sırasıyla konu ve kelimeler için çok terimli dağılımın parametreleri
$z_{g,c}$	g. dokümandaki c. kelimenin gizli konusu
$w_{g,c}$	g. dokümandaki c. kelime

Verilen modele göre, θ Dirichlet hiperparametresi α ve bilinen temel ölçüt u'ya göre asimetrik bir şekilde elde edilmiştir. z, α ve u verilmişken g. dokümandaki c. kelime için k. konunun koşullu sonsal olasılığı Eşitlik (3.1)'deki gibidir.

$$p\left(z_{g,c} = k \mid z_{g,-c}, \alpha u\right) = \int d\theta_g p\left(k \mid \theta_g\right) p\left(\theta_g \mid z_{g,-c}, \alpha u\right) = \frac{c_{g,k} + \alpha u}{C_g - 1 + K\alpha u} \tag{3.1}$$

$p(z_{g,c}=k|\ z_{g,-c},\alpha u)$ doküman g'deki örnekleme yapılacak kelime $w_{g,c}$'nin her konuya diğer tüm kelimeler $w_{g,-c}$ üzerinden koşullandırılmış atanma olasılığını temsil etmektedir. Eğer k. konu g. dokümanda yer almıyorsa $c_{g,k}$ yani g. dokümanda k. konuya atanan kelime sayısı sıfır olacaktır. αu bu durumda $c_{g,k}$'yı yumuşatmış olacaktır. Paydada g. dokümandaki toplam kelime sayısı C_g'den 1 çıkartılmasının nedeni ise örnekleme yapılacak kelime $w_{g,c}$'nin dahil edilmemesinden kaynaklanmaktadır. Eşitlik (3.1)'e dayalı olarak elde edilen tam koşullu sonsal olasılık $p(z_{g,c}=k|z_{g,-c},w)$ Eşitlik (3.2)'ye göre hesaplanmaktadır.

$$p\left(z_{g,c}=k\mid z_{g,-c},w\right)=\frac{c_{g,k}+\alpha u}{C_g-1+K\alpha u}\frac{c_{w,k}+\beta}{\sum_{w^{'}\in V}c_{w^{'},k}+V\beta} \tag{3.2}$$

Eşitlik (3.2)'nin sağ tarafında yer alan ikinci çarpan k. konunun g. dokümanda bulunma olasılığını vermektedir. $c_{w,k}$, tüm koleksiyonda w. kelimenin konu k'ya kaç kere atandığının sayısını vermektedir. $c_{w`,k}$ ise k. konunun w kelimesi hariç tüm koleksiyonda kaç kere kullanıldığını ifade etmektedir.

4. DENEYSEL ÇALIŞMA

Tez çalışmasının bu bölümünde Concept-LDA'yı ve NET-LDA'yı hem niceliksel hem de niteliksel değerlendirmek ve üç temel yöntemle karşılaştırmak amacıyla yapılan deneysel çalışma ayrıntılı bir şekilde anlatılmıştır. Modellerin niceliksel değerlendirilmesi adımında konu uyumluluğu ve F-skor bir ölçüt olarak kullanılırken, niteliksel değerlendirme adımında konular altında üretilen ürün özelliklerinin anlamsal olarak ilişkilerini ve detayları yakalayabilme yetenekleri incelenmiştir. Alt bölümlerde kullanılan veri kümeleri ve bu kümelere ait özet bilgiler, karşılaştırmada kullanılan temel yöntemler, parametre değerleri, değerlendirme ölçütleri ve tüm bunlar ışığında elde edilen sonuçlar sırasıyla verilmiştir.

4.1. Veri Kümeleri

Concept-LDA sadece İngilizce veri kümeleri için geliştirilen bir yöntem iken, NET-LDA dilden bağımsız bir yöntemdir. Concept-LDA'nın değerlendirilmesinde 12 farklı İngilizce veri kümesi kullanılmakta olup, NET-LDA için bu on iki veri kümesine ek bir adet Türkçe veri kümesi kullanılmıştır. Türkçe veri kümesi bir turizm web sitesi olan www.otepuan.com'dan elde edilmiştir. Belirli bir otel için yapılan kullanıcı yorumlarını içeren bu veri kümesinde toplam 1517 adet yorum yer almaktadır (Ekinci ve İlhan Omurca, 2018c). İngilizce veri kümeleri ise iki farklı kaynaktan sağlanmıştır. Bu veri kümelerinden ilki Jo ve Oh (2011) tarafından kendi çalışmalarında kullanılan Restoran ile ilgili yorumlar olup, ünlü web sitesi www.yelp.com'dan elde edilmiştir (URL-5, 2018). Buradaki veri kümesinin tamamı (25459 yorum) kullanılmamış olup, American (New) kategorisinde yer alan 2647 kullanıcı yorumu kullanılmıştır. Geri kalan on bir tanesi ise www.amazon.com'dan elde edilmiş, elektronikle ilgili farklı alan ve sayıdaki kullanıcı yorumlarını içermektedir (Chen ve Liu, 2014a). Veri kümelerine ait alan bilgisi, hangi dilde oldukları bilgisi, yorum sayısı, cümle sayısı, farklı kelime ve eşdizim sayısı, toplam kelime ve eşdizim sayısı, yorum başına düşen ortalama kelime ve eşdizim sayısı ile ortalama cümle sayısı ve cümle başına düşen ortalama kelime sayısı şeklindeki özet bilgiler Tablo 4.1'de verilmiştir.

Tablo 4.1. Veri kümelerine ait özet bilgiler

Alan	**Dil**	**Yorum Sayısı**	**Cümle Sayısı**	**Kelime Sayısı**	**Eşdizim Sayısı**	**Toplam Kelime ve Eşdizim Sayısı**	**Yorum Başına Ortalama Kelime ve Eşdizim Sayısı**	**Yorum Başına Ortalama Cümle Sayısı**	**Cümle Başına Düşen Ortalama Kelime ve Eşdizim Sayısı**
Otel	Türkçe	1517	6780	505	421	15924	10,5	4,47	2,35
Restaurant	İngilizce	2647	33653	5451	1982	88779	33,54	12,71	2,64
Alarm Clock	İngilizce	5113	5113	694	385	12501	2,44	1	2,44
Amplifier	İngilizce	5731	5731	923	712	17404	3,04	1	3,04
Battery	İngilizce	4056	4056	549	251	9592	2,36	1	2,36
Blu Ray Player	İngilizce	9170	9170	1041	883	26991	2,94	1	2,94
Cable Modem	İngilizce	5754	5754	691	457	15282	2,66	1	2,66
Camcorder	İngilizce	8361	8361	1049	898	25299	3,03	1	3,03
Camera	İngilizce	8958	8958	1218	980	27290	3,05	1	3,05
Car Stereo	İngilizce	5587	5587	761	518	15928	2,85	1	2,85
Cell Phone	İngilizce	5713	5713	924	608	15277	2,67	1	2,67
Computer	İngilizce	9090	9090	1251	1140	27764	3,05	1	3,05
DVD Player	İngilizce	6256	6256	871	647	16473	2,63	1	2,63

Veri kümelerini oluşturan kullanıcı yorumlarından bazılarının orijinal hali Şekil 4.1'de verilmiştir.

Şekil 4.1. Otel veri kümesi yorumları (a), Restaurant veri kümesi yorumları (b), Computer veri kümesi yorumları (c)

Concept-LDA'nın orijinal veri kümesine uygulanabilmesi için veri kümesinin kavram ve adlandırılmış varlıklar ile genişletilmesi gerekmektedir. Kavram ve adlandırılmış varlıklar ile genişletilen veri kümelerine ait özet bilgiler ise Tablo 4.2'de verilmiştir.

Tablo 4.2. Concept-LDA için genişletilen veri kümelerine ait özet bilgiler

Alan	Veri Kümesi	Kelime Sayısı	Eşdizim Sayısı	Toplam Kelime ve Eşdizim Sayısı
Restaurant	Orijinal	5451	1982	88779
	Genişletilmiş	14025	6784	383126
Alarm Clock	Orijinal	694	385	12501
	Genişletilmiş	2403	1186	44797
Amplifier	Orijinal	922	712	17404
	Genişletilmiş	3235	2081	71100
Battery	Orijinal	548	251	9592
	Genişletilmiş	1901	839	35860
Blu Ray Player	Orijinal	1040	883	26991
	Genişletilmiş	3296	2373	116180
Cable Modem	Orijinal	690	457	15282
	Genişletilmiş	2375	1376	61787
Camcorder	Orijinal	1049	898	25299
	Genişletilmiş	3420	2328	97990
Camera	Orijinal	1217	980	27290
	Genişletilmiş	3904	2601	103623
Car Stereo	Orijinal	760	518	15928
	Genişletilmiş	2640	1554	64915
Cell Phone	Orijinal	923	608	15277
	Genişletilmiş	3063	1855	59967
Computer	Orijinal	1251	1140	27764
	Genişletilmiş	3893	2920	113967
DVD Player	Orijinal	871	647	16473
	Genişletilmiş	2980	1879	64611

Her kelime için en az bir tane kavram ya da adlandırılmış varlık olduğunu göz önünde bulundursak Tablo 4.2'de de görüldüğü üzere kavram ve adlandırılmış varlıklar ile genişletilen dokümanlardaki toplam kelime sayısı üç ya da dört katına çıkmaktadır.

NET-LDA'da ise kavram ve adlandırılmış benzerliğine dayalı olarak orijinal dokümanlar birleştirilip koleksiyonların boyutu azaltılmaktadır. Dokümanların birleştirilmesi için anlamsal benzerliklerinin 0,5'ten büyük olması gerekmektedir. Anlam tabanlı birleşme sonucu koleksiyonlardaki toplam doküman sayıları Tablo 4.3'te verilmiştir.

Tablo 4.3. NET-LDA için genişletilen veri kümelerine ait özet bilgiler

Alan	Yorum Sayısı (Orijinal)	Yorum Sayısı (Birleştirilmiş)
Otel	1517	494
Restaurant	2647	2346
Alarm Clock	5113	1268
Amplifier	5731	1341
Battery	4056	1004
Blu Ray Player	9170	2052
Cable Modem	5754	1343
Camcorder	8361	1716
Camera	8958	2112
Car Stereo	5587	1347
Cell Phone	5713	1435
Computer	9090	1941
DVD Player	6256	1560

Veri kümeleri için maksimum ve minimum birleşen doküman sayıları Tablo 4.4'te verilmiştir.

Tablo 4.4. NET-LDA'da veri kümelerindeki maksimum ve minimum birleşen doküman sayıları

Alan	Maksimum Birleşme Sayısı	Minimum Birleşme Sayısı
Otel	28	1
Restaurant	34	1
Alarm Clock	81	1
Amplifier	166	1
Battery	79	1
Blu Ray Player	104	1
Cable Modem	62	1
Camcorder	246	1
Camera	88	1
Car Stereo	165	1
Cell Phone	72	1
Computer	245	1
DVD Player	84	1

4.2. Karşılaştırma Amaçlı Kullanılan Konu Modelleri

Geliştirilen yöntemleri karşılaştırmak amacıyla LDA, LTM ve AMC olmak üzere üç temel konu modeli kullanılmıştır. Bu modellerin seçilme nedeni ise başka çalışmalarda da karşılaştırma amaçlı kullanılmış olmalarıdır (Shams ve Baraani-Dastjerdi, 2017).

- LDA: Blei ve diğ. tarafından (2003) geliştirilen LDA, konuların dokümanlardaki dağılımını temsil eden parametrelerini bir Dirichlet dağılımından gelen değişkenler olarak ele almaktadır. En temel konu modellerinden biri olan LDA dokümanlardan konuları çıkartırken herhangi bir önbilgiye ihtiyaç duymamaktadır.
- LTM: LDA tabanlı olan bu konu modeli bir "yaşam boyu öğrenme" algoritması olup, büyük veriler ile çalışmaktadır. Klasik makine öğrenmesi yöntemleri tek bir görevi tek başına öğrenirken; yaşam boyu öğrenme algoritmaları öğrenme süreci boyunca pek çok öğrenme görevi ile karşı karşıya kalmaktadır (Thrun, 1998). Buradaki yaşam boyu öğrenme must-link (ML)'lerin otomatik olarak öğrenilmesi ve öğrenilen bu ML'leri önbilgi olarak kullanarak uyumlu ve anlamlı konuların elde edilmesi şeklinde işlemektedir (Chen ve Liu, 2014a).
- AMC: AMC'de LTM gibi bir "yaşam boyu öğrenme" algoritması olup, ML'lere ek olarak otomatik olarak çıkartılan cannot-link (CL)'leri de önbilgi olarak kullanarak büyük veriler üzerinden başarılı konular elde etmeyi amaçlamaktadır (Chen ve Liu, 2014b).

4.3. Modelleri Değerlendirmede Kullanılan Parametre Değerleri

Literatürdeki çalışmalar incelendiğinde model parametrelerinin belirlenmesi adımında bir gerekçe sunulmadığı gözlemlenmiştir (Wei ve Croft, 2006; Lu ve diğ., 2011). Bu nedenle beş modelin hepsi için CGS 50, 100, 200, 500 ve 1000 iterasyon olarak gerçeklenmiştir. Concept-LDA, LDA, LTM ve AMC'de simetrik Dirichlet parametreleri kullanılmışken, NET-LDA'da doküman-konu dağılımında asimetrik Dirichlet parametreleri kullanılmıştır. α değeri 50/K olarak belirlenmiş olup, β değeri 0.01 olarak belirlenmiştir. NET-LDA'da α parametresinin asimetrik olarak kullanılması adımında ise α bilinen temel ölçüt u ile çarpılmıştır. Yani α 50u/K olarak kullanılmıştır. α ve β için belirlenen bu değerler başka doküman koleksiyonlarında da kullanılmış ve bu parametre değerleri ile başarılı sonuçların elde edildiği gözlemlenmiştir (Steyvers ve Griffiths, 2007; Zheng ve diğ., 2014). Tez çalışması kapsamında yapılan deneylerde, konu sayısı 100 olarak tanımlanmış olup, her konu ilk 10 kelimesi ile temsil edilmiştir.

4.4. Değerlendirme Ölçütleri

NET-LDA ve Concept-LDA ile çıkartılan konuların kendi içlerindeki anlamsal olarak uyumluluğunu ölçmek ve bu modelleri LDA, LTM ve AMC ile karşılaştırmak için konu uyumluluğu (Chen ve Liu, 2014b) ve kesinlik (p), duyarlılık (r) ve F-skoru kullanılmıştır.

Mimno ve diğ. (2011) tarafından önerilen konu uyumluluğu çıkartılan konuların uzmanların verdiği kararlar ile ne kadar ilişkili olduğunu göstermektedir. Konu uyumluluğu Eşitlik (4.1) ile verilmiştir.

$$C\left(k; V^{(k)}\right) = \sum_{n=2}^{N} \sum_{l=1}^{n-1} \log \frac{D\left(v_n^{(k)}, v_l^{(k)}\right) + 1}{D\left(v_l^{(k)}\right)} \tag{4.1}$$

Eşitlik (4.1)'de $V^{(k)}=(v_1^k, v_2^k, \ldots, v_S^k)$ k. konudaki en olası S kelimeyi temsil etmektedir. Bu tez çalışmasında S 10 olarak belirlenmiştir. $D(v_n^{(k)}, v_l^{(k)})$; v_n ve v_l kelimelerinin kaç adet dokümanda birlikte geçtiklerinin sayısını göstermektedir. 1 ise yumuşatma amaçlı olarak paya eklenmiştir. $D(v_l^{(k)})$; v_l kelimesinin kaç adet dokümanda bulunduğunun sayısını vermektedir. Konu uyumluluğunun yüksek olması, çıkartılan konulardaki kelimelerin birbirleri ile olan anlamsal ilişkilerinin yüksek olduğunu göstermektedir.

Modelin performansını kesinlik, duyarlılık ve F-skoru açısından değerlendirebilmek için dokümanlardan uzmanlar tarafından çıkartılan ürün özellikleri kullanılmaktadır. Kesinlik; uzmanlar tarafından belirlenen ürün özellikleri ile ilgili konu modeli tarafından çıkartılan tüm kelimelerin kesişimi şeklinde hesaplanmaktadır. Duyarlılık; uzmanlar tarafından belirlenen ürün özellikleri ile ilgili konu modeli tarafından çıkartılan ürün özelliklerinin kesişimi şeklinde hesaplanmaktadır. F-skoru ise kesinlik ve duyarlılığın harmonik ortalaması olarak hesaplanmaktadır. Kesinlik (p), duyarlılık (r) ve F-skoru hesaplanmasında kullanılan formüller sırasıyla Eşitlik (4.2), (4.3) ve (4.4)'te verilmiştir.

$$p = \frac{t_a}{t_w} \tag{4.2}$$

$$r = \frac{t_a}{a_w} \tag{4.3}$$

$$F - skoru = \frac{2 \times p \times r}{p + r} \tag{4.4}$$

Yukarıdaki eşitliklerde a_w, veri kümelerinden uzmanlar tarafından çıkartılan ürün özelliklerini temsil etmektedir. t_w, konu modelleri tarafından çıkartılan kelimelerdir. t_a ise a_w ile t_w'nin kesişiminden elde edilen ürün özellikleridir.

Modelin niteliksel değerlendirilmesi adımında ise çıkartılan her konunun ilk 10 kelimesi değerlendirilmeye alınmıştır.

4.5. Deneysel sonuçlar

Anlamsal olarak ilişkili, uyumlu, tutarlı, detayları yakalayabilen ve daha başarılı konular elde edebilmek için geliştirilen Concept-LDA ve NET-LDA'nın niceliksel ve niteliksel olarak performanslarının değerlendirilmesi ve üç temel yöntemle karşılaştırılabilmesi için yöntemler 13 farklı veri kümesine uygulanmıştır. Elde edilen konu uyumluluğu 5 model üzerinden İngilizce veri kümeleri ve 4 model üzerinden Türkçe veri kümesi için Tablo 4.5'te verilmiştir.

Tablo 4.5. Her bir veri kümesi için her bir yöntemden farklı iterasyon sayıları ile elde edilen konu uyumluluğu değerleri

Alan	**Yöntem**	**İterasyon Sayısı**				
		50	**100**	**200**	**500**	**1000**
Otel	NET-LDA	-84,346	-84,716	-86,201	-81,922	-81,88
	LDA	-108,94	-102,205	-101,992	-102,315	-102,378
	LTM	-111,978	-112,612	-112,864	-106,925	-100,223
	AMC	-112,324	-106,187	-100,598	-90,757	-83,925
Restaurant	NET-LDA	-103,815	-103,23	-102,49	-102,169	-102,990
	Concept-LDA	-81,633	-81,989	-82,401	-84,012	-82,783
	LDA	-104,369	-106,981	-105,608	-105,91	-108,606
	LTM	-110,478	-107,817	-106,342	-102,933	-101,676
	AMC	-105,673	-102,569	-100,972	-99,295	-99,514

Tablo 4.5. (Devam) Her bir veri kümesi için her bir yöntemden farklı iterasyon sayıları ile elde edilen konu uyumluluğu değerleri

Alan	**Yöntem**	**İterasyon Sayısı**				
		50	**100**	**200**	**500**	**1000**
Alarm Clock	NET-LDA	-72,258	-70,537	-71,404	-70,816	-71,314
	Concept-LDA	-108,63	-105,57	-104,332	-104,254	-104,231
	LDA	-113,197	-113,22	-110,995	-111,589	-112,202
	LTM	-116,746	-116,789	-120,093	-127,771	-134,342
	AMC	-119,468	-120,97	-121,479	-123,729	-120,890
Amplifier	NET-LDA	-79,711	-78,123	-77,911	-77,83	-77,37
	Concept-LDA	-114,52	-113,892	-113,301	-112,524	-113,993
	LDA	-116,288	-115,117	-115,98	-115,884	-115,491
	LTM	-118,264	-117,795	-120,297	-126,483	-129,763
	AMC	-114,26	-116,792	-119,676	-124,074	-123,822
Battery	NET-LDA	-68,11	-66,991	-66,068	-65,522	-67,623
	Concept-LDA	-102,016	-99,558	-98,3	-99,448	-97,886
	LDA	-109,043	-108,124	-106,723	-106,633	-107,593
	LTM	-116,199	-114,49	-122,403	-129,426	-133,088
	AMC	-118,037	-121,252	-125,023	-130,177	-127,348
Blu Ray Player	NET-LDA	-88,319	-87,126	-85,495	-84,827	-85,354
	Concept-LDA	-129,483	-126,421	-126,61	-126,552	-125,449
	LDA	-134,324	-130,874	-128,905	-127,428	-128,262
	LTM	-132,893	-129,346	-132,324	-132,71	-133,711
	AMC	-127,572	-129,481	-129,602	-128,264	-133,507
Cable Modem	NET-LDA	-78,075	-79,058	-77,423	-76,776	-75,29
	Concept-LDA	-112,895	-110,31	-110,358	-109,225	-108,694
	LDA	-118,471	-117,214	-118,26	-118,019	-115,123
	LTM	-121,718	-119,659	-124,934	-128,118	-129,965
	AMC	-119,46	-119,748	-120,969	-124,071	-124,181

Tablo 4.5. (Devam) Her bir veri kümesi için her bir yöntemden farklı iterasyon sayıları ile elde edilen konu uyumluluğu değerleri

Alan	**Yöntem**	**İterasyon Sayısı**				
		50	**100**	**200**	**500**	**1000**
Camcorder	NET-LDA	-87,156	-86,892	-85,523	-84,18	-84,747
	Concept-LDA	-123,24	-124,013	-122,526	-121,558	-122,164
	LDA	-127,186	-126,434	-125,202	-125,473	-122,937
	LTM	-129,62	-126,313	-127,899	-128,89	-129,894
	AMC	-119,945	-119,4	-122,108	-123,392	-125,799
Camera	NET-LDA	-89,236	-89,119	-88,089	-87,608	-85,755
	Concept-LDA	-126,398	-124,308	-124,245	-122,64	-123,314
	LDA	-130,377	-129,095	-126,608	-126,216	-126,188
	LTM	-130,338	-126,62	-129,139	-130,399	-131,623
	AMC	-118,395	-116,867	-118,363	-117,656	-121,324
Car Stereo	NET-LDA	-78,95	-78,676	-78,093	-76,852	-76,081
	Concept-LDA	-112,797	-111,2963	-111,28	-110,224	-109,779
	LDA	-119,614	-117,808	-118,229	-118,776	-115,997
	LTM	-122,314	-120,23	-125,191	-126,797	-131,007
	AMC	-118,358	-120,91	-124,64	-125,851	-124,864
Cell Phone	NET-LDA	-76,657	-75,583	-74,18	-73,846	-74,454
	Concept-LDA	-111,051	-108,835	-106,954	-107,773	-107,162
	LDA	-111,755	-112,42	-111,466	-110,767	-112,471
	LTM	-116,708	-115,631	-122,242	-125,053	-127,648
	AMC	-112,018	-113,078	-114,56	-119,892	-123,56
Computer	NET-LDA	-91,759	-89,962	-88,934	-89,28	-87,601
	Concept-LDA	-128,001	-125,429	126,122	127,39	-125,416
	LDA	-129,215	-127,61	-125,15	-126,09	-125,066
	LTM	-130,246	-126,941	-128,646	-128,43	-131,037
	AMC	-122,581	-124,344	-126,229	-129,993	-131,107

Tablo 4.5. (Devam) Her bir veri kümesi için her bir yöntemden farklı iterasyon sayıları ile elde edilen konu uyumluluğu değerleri

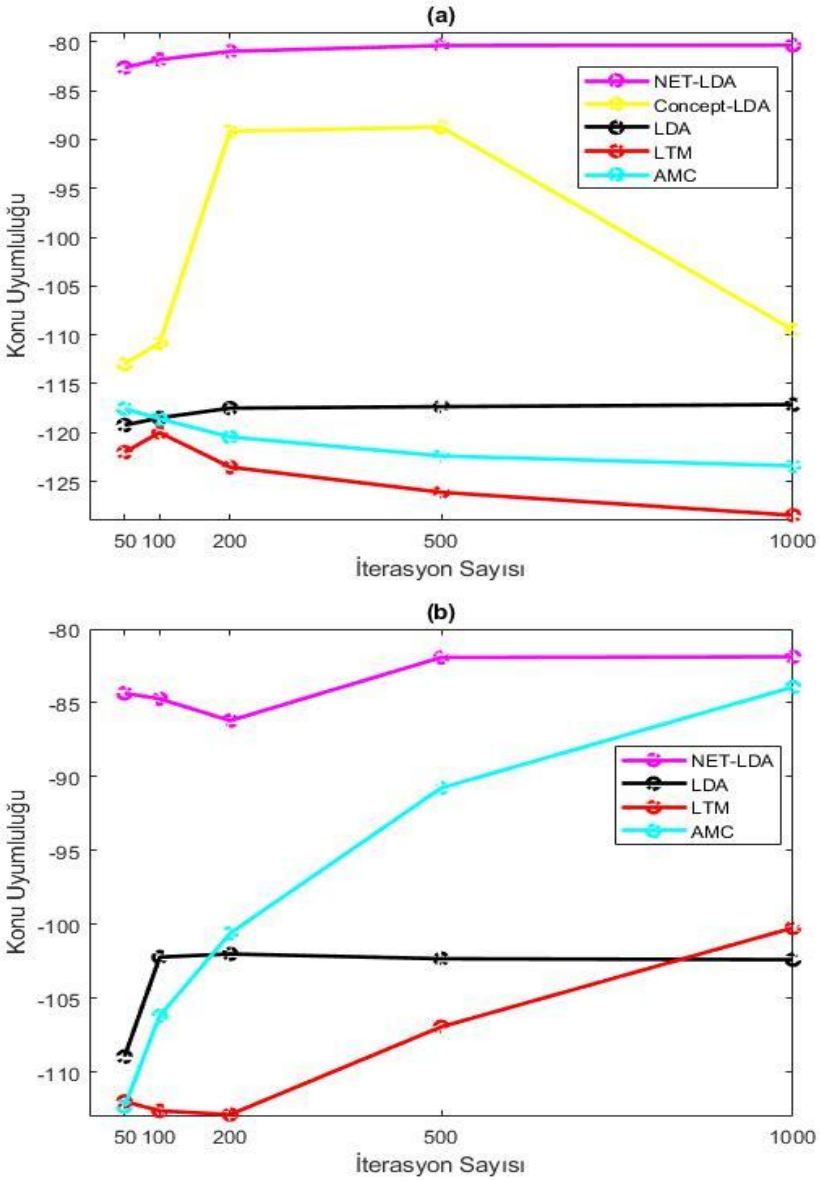
Alan	**Yöntem**	**İterasyon Sayısı**				
		50	**100**	**200**	**500**	**1000**
DVD Player	NET-LDA	-77,579	-76,424	-76,004	-74,783	-75,533
	Concept-LDA	-105,187	-98,183	-95,669	-93,76	-92,93
	LDA	-117,019	-116,981	-116,848	-115,366	-115,671
	LTM	-118,663	-118,291	-122,889	-126,343	-127,845
	AMC	-114,856	-117,736	-121,700	-122,303	-124,966

İngilizce 13 veri kümesi üzerinden için ortalama konu uyumlulukları iterasyonlara göre Tablo 4.6'te verilmiştir.

Tablo 4.6. İngilizce veri kümeleri için yöntemler üzerinden ortalama konu uyumluluğu

		İterasyon Sayısı				
		50	100	200	500	1000
Yöntem	**NET-LDA**	-82,635	-81,810	-80,968	-80,374	**-80,343**
	Concept-LDA	-112,988	-110,817	-89,155	-88,715	-109,483
	LDA	-119,238	-118,49	-117,498	-117,346	-117,134
	LTM	-122,016	-119,994	-123,533	-126,113	-128,467
	AMC	-117,552	-118,596	-120,443	-122,391	-123,407

İngilizce veri kümeleri için NET-LDA, Concept-LDA, LDA, LTM ve AMC olmak üzere 5 model üzerinden ortalama konu uyumluluğu, Türkçe veri kümesi için ise NET-LDA, LDA, LTM ve AMC olmak üzere 4 model üzerinden konu uyumluluğu Şekil 4.2'de verilmiştir.



Şekil 4.2. Her LDA modeli için İngilizce veri kümeleri üzerinden ortalama konu uyumluluğu (a) Türkçe veri kümesi üzerinden konu uyumluluğu (b)

Tablo 4.5 ve 4.6 ile Şekil 4.2 incelendiğinde aşağıdaki sonuçlara varılmıştır:

- NET-LDA yöntemi hem İngilizce hem de Türkçe veri kümeleri için konu uyumluluğu açısından değerlendirildiğinde; yöntemin diğer yöntemlere kıyasla oldukça başarılı olduğu görülmektedir. Bu da kelimelerin birlikte geçme durumlarını anlamsal olarak güçlendirmenin konu modelleri üzerindeki etkisinin önemli olduğunu ve modelin performansını arttırdığını göstermektedir. Elde edilen sonuçlara dayanarak, NET-LDA ile diğer yöntemlere göre anlamsal olarak ilişkili, daha uyumlu, tutarlı, detayları yakalayabilen ve daha başarılı konuların elde edildiği görülmektedir.

- Concept-LDA sadece İngilizce veri kümelerine uygulanmış olup, NET-LDA'dan sonra başarılı olan ikinci yöntemdir. Bu da kelime torbası yerine {kelime+kavram+adlandırılmış varlıklar} torbasının kullanılmasının modelin genelleştirme performansına olumlu yönde etkisi olduğunu göstermektedir.
- İngilizce veri kümesi için en başarısız yöntemlerin LTM ve AMC olduğu gözlenmiştir. Bunun nedeni; bu iki yöntemin önbilgi olarak sık geçen öğe kümelerinden yararlanmalarıdır. Sık geçen öğelerin önbilgi olarak kullanılmasının sonucunda; anlamsal olarak ilişkili konuların elde edilmesi yerine sık geçen kelimelerin hemen hemen her konuda yer alması problemi ortaya çıkmaktadır. Dolayısıyla konuların ayırt ediciliği ve kalitesi düşmektedir. Computer veri kümesinde yer alan "charge" kelimesi bu duruma örnek olarak verilebilir. "charge" kelimesi hem LTM hem de AMC ile üretilen konuların büyük çoğunluğunda yer almaktadır.
- Türkçe veri kümesi için konu uyumluluğu incelendiğinde LTM ve AMC'nin iterasyon sayısı arttıkça konu uyumluluğu performanslarının arttıkları gözlemlenmiştir. Bu amaçla Türkçe veri kümesi için NET-LDA'da CGS 2000 iterasyonda da çalıştırılmıştır. 2000 iterasyon sonucunda NET-LDA'dan elde edilen konu uyumluluğu -84,135'e düşerken; LTM -90,285'e, AMC ise -76,519'a yükselmiştir. LDA ise -103,154'e düşmüştür. Bu durumda 2000 iterasyon sonucunda AMC'nin NET-LDA'yı geçtiği görülmüştür. Ancak literatürde geliştirilen diğer LDA modellerinde de genelde maksimum 1000 iterasyon üzerinden değerlendirme yapıldığı için bu çalışmada da 1000 iterasyon üzerinden değerlendirme yapılmıştır (Bagheri ve diğ., 2014; Shams ve Baraani-Dastjerdi, 2017; Fu ve diğ., 2018).
- Tükçe veri kümesi için 1000 iterasyon üzerinden değerlendirme yapıldığında; AMC'nin LTM'den daha başarılı ve hızla yükselen bir performansa sahip olduğu gözlemlenmiştir. AMC LTM ile kıyaslandığında ML'lerin yanı sıra CL'leri de kullanmaktadır ve bu durumda CL'lerin Türkçe veri kümesi üzerinde başarıyı arttıran bir etkiye sahip olduğu söylenebilir.
- NET-LDA'nın ve Concept-LDA'nın LDA'ya göre daha başarılı olduğu gözlenmiştir, çünkü her iki yöntemde anlamsal olarak uyumlu konuları çıkartabilmektedir. LDA içinse bu durum bir dezavantajdır.

Modellerin niceliksel değerlendirilmesinde kullanılan diğer ölçütler ise kesinlik, duyarlılık ve F-skorudur. 1000 iterasyon sonucunda elde edilen konular altındaki

kelimeler uzmanlar tarafından belirlenen ürün özellikleri ile karşılaştırılmış ve bunun sonucunda elde edilen değerler Tablo 4.7'de verilmiştir.

Tablo 4.7. 1000 iterasyon sonucu elde edilen konu kelimeleri üzerinden kesinlik, duyarlılık ve F-skor değerleri

Alan	**Yöntem**	**Kesinlik**	**Duyarlılık**	**F-skoru**
Otel	NET-LDA	0,9	1	0,95
	LDA	0,84	0,94	0,89
	LTM	1	0,22	0,36
	AMC	1	0,12	0,21
Restaurant	NET-LDA	0,57	0,94	0,71
	Concept-LDA	0,55	1	0,71
	LDA	0,54	0,93	0,68
	LTM	0,57	0,90	0,7
	AMC	0,58	0,85	0,69
Alarm Clock	NET-LDA	0,81	1	0,9
	Concept-LDA	0,74	0,68	0,71
	LDA	0,75	0,78	0,76
	LTM	0,78	0,24	0,37
	AMC	0,78	0,15	0,25
Amplifier	NET-LDA	0,86	1	0,92
	Concept-LDA	0,86	0,98	0,92
	LDA	0,82	0,78	0,8
	LTM	0,88	0,33	0,48
	AMC	0,88	0,23	0,36
Battery	NET-LDA	0,77	1	0,87
	Concept-LDA	0,77	1	0,87
	LDA	0,76	0,85	0,8
	LTM	0,8	0,28	0,41
	AMC	0,79	0,16	0,27

Tablo 4.7. (Devam) 1000 iterasyon sonucu elde edilen konu kelimeleri üzerinden kesinlik, duyarlılık ve F-skor değerleri

Alan	Yöntem	Kesinlik	Duyarlılık	F-skoru
Blu Ray Player	NET-LDA	0,79	1	0,88
	Concept-LDA	0,78	0,95	0,86
	LDA	0,8	0,84	0,82
	LTM	0,87	0,48	0,62
	AMC	0,88	0,31	0,46
Cable Modem	NET-LDA	0,8	1	0,89
	Concept-LDA	0,98	1	0,99
	LDA	0,81	0,86	0,83
	LTM	0,83	0,37	0,51
	AMC	0,86	0,25	0,39
Camcorder	NET-LDA	0,63	1	0,77
	Concept-LDA	0,6	0,91	0,72
	LDA	0,64	0,85	0,73
	LTM	0,68	0,49	0,57
	AMC	0,76	0,33	0,46
Camera	NET-LDA	0,53	1	0,69
	Concept-LDA	0,55	1	0,71
	LDA	0,56	0,86	0,68
	LTM	0,64	0,51	0,57
	AMC	0,76	0,33	0,46
Car Stereo	NET-LDA	0,49	1	0,66
	Concept-LDA	0,5	1	0,67
	LDA	0,49	0,81	0,61
	LTM	0,58	0,39	0,47
	AMC	0,71	0,26	0,38

Tablo 4.7. (Devam) 1000 iterasyon sonucu elde edilen konu kelimeleri üzerinden kesinlik, duyarlılık ve F-skor değerleri

Alan	**Yöntem**	**Kesinlik**	**Duyarlılık**	**F-skoru**
Cell Phone	NET-LDA	0,49	1	0,66
	Concept-LDA	0,49	0,99	0,66
	LDA	0,5	0,84	0,63
	LTM	0,85	0,43	0,57
	AMC	0,68	0,24	0,35
Computer	NET-LDA	0,56	1	0,72
	Concept-LDA	0,53	0,77	0,63
	LDA	0,54	0,81	0,65
	LTM	0,61	0,5	0,55
	AMC	0,75	0,35	0,48
DVD Player	NET-LDA	0,47	1	0,64
	Concept-LDA	0,32	0,75	0,42
	LDA	0,46	0,78	0,58
	LTM	0,54	0,42	0,47
	AMC	0,62	0,25	0,36

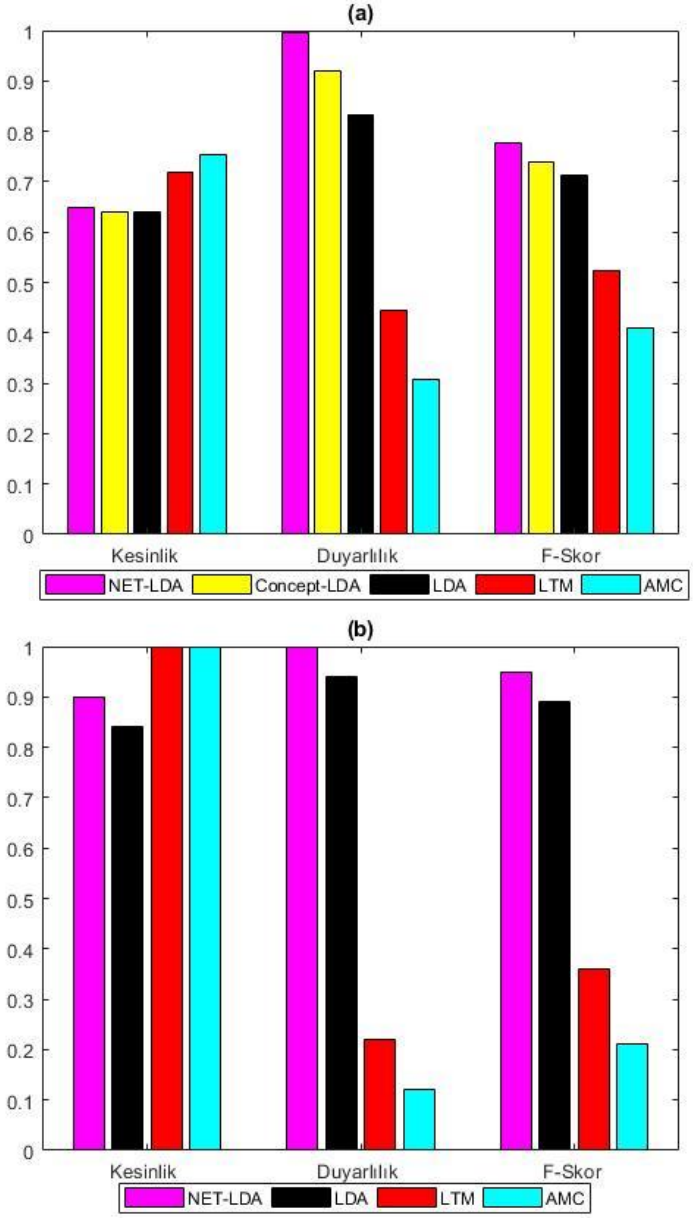
İngilizce veri kümeleri için ortalama kesinlik, duyarlılık ve F-skoru Tablo 4.8'de verilmiştir.

Tablo 4.8. İngilizce veri kümeleri için yöntemler üzerinden ortalama kesinlik, duyarlılık ve F-skoru

		Kesinlik	Duyarlılık	F-skoru
Yöntem	**NET-LDA**	0,648	0,995	**0,776**
	Concept-LDA	0,639	0,919	0,739
	LDA	0,639	0,833	0,714
	LTM	0,719	0,445	0,524
	AMC	0,754	0,309	0,409

İngilizce veri kümeleri için NET-LDA, Concept-LDA, LDA, LTM ve AMC olmak üzere 5 model üzerinden ortalama kesinlik, duyarlılık ve F-Skoru; Türkçe veri kümesi

için ise NET-LDA, LDA, LTM ve AMC olmak üzere 4 model üzerinden kesinlik, duyarlılık ve F-Skoru Şekil 4.3'te verilmiştir.



Şekil 4.3. Her LDA modeli için İngilizce veri kümeleri üzerinden ortalama kesinlik, duyarlılık ve F-Skoru (a) Türkçe veri kümesi üzerinden kesinlik, duyarlılık ve F-Skoru (b)

Modeller F-skora göre değerlendirildiklerinde İngilizce veri kümesi için NET-LDA; Concept-LDA'ya göre %3,7 oranında daha başarılı iken LDA'ya göre %6,2 daha başarılı olmuştur. NET-LDA LTM'ye göre %25,2 daha başarılı iken AMC'ye göre %36,7 daha başarılı olmuştur.

NET-LDA'nın ve Cocept-LDA'nın özellikle LTM ve AMC'ye oranla çok daha başarılı olmasının nedeni LTM ve AMC ile elde edilen konuların daha genel olmasıdır. Bunun yanında NET-LDA ve Concept-LDA ile daha ayrıntılı konular elde edilmektedir. Yine LTM ve AMC ile elde edilen konularda sık geçen kelimeler daha baskın şekilde yer almaktadır ancak NET-LDA ve Concept-LDA için böyle bir durum söz konusu değildir. NET-LDA ve Concept-LDA ile "brightness level", "battery power", "bluetooth keyboard" gibi ürün özellikleri elde edilirken LTM ve AMC ile bu ürün özellikleri çıkartılamamaktadır. Bunun nedeni NET-LDA kelimelerin birlikte geçme durumlarını anlamsal olarak güçlendirmekte olması; Concept-LDA'nın ise {kelime+kavram+adlandırılmış varlıklar} torbasını kullanması iken LTM ve AMC'nin önbilgi olarak yalnızca ML ve CL bilgileri kullanmalarıdır. Yani LTM ve AMC ile asıl amaçlanan sadece sık geçen ürün özelliklerinin yakalanmasıdır, ürün özellikleri üzerinde budama yapılmaktadır. Ancak sık geçen özellikler yanında sık geçmeyen ve birbiri yerine kullanılan ürün özelliklerinin de yakalanması doğru bir analiz için oldukça önemlidir. Ayrıca LTM ve AMC birlikte geçmesi ve geçmemesi kelimeleri sık geçen öğeler algoritması üzerinden belirleyip bunu modele vermekte iken önerilen yöntemlerde birlikte geçme durumu doküman benzerliği ve kavram ve adlandırılmış varlıklar üzerinden otomatik olarak yakalanmaktadır.

Türkçe veri kümesi üzerinden modellerin F-skor değerleri karşılaştırıldığında NET-LDA'nın LDA'dan %6 oranında, LTM'den %59 ve AMC'den %74 daha başarılı olduğu görülmüştür. LTM ve AMC ile personel ile ilgili sadece "personel", "personel davranışı" ve "personel ilgisi" ürün özellikleri yakalanmakta iken NET-LDA ile bunlara ek olarak "personel hizmeti", "personel kalitesi", "personel sayısı", "personel tavrı" ve "personel yaklaşımı" gibi ayrıntılı ürün özellikleri de yakalanmaktadır.

LDA NET-LDA ve Concept-LDA LDA ile her iki veri kümesi için karşılaştırıldığında F-skor açısından aralarında yüksek bir performans farkı olmadığı gözlemlenmektedir. Çünkü NET-LDA ve Concept-LDA anlamsal konuları bulurken bir budama işlemi gerçekleştirmemektedir. LDA da ürün özelliklerini çıkartırken herhangi bir budama gerçekleştirmemektedir.

Modelin niteliksel değerlendirilmesi adımında ise tüm yöntemlerden 1000 iterasyon sonucunda elde edilen konular altındaki kelimeler uzmanlar tarafından incelenmiş ve

elde edilen konulardan Otel veri kümesi ile ilgili olanlar Tablo 4.9'da, Restaurant ile ilgili olanlar Tablo 4.10'da, Cell Phone ile ilgili olanlar Tablo 4.11'de ve Computer ile ilgili olanlar Tablo 4.12'de verilmiştir. Konu ile ilişkisi olmayan kelimeler kırmızı ve italik olarak belirtilmiştir.

Tablo 4.9. Otel veri kümesinden elde edilen konulardan örnekler

Yeme İçme				**Yüzme**			
NET-LDA	**LDA**	**LTM**	**AMC**	**NET-LDA**	**LDA**	**LTM**	**AMC**
a la carte	restoran	restoran	içecek	*otel*	deniz	deniz	*personel*
restoran	a la carte	a la carte	*personel*	plaj	havuz	havuz	*yemek*
snack	*tadilat*	rezervasyon	*oda*	çim alan	*a la carte*	iskele	havuz
Rum a la carte	mini bar	şef	*konum*	aqua	şezlong	*et*	konum
gıda	ela ela	pastane	yiyecek	*personel*	konum	temizlik	*hizmet kalitesi*
balık restorantı	*oda hizmeti*	*ilgi*	*tesis*	manzara	aquapark	*hizmet*	*çalışan*
balık	ana yemek	çeşit	restoran	havuz olanakları	*pastane*	sahil	*animasyon*
stajyer	*komidin*	*ilgi alaka*	*hizmet kalitesi*	çim	valf	ağaçlık	imkan
yenilik	*otel doğası*	snack	yemek	konsept	beach club	konum	iskele
Rum eğlencesi	çocuk mutfağı	mutfak	a la carte	*Türk gecesi*	*çocuk girişimcisi*	*animasyon*	yüzme olanakları

Tablo 4.10. Restaurant veri kümesinden elde edilen konulardan örnekler

Salad					**Sandwich**				
NET-LDA	**Concept-LDA**	**LDA**	**LTM**	**AMC**	**NET-LDA**	**Concept-LDA**	**LDA**	**LTM**	**AMC**
salad	salad	salad	salad	salad	*lunch*	burger	sandwich	*lunch*	*lunch*
beet	moist	green	green	green	sandwich	sandwich	*soup*	sandwich	sandwich
vinaigrette	food	*tender*	*tender*	*tender*	lettuce	ingredient	salad	turkey	turkey
protein	*hit*	mashed potato	mashed potato	mashed potato	turkey	bun	cauliflower	lettuce	lettuce
combination	mixed greens	tuna	vegetable	chicken	*menu*	cake	turkey	*split*	*ok*
cucumber	eater	vegetable	*culver city*	vegetable	*fare*	beef	*lemonade*	steak sandwich	*work*
lemonade	flank	*culver city*	*worker*	*line*	steak sandwich	hamburger	*split*	*busboy*	steak sandwich
leaf	crouton	*chunk*	chipotle	*culver city*	*fill*	japan	*combo*	*ok*	*busboy*
crouton	spinach salad	*downtown*	flat iron steak	chipotle	sprout	onion	element	*split pea soup*	*tray*
baguette	vinaigrette	vegan	veggie	*chunk*	poblano	*medium*	*iced tea*	ahi tuna	*split*

Tablo 4.11. Cell Phone veri kümesinden elde edilen konulardan örnekler

Connection					**Camera**				
NET-LDA	**Concept-LDA**	**LDA**	**LTM**	**AMC**	**NET-LDA**	**Concept-LDA**	**LDA**	**LTM**	**AMC**
usb cable	usb cable	card	card	card	camera	picture	camera	image	*sound*
usb	usb	worth	micro	micro	image	image	picture	picture	picture
micro	usb port	micro	micro sd card	micro sd card	*hole*	movie	*case*	charge	*point*
sd card	usb cord	*metal*	*black*	*music*	*today*	video	battery life	*hour*	option
mode	usb plug	*buck*	cell	*number*	mp	*smartphone*	bug	crisp	*headset*
usb port	micro usb	error	*htc*	*charge*	megapixel	*radio*	*result*	*motorola*	*order*
micro sd card	usb device	micro sd card	speed	hardware	night	*antenna*	*buyer*	*iphone*	image
usb cord	*rock*	protection	cord	*windows*	tv	*audio*	*mark*	*longer*	*noise*
mind	*light weight*	*note*	functionality	tech support	*space*	*playback*	*internet access*	camera	*access*
minimal	device firmware upgrade	*default*	*ringtone*	*light*	*high*	video recording	*significant*	cable	*volume*

Tablo 4.12. Computer veri kümesinden elde edilen konulardan örnekler

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

Geliştirilen yöntemler her iki dile ait veri kümeleri için değerlendirildiğinde; niteliksel olarak NET-LDA ile elde edilen konuların, birbirlerine anlamca uyumlu oldukları, tutarlı oldukları, kolay etiketlenebilir oldukları ve ilgili özelliği başarılı bir şekilde temsil edebilme açısından daha başarılı oldukları görülmektedir.

Sadece İngilizce veri kümelerine uygulanan Concept-LDA'nın da niteliksel açıdan en az NET-LDA kadar başarılı olduğu yine elde edilen sonuçlardan çıkarılabilmektedir.

LDA, LTM ve AMC ise NET-LDA ve Concept-LDA'ya göre daha başarısız konular üretmektedir.

Niteliksel ve niceliksel değerlendirme yanında modeller çalışma süreleri açısından da değerlendirilmişlerdir. Modellerin çalışma sürelerine göre değerlendirilmesinde Intel(R) Xeon(R) CPU E5-1620 v3 .350 GHz işlemcili, 64 bit, 16 GB RAM'e sahip bir masaüstü bilgisayar kullanılmıştır. Elde edilen sonuçların saniye cinsinden değerleri Tablo 4.13'te verilmiştir.

Tablo 4.13. Yöntemlerin saniye cinsiden çalışma süreleri

		Çalışma Süreleri (saniye)				
		NET-LDA	Concept-LDA	LDA	LTM	AMC
Veri Kümesi	Otel	**13**	-	14	32	91
	Restaurant	**117**	555	125	321	1565
	Alarm Clock	**17**	65	18	42	146
	Amplifier	**23**	103	25	66	204
	Battery	**13**	52	14	35	108
	Blu Ray Player	**36**	169	45	103	359
	Cable Modem	**20**	87	23	57	189
	Camcorder	**34**	142	38	94	328
	Camera	**36**	147	41	104	366
	Car Stereo	**21**	91	23	61	195
	Cell Phone	**20**	85	22	58	182
	Computer	**37**	167	43	105	350
	DVD Player	**23**	93	25	62	196

Tablo 4.13 incelendiğinde her veri kümesi için en kısa çalışandan en uzun çalışana konu modellerinin sıralaması: NET-LDA, LDA, LTM, Concept-LDA, AMC olacak şekildedir. AMC'nin en yavaş çalışan model olması veri kümesi üzerinden hem ML

hem de CL'lere göre konuları elde etmesidir. Concept-LDA'nın yavaş çalışmasının nedeni ise koleksiyondaki kelime sayısının üç dört katına çıkmasıdır. LTM sadece ML'ler ile işlem yapmasına rağmen oldukça hızlı çalışan bir modeldir. Yöntemler içerisinden en hızlı çalışan ise NET-LDA'dır. Dokümanlarını benzerliklerine göre birleştirip, modele dahil edilen adaptif parametre modeli yavaşlatmak yerine diğer tüm yöntemlere kıyasla hızlandırmaktadır. Veri kümeleri üzerinden modellerin çalışma hızları değerlendirildiğinde ise veri kümesindeki kelime sayısı arttıkça yöntemlerin çalışma sürelerinin de o hızla arttığı görülmektedir.

Tablo 4.13'te verilen saniye cinsinden çalışma sürelerinin Eşitlik (4.5)'teki Doğrusal normalizasyona göre normalize edilmiş halleri Tablo 4.14'te, normalize edilmiş çalışma sürelerinin grafiksel temsili ise İngilizce ve Türkçe veri kümeleri için Şekil 4.4'te verilmiştir. Her veri kümesi için çalışma süresi kendi içerisinde normalize edilmiştir.

$$v' = \frac{v}{\max_A} \tag{4.5}$$

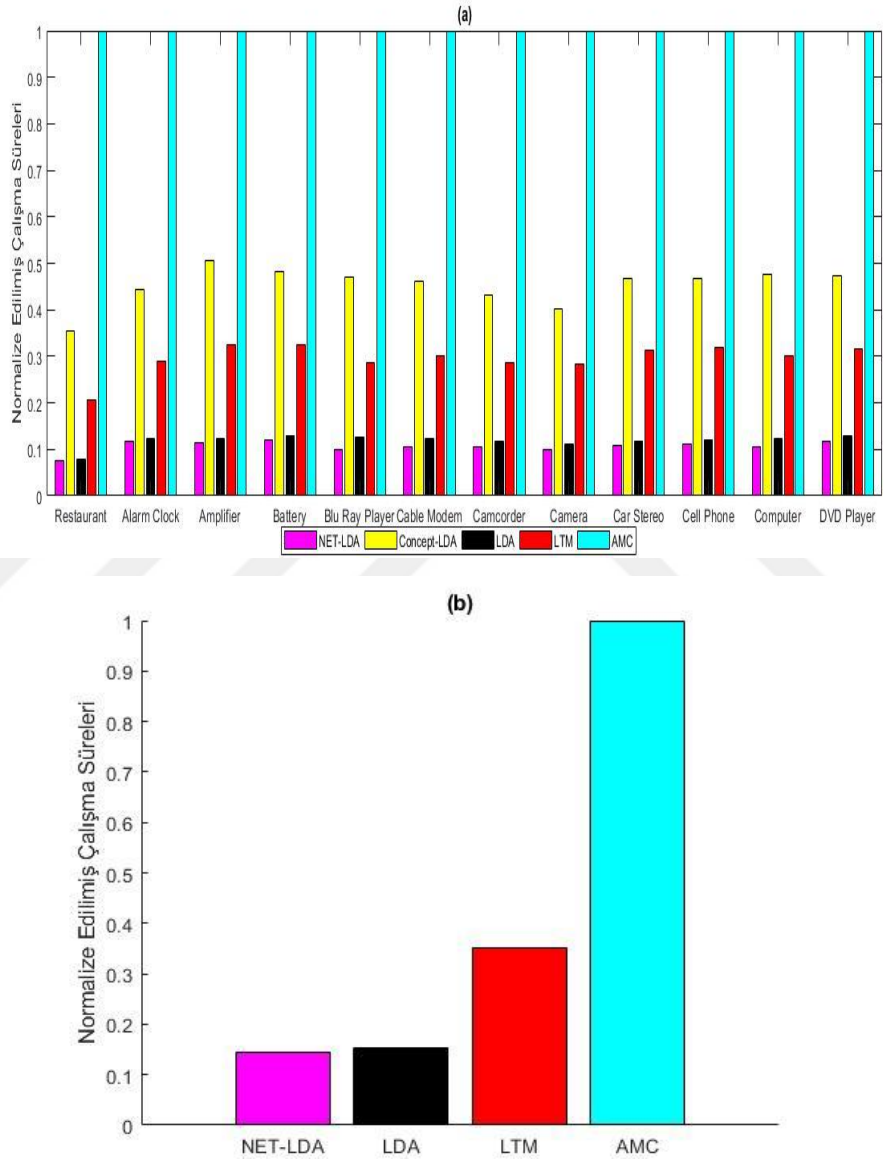
Burada v; bir veri kümesi için ilgili modelin saniye cinsinden çalışma süresi iken v' v değerinin normalize edilmiş halidir. $\max_A$ bir veri kümesi üzerine uygulanan farklı konu modellerinin çalışma süreleri içerisinden en uzun olanıdır.

Tablo 4.14. Yöntemlerin saniye cinsiden çalışma sürelerinin normalize edilmiş hali

		Çalışma Sürelerinin Normalize Edilmiş Hali				
		NET-LDA	Concept-LDA	LDA	LTM	AMC
Veri Kümesi	Otel	0,143	-	0,154	0,352	1
	Restaurant	0,075	0,355	0,08	0,205	1
	Alarm Clock	0,116	0,445	0,123	0,288	1
	Amplifier	0,113	0,505	0,123	0,324	1
	Battery	0,12	0,481	0,13	0,324	1
	Blu Ray Player	0,1	0,471	0,125	0,287	1
	Cable Modem	0,106	0,46	0,122	0,302	1
	Camcorder	0,104	0,433	0,116	0,287	1
	Camera	0,098	0,402	0,112	0,284	1
	Car Stereo	0,108	0,467	0,118	0,313	1
	Cell Phone	0,11	0,467	0,121	0,319	1

Tablo 4.14. (Devam) Yöntemlerin saniye cinsiden çalışma sürelerinin normalize edilmiş hali

		Çalışma Sürelerinin Normalize Edilmiş Hali				
		NET-LDA	Concept-LDA	LDA	LTM	AMC
Veri Kümesi	Computer	0,106	0,477	0,123	0,3	1
	DVD Player	0,117	0,474	0,128	0,316	1



Şekil 4.4. İngilizce (a) ve Türkçe (b) veri kümeleri için normalize edilmiş çalışma süreleri

Normalizasyon sonuçlarına bakıldığında daha rahat bir şekilde modellerin çalışma zamanları arasındaki fark görülmektedir. Normalizasyon ile en hızlı çalışan NET-LDA'nın çalışma zamanı sıfır, en yavaş çalışan AMC'nin ise çalışma zamanı bire denk

gelmiştir. NET-LDA; Concept-LDA'ya göre ortalama %0,04, LDA'ya göre ortalama %0,001, LTM'ye göre ortalama %0,02, AMC'ye göre ise ortalama %1 daha hızlı çalışmaktadır.

Niceliksel ve niteliksel değerlendirme yanında önerilen NET-LDA yönteminin hız açısından da diğer yöntemlerden hızlı olması oldukça önemlidir ve bu durum yönteme avantaj sağlamaktadır. Ayrıca NET-LDA'nın her açıdan oldukça başarılı olduğunun da bir kanıtıdır. Önerilen diğer bir yöntem olan Concept-LDA ise doküman koleksiyonunu kavram ve adlandırılmış varlıklar ile genişletilmesinden ötürü yavaş çalışmaktadır. Bu durum her ne kadar dezavantaj olarak gözükse de niceliksel ve niteliksel olarak Concept-LDA'da en az NET-LDA kadar başarılı bir yöntemdir.

5. SONUÇLAR VE ÖNERİLER

Günümüz teknolojileri ile birlikte haberler, forumlar, sosyal ağlar, elektronik alışveriş siteleri, gibi çevrimiçi ortamlar üzerinden dijitalleşen ve depolanan veri miktarında yaşanan büyük artış biz kullanıcıların aradığı ve keşfetmek istediği bilgiye erişimi zor hale getirmektedir. Bu büyük veriyi organize etmemize, anlamamıza, özetlememize ve istediğimiz bilgiyi içerisinden elde etmemize yarayacak otomatik yöntemlere ise her zaman ihtiyaç duyulmaktadır. Konu modelleri; büyük hacimli verileri organize etme, anlama, özetleme ve istenilen bilginin elde edilmesini sağlayan denetimsiz algoritmaları biz kullanıcılarına sunmaktadır.

İlk olarak 1990'da ortaya çıkan konu modelleri, yapısal olmayan doküman koleksiyonlarındaki gizli tematik bilgiyi küçük boyutlu uzaya çevirerek ortaya çıkarmaktadır. Çıkarılan bu tematik bilgi; özetlemeden, etiketlemeye, tahminlemeden, anahtar kelime çıkarımına, özellik çıkarımına kadar pek çok görev için kullanılabilmekte ve başarılı sonuçlar elde edilmektedir. Ancak konu modelleri üzerine literatürde pek çok başarılı çalışma olmakla birlikte, hala daha teorikte anlaşılması güç bir konu olarak karşımıza çıkmaktadır.

Geçmişten günümüze geliştirilen konu modellerinden en yaygın olanı ise LDA'dır. LDA'nın başarılı bir konu modeli olması ile birlikte araştırmacılar daha sonraki yıllarda yeni konu modelleri tasarlamak yerine LDA tabanlı modeller geliştirmeye ve pek çok farklı alana uygulamaya başlamışlardır.

LDA, doküman gibi ayrık verileri modellemek ve dokümanı meydana getiren konuları ortaya çıkarmak için kullanılan hem üretici hem de grafiksel bir modeldir. Herhangi bir önbilgiye ihtiyaç duymayan ve tamamen denetimsiz bir yöntem olan LDA ile kelime torbası yaklaşımına dayalı olarak gizli uzaydaki konular elde edilmektedir. Kelimelerin doküman içerisindeki yerleşimi göz ardı edilirken, kelimelerin birlikte bulunması bu yöntemde kullanılmaktadır. Dokümanı oluşturan kelimeler arasındaki anlamsal ilişkiyi veya dokümanlar arasındaki anlamsal bilgi bu yöntemde hiçbir şekilde dikkate alınmamaktadır. Dolayısıyla elde edilen konular kendi içlerinde

anlamsal yönden bir uyum içermemektedir ve bu durum LDA için bir dezavantaj oluşturmaktadır. LDA'nın bu dezavantajının üstesinden gelebilmek için ise metinlerin doğru analizi, doğru analiz için ise kelime vektörlerinin anlamsal olarak başarılı bir şekilde temsil edilmesi oldukça önemlidir. Bu amaçla tez çalışması kapsamında anlamsal bilgiyi farklı şekillerde modele dahil eden Concept-LDA ve NET-LDA olmak üzere iki farklı konu modeli geliştirilmiştir.

Concept-LDA ile LDA'nın temel varsayımı olan kelimelerin birlikte geçme durumları yani kelime torbası yaklaşımı {kelime+kavram+adlandırılmış varlıklar} torbası olarak genişleterek kendi içerisinde anlamsal açıdan uyumlu konuların elde edilmesi hedeflenmiştir. Böyle bir yöntemin esin kaynağı ise aynı kavram veya adlandırılmış varlığı paylaşan kelimelerin aynı konu altında yer alması gerekliliği ve bunun da doküman uzayının kavram ve adlandırılmış varlıklar ile genişletilmesi ile sağlanabileceğidir. Kavram ve adlandırılmış varlıklar graf tabanlı bir yaklaşım olan Babelfy ile elde edilmiştir. Sonuç olarak; model farklı ürünler için İngilizce yorumları içeren on iki veri kümesine uygulanmış olup, hem niceliksel hem de niteliksel olarak başarılı konular elde edilmiştir.

NET-LDA ise anlamsal olarak benzer olan dokümanların içerdikleri konuların da benzer olacağı fikrinden yola çıkarak tasarlanmıştır. Bu amaçla, dokümanlar arasındaki anlamsal benzerlik kavram ve adlandırılmış varlıklar üzerinden hesaplanmış; benzer dokümanlar birleştirilerek bu adımda elde edilen bilinen temel ölçüt doküman-konu dağılımına etki ettirilerek asimetrik önseller üzerinden konular elde edilmiştir. Önerilen yöntem ile LDA'nın temel varsayımı olan kelimelerin birlikte geçme durumlarını anlamsal olarak güçlendirilmiştir. Bu yöntemde de kavram ve adlandırılmış varlıklar Babelfy ile elde edilmiştir. Yöntem biri Türkçe geri kalan on iki tanesi İngilizce olmak üzere farklı ürünler için kullanıcı yorumlarını içeren on üç farklı veri kümesine uygulanmıştır.

Tez çalışmasının literatüre yaptığı temel katkılar aşağıdaki gibi maddeler halinde sıralanabilir:

- Literatürde yer alan LDA modelleri derinlemesine incelendiğinde; konu modellerinin ve dokümanların anlamsal benzerliklerinin avantajlarının ilk olarak

NET-LDA'da birleştirildiği ve böylece anlamsal olarak tutarlı ve başarılı konuların elde edildiği görülmüştür.

- Dokümanların anlamsal benzerliğe dayalı olarak birleştirilmesi, birleştirilen dokümanların ve birleştirme sonucu elde edilen adaptif parametrenin modele dahil edildiği ilk çalışma NET-LDA'dır.
- Doküman uzayının kavramlar ve adlandırılmış varlıklar ile genişleterek anlamsal bir zenginleştirme ile LDA'nın temel varsayımı olan kelime torbası yaklaşımının yerine {kelime+kavram+adlandırılmış varlık} torbasının kullanıldığı ilk model Concept-LDA'dır.
- Concept-LDA'nın kullandığı {kelime+kavram+adlandırılmış varlık} torbası ile anlamsal olarak benzer olan kelimeler aynı konu altında yer almış ve anlamsal olarak tutarlı ve başarılı konuların elde edildiği görülmüştür.
- Hem NET-LDA hem de Concept-LDA'da kavram ve adlandırılmış varlıkların elde edilmesinde Babelfy kullanılmıştır. Babelfy'dan elde edilen kavram ve adlandırılmış varlıkların kullanıldığı ilk çalışmalar sırasıyla Concept-LDA ve NET-LDA'dır.
- Her iki yöntemde kavram ve adlandırılmış varlıkların kullanılması ile DDİ'deki önemli problemlerden birisi olan WSA problemi çözümlenmiştir.
- Kelime şeklindeki ürün özellikleri yanı sıra eşdizim şeklindeki ürün özellikleri de kullanıcı yorumlarından çıkartılarak daha doğru bir analiz yapılması sağlanmıştır. Eşdizim şeklindeki ürün özellikleri çıkartılırken yine Babelfy'dan yararlanılmıştır. Babelfy eşdizimleri çıkartmak için yine ilk defa Concept-LDA ve NET-LDA'da kullanılmıştır.
- Concept-LDA ve NET-LDA alandan bağımsız olarak geliştirilen iki konu modeli yöntemidir. Yapılan deneyler ile her iki yöntemden farklı ürün özellikleri için başarılı sonuçlar elde edildiği görülmüştür.
- Concept-LDA dil bağımlı bir yöntemdir ancak NET-LDA dilden bağımsız bir yöntemdir. NET-LDA Babelfy tarafından sağlanan 284 farklı dil için uygulanabilirdir. Modelin dilden bağımsız olduğunu ispatlamak için İngilizce ve Türkçe veri kümeleri deneysel çalışmalarda kullanılmıştır.
- Ayrıca modellerin herhangi bir önsel bilgiye ihtiyaç duymaması da önemli bir üstünlük ve avantajdır.
- Modeller çalışma süreleri açısından değerlendirildiğinde en hızlı olan yöntemin NET-LDA olduğu görülmüştür. Concept-LDA doküman uzayını genişleterek çalıştığı

için çalışma süresi NET-LDA, LDA ve LTM'ye göre oldukça yavaş, AMC'ye göre ise hızlıdır.

- Tez çalışması kapsamında önerilen bu iki yöntem niceliksel ve niteliksel değerlendirilmiş ve temel yöntemlerden LDA, LTM ve AMC ile karşılaştırılmıştır. Niceliksel değerlendirme adımında konu uyumluluğu ve F-skoru ölçütlerinden yararlanılmıştır. Niteliksel değerlendirme adımında ise elde edilen konular altındaki ürün özellikleri uzmanlar tarafından hem anlamsal uyumluluk hem de detayları yakalayabilme açısından değerlendirilmiştir. Sonuç olarak önerilen yöntemlerin hem niceliksel hem de niteliksel açıdan başarılı olduğu tespit edilmiştir.

Geliştirilen yöntemlerden Concept-LDA'nın en önemli dezavantajı dile bağımlı bir yöntem olmasıdır. Çünkü Babelfy ile elde edilen kavram ve adlandırılmış varlıklar İngilizcedir. Dolayısıyla bir Türkçe olarak yazılmış dokümanların İngilizce kelimeler ile genişletilmesi modelin başarısız olmasına ve kelimeler arasındaki anlamsal ilişkilerin yakalanmasının engellenmesine neden olmaktadır. Bu yöntemin dilden bağımsız olması için ara bir adım olarak bir çeviriciden yararlanılması çözüm olabilir.

Babelfy ile elde edilen kavram ve adlandırılmış varlıkların doğru anlamsal ilişkileri yakalamadaki başarısı hem Concept-LDA hem de NET-LDA için bu tez çalışmasında ispatlanmıştır. Yapılacak gelecek çalışmalarda anlamsal ilişkilerin yakalanması adımında günümüz popüler ve başarılı yöntemlerinden kelime vektörleştirme yaklaşımları denenebilir.

KAYNAKLAR

Akın M. D., Akın A. A., Türk Dilleri için Açık Kaynaklı Doğal Dil İşleme Kütüphanesi : Zemberek, *Elektrik Mühendisliği,* 2007, **431**, 38–44.

Al-Thubaity A., Baazeem I., A collocation extraction tool and two language resources for MSA, *Procedia Computer Science,* 2017, **117**, 23–29.

Alam Md. H., Ryu W. J., Lee S., Joint multi-grain topic sentiment: modeling semantic aspects for online reviews, *Information Sciences,* 2016, **339**, 206-223.

Atıcı B., İlhan 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.

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 Sciences,* 2014, **40**, 621-636.

Bahmanyar R., Espinoza-Molina D., Datcu M., Multisensor Earth Observation Image Classification Based on a Multimodal Latent Dirichlet Allocation Model, *IEEE Geoscience and Remote Sensing Letters,* 2018, **15**(3), 459-463.

Banko M., https://pdfs.semanticscholar.org/fe15/eea496a86ddee160fc23ea8c7e2d3f88aa44.pdf (Ziyaret Tarihi: 4 Aralık 2018)

Bao S., Xu S., Zhang L., Yan R., Su Z., Han D., Yu Y., Joint Emotion-Topic Modeling for Social Affective Text Mining, *Ninth IEEE International Conference on Data Mining,* Miami, FL, USA, 6-9 December 2009.

Bao S., Xu S., Zhang L., Yan R., Su Z., Han D., Yu Y., Mining Social Emotions from Affective Text, *IEEE Transactions on Knowledge and Data Engineering,* 2012, **24**(9), 1658-1670.

Barnard K., Duygulu, P., Forsyth D., de Freitas N., Blei D. M., Jordan M. I., Matching Words and Pictures, *Journal of Machine Learning Research,* 2003, **3**, 1107-1135.

Bishop C. M., *Pattern Recognition and Machine Learning,* 1st ed., Springer, Berlin, 2006.

Bissacco A., Yang M. H., Soatto S., Detecting Humans via Their Pose, *20th Annual Conference on Neural Information Processing Systems,* Vancouver, B.C., Canada, 4-7 December 2006.

Blei D. M., Probabilistic Topic Models, *Communications of the ACM*, 2012, **55**(4), 77-84.

Blei D. M., Probabilistic Topic Models, http://www.cs.columbia.edu/~blei/talks/Blei_MLSS_2012.pdf (Ziyaret tarihi: 4 Ağustos 2018).

Blei D. M., Jordan M. I., Modeling Annotated Data, *26th annual international ACM SIGIR conference on Research and development in informaion retrieval,* Toronto, Canada, 28 July-1 August 2003.

Blei D. M., Lafferty J. D., Correlated Topic Models, *23rd international conference on Machine learning,* Pittsburgh, Pennsylvania, USA, 25-29 June 2006a.

Blei D. M., Lafferty J. D., Dynamic Topic Models, *23rd International Conference on Machine Learning,* Pittsburgh, Pennsylvania, USA, 25-29 June 2006b.

Blei D. M., McAuliffe J. D., Supervised Topic Models, *21st Annual Conference on Neural Information Processing Systems,* Vancouver, B.C., Canada, 3-6 December 2007.

Blei D. M., Ng A.Y., Jordan M. I., Latent dirichlet allocation, *The Journal of Machine Learning Research*, 2003, **3**, 993-1022.

Boyd-Graber J. L., Blei D. M., Syntactic Topic Models, *22nd Annual Conference on Neural Information Processing Systems,* Vancouver, B.C., Canada, 8-11 December 2008.

Brody S., Elhadad N., An Unsupervised Aspect-Sentiment Model for Online Reviews, 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, USA, 1-6 June 2010.

Bundschus M., Yu S., Tresp V., Rettinger A., Dejori M., Kriegel H. P., Hierarchical bayesian models for collaborative tagging systems, *9th IEEE International Conference on Data Mining,* Miami, USA, 6-9 December 2009.

Cao J., Li D., Huang D., A Three-layered Collocation Extraction Tool and its Application in China English Studies, Editors: Sun M., Liu Z., Zhang M., Liu Y., *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data,* Springer, Cham, 38-49, 2015.

Chang J., Gerrish S., Wang C., Blei D. M., Reading tea leaves: How humans interpret topic models, *22nd International Conference on Neural Information Processing Systems,* Vancouver, B.C., Canada, 7-10 December 2009.

Chemudugunta C., Holloway A., Smyth P., Steyvers M., Modeling documents by combining semantic concepts with unsupervised statistical learning, Editörs: Sheth A. et al., *The Semantic Web - Lecture Notes in Computer Science,* Springer, Berlin, Heidelberg, 229-244, 2008.

Chen Y. S., Chen L. H., Takama Y. Proposal of LDA-based Sentiment Visualization of Hotel Reviews, 2015 IEEE 15th International Conference on Data Mining Workshop, Atlantic City, NJ, USA, 14-17 November 2015.

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 2014a.

Chen Z., Liu B., Mining Topics in Documents: Standing on the Shoulders of Big Data, *20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* New York, NY, USA, 24-27 August 2014b.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R., Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, 1990, **41**, 1-30.

Delac D., Krleza Z., Snajder J., TermeX: A Tool for Collocation Extraction, Editor: Gelbukh A., *Computational Linguistics and Intelligent Text Processing*, Springer, Berlin, Heidelberg, 149-157, 2009.

Ehrmann M., Cecconi F., Vannella D., McCrae J., Cimiano P., Navigli R., Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0, *Ninth International Conference on Language Resources and Evaluation,* Reykjavik, Iceland, 26-31 May 2014.

Ekinci E., İlhan Omurca S., Ürün Özelliklerinin Konu Modelleme Yöntemi ile Çıkarılması, *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi,* 2017a, **9**(1), 51-58.

Ekinci E., İlhan Omurca S., Extracting Implicit Aspects based on Latent Dirichlet Allocation, *9th International Conference on Agents and Artificial Intelligence-Doctoral Consortium,* Porto, Portugal, 24-26 February 2017b.

Ekinci E., İlhan Omurca S., An Aspect-Sentiment Pair Extraction Approach Based on Latent Dirichlet Allocation, *International Journal of Intelligent Systems and Applications in Engineering,* 2018a**, 6**(3), 209-213.

Ekinci E., İlhan Omurca S., Babelfy-Based Extraction of Collocations from Turkish Hotel Reviews, *International Conference on Artificial Intelligence and Data Processing,* Malatya, Türkiye, 28-30 Eylül 2018b.

Ekinci E., İlhan Omurca S., https://www.researchgate.net/publication/325170185_Topic_Extraction_Dataset-Turkish_Hotel_Reviews (Ziyaret Tarihi: 9 Kasım 2018c)

Ekinci E., Türkmen H., İlhan Omurca S., Multi-word Aspect Term Extraction Using Turkish User Reviews, *International Journal of Computer Engineering and Information Technology*, 2017a, **9**(1), 15-23.

Ekinci E., Türkmen H., İlhan Omurca S., Determining Multi Word Aspects by Using Apriori Algorithm and Syntactic Rules for Turkish Hotel Reviews, *8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics,* Poznan, Poland, 17-19 November 2017b.

Fu X., Sun X., Wu H., Cui L., Huang J. Z., Weakly supervised topic sentiment joint model with word embeddings, *Knowledge-Based Systems,* 2018, **147**, 43–54.

Godin F., Slavkovikj V., De Neve W., Schrauwen B., Van de Walle R., Using topic models for twitter hashtag recommendation, *22nd International Conference on World Wide Web,* Rio de Janeiro, Brazil, 13-17 May 2013.

Griffiths T. L., Steyvers M., A probabilistic approach to semantic representation, *24th Annual Conference of the Cognitive Science Society,* Virginia, USA, 7-10 August 2002a.

Griffiths T. L., Steyvers M., Prediction and semantic association, *16th Annual Conference on Neural Information Processing Systems,* Vancouver, B.C., Canada, 9-14 December 2002b.

Griffiths T. L., Steyvers M., Finding Scientific Topics, *National Academy of Sciences,* 2004, **101**(1), 5228-5235.

Griffiths T. L., Steyvers M., Tenenbaum J., Topics in semantic representation, Psychological Review, 2017, **114**(2), 211-244.

Hanson A. J., Geometry for N-dimensional Graphics, Editor: Heckbert P. S ., Graphics Gems IV, Academic Press Professional, Inc., USA, 149-170, 1994.

Heid U., Weller M., Tools for collocation extraction: preferences for active vs. passive, *Sixth International Conference on Language Resources and Evaluation,* Morocco, 28-30 May 2008.

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.

Hockenmaier J., Conjugate priors, Lecture Notes, https://courses.engr.illinois.edu/cs598jhm/sp2010/Slides/Lecture02.pdf, (Ziyaret Tarihi: 10 Ağustos 2018)

Hofmann T., Probabilistic Latent Semantic Analysis, *Fifteenth Conference on Uncertainty in Artificial Intelligence,* Stockholm, Sweden, 30 July-1 August 1999.

Hofmann T., Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning,* 2001, **42**(1), 177-196.

Hu Y., John A., Wang F., Seligmann D. D., Kambhampati S., ET-LDA: Joint Topic Modeling For Aligning, Analyzing and Sensemaking of Public Events and Their Twitter Feeds, *26th Conference on Artificial Intelligence*, Toronto, Ontario, Canada, 22-26 July 2012.

Jadhav N., http://www.cfilt.iitb.ac.in/resources/surveys/Topic-Models-For-Sentiment-Analysis-2014-Nikhilkumar-Jadhav.pdf (Ziyaret Tarihi: 5 Ağustos 2018)

Jelodar H., Wang Y., Yuan C., Feng X., https://arxiv.org/ftp/arxiv/papers/1711/1711.04305.pdf (Ziyaret Tarihi: 1 Ağustos 2018)

Jimènez-Zafra, S. M., Martin-Valdivia M. T., Martìnez-Càmara E., Ureña-Lòpez L. A., Combining resources to improve unsupervised sentiment analysis at aspect-level, *Journal of Information Science,* 2016, **42**(2), 213-229.

Jo Y, Oh A., Aspect and Sentiment Unification Model for Online Review Analysis, *4th ACM International Conference on Web Search and Data Mining,* Kowloon, Hong Kong, 9-12 February 2011.

Kim S. M., Hovy E., Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text, *Workshop on Sentiment and Subjectivity in Text,* Sydney, Australia, 22 July 2006.

Krestel R., Fankhauser P., Tag recommendation using probabilistic topic models, *2009th International Conference on ECML PKDD Discovery Challenge,* Bled, Slovenia, 7-11 September 2009.

Krestel R., Fankhauser P., Nejdl W., Latent dirichlet allocation for tag recommendation, *3rd ACM conference on Recommender systems,* New York, USA, 23-25 October 2009.

Kumova Metin S., Karaoğlan B., Collocation Extraction in Turkish Texts Using Statistical Methods, Editors: Loftsson H., Rögnvaldsson E., Helgadóttir S., *Advances in Natural Language Processing,* Springer, Berlin, Heidelberg, 238-249, 2010.

Kumova Metin S., Karaoğlan B., Identifying Collocations in Turkish Using Statistical Methods, *Bilig*, 2016, **78**, 253-284.

Lee A. J. T., Yang F. C., Chen C. H., Wang C. S., Sun C. Y., Mining perceptual maps from consumer reviews, *Decision Suppoprt Systems,* 2016, **82**, 12-25.

Lehmann F., Semantic Networks, *Computers & Mathematics with Applications,* 1992, **23**(2-5), 1-50.

Li D., Cao J., Huang D., A Hierachical Collocation Extraction Tool, *IEEE Fifth International Conference on Big Data and Cloud Computing,* Dalian, China, 26-28 August 2015.

Li H. C., Song M., Chang C. I., Simplex volume analysis for finding endmembers in hyperspectral imagery, *SPIE international symposium on SPIE sensing technology + applications,* Baltimore, MD, 20-24 April 2015.

Li W., McCallum A., Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations, *23rd International Conference on Machine learning,* Pittsburgh, Pennsylvania, USA, 25-29 June 2006.

Li W. Y., Lu Q., Liu J., TCtract-A Collocation Extraction Approach for Noun Phrases Using Shallow Parsing Rules and Statistic Models, *20th Pacific Asia Conference on Language, Information and Computation,* Wuhan, China, 1-3 November 2006.

Li W. Y., Lu Q., Liu J., Chinese Typed Collocation Extraction using Corpus-based Syntactic Collocation Patterns, *2007 International Conference on Natural Language Processing and Knowledge Engineering,* Beijing, China, 30 August-1 September 2007.

Liang W., Xieb H., Rao Y., Lau R. Y. K., Wang F. L., Universal Affective Model for Readers' Emotion Classification over Short Texts, *Expert Systems With Applications,* DOI: 10.1016/j.eswa.2018.07.027.

Lin C., He Y., Joint sentiment/topic model for sentiment analysis, *18th ACM Conference on Information and Knowledge Management,* Hong Kong, China, 2-6 November 2009.

Lin J. F., Li S., Cai Y., A New Collocation Extraction Method Combining Multiple Association Measures, *Seventh International Conference on Machine Learning and Cybernetics,* Kunming, China, 12-15 July 2008.

Lin J. F., Li S., Cai Y., Collocation Extraction Using Web Feedback Data, *Chinese Journal of Electronics,* 2009, **18**, 312-316.

Linstead E., Rigor P., Bajracharya S., Lopes C., Baldi P., Mining Concepts from Code with Probabilistic Topic Models, *22nd IEEE/ACM International Conference on Automated Software Engineering,* Atlanta, Georgia, USA, 5-9 November 2007.

Liu Z., Wang H., Wu H., Li S., Two-Word Collocation Extraction Using Monolingual Word Alignment Method, *ACM Transactions on Intelligent Systems and Technology,* 2011, **3**(1), 16:1–16:29.

Lu H. M., Lee C. H., The Topic-Over-Time Mixed Membership Model (TOT-MMM): A Twitter Hashtag Recommendation Model that Accommodates for Temporal Clustering Effects, *IEEE Intelligent Systems,* 2015, **30**(3), 18-25.

Lu Q., Li Y., Xu R., Improving xtract for chinese collocation extraction, *2003 International Conference on Natural Language Processing and Knowledge Engineering,* Beijing, China, 26-29 October 2003.

Lu Y., Mei Q., Zhai C. X., Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA, *Information Retrieval,* 2011, **14**, 178-203.

Lukins S. K., Kraft N. A., Etzkorn L. H., Source Code Retrieval for Bug Localization using Latent Dirichlet Allocation, *15th Working Conference on Reverse Engineering,* Antwerp, Belgium, 15-18 October 2008.

Lukins S. K., Kraft N. A., Etzkorn L. H., Bug localization using latent Dirichlet allocation, *Information and Software Technology*, 2010, **52**(9), 972-990.

Mahmoud A., Niu N., On the role of semantics in automated requirements tracing, *Requirements Engineering*, 2015, **20**, 281-300.

Manning C., Schütze H., *Foundations of Statistical Natural Language Processing,* 1st ed., MIT Press, Cambridge, MA, USA, 1999.

Manning C. D., Surdeanu M., Bauer J., Finkel J., Bethard S. J., McClosky D., The Stanford Corenlp Natural Language Processing Toolkit, *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations,* Baltimore, Maryland, USA, 22-27 June 2014.

Miller G., WordNet: A Lexical Database for English, *Communications of the ACM,* 1995, **38**(11), 39-41.

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A., Optimizing semantic coherence in topic models, 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, United Kingdom, 27-31 July 2011.

Moro A., Cecconi F., Navigli, R., Multilingual Word Sense Disambiguation and Entity Linking for Everybody, *13th International Semantic Web Conference, Posters and Demonstrations,* Riva del Garda, Italy, 19-23 October 2014a.

Moro A., Raganato A., Navigli R., Entity Linking meets Word Sense Disambiguation: A Unified Approach, *Transactions of the Association for Computational Linguistics,* 2014b, **2**, 231-244.

Nadkarni P. M., Ohno-Machado L., Chapman W. W., Natural Language Processing: An Introduction ", *Journal Of The American Medical Informatics Association,* 2011, **18**(5), pp. 544-551.

Navigli R., Multilinguality at Your Fingertips: BabelNet, Babelfy and Beyond!, http://sssw.org/2015/?page_id=379, (Ziyaret Tarihi: 31 Ağustos 2018)

Navigli R, Ponzetto S. P., BabelNet: Building a Very Large Multilingual Semantic Network, *48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11-16 July 2010.

Ng K. W., Tian G. L., Tang, M. L., *Dirichlet and Related Distributions: Theory, Methods and Applications,* 1st ed., Wiley, New York, 2011.

Pecina P., Lexical association measures and collocation extraction, *Language Resources and Evaluation,* 2009, **44**(1-2), 137-158.

Petrovic S., Snajder J., Basic B. J., *Computer Speech & Language,* 2010, **24**(2), 383-394.

Popescul A., Ungar L., Pennock D., Lawrence S., Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments, *17th Conference in Uncertainty in Artificial Intelligence,* Washington, USA, 2-5 August 2001.

Poria S., Chaturvedi I., Cambria E., Bisto F., Sentic LDA: Improving on LDA with Semantic Similarity for Aspect-Based Sentiment Analysis, *2016 International Joint Conference on Neural Networks,* Vancouver, Canada, 24-29 July 2016.

Ramage D., Hall D., Nallapati R., Manning C. D., Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, *2009 Conference on Empirical Methods in Natural Language Processing,* Singapore, 6-7 August 2009.

Rao D., McNamee P., Dredze M., Entity Linking: Finding Extracted Entities in a Knowledge Base, Editors: Poibeau T., Saggion H., Piskorski J., Yangarber R., *Multi-source, Multilingual Information Extraction and Summarization - Theory and Applications of Natural Language Processing,* Springer, Berlin, Heidelberg, 93-115, 2013.

Rasiwasia N., Vasconcelos N., Latent Dirichlet Allocation Models for Image Classification, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2013, **35**(11), 2665-2679.

Ritter A., Mausam, Etzioni O., Clark S., Open domain event extraction from twitter, *18th ACM SIGKDD international conference on Knowledge discovery and data mining,* Beijing, China, 12-16 August 2012.

Rosen-Zvi M., Griffiths T., Steyvers M., Smyth P., The author-topic model for authors and documents, *20th conference on Uncertainty in artificial intelligence,* Banff, Canada, 7-11 July 2004.

Rule W., Duan W., Prakash N., Zhuang N., Alvarado R. C., Brown D. E., Social Pressure Analysis of Local Events using Social Media Data, *2018 Systems and Information Engineering Design Symposium,* Charlottesville, VA, USA, 27 April 2018.

Sanderson M., Word Sense disambiguation and information retrieval, Editors: Croft B. W., van Rijsbergen C. J., *SIGIR'94,* Springer, London, 142-151, 1994.

Savage T., Dit B., Gethers M., Poshyvanyk D., TopicXP: Exploring Topics in Source Code using Latent Dirichlet Allocation, *2010 IEEE International Conference on Software Maintenance*, Timisoara, Romania, 12-18 September 2010.

Seretan V., Wehrli E., Accurate collocation extraction using a multilingual parser, *21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia, 17-18 July 2006.

Seretan V., Wehrli E., Multilingual collocation extraction with a syntactic parser, *Language Resources and Evaluation*, 2009, **47**(1), 71-85.

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.

Si X., Sun M., Tag-LDA for scalable real-time tag recommendation, *Journal of Computational Information Systems*, 2009, **6**(2), 1009-1016.

Speer R., Chin J., Havasi C., ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, *Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, 4-9 February 2017.

Steyvers M., Griffiths T. L., Probabilistic Topic Models, Editors: Landauer T., McNamara D. S., Dennis S., Kintsch W., *Handbook of Latent Semantic Analysis: A Road to Meaning*, Psychology Press, **427**(7), 424-440, 2007.

Steyvers M., Smyth P., Rosen-Zvi M., Griffiths T., Probabilistic Author-Topic Models for Information Discovery, *10th ACM SigKDD Conference Knowledge Discovery and Data Mining*, Seattle, WA, USA, 22-25 August 2004.

Steyvers M., Tenenbaum J. B., The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth, *Cognitive Science*, 2005, **29**(1), 41-78.

Suarez O. S., Sanchez-Berriel I., Aguiar J. P., Rodriguez V. G., Outlier Detection in Automatic Collocation Extraction, *Procedia - Social and Behavioral Sciences*, 2015, **198**, 433-441.

Thrun S., Lifelong Learning Algorithms, Editors: Thrun S., Pratt L., *Learning to Learn*, Springer, Boston, MA, 181-209, 1998.

Titov I., McDonald R., Modeling Online Reviews with Multi-grain Topic Models, *17th international conference on World Wide Web*, Beijing, China, 21-25 April 2008.

Todirascu A., Gledhill C., Stefanescu D., Extracting Collocations in Contexts, Editors: Vetulani Z., Uszkorei H., *Human Language Technology - Challenges of the Information Society*, Springer, Berlin, Heidelberg, 336-349, 2009.

URL-1: http://babelfy.org/about (Ziyaret Tarihi: 27 Ağustos 2018).

URL-2: http:// http://babelfy.org/about (Ziyaret Tarihi: 27 Ağustos 2018).

URL-3: http://wiki.languagetool.org/java-api (Ziyaret Tarihi: 08 Ekim 2018).

URL-4: http://nlp.stanford.edu/software/ (Ziyaret Tarihi: 08 Ekim 2018).

URL-5 http://web.uilab.kr/research/files/WSDM11/Yelp.zip (Ziyaret Tarihi: 07 Kasım 2016).

van Ravenzwaaij D., Cassey, P., Brown, S. D., A simple introduction to Markov Chain Monte–Carlo sampling, *Psychonomic Bulletin & Review,* 2018, **25**(1), 143-154.

Vossen P., Introduction to EuroWordNet, Editor: Vossen P., *EuroWordNet: A multilingual database with lexical semantic networks,* Springer Netherlands, 1-17, 1998.

Wallach H. M., Mimno D., McCallum A., Rethinking LDA: Why Priors Matter, *22nd International Conference on Neural Information Processing Systems,* Vancouver, British Columbia, Canada, 7-10 December 2009.

Wang W., Sentiment Analysis of Online Product Reviews with Semi-supervised Topic Sentiment Mixture Model, *7th International Conference on Fuzzy Systems and Knowledge Discovery,* Yandai, Shandong, China, 10-12 August 2010.

Wang T., Cai Y., Leung H., Lau R. Y. K., Li Q., Min H., Product aspect extraction supervised with online domain knowledge, *Knowledge-Based Systems,* 2014, **71**, 86-100.

Wang W., Feng Y., Dai Y., Topic analysis of online reviews for two competitive products using latent Dirichlet allocation, *Electronic Commerce Research and Applications,* 2018, **29**, 142-156.

Wei X., Croft, W. B., LDA-Based Document Models for Ad-hoc Retrieval, *29th annual international ACM SIGIR conference on research and development in information retrieval,* Seattle, WA, USA, 6-10 August 2006.

Xianghua F., Guo L., Yanyan G., Zhiqiang W., Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet Lexicon *Knowledge-Based Systems,* 2013, **37**, 186-195.

Xu R., Lu Q., A Multi-stage Chinese Collocation Extraction System, Editors: Yeung D. S., Liu Z. Q., Wang X. Z., Yan H., *Advances in Machine Learning and Cybernetics - Lecture Notes in Computer Science,* Springer, Berlin, Heidelberg, 3254-3259, 2005a.

Xu R., Lu Q., Improving Collocation Extraction by Using Syntactic Patterns, *2015 International Conference on Natural Language Processing and Knowledge Engineering,* Wuhan, China, 30 October-1 November, 2005b.

Xu R., Lu Q., Li Y., An automatic Chinese collocation extraction algorithm based on lexical statistics, *2003 International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China, 26-29 October 2003.

Yang Y., Chen C., Qiu M., Bao F. S., Aspect Extraction from Product Reviews Using Category Hierarchy Information, *15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, 3-7 April 2017.

Zhai Z., Liu B., Xu H., Jia P., Constrained LDA for Grouping Product Features in Opinion Mining, Editors: Huang J. Z., Cao L., Srivastava J., *Advances in Knowledge Discovery and Data Mining*, Springer, Berlin, Heidelberg, 448–459, 2011.

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.

Zhao F., Zhu Y., Jin H., Yang L. T., A personalized hashtag recommendation approach using LDA-based topic model in microblog environment, *Future Generation Computer Systems*, 2016, **65**, 196-206.

Zheng X., Lin Z., Wang X., Lin K. J., Song M., Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification, *Knowledge-Based Systems*, 2014, **61**, 29-47.

Zhu J., Ahmed A., Xing E. P., MedLDA: Maximum Margin Supervised Topic Models for Regression and Classification, *26th Annual International Conference on Machine Learning*, Montreal, Quebec, Canada, 14-18 June 2009.

KİŞİSEL YAYINLAR VE ESERLER

Ekinci E., İlhan Omurca S., A New Approach for A Domain-Independent Turkish Sentiment Seed Lexicon Compilation, International Arab Journal of Information Technology, 2019, **16**(5). Kabul edildi.

Ekinci E., İlhan Omurca S., An Aspect-Sentiment Pair Extraction Approach Based on Latent Dirichlet Allocation for Turkish, *International Journal of Intelligent Systems and Applications in Engineering*, 2018, **6**(3), 209-213.

Ekinci E., İlhan Omurca S., Babelfy-Based Extraction of Collocations from Turkish Hotel Reviews, *International Conference on Artificial Intelligence and Data Processing (IDAP 2018),* Malatya, Türkiye, 28-30 Eylül 2018.

Ekinci E., İlhan Omurca S., Acun N., A Comparative Study on Machine Learning Techniques Using Titanic Dataset, *7th International Conference on Advanced Technologies*, Antalya, Turkey, 28 April-1 May 2018.

Sevim S., İlhan Omurca S., **Ekinci E.,** An Ensemble Model using a BabelNet Enriched Document Space for Twitter Sentiment Classification, *International Journal of Information Technology and Computer Science*, 2018, **10**(1), 24-31.

Ekinci E., İlhan Omurca S., Ürün Özelliklerinin Konu Modelleme Yöntemi ile Çıkartılması, *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 2017, **9**(1), 51-58.

Ekinci E., Türkmen H., İlhan Omurca S., Multi-word Aspect Term Extraction Using Turkish User Reviews, *International Journal of Computer Engineering and Information Technology*, 2017, **9**(1), 15-23.

Ekinci E., İlhan Omurca S., Extracting Implicit Aspects based on Latent Dirichlet Allocation, *9th International Conference on Agents and Artificial Intelligence*, Porto, Portugal, 24-26 Şubat 2017.

Atıcı B., İlhan Omurca S., **Ekinci E.,** Product aspect detection in customer complaints by using latent dirichlet allocation, *2017 International Conference on Computer Science and Engineering (UBMK),* Antalya, Türkiye, 5-8 Ekim 2017.

Ekinci E., Türkmen H., İlhan Omurca S., Determining Multi Word Aspects by Using Apriori Algorithm and Syntactic Rules for Turkish Hotel Reviews, *8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan, Poland, 17-19 November 2017.

İlhan Omurca S., **Ekinci E.,** Türkmen H., An annotated corpus for Turkish sentiment analysis at sentence level. *2017 International Artificial Intelligence and Data Processing Symposium (IDAP 2017)*, Malatya, Turkey, 16-17 September 2017.

Türkmen H., İlhan Omurca S., **Ekinci E.,** An Aspect Based Sentiment Analysis on Turkish Hotel Reviews, *Girne American University Journal of Social and Applied Sciences*, 2016, **6**(2), 12-15.

Türkmen H., **Ekinci E.,** İlhan Omurca S., A Novel Method for Extracting Feature Opinion Pairs for Turkish, Editors: Dichev C., Agre G., *Artificial Intelligence: Methodology, Systems, and Applications - Lecture Notes in Computer Science*, Springer, Cham, 162-171, 2016.

Ekinci E., Türkmen H., İlhan Omurca S., Multi word Aspect Extraction from User Reviews, 6th World Conference on Innovation and Computer Scıence (INSODE-2016), Antalya, Turkey, 12-14 May 2016.

Aydın Keskin G., İlhan Omurca S., Aydın N., **Ekinci E.,** A comparative study of production inventory model for determining effective production quantity and safety stock level, *Applied Mathematical Modelling*, 2015, **39**(20), 6359-6374.

İlhan Omurca S., Baş S., **Ekinci E.,** An Efficient Document Categorization Approach for Turkish Based Texts. *International Journal of Intelligent Systems and Applications in Engineering*, 2015, **3**(1), 7-13

İlhan Omurca S., **Ekinci E.,** An alternative evaluation of post traumatic stress disorder with machine learning methods, *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, Madrid, Spain, 2-4 September 2015.

ÖZGEÇMİŞ

Ekin Ekinci 1987 yılında Trabzon'da doğdu. İlk, orta ve lise öğrenimini Trabzon'da tamamladı. 2005 yılında girdiği Çanakkale Onsekiz Mart Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü'nden 2009 yılında, bölüm üçüncüsü olarak mezun oldu. 2009 yılında başladığı Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Bölümü Bilgisayar Mühendisliği Anabilim Dalı'ndaki Yüksek Lisans eğitimini 2013 yılında tamamladı. 2013 yılından itibaren Kocaeli Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Bölümü Bilgisayar Mühendisliği Anabilim Dalı'nda doktora eğitimine devam etmektedir. Aynı zamanda 2010 yılından itibaren Kocaeli Üniversitesi Bilgisayar Mühendisliği Bölümü'nde Araştırma Görevlisi olarak çalışmaktadır.