



**HETEROJEN VERİ KAYNAKLARINDA FİKİR MADENCİLİĞİ,
OTOMATİK ÖZELLİK ÇIKARIMI VE DUYGU ANALİZİ**

Barış ÖZYURT

**DOKTORA TEZİ
BİLGİSAYAR BİLİMLERİ ANA BİLİM DALI**

**GAZİ ÜNİVERSİTESİ
BİLİŞİM ENSTİTÜSÜ**

HAZİRAN 2021

Barış ÖZYURT tarafından hazırlanan “HETEROJEN VERİ KAYNAKLARINDA FİKİR MADENCİLİĞİ, OTOMATİK ÖZELLİK ÇIKARIMI VE DUYGU ANALİZİ” adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi Bilişim Enstitüsü Bilgisayar Bilimleri Ana Bilim Dalında DOKTORA TEZİ olarak kabul edilmiştir.

Danışman: Prof. Dr. M. Ali AKCAYOL

Bilgisayar Mühendisliği Bölümü, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum.

.....

Başkan: Prof. Dr. Erdoğan DOĞDU

Bilgisayar Bilimleri Bölümü, Angelo State Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum.

.....

Üye: Doç. Dr. Hacer KARACAN

Bilgisayar Mühendisliği Bölümü, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum.

.....

Üye: Prof. Dr. Hasan Şakir BİLGE

Elektrik Elektronik Mühendisliği Bölümü, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum.

.....

Üye: Doç. Dr. Sevil ŞEN

Bilgisayar Mühendisliği Bölümü, Hacettepe Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum.

.....

Tez Savunma Tarihi: 22/06/2021

Jüri tarafından kabul edilen bu tezin Doktora Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

.....

Prof. Dr. Aslıhan TÜFEKÇİ

Bilişim Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Bilişim Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Barış ÖZYURT

22/06/2021

HETEROJEN VERİ KAYNAKLARINDA FİKİR MADENCİLİĞİ, OTOMATİK ÖZELLİK ÇIKARIMI VE DUYGU ANALİZİ

(Doktora Tezi)

Barış ÖZYURT

GAZİ ÜNİVERSİTESİ

BİLİŞİM ENSTİTÜSÜ

Haziran 2021

ÖZET

Günümüzde sosyal ağların, blogların, forumların ve e-ticaret sitelerinin kullanımının yaygınlaşmasıyla, kullanıcılar tarafından üretilen metinlerin hacmi her geçen gün katlanarak artmaktadır. Ürün yorumlarındaki ve diğer metinlerdeki kullanıcı görüşleri ve fikirleri bu ürünlerin üreticileri, satıcıları ve pazarlayıcıları için hayati öneme sahiptir. Bu yüzden, duygu analizi ve fikir madenciliği önemli bir araştırma alanı haline gelmiştir. Kullanıcı yorumları üzerindeki fikir madenciliğinde konu modelleme tabanlı yaklaşımlar ve Latent Dirichlet Allocation (LDA) özellik tabanlı düzeydeki duygu analizinde özellik çıkarımında önemli yöntemler haline gelmişlerdir. Fakat LDA doğrudan kullanıcı yorumlarında ve kısa metinlerde metin seyrekliği sorunu yüzünden doğrudan kullanılamamaktadır. LDA'nın kısa metinlerde kullanılabilir şekilde uyarlanması ile ilgili çok sayıda çalışma yapılmıştır. Bu çalışmada özellik tabanlı düzeyde duygu analizi ve özellik çıkarımı için özgün bir yöntem olan Sentence Segment LDA (SS-LDA) önerilmiştir. SS-LDA, kısa metinler ve yorumlardan özellik çıkarımı için geliştirilmiştir. Bu çalışmada veri seti olarak akıllı telefonlar hakkındaki müşteri yorumları kullanılmıştır. Yorumlar www.hepsiburada.com e-ticaret sitesinden alınmıştır. Deneysel sonuçlar SS-LDA'nın özellik tabanlı düzeydeki duygu analizinde ürün özellik çıkarımında başarılı olduğunu ortaya koymuştur.

Bilim Kodu : 92429

Anahtar Kelimeler : Duygu analizi, özellik tabanlı duygu analizi, özellik çıkarımı

Sayfa Adedi : 107

Danışman : Prof. Dr. M. Ali AKCAYOL

OPINION MINING IN HETEROGENEOUS DATA SOURCES, AUTOMATIC ASPECT
EXTRACTION AND SENTIMENT ANALYSIS

(Ph. D. Thesis)

Barış ÖZYURT

GAZİ UNIVERSITY
INSTITUTE OF INFORMATICS

June 2021

ABSTRACT

Today while social networks, forums, e-commerce web sites and blogs are covering most of our lives, textual data generated by users is exponentially growing. User opinions in user reviews or in other textual data are very important for retailers, manufacturers and providers of these products and services. Therefore, opinion mining and sentiment analysis have emerged as important research areas recently. In mining user reviews, Latent Dirichlet Allocation (LDA) which is the most popular topic modeling algorithm is a significant method that is used in extracting product aspects in aspect based sentiment analysis. However, LDA algorithm is not so effective on short texts like user reviews or social media messages because of lack of co-occurrence patterns and data sparsity problem. Adaptation of LDA algorithm for short texts is an hot research area in literature. In this study, Sentence Segment LDA (SS-LDA), which is a novel method for aspect based sentiment analysis and aspect extraction is proposed. SS-LDA has been proposed for aspect extraction from short texts and product reviews. In this study, user reviews about smartphones are used as the dataset. The user reviews have been collected from www.hepsiburada.com, which is a popular e-commerce site in Turkey. Experimental results prove that SS-LDA is quite competitive in extracting products aspects.

Science Code : 92429

Key Words : Sentiment analysis, aspect based sentiment analysis, aspect extraction

Page Number : 107

Supervisor : Prof. Dr. M. Ali AKCAYOL

TEŞEKKÜR

Bu doktora tez çalışmasını yerine getirmemde öncelikle en başından sonuna kadar bana yol gösteren ve her türlü desteğini esirgemeyen danışman hocam Prof. Dr. M. Ali AKCAYOL'a, ikinci olarak doktora eğitimim boyunca bana her türlü desteği veren eşim Sinem ÖZYURT'a teşekkür ediyorum.



İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT.....	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	ix
ŞEKİLLERİN LİSTESİ.....	x
SİMGELER VE KISALTMALAR.....	xi
1. GİRİŞ.....	1
1.1. Doküman Düzeyinde Duygu Analizi Çalışmaları.....	3
1.2. Özellik Tabanlı Düzeyde Duygu Analizi Çalışmaları	11
2. FİKİR MADENCİLİĞİ VE DUYGU ANALİZİ PROBLEMİ	21
2.1. Duygu Analizi Çalışma Düzeyleri	22
2.1.1. Doküman düzeyinde duygu analizi.....	22
2.1.2. Cümle düzeyinde duygu analizi	23
2.1.3. Özellik tabanlı düzeyde duygu analizi.....	23
3. DUYGU ANALİZİ VE FİKİR MADENCİLİĞİ ALGORİTMALARI	25
3.1. Duygu Polaritesinin Tespitinde Kullanılan Yöntemler.....	25
3.1.1. Makine öğrenmesine dayalı yöntemler.....	25
3.1.2. Sözcük tabanlı yöntemler	27
3.1.3. Yöntemlerin karşılaştırmalı analizi.....	28
3.1.4. Özellik çıkarımında kullanılan yöntemler	29
4. TEZ İLE İLGİLİ ÖN ÇALIŞMALAR VE ÖN BİLGİLER	33
4.1. Türkçe'deki Doğal Dil İşleme Araçları ve Mevcut Durum	33

	Sayfa
4.2. Morfolojik Analiz ve Morfolojik Belirsizlik	35
4.3. Latent Dirichlet Allocation Algoritması	37
5. GELİŞTİRİLEN YÖNTEM.....	47
5.1. Metin Ön İşleme	54
5.1. Zemberek ile yazım yanlış düzeltme yöntem önerisi	57
5.2. Cümle Polarite Tespiti.....	60
5.3. Cümle Parçalama.....	61
5.4. Gruplama – Sentence Segment LDA (SS-LDA).....	71
5.4.1. LDA’nın kısa metinlerde uygulanması.....	72
5.4.2. Sentence Segment LDA (SS-LDA)	73
5.5. Özellik Birleştirme	77
6. TESTLER VE DENEYSEL SONULAR	79
7. SONUÇ VE ÖNERİLER.....	93
KAYNAKLAR	97
ÖZGEÇMİŞ	106

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 1.1. Doküman düzeyinde yapılmış duygu analizi çalışmaları.....	9
Çizelge 1.2. Özellik tabanlı düzeyde yapılmış duygu analizi çalışmaları.....	19
Çizelge 3.1. Makine öğrenmesine dayalı - Sözcük tabanlı yöntemlerin karşılaştırması	29
Çizelge 4.1. LDA notasyonları	39
Çizelge 5.1. Cümle parçacıklarının ürün özellikleri altında gruplanması	54
Çizelge 5.2. Zemberek'in yazım yanlışlarını düzeltme istatistikleri	58
Çizelge 5.3. Yazım yanlışlarını önerilen yöntemle düzeltme istatistikleri	58
Çizelge 5.4. Veri setinden elde edilmiş örnek confidence değerleri	65
Çizelge 5.5. SS-LDA notasyonları	76
Çizelge 5.6. Akıllı telefon özellikleri	77
Çizelge 6.1. Ürün özellikleri etiketleme sayıları	79
Çizelge 6.2. Kelimelerin ürün özellikleri altında gruplanması	80
Çizelge 6.3. Önerilen yöntemin ürün özelliklerini çıkarmadaki başarı ölçüm sonuçları	83
Çizelge 6.4. Önerilen yöntemin özellik çıkarımında başarı ölçüm sonuçları	85
Çizelge 6.5. Önerilen yöntemin her bir ürün özelliği için başarı ölçümleri	86
Çizelge 6.6. SS-LDA ve SemEval-2016 çalışmaları başarı ölçümü	88
Çizelge 6.7. SS-LDA, Sentence-LDA, Biterm Topic Model ve LDA başarı ölçüm sonuçları	89

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 4.4. Konu Modelleme – LDA grafiksel gösterim	38
Şekil 4.5. Farklı dirichlet parametrelerine göre yoğunluk dağılımları	40
Şekil 4.6. LDA plaka gösterimi	42
Şekil 5.1. Veri yapısı UML sınıf diyagramı	48
Şekil 5.2. Örnek yorum 1'in sınıf yapısı ile temsil edilmesi	49
Şekil 5.3. Örnek yorum 2'nin sınıf yapısı ile temsil edilmesi	50
Şekil 5.4. Sistemin ana hatları	55
Şekil 5.5. Cümle parçalama algoritması akış diyagramı.....	67
Şekil 5.6. SS-LDA ve LDA'nın plaka notasyonları.....	74
Şekil 6.1. Veri setinden etiketlenmiş örnek bir yorum.....	79
Şekil 6.2. K-Fold Cross Validation	84
Şekil 6.3. Test veri setleri elde etme yöntemi.....	84
Şekil 6.4. Önerilen yöntemin her bir ürün özelliği için başarı ölçümleri.....	86
Şekil 6.5. SS-LDA ve SemEval-2016 çalışmaları başarı ölçümü	88
Şekil 6.6. SS-LDA, Sentence-LDA, Bitern Topic Model ve LDA başarı ölçüm sonuçları	90

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar

Açıklama

A-TF	Arttırılmış Terim Frekansı (Augmented Term Frequency)
BoW	Kelime Torbası (Bag of Words)
B-TF	Mantıksal Terim Frekansı (Boolean Term Frequency)
CRF	Koşullu Rastgele Alanlar (Conditional Random Fields)
DDİ	Doğal Dil İşleme
DVM	Destek Vektör Makinesi
FN	Yanlış Negatif (False Negative)
FP	Yanlış Pozitif (False Positive)
KNN	K-En Yakın Komşu (K-Nearest Neighbour)
LA-TF	Logaritmik Ortalama Terim Frekansı (Logarithmic Average Term Frequency)
LDA	Gizli Dirichlet Ataması (Latent Dirichlet Allocation)
LR	Lojistik Regresyon
L-TF	Logaritmik Terim Frekansı (Logarithmic Term Frequency)
ME	Maksimum Entropi
NB	Naive Bayes
N-TF	Normalize Terim Frekansı (Normalized Term Frequency)
PMI	Nokta Tabanlı Müşterek Bilgi (Pointwise Mutual Information)
POS	Sözcük Türü (Part Of Speech)

Kısaltmalar**Açıklama****SMM**

Saklı Markov Modeli

SS

Cümle Parçası (Sentence Segment)

SS-LDA

Cümle Parçası LDA (Sentence Segment LDA)

TF-IDFTerim Frekansı - Ters Doküman Frekansı
(Term Frequency - Inverse Document Frequency)**TN**

Doğru Negatif (True Negative)

TP

Doğru Pozitif (True Positive)



1. GİRİŞ

Günümüzde artık internet hayatımızın her alanında kullanılmaktadır. İnternette insanların duygularını, fikirlerini ve görüşlerini ifade edebildikleri e-ticaret siteleri kullanıcı yorumları, sosyal medya platformları, forumlar gibi web platformları bulunmaktadır. Bu web platformlarında insanların görüşlerini, duygularını ifade ettikleri çok yüksek boyutta metin verisi birikmiş durumdadır. Bu metinsel veriler, başta bu ürün ve hizmetleri üreten ve tedarik eden şirketler olmak üzere çok sayıda kişinin, kuruluşun ve araştırmacının ilgi alanına girmektedir. Bundan ötürü son yıllarda duygu analizi ve fikir madenciliği oldukça önemli bir araştırma alanı haline gelmiştir.

İnsanların web platformlarında ürünler, organizasyonlar, hizmetler, siyasi fikirler ve olaylar, vs. gibi konular ile ilgili görüşlerini bildirdikleri metinler içindeki saklı olan düşünce, fikir ve duyguları ortaya çıkarmayı amaçlayan çalışma alanına duygu analizi ve fikir madenciliği denilmektedir. Duygu analizi çalışmaları metinlerdeki saklı olan fikir ve düşünceleri ortaya çıkarmayı amaçladığı için doğal dil işleme ile çok yakında ilişkili bir disiplindir. Duygu analizi, doğal dil işleme dışında metin madenciliği, veri madenciliği ve web madenciliği çalışma alanları altında da yapılmaktadır [1].

Duygu analizi çalışmaları doküman düzeyi duygu analizi, cümle düzeyi duygu analizi ve özellik tabanlı düzeydeki duygu analizi olmak üzere üç farklı düzeyde yapılmaktadır. Doküman düzeyinde duygu analizi dokümanın bütün olarak nötr/pozitif/negatif olarak sınıflandırılması çalışmasıdır. Bu şekilde nötr/pozitif/negatif şeklinde sınıflandırmaya “duygu polarite” tespiti denilmektedir. Cümle düzeyinde duygu analizinde dokümandaki her bir cümlenin polaritesi nötr/pozitif/negatif şeklinde tespit edilmektedir. Özellik tabanlı düzeydeki duygu analizi çalışmasında ise kullanıcıların ifade ettikleri görüşler ve bu görüşlerin hedefi olan ürün özellikleri, birlikte ortaya çıkarılmaktadır. Kullanıcı görüşlerinin hedefi olan ürün özelliklerinin ortaya çıkarılmasına “özellik çıkarımı” denilmektedir. Özellik tabanlı duygu analizi çalışmalarında özellik çıkarımı duygu polarite tespitine göre çoğu zaman daha zorlu bir görevdir.

Bu tezdeki çalışma özellik tabanlı düzeyde yapılmış bir duygu analizi çalışmasıdır ve özgün bir özellik çıkarımı yöntemi önerilmiştir. Yöntem adı SS-LDA (Sentence Segment

LDA) olarak adlandırılmıştır. SS-LDA, en çok bilinen konu modelleme algoritması olan Latent Dirichlet Allocation (LDA) temelli bir yöntemdir. LDA, 2003 yılında Blei, Ng ve Jordan tarafından önerilmiştir [2]. Önerildiği 2003 yılından bu yana, LDA doküman ve metinlerdeki saklı olan konuları keşfetmek için kullanılmaktadır. LDA uzun metinlerde ve büyük metin yığınlarında başarılı olan bir yöntemdir. Ürün yorumları, Twitter mesajları, sosyal medya mesajları gibi kısa metinlerde etkili olamamaktadır. Son yıllarda kısa metin madenciliği ve duygu analizi önemli araştırma alanları haline geldikçe LDA algoritmasını kısa metinlerde başarılı olacak şekilde uyarlanması önemli bir araştırma alanı haline gelmiştir. Bu çalışmada önerilen SS-LDA da LDA'nın kısa metinlerde çalışacak şekilde özgün bir uyarlamasıdır.

Özgün bir yöntem önerisi dışında, bu çalışmanın literatüre önemli bir diğer katkısı da Türkçe dili üzerinde yapılmış olmasıdır. Duygu analizi ve fikir madenciliği çalışmaları doğal dil işleme ile çok yakından ilişkilidir ve dile bağımlılığı yüksektir. Dünyada duygu analizi çalışmalarının çok büyük bir kısmı İngilizce dili üzerinde yapılmaktadır. Türkçe dilinde yapılmış duygu analizi çalışmalarının sayısı oldukça azdır, özellikle de özellik tabanlı düzeydeki duygu analizi çalışmaları çok azdır. Literatürde Türkçe dilinde özellik tabanlı düzeyde duygu analizi çalışmaları alanında önemli bir boşluk bulunmaktadır. Bu çalışma literatürdeki bu eksikliği doldurması bakımından önemlidir.

Bu paragraftan sonraki bölümde literatürdeki duygu analizi ve fikir madenciliği çalışmaları Doküman Düzeyinde Duygu Analizi Çalışmaları ve Özellik Tabanlı Düzeydeki Duygu Analizi Çalışmaları başlıkları altında literatür taraması şeklinde özetlenmişlerdir.

Literatür taramasında 2010 yılından günümüze kadar olan çalışmalar incelenmiştir. Literatürde 2010 yılından günümüze kadar olan periyotta çok sayıda duygu analizi çalışması vardır. Bu çalışmalar seçilirken mümkün mertebede özgün yöntemlerin önerildiği çalışma örnekleri dâhil edilmeye çalışılmıştır. 2010 yılından sonraki çalışmaların yanında daha önceki tarihlerde yapılmış yenilikçi olan ve çok sayıda araştırmacı tarafından referans verilen bazı çalışmalar da taramaya dâhil edilmişlerdir.

Tetsuya ve Jeonghee [3] tarafından 2003 yılında ilk defa duygu analizi (sentiment analysis) ifadesi kullanılmıştır. Kushal, Steve ve Pennock [4] tarafından ise yine 2003 yılında ilk defa “opinion mining” (fikir madenciliği) tabiri ise kullanılmıştır. Aslında bu konudaki çalışmalar 2003 yılından daha önce yapılmaya başlanmıştır. Bu alandaki ilk çalışmalar Vasileios ve Janyce [5], Tong [6], Turney [7], Pang, Lee ve Vaithyanathan [8] tarafından 2000’li yılların hemen başında yapılmışlardır.

1.1. Doküman Düzeyinde Duygu Analizi Çalışmaları

Duygu analizi ve fikir madenciliği bilimi ilk olarak doküman düzeyinde sınıflandırma çalışmaları olarak ortaya çıkmıştır. Önceki bölümde, bu alandaki ilk çalışmalar olarak verilen [5-8] çalışmalarının hepsi doküman düzeyinde yapılmış araştırmalardır.

McDonald, Hannan, Neylon, Wells ve Reynar koşullu rastgele (CRF) alanlar yönteminin benzeri hiyerarşik bir öğrenme algoritmasını ortaya koydukları bir çalışma yapmışlardır [9]. Önerdikleri yöntemde doküman düzeyinde sınıflandırma, cümle düzeyinde sınıflandırma birbirleriyle paralel bir şekilde yapılmaktadır. McDonald ve diğerleri duygu sınıflandırmasını iki seviyede birlikte paralel yapmanın başarıyı arttırdığını ortaya koymuşlardır.

Nakagawa, Inui ve Kurohashi Japonca ve İngilizce dilleri üzerinde yaptıkları çalışmada CRF ve bağlılık ağacı algoritmalarından faydalandıkları bir yöntem önerisinde bulunmuşlardır [10]. Önerdikleri yöntemin test ve başarı ölçümünde, literatürdeki altı tane yaygın yöntemle karşılaştırmasını yapmışlardır. Diğer yöntemler ile %62 ile %83 aralığında değişen başarı seviyeleri elde edilmiştir. Önerdikleri yöntem %76 ile %85 aralığında değişen başarı ortaya koymuştur.

Davidov, Tsur ve Rappoport yaptıkları çalışmada Twitter mesajları üzerinde duygu sınıflandırması yapmışlardır [11]. Denetimli makine öğrenmesi yöntemini uyguladıkları çalışmalarında özellik vektörünü oluştururken metin madenciliğinde yaygın şekilde kullanılan özellikler haricinde diğer metinlerde olmayıp web platformlarında ve Twitter’da olan smiley karakterleri ve hashtag’ler gibi özelliklerden de faydalanmışlardır. Twitter’ın bu kendine özgü özelliklerinden faydalanmanın başarı seviyesini yükselttiğini ortaya koymuşlardır.

Taboada, Brooke, Tofiloski, Voll ve Stede yaptıkları duygu analizi çalışmasında ikili sınıflandırma değil, kullanıcı yorumlarını derecelendirme ile ilgili bir yöntem önermişlerdir [12]. Kendi oluşturdukları duygu terimleri sözlüğünü kullandıkları çalışmada sözlüğü Amazon'un bir servisi olan Mechanical Turk ile oluşturmuşlardır. Mechanical Turk kullanarak oluşturulan duygu terimleri sözlüğü ile elde edilen başarı seviyesini, SentiWordNet, Google-PMI, Maryland gibi duygu terimleri sözlükleriyle elde edilen başarı seviyeleriyle karşılaştırılmış ve onlardan daha başarılı olduğu kanıtlanmıştır.

Kang, Yoo ve Han veri seti olarak restoran yorumlarını kullandıkları çalışmada duygu polarite tespiti için bir yöntem önermişlerdir [13]. Çalışmada sadece yöntem önerilmemiş, aynı zamanda farklı denetimli makine öğrenmesi yöntemleri ve farklı makine öğrenmesi özellikleri karşılaştırılmıştır. Çalışmada duygu terimleri sözlüğü kullanmışlar ama hazır bir sözlük kullanmamışlar, sözlüğü kendileri oluşturmuşlardır. Oluşturdukları sözlük sadece tek kelimelik duygu terimleri değil, iki kelimededen oluşan kalıpları da içermektedir. Yaptıkları testlerde başarı oranının negatif yorumlarda, pozitif yorumlara göre %10 daha düşük olduğunu gözlemlemişler ve bu %10'luk farkı düşürmek için Naive Bayes (NB) sınıflandırıcısı üzerinde geliştirme çalışması yapmışlardır. Test ve başarı ölçümü aşamasında standart duygu terimleri sözlüğü ile kendi oluşturdukları duygu terimleri sözlüğünü; NB, geliştirilmiş NB ve Destek Vektör Makinesi (DVM) sınıflandırıcılarını kıyaslamışlardır. Ayrıca unigram, bigram ve unigram + bigram özelliklerini karşılaştırmışlardır.

Mejova ve Srinivasan yaptıkları duygu analizi çalışmasında, denetimli öğrenmede kullanılan özellik vektörü özelliklerinin etkinliklerini karşılaştırmışlardır [14]. “kelime - kelime kökü”, “ikili - terim frekans ağırlığı”, “olumsuzluk kelime ve ekleri”, “n-gram” gibi özelliklerin duygu sınıflandırmada başarı oranı üzerindeki etkisi ilgili bir çalışmışlardır. Yaptıkları deneysel sonuçlar özellikle kelime n-gram tercihinin başarı seviyesi üzerinde önemli bir etkiye sahip olduğunu göstermiştir. “Kelime - kelime kökü”, “ikili - terim frekans ağırlığı”, “olumsuzluk kelime ve ekleri” şeklindeki kıyaslamalarda, bu kriterlerin başarı seviyesinde kayda değer bir etkilerinin olmadığını ortaya koymuşlardır. Bespalov, Bai, Qi ve Shokoufandeh de kelime n-gramları ile ilgili benzer bir çalışma yapmışlar ve bazı durumlarda daha iyi bir başarı yakalayabilmek için daha yüksek seviyeli kelime n-gramları kullanmanın gerektiğini ortaya koymuşlardır [15]. Fakat kelime n-gramlarında seviye arttıkça hesaplama maliyeti geometrik olarak artmakta ve özellikle de veri seti

büyükse hesaplama maliyeti çok yükselmektedir. Bu durumda yüksek seviye n-gram kullanımı mümkün olamamaktadır. Bespalov ve diğerleri bu gerçekleri ortaya koymuşlar ve yüksek seviyeli kelime n-gramları ile ilgili hesaplama maliyetini düşürme amacıyla “supervised embedding strategy” isimlendirdikleri bir yöntem önermişlerdir. Önerdikleri yöntem yapay sinir ağları yaklaşımına dayalıdır. Yaptıkları testlerde “supervised embedding strategy” yöntemiyle yüksek seviyeli n-gram’ları ile duygu sınıflandırmasını yapmışlardır. Yüksek seviyeli kelime n-gram’larının başarıyı yükselttiğini ortaya koymuşlardır.

Balahur, Hermida ve Montoyo yaptıkları çalışmada pozitif/negatif şeklinde ikili sınıflandırma yerine mutlu, mutsuz, sinirli, suçluluk hissi gibi psikolojik bir takım duygulara göre sınıflandırma yapmışlardır [16]. Çalışmalarında Appraisal Theory'den faydalanmışlardır. Psikolojide Appraisal Theory insanların varlıklar ve olaylar ile ilgili fikir oluşturma süreçlerini analiz eden bir çalışma alanıdır. Ayrıca insanların farklı olay ve durumlar karşısındaki tutumları, davranışları gibi tepkileri hakkında geliştirilmiş bir bilgi bankası olan ISEAR'dan [17] istifade etmişlerdir. Balahur ve diğerleri önerdikleri yöntemlerini, dayalı makine öğrenmesine dayalı yöntemlerle kıyaslamışlar ve kendi yöntemlerinin daha başarılı olduğunu ortaya koymuşlardır.

Zhang, Hua, Wang, Qian ve Zhang AppleStore'daki mobil uygulama yorumları üzerinde duygu analizi çalışması yapmışlardır. Denetimli makine öğrenmesi metotlarını ve bu metotların farklı konfigürasyonlara göre etkinliklerinin kıyaslamasını yapmışlardır [18]. Çalışmalarında, DVM ve NB sınıflandırıcılarının, farklı kelime n-gramlarının, farklı uzunluklardaki yorumlardaki başarı seviyesini nasıl etkilediğinin testlerini gerçekleştirmişlerdir. Elde edilen test sonuçlarında, sınıflandırıcı olarak NB DVM’den daha başarılı çıkmıştır. Bigram, kelime n-gramları arasında en başarılısı olarak ortaya çıkmıştır. Kısa yorumlarda başarı seviyesinin uzun yorumlara kıyasla ise daha yüksek olduğunu gözlemlemişlerdir.

Zhang ve diğerleri, veri seti olarak ürün yorumları ve Facebook mesajlarını kullandıkları fikir madenciliği çalışması yapmışlardır [19]. Yaptıkları çalışmada önerdikleri yöntem cümlelerin linguistik analizine dayanmaktadır. Önerdikleri yöntemde CRF yaklaşımıyla cümlelerin kendinden sonra ve kendinden önce gelen cümlelerle olan ilişkilerini, cümle bağlaçlarını detaylı şekilde analiz etmişlerdir. Önerdikleri yöntemi, DVM, Saklı Markov

Modeli (HMM), Lojistik Regresyon (LR), ve Choi ve Cardie tarafından önerilmiş Compositional Semantic Rules (CSR) [20] yöntemleriyle kıyaslamalarını gerçekleştirmişlerdir. Önerdikleri yöntem ürün yorumlarında, en yüksek başarı seviyesini ortaya koymuştur. Fakat Facebook mesajlarında en yüksek başarı seviyesi CSR yöntemi ile elde edilmiştir, önerilen yöntemin diğer yöntemlere göre bir üstünlüğü gözlemlenmemiştir. Facebook mesajları çoğunlukla 1-2 cümlelik kısa metinlerdir. Bundan ötürü kullandıkları CSR yöntemi çok etkin olamamıştır.

Habernal, Ptacek ve Steinberger, Çek diliyle sinema yorumları, ürün yorumları ve Facebook mesajları veri setiyle duygu analizi çalışması yapmışlardır [21]. Bu çalışmada DVM ve Maksimum Entropi sınıflandırıcılardan faydalanarak denetimli makine öğrenmesine dayalı bir yöntemin önermişlerdir. Sözcük türleri ile ilgili özellikler, kelime unigram ve bigramı, TF-IDF, karakter n-gramı ve smiley karakterleri özelliklerini kullanmışlardır. Facebook mesajları genellikle kısa mesajlardır. Ayrıca web platformlarında yaygın olan birçok jargon ve kısaltma içermektedirler. Bundan ötürü için karakter n-gram kullanımının Facebook mesajları üzerinde oldukça etkin olduğunu gözlemlemişlerdir. Facebook mesajlarına kıyasla daha uzun metinler olan sinema yorumları ve ürün yorumlarında ise en başarılı sonuçların kelime bigramı ile elde edildiğini gözlemlemişlerdir.

Park, Lee ve Moon yaptıkları çalışmada duygu terimleri sözlüğü oluşturmak için bir yöntem önerisinde bulunmuşlardır [22]. Çalışmalarında, eğitim veri setinden, duygu terimleri çıkarımı yapma ile ilgili için bir yöntem önerisinde bulunmuşlardır. Önerdikleri yöntem kullanılarak oluşturulan sözlük ile SentiWordNet duygu terimleri sözlüğünün kıyaslamasını gerçekleştirmişlerdir. Eğitim verisi az olduğunda önerdikleri yöntem ile oluşturulmuş sözlüğün başarı seviyesi daha düşük çıkmıştır. Ancak, eğitim verisi yeteri miktarda olduğu zaman SentiWordNet'e göre daha yüksek bir başarı elde edildiğini gözlemlemişlerdir.

Fernandez-Gavilanes, Alvarez-Lopez, Juncal-Martinez, Costa-Montenegro ve Gonzalez-Castano Twitter mesajları ve sinema yorumları veri setlerini kullanarak duygu sınıflandırma çalışması yapmışlardır [23]. Çalışmalarında sözcük tabanlı ve cümlelerin linguistik analizine dayalı bir yöntem uygulamışlardır. Linguistik analiz için bağıllık ağacı veri yapısını kullanmışlardır. Zıtlık bağlaçlarını, pekiştiricileri, olumsuzluk kelime ve

eklerini detaylı şekilde analiz etmişlerdir. Çok farklı veri setleri üzerinde yaptıkları testlerde %60 - %75 aralığında değişen başarı seviyeleri elde etmişlerdir.

Khan, Qamar ve Bashir ürün yorumları, sinema yorumları veri setlerini kullanarak DVM sınıflandırıcısı ile denetimli makine öğrenmesine dayalı bir yöntem ile duygu analizi çalışması yapmışlardır [24]. Bu çalışmada özgün bir İngilizce duygu terimleri sözlüğü oluşturma yöntemi önerilmiştir. İngilizce duygu terimleri sözlüğü olan SentiWordNet'in duygu analizi çalışması yapmak için çok etkin olmadığını ortaya koymuşlardır. Daha sonra SentiWordNet'ten faydalanarak yeni bir İngilizce duygu terimleri sözlüğü oluşturma yöntemi önermişlerdir. Önerdikleri yöntemle oluşturdukları sözlük ve SentiWordNet sözlüğü ile karşılaştırmalı testleri gerçekleştirmişlerdir. Yaptıkları testlerde önerdikleri yöntemler kullanılarak oluşturulan sözlüğün SentiWordNet'e göre daha etkin olduğunu ortaya koymuşlardır.

Chen, Xu, He ve Wang ürün yorumları, sinema yorumları veri setleri üzerinde yaptıkları duygu analizi çalışmasında özgün yöntem önerisinde bulunmuşlardır [25]. Çalışmada ilk olarak cümleler duygu ifadelerinin hedeflerine göre hedefsiz cümleler, tek hedefli cümleler ve çok hedefli cümleler olarak üç gruba ayrılmışlardır. Böyle gruplara ayırmalarının nedeni olarak da bu üç grup altındaki cümlelerin cümle yapılarının birbirlerinden oldukça farklı olduğunu ortaya koymuşlar ve her bir grubun ayrı ayrı ele alınmaları gerektiğini belirtmişlerdir. Cümleleri bu şekilde üç gruba göre sınıflandırmayı CRF tabanlı bir yöntem ile gerçekleştirmişlerdir. Daha sonra evrişimli sinir ağlarına dayalı bir yöntem uygulayarak bu üç grup altındaki cümlelerin duygu polaritelerini tespit etmişlerdir.

Türkçe dili ile duygu analizi çalışmalarına İngilizce'den çok daha sonraları başlanmıştır. Eroğul tez çalışması olarak sinema yorumları veri seti üzerinde makine öğrenmesine dayalı bir yöntem kullanarak duygu analizi çalışması yapmıştır [26]. Çalışmasında özellik vektörü için Bag of Words (BoW) modelini, sözcük türü ile ilgili özellikleri kullanmıştır. Sınıflandırma algoritması olarak ise DVM'yi tercih etmiştir. Farklı parametrelerle yaptığı test ve başarı ölçümlerinde %72 - %85 aralığında değişen başarı seviyeleri elde etmiştir.

Kaya, Fidan ve Toroslu yaptıkları duygu analizi çalışmasında Türkçe siyasi haberler üzerinde çalışmışlardır [27]. Ürün yorumları gibi metinler, düşünce bakımından yoğun ve konu odaklı metinler iken, siyasi haberler düşünce bakımından daha düşük yoğunluktadır.

Ayrıca siyasi haberler ürün yorumları kadar konu odaklı değildirler. Kaya ve diğerleri makine öğrenmesine dayalı yöntemler dışında sözcük tabanlı yöntemleri de uygulamışlardır. Siyasi haberler için duygu terimleri listesi oluşturmuşlardır. Yaptıkları test ve ölçümlerde %66 ile %77 aralığında değişen başarı düzeylerine ulaşmışlardır.

Şimşek ve Özdemir Twitter mesajlarını mutlu ve mutsuz şeklinde sınıflandırdıkları bir duygu analizi çalışması yapmışlardır [28]. Bu çalışmada Amazon'un sunduğu Mechanical Turk servisini kullanılarak üzüntü ve mutluluk ifadelerinde kullanılan 113 kelimeden oluşan Türkçe duygu terimleri sözlüğü oluşturmuşlardır. Daha sonra bu kelimelerin Twitter mesajlarındaki frekanslarına dayalı olarak sınıflandırmayı gerçekleştirmişlerdir.

Çetin ve Amasyalı Türkçe Twitter mesajları veri seti ile duygu sınıflandırma çalışması yapmışlardır. Çalışmalarında WEKA sınıflandırma aracını kullanmışlardır. Sınıflandırma algoritması olarak ise NB, DVM, karar ağacı, rastgele orman ve 1-en yakın komşu sınıflandırıcılarını uygulamışlardır [29]. Özellik vektöründe TF, TF-IDF ve Delta TF-IDF [30] terim ağırlıklandırma yöntemlerinden faydalanmışlardır. Yapılan testlerde sonuçlarda Delta TF-IDF yönteminin diğer iki yöntemden daha başarılı olduğu gözlemlenmiştir.

Aytekin yaptığı duygu sınıflandırma çalışmasında veri seti olarak Türkçe blog sitelerindeki beyaz eşya yorumlarını kullanmıştır [31]. Çalışmasında makine öğrenmesine dayalı yöntem ile sözcük tabanlı yöntemin karışımı yarı-denetimli bir yöntem uygulamıştır. Aytekin, İngilizce'de hazırlanmış bir duygu terimleri sözlüğünü Türkçe'ye tercüme ederek Türkçe duygu terimleri sözlüğü oluşturmuştur. Çalışmasında 350 tane pozitif, 350 tane negatif yorum üzerinde yaptığı testlerde negatif yorumlarda %73, pozitif yorumlarda %72 başarı seviyesi elde etmiştir.

Sevindi tez çalışmasında, duygu analizinde kullanımı yaygın olan terim ağırlıklandırma yöntemlerinin ve sınıflandırma yöntemlerinden NB, DVM ve KNN'nin karşılaştırmasını yapmıştır [32]. Sevindi sinema yorumlarını kullandığı çalışmada TF-IDF, A-TF, B-TF, LA-TF, L-TF ve N-TF terim ağırlıklandırma yöntemlerinin etkinliklerin, karşılaştırmıştır. Bu ağırlıklandırma yöntemlerinin hepsi unigram, bigram ve 3-gram BoW modelleri ile ayrı ayrı test edilmiştir. Test sonuçlarına göre TF-IDF ağırlıklandırma yönteminin genel olarak başarı seviyesi diğerlerinden daha yüksek ölçülmüştür.

Çizelge 1.1. Doküman düzeyinde yapılmış duygu analizi çalışmaları

Ref. No	Yıl	Dil	Çalışma Türü	Veri Seti	Uygulanan Yöntem/Algoritma	Sonuçlar
[9]	2007	İngilizce	Yöntem önerilmiştir	Ürün yorumları	CRF	%62 ve %70 aralığında değişen başarı düzeyleri elde edilmiştir
[10]	2010	Japonca, İngilizce	Yöntem önerilmiştir	Ürün yorumları, Sinema yorumları	CRF, Bağlılık Ağacı	%77 - %86 aralığında değişen başarı düzeyleri elde edilmiştir
[11]	2010	İngilizce	Yöntem önerilmiştir	Twitter mesajları	Kısaltmalar ve Hashtag gibi Twitter'a özgü metinleri kapsayan BoW modeli	%57 - %86 aralığında değişen başarı seviyeleri elde edilmiştir.
[12]	2011	İngilizce	Duygu derecelendirme	Ürün yorumları	Sözcük tabanlı. Duygu terimleri sözlüğü "Mechanical Turk" servisi ile oluşturulmuştur.	Farklı duygu terimleri sözlükleriyle %58 - %78 arası değişen başarı düzeyleri elde edilmiştir.
[13]	2012	İngilizce	Kıyaslamalı Analiz	Restoran yorumları	Unigram, Bigram, Unigram+Bigram modelleri, NB, DVM kıyaslanmıştır. Improved NB yöntem önerisi yapılmıştır.	Önerdikleri Improved NB'nin en iyi başarı seviyelerini elde ettiğini ortaya koymuşlardır.
[14]	2011	İngilizce	Kıyaslamalı Analiz	Ürün yorumları	"kelime - kelime kökü", "olumsuzluk kelime ve ekleri", "ikili - terim frekans ağırlığı", "n-gram" (unigram, bigram, 3-gram) özellikleri karşılıklı karşılaştırılmıştır	Farklı özellik seçim yöntemlerine göre çok detaylı ve kapsamlı sonuçlar elde etmişlerdir.
[15]	2011	İngilizce	Yüksek seviyeli n-gramlar için yöntem önerisi	Ürün yorumları	Yüksek seviye n-gramlarda hesaplama süresini azaltmak için "Neural Network" tabanlı bir yöntem önerilmiştir.	Yüksek seviye n-gramların hesaplama sürelerini düşürmüşlerdir.
[16]	2012	İngilizce	Yöntem önerilmiştir	Sosyal mesajlar	Psikolojideki Appraisal Theory ve ISEAR bilgi bankasından faydalanılmıştır.	%22 - % 62 aralığında değişen başarı seviyeleri elde etmişlerdir.
[18]	2014	İngilizce	Kıyaslamalı Analiz	Apple Store yorumları	Farklı n-gram modelleri, DVM, NB sınıflandırıcıları karşılaştırılmıştır	NB'nin DVM'den biraz daha iyi olduğu, n-gram'lardan 2-gramın en iyi olduğu sonuçlarını elde etmişlerdir.
[19]	2014	İngilizce	Yöntem önerilmiştir	Facebook mesajları, Ürün yorumları	Linguistik analiz, CRF	Facebook mesajları veri setinde %61, ürün yorumları veri setinde %66 - %72 aralığında başarı seviyeleri elde etmişlerdir.
[21]	2015	Çek Dili	Yöntem önerilmiştir	Facebook mesajları	Kelime n-gram, Karakter n-gram, TF-IDF	Testlerde karakter n-gram'ın başarı oranını arttırdığını gözlemlemişlerdir.
[22]	2015	İngilizce	Sözlük oluşturma yöntem önerisi	Sosyal medya mesajları	Etiketlenmiş eğitim veri seti kullanarak duygu terimleri sözlüğü oluşturmuşlardır	Yeterli miktarda eğitim verisi olduğunda SentiWordNet'ten daha başarılı olduğunu göstermişlerdir

Çizelge 1.1. (devam) Doküman düzeyinde yapılmış duygu analizi çalışmaları

[23]	2016	İngilizce	Yöntem önerilmiştir	Sinema yorumları Twitter mesajları	Bağlılık ağacıyla detaylı linguistik analiz	%60 - %75 aralığında başarı seviyeleri elde etmişlerdir.
[24]	2016	İngilizce	Duygu terimleri sözlüğü oluşturma yöntem önerisi	Sinema yorumları Ürün yorumları	SentiWordNet sözlüğünden faydalanarak duygu analizi için daha kullanışlı sözlük oluşturmuşlardır	Oluşturdukları sözlüğün duygu analizinde SentiWordNet'ten daha başarılı olduğunu ortaya koymuşlardır
[25]	2017	İngilizce	Yöntem önerilmiştir	Sinema yorumları Ürün yorumları	CRF, Evrişimli sinir ağları	Birçok farklı yöntemle kıyaslaması yapılmış ve %85'lik başarı oranı ile en yüksek sonucu elde edilmiştir.
[26]	2009	Türkçe	Yöntem önerilmiştir	Sinema Yorumları	BoW, sözcük türü, DVM	Farklı konfigürasyonlarla yapılan testlerde %75 - %89 aralığında değişen başarı oranları elde edilmiştir
[27]	2012	Türkçe	Yöntem önerilmiştir	Politika haberleri	Hibrit yaklaşım. Ayrıca İngilizce ve Türkçe metinlerde performans kıyaslaması yapmışlardır.	Farklı parametrelerle yapılan testlerde %62 - %82 aralığında değişen başarı oranları elde edilmiştir.
[28]	2012	Türkçe	Yöntem önerilmiştir	Twitter mesajları	Mechanical Turk servisi ile sözlük oluşturulmuştur. WEKA aracı ile sınıflandırma	Tweetlerdeki mutluluk düzeyinin borsa endeksi ile ilişkisini analiz etmişlerdir. %45'lik bir korelasyon gözlemişlerdir.
[29]	2013	Türkçe	Yöntem önerilmiştir	Twitter mesajları	TF, TF-IDF, NB, 1-en yakın komşu, DVM, Karar Ağacı, Rastgele Orman	Farklı parametrelerle karşılaştırmalı testler yapılmış ve sonuçlar elde edilmiştir.
[31]	2013	Türkçe	Yöntem önerilmiştir	Ürün yorumları	Sözcük tabanlı yöntem ve NB sınıflandırıcısı ile dayalı makine öğrenmesi bir arada.	Yapılan testlerde %73 oranında başarıya ulaşılmıştır.
[32]	2013	Türkçe	Kıyaslamalı analiz	Sinema Yorumları	Özellik vektörü özellikleri ve DVM, NB, K-NN sınıflandırma algoritmaları karşılıklı kıyaslanmıştır.	Farklı n-gramlar ve DVM, NB, K-NN gibi farklı sınıflandırıcılarla çok çeşitli sonuçlar elde etmişlerdir.
[33]	2014	Türkçe	Yöntem önerilmiştir	Twitter mesajları	Karakter 2-gram, 3-gram. NB, DVM, Rastgele Orman	Farklı parametrelerle yaptıkları testlerde %59 - %79 aralığında değişen başarı oranları elde etmişlerdir.
[34]	2014	Türkçe	Kıyaslamalı analiz	Sinema yorumları	“Information Gain” ve “Chi Square” özellik algoritmaları DVM ve NB kullanarak kıyaslanmıştır	“Information Gain” ve “Chi Square” birbirlerine benzer sonuçlar ortaya koymuştur. DVM ise NB'den daha başarılı sonuçlar ortaya koymuştur.

Meral ve Diri siyaset, spor, sađlık gibi konular hakkındaki Twitter mesajları veri seti üzerinde duygu sınıflandırma çalışması yapmışlardır [33]. Çalışmada sosyal ağlarda kullanılan kısaltma ve jargonları özellik olarak kapsayabilmek için karakter n-gramı kullanmışlardır. Karakter 2-gram ve 3-gramları kullanmışlardır. Sınıflandırıcı olarak NB, DVM ve rastgele orman ile duygu sınıflandırmasını yapmışlardır. Bu sınıflandırıcılardan en iyi sonuçları DVM ile elde etmişlerdir ama diğler sınıflandırıcıların da başarı oranları çok yakın olarak ölçülmüştür.

Akba, Uçan, Sezer ve Sever sinema yorumları üzerinde NB ve DVM sınıflandırıcıları ve “Chi Square” ile “Information Gain” özelliklerinin başarılarını kıyasladıkları duygu analizi çalışması yapmışlardır [34]. Test sonuçlarında “Chi Square” ile “Information Gain” özelliklerinin başarı seviyeleri arasında belirgin bir fark gözlemlenmemiştir. Sınıflandırıcılarda DVM’nin NB’den daha başarılı olduğunu gözlemlemişlerdir.

Çizelge 1.1’de doküman düzeyinde duygu analizi çalışmalarının kısa özeti tablo şeklinde sunulmuştur.

1.2. Özellik Tabanlı Düzeyde Duygu Analizi Çalışmaları

Duygu analizi çalışmaları literatürde ilk olarak doküman düzeyinde yapılmaya başlanmış, özellik tabanlı düzeydeki duygu analizi çalışmaları daha sonra başlamıştır. 2004 yılında Hu ve Liu’nun ürün yorumları veri seti ile yaptıkları özellik tabanlı düzeyde çalışma, bu çalışmaların ilk örneğidir [35]. Hu ve Liu bu çalışmalarında sözcük tabanlı ve duygu terimleri sözlüğü kullandıkları bir yöntem uygulamışlardır. Sözcük tabanlı yöntemde duygu terimleri sözlüğünü oluşturmak için önce belli bir sayıda tohum (seed) duygu terimleri belirlemişler, ardından WordNet’in zıt ve eş anlamlı sözcükler fonksiyonuyla bu terimleri genişletmişler ve duygu terimleri sözlüğünü oluşturmuşlardır. Özellik çıkarımı için ise detayları sonraki bölümlerde anlatılacak olan “sık geçen isimler” ismini verdikleri yöntemi önermişlerdir. “Sık geçen isimler” yöntemi daha sonra birçok araştırmacı tarafından özellik çıkarımında kullanılmıştır.

Ding, Liu ve Yu sözcük tabanlı bir yöntem uyguladıkları duygu analizi çalışmalarında NLProcessor aracını kullanarak linguistik analiz yapmışlardır. Linguistik analizde zıtlık

bağlaçları ile özel olarak ilgilenmişlerdir [36]. Ayrıca özellik çıkarımı için cümlede kelimelerin arasındaki mesafeyi dikkate aldıkları bir yöntem geliştirmişlerdir. Ding ve diğerlerş yöntemlerini Hu ve Liu'nun [35]'da kullandıkları veri seti üzerinde [35] ve [37]'de önerilen yöntemlerle kıyaslamalı olarak test etmişler ve daha kendi önerdikleri yöntemin daha başarılı olduğunu ortaya koymuşlardır.

Wu, Zhang, Huang ve Wu ürün yorumları üzerinde bağıllık ağacı tabanlı bir yöntem uyguladıkları özellik tabanlı düzeyde duygu analizi çalışması yapmışlardır [38]. Önerdikleri yöntemde “phrase dependency tree” olarak isimlendirdikleri ve kendi geliştirdikleri daha farklı bir yapıdaki bağıllık ağacı kullanmışlardır. Normalde cümle bağıllık ağacına dönüştürüldüğü zaman her bir düğümde bir kelime tutulur ve düğümler arasındaki bağlantılarla cümledeki kelimeler arası bağımlılıklar temsil edilir. Phrase dependency tree yapısında düğümlerde cümlelerin anlam bütünlüğü oluşturan alt parçacıkları tutulmaktadır. Buradaki amaç isim tamlaması gibi anlamlı cümle alt parçacıklarının parçalanmaması ve anlam kaybı oluşmamasıdır. Önerilen yöntemde Phrase dependency tree'de düğümler ve düğümler arasındaki ilişkiler analiz edilerek duygu ifadelerinin hedefi olan ürün özellikleri ve duygu polariteleri belirlenmeye çalışılmıştır. Wu ve diğerleri literatür kıyaslamasını yöntemlerini yaygın kullanılan diğer yöntemlerle karşılaştırarak yapmışlardır. Karşılaştırmada kullandıkları veri seti çok karmaşık bir veri seti olması sebebiyle karşılaştırma yapılan yöntemlerin başarı düzeyleri %26, %29'lara kadar düşmüştür. Önerdikleri yöntemde %54 ile %69 aralığında değişen başarı seviyesi elde edilmiştir.

Wei ve Gulla yaptıkları özellik tabanlı düzeydeki duygu analizi çalışmasında ontolojilerden faydalanmışlardır [39]. Çalışmalarında ontoloji biliminden faydalanmışlar ve Sentiment Ontology Tree (SOT) olarak adlandırdıkları bir ağaç yapısı kullanmışlardır. SOT'da ürünün kendisini, hiyerarşik yapıda ürünün tüm özellikleri ve bunların duygu polaritesi tutulmaktadır. We ve Gulla yöntemlerini Cesa-Bianchi, Gentile ve Zaniboni'nin geliştirdikleri H-RLS olarak adlandırılmış olan hiyerarşik sınıflandırma yöntemi [40] üzerine inşaa etmişlerdir. HL-SOT olarak isimlendirdikleri hiyerarşik bir sınıflandırma yöntemi ile ürünün özelliklerine göre duygu polarite tespiti yapmışlardır.

Brody ve Elhadad yaptıkları özellik tabanlı düzeydeki duygu analizi çalışmasında restoran yorumları veri seti üzerinde çalışmışlardır [41]. Ürün özellik çıkarımı için geliştirdikleri

yöntem LDA algoritmasının kısa metinler için adapte edilmesi ve cümlelerden özellik çıkarımı yapacak şekilde bir uyarlanmasıdır. Brody ve Elhadad yöntemlerinde özellik çıkarımı yaptıktan sonra, o özellik ile ilgili duygu polaritesini tespit etmek için, bir ayrıştırıcı kullanmışlar ve bir bağıllık graf yapısı kullanarak duygu polarite tespitini yapmışlardır.

Moghaddam ve Ester özellik çıkarımı için sık geçen isimler yöntemini uyguladıkları, veri seti olarak ürün yorumlarını kullandıkları özellik tabanlı düzeyde duygu analizi çalışması yapmışlardır [42]. Moghaddam ve Ester sık geçen isimler yöntemi ile özellik çıkarımından sonra duygu polarite tespiti linguistik analize başvurmuşlardır. Linguistik analizde cümle içinde ürün özelliğini ile ilgili kelimeye en yakın olan sıfat türündeki kelimeyi duygu terimi olarak ele almışlar ve WordNet'in sunduğu bazı fonksiyonları kullanarak duygu polarite tespitini gerçekleştirmişlerdir. Yapılan testlerde %70 ile %90 aralığında değişen başarı düzeyleri elde edilmiştir.

Jiang, Yu, Zhou, Liu ve Zhao yaptıkları özellik tabanlı duygu analizi çalışmasında veri seti olarak Twitter mesajlarını kullanmışlardır [43]. Yapılan çalışmada, belli bir konu hakkında, tweetlerdeki duygu polaritelerini bulmak amaçlanmıştır. Duygu polarite tespiti için de doğal dil işleme yöntem ve araçlarından, Pointwise Mutual Information (PMI) skorlamasından faydalanmışlardır. Sözcük türleri, kelime kökleri, cümlelerin linguistik olarak analiz edilmesi ve kural tabanlı bir takım yöntemler kullanmışlardır. İlave olarak yöntemin başarı oranını yükseltmek için kullanıcının daha önce retweet ettiği, cevap yazdığı tweetlerden bir graf oluşturarak, o tweetlerdeki duygu polaritelerini de göz önüne almışlardır. Jiang ve diğerleri testlerde ve deneysel sonuçlarda yöntemlerinin %67 oranında bir başarı seviyesi elde edildiğini, kıyasladıkları diğer yöntemlerin ise başarı seviyelerinin ise %59 - %62 arasında ölçüldüğünü belirtmişlerdir.

Qiu, Liu, Bu ve Chen veri seti olarak ürün yorumlarını kullandıkları ve sözcük tabanlı bir yöntem uyguladıkları özellik tabanlı düzeyde duygu analizi çalışması gerçekleştirmişlerdir [44]. "Double Propagation" olarak isimlendirdikleri yöntem geliştirmişlerdir, Bu yöntemde önce az sayıda pozitif/negatif tohum (seed) duygu terimleri kümesi oluşturulmaktadır. Daha sonra cümlelerin bağımlılık ayrıştırıcı ile linguistik analizi yapılmakta, daha sonra en başta belirlenmiş olan az sayıdaki tohum duygu ifadelerinden ürün özelliği olan hedefler keşfedilmektedir. Daha sonra da keşfedilmiş hedef ve mevcut duygu ifadelerinden yeni

duygu ifadesi terimlerinin keşfedilmesi şeklinde döngüsel olarak ilerlenmektedir. Testlerde ve deneysel sonuçlarda önerdikleri yöntemin %72 - %88 aralığında başarı düzeyi ortaya koyduğunu gözlemlemişlerdir.

Eirinaki, Pisal ve Singh yaptıkları ürün yorumları veri seti üzerinde yaptıkları duygu analizi çalışmasında duygu derecelendirme yapmışlardır [45]. Çalışmalarında, özellik çıkarımında sık geçen isimler yöntemini geliştirerek kullanmışlardır. Direkt olarak en sık geçen isim türü kelimeleri almak yerine sadece duygu terimleri ile bağlantılı olan isim türü kelimelerden sık geçenleri almışlardır. Bu yöntemle sık geçen isimlerin arasında gelen ve ürün özelliği olmayan isim türü sözcükleri azaltmayı amaçlamışlardır. -4 ile +4 aralığında duygu polarite derecelendirmesi için sözcük tabanlı bir yöntem kullanmışlardır. Eirinaki ve diğerleri, önerdikleri yöntemin testlerde başarı düzeyinin %87 olduğunu ortaya koymuşlardır.

Kontopoulos, Berberidis, Dergiades ve Bassiliades özellik tabanlı düzeyde duygu analizi çalışması yapmışlar ve veri seti olarak Twitter mesajlarını kullanmışlardır [46]. Yaptıkları çalışma duygu polaritesinin tespit edilmesinden ziyade mesajlardan varlık ve varlıkların özelliklerini çıkarma üzerinedir. Çalışmalarında duygu polaritesini tespit etmek için bir yöntem geliştirmemişler, OpenDover web servisinden faydalanmışlardır. Kontopoulos ve diğerleri tweetlerden varlıklar ve bu varlıkların özelliklerini çıkarmak için ontoloji tabanlı Formal Concept Analysis yöntemini kullanmışlardır.

Xianghua, Guo, Yanyan ve Zhiqiang Çince dilindeki bloglar veri seti üzerinde duygu analizi çalışması gerçekleştirmişlerdir [47]. Özellik çıkarımı için LDA algoritmasını uygulamışlardır. Duygu polaritesini tespit etmek için "Çince WordNet" olan HowNet aracından faydalanmışlardır. Cümlelerdeki duygu terimlerinin ile HowNet'teki duygu terimleri arasındaki anlamsal yakınlığa dayalı bir metotla duygu polarite tespitini yapmışlardır. Test ve deneysel sonuçlarda özellik çıkarımında %90, polarite tespitinde %91'e varan başarı seviyeleri ölçmüşlerdir.

Bagheria, Saraee ve Jong yaptıkları duygu analizi çalışmasında özellik çıkarımı için özgün bir yöntem geliştirmişlerdir [48]. Sözcük tabanlı bir yöntem önermişlerdir. Yöntemlerinde önce cümlelerde sözcük türü tespiti yapmışlar, sonra "isim", "sıfat ve onu takip eden isim" gibi sözcük türü örüntüleri tanımlamışlar ve bu örüntülere uyan kelime ve kelime

gruplarını "aday özellikler" olarak ortaya çıkarmışlardır. Bu aday özelliklerin arasında ürün özelliği olmayanları elemek için önce Nakagawa ve Mori'nin [49] yaptıkları çalışmada önerilen FLR yöntemini uygulamışlardır.

Kang ve Park yaptıkları özellik tabanlı düzeyde duygu analizi çalışmasında veri seti olarak mobil hizmetler ile ilgili yorumları kullanmışlardır [50]. Mobil hizmetlerin farklı yönleri hakkındaki kullanıcıların fikirlerinin tespit edilmeye çalışıldığı çalışmada özellik çıkarımı otomatik olarak gerçekleştirilmemiş, manuel olarak belirlenen özelliklere göre duygu polarite sınıflandırması yapılmıştır.

Isidro, Rafael ve Francisco sinema yorumları veri seti ile yaptıkları duygu analizi çalışması yapmışlardır [51]. Çalışmalarında özellik çıkarımı için IMDb kaynaklı verilerden oluşturulmuş olan "<http://www.movieontology.org>" sitesindeki sinema ontolojisi veri tabanından faydalanarak ontoloji tabanlı özellik çıkarımı yapmışlardır. Duygu polarite tespitinde denetimsiz makine öğrenmesi yöntemini uygulamışlar ve SentiWordNet sözlüğünü kullanmışlardır. Testlerde özellik çıkarımında %70 ile %77 aralığında, duygu polarite tespitinde ise %56 ile %89 aralığında değişen başarı seviyeleri elde edilmiştir.

Quan ve Ren ürün yorumları veri seti ile yaptıkları duygu analizi çalışmasında özellik çıkarımı için öncelikle yorumdaki tüm isim kelimeleri seçmişler ve bu kelimeleri aday özellik olarak belirlemişlerdir. İkinci adım olarak sonra bu aday özellik olarak belirlenmiş kelimelerin ürünle olan anlamsal yakınlığını belirlemek için TF-IDF ile PMI yöntemlerinden faydalanmışlardır [52]. Duygu polarite tespiti için ise bağıklık ayrıştırıcısı ile linguistik analiz gerçekleştirmişlerdir. Önerdikleri yöntemi akıllı telefonlar, mp3 çalarlar ve dijital kameralar hakkındaki derledikleri yorumlardan oluşan veri seti ile test etmişlerdir. Yaptıkları başarı ölçümlerinde özellik çıkarımında %60 ile %88 aralığında, duygu polarite tespitinde ise %65 ile %76 aralığında başarı oranları elde etmişlerdir.

Kansal ve Toshniwal ürün yorumlarından oluşan veri seti ile özellik tabanlı düzeyde duygu analizi çalışması yapmışlardır [53]. Sık geçen isimler yönteminden faydalanarak özellik çıkarımı görevini yerine getirmişlerdir. Sık geçen isim türü kelimelerden ürün özelliği olmayanları belirlemek için olasılık biliminden faydalanmışlardır. Ürün özelliklerine göre duygu polarite tespitinde diğer çoğu çalışmada göz ardı edilen durumları ele almışlardır. Örneğin "Telefonun açılma süresi çok uzun" şeklinde bir yorumda "uzun" kelimesinin

polaritesi negatif, "Bu telefonun pil ömrü uzun" şeklindeki bir yorumda ise polaritesi pozitifdir. Ama normalde uzun kelimesi birçok duygu terimleri sözlüğünden nötr polariteye sahiptir ama örnek cümlelerde görüldüğü gibi bağlama göre polaritesi değişebilmektedir. Kansal ve Toshniwal cümleleri ve özellikle bağlaçları detaylı şekilde linguistik analiz ederek "bağlama bağlı duygu terimlerini" belirlemeye yönelik çalışmışlardır.

Bhadane, Dalal ve Doshi veri seti olarak ürün yorumlarını kullanarak DVM sınıflandırıcısı ile denetimli makine öğrenmesine dayalı özellik tabanlı düzeydeki duygu analizi çalışması yapmışlardır [54]. Önerdikleri yöntemde ilk aşamada özellik çıkarımı görevi, daha sonra o ürün özelliği ile ilgili duygu polaritesinin belirlenmesi için iki aşamadan oluşan denetimli makine öğrenmesine dayalı sınıflandırma yapmışlardır. Yapılan başarı ölçümlerinde %78'lik gibi bir başarı seviyesi ölçmüşlerdir.

Maharani, Widyanoro ve Khodra yaptıkları özellik tabanlı düzeydeki duygu analizi çalışmasında özellik çıkarımı için özgün bir yöntem önermişlerdir [55]. Önerdikleri özellik çıkarım yöntemi doğal dil işleme yöntem ve araçları ile linguistik analize dayalı bir yöntemdir. Maharani ve diğerleri sözcük türü örüntülerine dayalı olarak kural tabanlı bir yaklaşım uygulamışlardır. Yaptıkları başarı ölçümlerinde özellik çıkarımında %40 ile %79 aralığında değişen başarı oranları yakalamışlardır.

Li, Zhou ve Li Çince dilindeki kullanıcı yorumları veri setinde özellik çıkarımı ile ilgili bir çalışma gerçekleştirmişlerdir [56]. Çalışmada sık geçen isimler yöntemi üzerinde geliştirmeler yaparak kullanmışlardır. Li ve diğerleri ürün özelliği olmayan isimleri eleme görevi için cümlelerdeki kelimelerin sıralanışını analiz etmişler, PMI skorlarını göz önüne almışlar ve kelimeler arası anlamsal yakınlıkları ölçerek sık geçen isimler yöntemi üzerinde iyileştirmeler yapmışlardır. Yaptıkları testler ve başarı ölçümlerinde özellik çıkarımı görevinde %67 ile %82 aralığında değişen başarı düzeyleri ölçmüşlerdir.

Liao, Fengn, Yang ve Huang gerçekleştirdikleri duygu analizi çalışmasında Çince sosyal medya mesajları üzerinde çalışmışlardır [57]. Çalışmada graf veri yapısını kullandıkları özgün bir yöntem geliştirmişlerdir. Oluşturdıkları grafda düğümlerde konuyla (ürün özelliği) ilgili kelimeler tutulmakta, kenarlarda ise iki düğüm (kelime) arasında anlamsal yakınlığı temsil eden bir ağırlık değeri tutulmaktadır. Bu ağırlık değerlerini TF-IDF terim ağırlıklandırma metoduna benzeyen, kendi önerileri olan bir metot ile hesaplamışlardır.

Daha sonra son zamanlarda kullanımı yaygın hale gelmiş olan kelime yerleştirme (word embeddings) yöntemi ile bu ağırlıklandırılmış grafi birleştirerek özellik çıkarımını gerçekleştirmişlerdir. Yaptıkları test ve başarı ölçümlerinde elde ettikleri başarı oranları %72 ile %84 aralığında ölçülmüştür.

Liu, Gao, Liu ve Zhang gerçekleştirdikleri duygu analizi çalışmasında duygu polaritesinden ziyade özellik çıkarımı görevi üzerine odaklanmışlardır [58]. Kural tabanlı bir yöntem uygulamışlar fakat uyguladıkları yöntem diğer kural tabanlı yöntemlerden oldukça farklıdır. Kural tabanlı yöntemlerde önce kurallar tanımlanır ve metin içinde bu kurallara uyan cümle yapıları tespit edilir. Liu ve diğerleri kendileri kurallar tanımlayıp uygulamak yerine olan kural tabanlı kendilerinden önce yapılmış kural tabanlı çalışmaların kurallarını aynen almışlardır. Bu kuralları doğrudan kullanmamışlar, onun yerine kurallar içinde en yüksek başarı seviyesini sunacak olan en ideal kurallar kümesini seçme üzerine çalışmışlardır. En ideal kurallar kümesini seçmenin önce NP-Hard bir problem olduğunun ispatını yapmışlar, daha sonra da sezgisel bir takım arama yöntemlerine başvurmuşlardır. Açgözlü (greedy search) arama ve benzetimli tavlama (simulated annealing) sezgisel arama algoritmalarını uygulayarak en ideal kurallar kümesini belirlemeye çalışmışlardır. Önerdikleri yöntemle belirledikleri kurallar kümesi ile yaptıkları başarı ölçümlerini, kuralların hepsini uygulayarak yaptıkları başarı ölçümleri ile karşılaştırmışlardır. Test ve ölçümler sonucunda önerdikleri sezgisel kurallar kümesi belirleme yönteminin başarı seviyesini yükselttiğini ortaya koymuşlardır.

Akhtar, Gupta, Ekbal ve Bhattacharyya restoran ve dizüstü bilgisayar yorumları veri seti ile yaptıkları duygu analizi çalışmasında hem özellik çıkarımı hem de duygu polaritesinin tespiti için denetimli makine öğrenmesi yöntemini uygulamışlardır [59]. Özellik vektörü için özellik seçiminde oldukça farklı bir yol izlemişlerdir. DVM, CRF ve Maksimum Entropi (ME) sınıflandırma algoritmalarını kullandıkları yöntemlerinde toplamda 19 farklı özellik seçimi yöntemi belirlemişler, daha sonra parçacık sürü optimizasyonu ile bu özellik seçimi yöntemlerinden her sınıflandırma algoritması için en optimum kombinasyonu belirlemeye çalışmışlardır. Yaptıkları testlerde, parçacık sürü optimizasyonu ile belirledikleri optimum özellik seçimi yöntemler kombinasyonu ile elde ettikleri sonuçların, tek tek özellik seçimi yöntemleri ile elde ettikleri sonuçlardan belirgin şekilde daha başarılı olduğunu gözlemlemişler ve önerdikleri parçacık sürü optimizasyonuna dayalı özellik seçimlerini optimum şekilde kombine etme yönteminin başarılı olduğunu göstermişlerdir.

Son yıllarda özellik tabanlı duygu analizi çalışmalarında özellik çıkarımı LDA algoritmasının kullanımı yaygınlaşmış ve bu konuda çok sayıda çalışma yapılmıştır.

Xiong, Wang, Ji ve Wang yaptıkları çalışmada WSTM (Word-pair Sentiment-Topic Model) yöntemini önermişlerdir [60]. WSTM’de duygu terimleri ve özellik terimlerinin eşzamanlı üretimsel sürecini modellemişlerdir. Yöntemlerinde kayan pencere kullanmışlar ve kayan pencerenin her adımında kelime çiftleri oluşturmuşlar ve daha sonra bu kelime çiftlerini konu modellemede kullanmışlardır. Xiong W. ve arkadaşları veri seti olarak dizüstü bilgisayar ve akıllı telefon yorumlarını kullanmışlardır.

García-Pablosa, Cuadros ve Rigau yaptıkları çalışmada özellik çıkarımı ve duygu polarite tespiti için W2VLDA yöntemini önermişlerdir [61]. W2VLDA’da LDA algoritmasını Word Embeddings, word2vec ve maksimum entropi sınıflandırıcı yöntemi ile kombine bir şekilde kullanmışlardır. Yöntemlerini restoran yorumları ve elektronik cihaz (dizüstü bilgisayar, profesyonel fotoğraf makinesi) yorumları üzerinde test etmişlerdir.

Türkçe dilinde özellik tabanlı düzeyde yapılmış ilk duygu analizi çalışması Akbaş’ın Twitter mesajları üzerinde yaptığı tez çalışmasıdır [62]. Akbaş çalışmasında önce Twitter mesajlarının hangi konu ile ilgili olduğunu belirlemiş, daha sonra duygu polaritesini tespit etmiştir. Çalışmada özellik çıkarımı için LDA algoritması uygulanmıştır. Duygu polarite tespiti için ise sözcük tabanlı bir yöntem ile makine öğrenmesi yöntemleri birleştirilerek uygulamıştır.

Ayrıca Dehkharghani, Saygın, Yanıkoğlu ve Oflazer Türkçe diliyle yaptıkları duygu analizi çalışmasında özellik tabanlı düzeyde de duygu analizi çalışması yapmışlardır [63]. Fakat çalışmalarında otomatik özellik çıkarımı yapmamışlar, onun yerine ürün özellikler ile ilgili manuel olarak bir kelime listesi oluşturmuşlar ve bu kelimeleri kullanıcı yorumlarında aramışlardır.

Özellik tabanlı düzeyde duygu analizi ile ilgili yapılmış olan çalışmalar özet halinde Çizelge 1.2.’de sunulmuştur.

Çizelge 1.2. Özellik tabanlı düzeyde yapılmış duygu analizi çalışmaları

Ref. no	Yıl	Dil	Çalışma türü	Veri kümesi	Uygulanan yöntem/algorithm	Sonuçlar
[38]	2009	İngilizce	Yöntem önerisi	Ürün yorumları	"Phrase Dependency Tree" olarak adlandırdıkları bir veri yapısı kullanmışlardır	%53 - %68 aralığında değişen başarı oranları elde edilmiştir
[39]	2010	İngilizce	Yöntem önerisi	Ürün yorumları	Ontoloji tabanlı "Sentiment Ontoloji Tree" ismini verdikleri veri yapısını kullanmışlardır	Yapılan testlerde %84 oranında başarıya ulaşılmıştır
[41]	2010	İngilizce	Yöntem önerisi	Restoran yorumları	Özellik çıkarımı: LDA Duygu tespiti: Linguistik analiz.	%80'lik recall ve %50 - %85 aralığında precision ölçülmüştür.
[42]	2010	İngilizce	Yöntem önerisi	Ürün yorumları	Özellik çıkarımı: "Sık Geçen İsimler" yöntemi Duygu polaritesi tespiti: WordNet'ten faydalanmışlardır	%70 ile %90 aralığında değişen başarı düzeyleri elde edilmiştir
[44]	2011	İngilizce	Yöntem önerisi	Ürün yorumları	Dependency Tree ile cümlelerin linguistik analizine dayalı "Double Propagation" ismini verdikleri yöntem	%73 - %89 aralığında değişen başarı düzeyleri elde edilmiştir.
[43]	2011	İngilizce	Yöntem önerisi	Twitter mesajları	Özellik çıkarımı: Co-reference çözümü ve PMI Polarite tespiti: Linguistik analize dayalı kural tabanlı.	%68'lik bir başarı elde etmişlerdir.
[45]	2012	İngilizce	Duygu derecelendirme yöntem önerisi	Ürün yorumları	Özellik Çıkarımı: Linguistik analiz ve sık geçen isimler üzerinde iyileştirme. Duygu polarite tespiti: Linguistik analiz ile (-4, +4) aralığında duygu derecelendirme.	Farklı veri setleri üzerinde yaptıkları çok sayıda testlerde %40 - %80 aralığında değişen başarı düzeyleri elde etmişlerdir.
[46]	2013	İngilizce	Özellik çıkarımı yöntem önerisi	Twitter mesajları	Özellik çıkarımı: Ontoloji tabanlı Formal Concept Analysis yöntemi	Çok farklı parametrelere göre sonuçlar yayınlanmıştır. %53 - %72 aralığında değişen başarı oranları elde edilmiştir.
[47]	2013	Çince	Yöntem önerisi	Sosyal mesajlar	Özellik çıkarımı: LDA Duygu polarite tespiti: HowNet aracından faydalanmışlardır.	%92'ye varan başarı düzeyleri elde edilmiştir.
[48]	2013	İngilizce	Özellik çıkarımı yöntem önerisi	Ürün yorumları	Özellik çıkarımı: Linguistik analiz ve PMI skorları ile kural tabanlı yöntem.	Farklı veri setlerinde yapılan testlerde %79 - %90 aralığında değişen başarı seviyeleri elde etmişlerdir.
[50]	2014	İngilizce	Duygu derecelendirme yöntem önerisi	Ürün yorumları	WordNet kullanılarak linguistik analiz yöntemi ile -2, +2 aralığında duygu derecelendirme	Çok farklı parametrelere göre çok farklı sonuçlar yayınlanmıştır
[51]	2011	İngilizce	Yöntem önerisi	Sinema yorumları	Özellik çıkarımı: Sinema ontolojisi ile ilgili "http://www.movieontology.org" sitesindeki Ontolojilerden faydalanarak yapılmıştır. Duygu polarite tespiti: SentiWordNet kullanılmıştır.	Özellik çıkarımında %71 ile %78 aralığında değişen, duygu polaritesinde ise %56 ile %89 aralığında değişen başarı etmişlerdir
[52]	2014	İngilizce	Yöntem önerisi	Ürün yorumları	Özellik çıkarımı: PMI ve TF-IDF Duygu polarite tespiti: Bağlılık ağacı kullanmışlardır	Farklı veri setlerinde %61 - %89 arasında başarı düzeylerine ulaşılmıştır.

Çizelge 1.2. (devam) Özellik tabanlı düzeyde yapılmış duygu analizi çalışmaları

[53]	2014	İngilizce	Yöntem önerisi	Ürün yorumları	Özellik çıkarımı: Sık geçen isimler Duygu polarite tespiti: Linguistik analiz, duygu terimleri sözlüğü	Farklı veri setleri kullanılarak yapılan testlerde %78 - %85 aralığında değişen başarı seviyeleri elde etmişlerdir.
[54]	2015	İngilizce	Yöntem önerisi	Ürün yorumları	DVM sınıflandırıcısı ile denetimli makine öğrenmesine dayalı yöntem	%78'e varan başarı seviyeleri elde etmişlerdir.
[55]	2015	İngilizce	Özellik çıkarımı yöntem önerisi	Ürün yorumları	Sözcük türü örüntülerine göre kural tabanlı yöntem	Farklı veri setleri üzerinde farklı parametrelerle yapılan testlerde %39 - %80 aralığında değişen başarı oranları elde edilmiştir.
[56]	2015	Çince	Özellik çıkarımı yöntem önerisi	Ürün yorumları	Sık geçen isimler, PMI skorlaması ve CRF yaklaşımının bir arada olduğu bileşik yöntem	Farklı veri setlerinde yapılan başarı ölçümlerinde %73 - %79 aralığında değişen başarı seviyelerine ulaşılmıştır.
[57]	2016	Çince	Yöntem önerisi	Sosyal medya mesajları	Kendi tasarladıkları ağırlıklandırılmış bir graf yapısı ile Word Embeddings yöntemini kombine etmişlerdir.	%72 ile %84 aralığında değişen başarı düzeyleri elde edilmiştir.
[58]	2016	İngilizce	Yöntem önerisi	Ürün yorumları	Kural tabanlı bir yöntem uygulamışlardır. Ideal kurallar kümesini belirlemek için sezgisel arama yöntemlerine başvurmışlardır	Aynı kurallarla normal kural tabanlı yöntemlere göre daha yüksek başarı seviyeleri elde edilmiştir.
[59]	2017	İngilizce	Yöntem önerisi	Restoran ve ürün yorumları	Özellik vektörü için 19 farklı özellik seçim yönteminden en iyi kombinasyonu belirlemek için parçacık sürü optimizasyonunu kullanmışlardır	19 farklı özellik seçim yönteminin doğrudan uygulanmasından daha başarılı sonuçlar elde edilmiştir.
[60]	2018	İngilizce	Yöntem önerisi	Bilgisayar ve akıllı telefon yorumları	Özellik çıkarımı: LDA	Özellik çıkarımında %69 ile %75 aralığında değişen başarı düzeyleri elde edilmiştir.
[61]	2019	İngilizce	Yöntem önerisi	Restoran ve elektronik cihaz yorumları	Özellik çıkarımı: LDA yöntemini Word Embeddings, word2vec yöntemi ile kombine şekilde uygulamışlardır.	Farklı veri setlerinde ve farklı parametrelerle %71 ile %81 aralığında başarı oranları elde etmişlerdir.
[62]	2012	Türkçe	Duygu derecelendirme yöntem önerisi	Twitter mesajları	Özellik çıkarımı: LDA Duygu polarite tespiti: Sözcük tabanlı ve denetimli makine öğrenmesi yöntemi ile hibrit yaklaşım	Farklı parametrelerle yaptıkları testlerde %45 - %88 aralığında başarı seviyeleri elde etmişlerdir.
[63]	2016	Türkçe	Yöntem önerisi	Ürün yorumları	Özellik çıkarımı: Manuel ürün özellik kelime listesi oluşturma ve bu kelimeleri yorumlarda arama	Farklı veri setlerinde ve farklı konfigürasyonlarla %60 ile %73 aralığında başarı oranları elde etmişlerdir.

2. FİKİR MADENCİLİĞİ VE DUYGU ANALİZİ PROBLEMİ

Günümüzde sosyal ağlarda ve web platformlarında insanlar satın aldıkları ürün ve hizmetler, şahit oldukları olaylar, katıldıkları organizasyonlar gibi birçok konu hakkında görüşlerini, fikirlerini ve duygularını ifade edebilmektedirler. Bu platformlarda insanların görüşlerini içeren çok büyük miktarlarda metinler birikmiş durumdadır. Duygu analizi ve fikir madenciliği bu metinler içinde saklı olan fikir, duygu ve düşünceleri ortaya çıkaran çalışmalara denilmektedir. Duygu analizinde, üzerinde fikir ifade edilen ürün, hizmet, olay gibi şeylere *varlık* denilmektedir. Duygu ifadesi bazen varlığın kendisi hakkında bazen de varlığın bir yönü veya bir özelliği hakkında olabilir. Örnek olarak bir kullanıcı diz üstü bilgisayar hakkında fikir ifade ettiğinde, genel anlamda diz üstü bilgisayarı beğendiğini ifade edebilir. Bu doğrudan varlığın kendisi hakkında bir duygu ifadesidir. Bir başka kullanıcı ise diz üstü bilgisayarın ekranını beğendiğini ama bataryasını beğenmediğini söyleyebilir. Bu durumda duygu ifadesinin hedefleri diz üstü bilgisayarın ekran ve batarya özellikleridir.

Varlık ve özellik dışında, yapılan duygu analizi çalışmasının türüne göre görüşün ifade eden kişi yani duygu ifadesinin *öznesi* de önemli olabilmektedir. Sonuç itibarıyla Liu duygu ifadesini aşağıdaki gibi dört bileşenden oluşan bir model şeklinde tanımlamıştır [1]:

(*v*, *ö*, *d*, *n*)

v: varlık, *ö*: özellik, *d*: duygu, *n*: özne

Duygu analizi çalışmalarındaki görevler, duygu ifadesini oluşturan bu bileşenlerin hepsinin ya da bir kısmının, çalışmanın düzeyine ve kapsamına göre metinlerden çıkartılmalarıdır. Bu bileşenlerden birincisi varlıktır. Metinlerden varlık çıkarılma görevi bütün fikir madenciliği çalışmalarında yapılmamaktadır. Örnek olarak elektronik ticaret sitelerinde kullanıcıların yazdıkları ürün yorumlarında belli bir ürün (varlık) ile ilgili görüş ifade edilmektedir, bu yüzden ürün yorumlarında genelde metinden varlık çıkarımı yapılmasına gerek olmamaktadır. Diğer taraftan mesela siyaset ile ilgili bir gazete yazısında, yazar farklı siyasi oluşumlar ve kişiler hakkında görüş bildirebilir. Böyle bir durumda metinden varlık çıkarımı önemli olur. Varlık çıkarım görevinde kural tabanlı varlık çıkarımındaki (Named Entity Recognition) yöntemlerden faydalanılır. Bu konuyla ilgili Dalkılıç, Gelişli

ve Diri'nin [64], Akın ve Eryiğit'in [65], Küçük ve Yazıcı'nın [66] Türkçe diliyle yapmış oldukları çalışmalar bulunmaktadır.

2.1. Duygu Analizi Çalışma Düzeyleri

Giriş bölümünde duygu analizi çalışma düzeyleri hakkında kısaca bilgi verilmişti. Bu bölümde bu konu daha detaylı bir şekilde ele alınacaktır. Duygu analizi ve fikir madenciliği alanındaki çalışmalar, metinden çıkardıkları bilgilere ve kapsamlarına göre üç ayrı düzey altında yürütülmektedir: 1. Doküman düzeyinde duygu analizi. 2. Cümle düzeyinde duygu analizi. 3. Özellik tabanlı düzeyde duygu analizi.

2.1.1. Doküman düzeyinde duygu analizi

Doküman düzeyi duygu sınıflandırmasında görüş ifade edilen bir metnin detaylarına inilmeden bir bütün olarak negatif ya da pozitif olarak sınıflandırılır. Doküman düzeyinde sınıflandırma çalışmalarında iki varsayım yapılmış ve çalışmalar bu bağlamda yürütülmüştür [1]. Birincisi yorum veya inceleme yazısının tek bir varlık hakkında olduğu varsayılmıştır. İkincisi ise yorum veya inceleme tek bir öznenin görüşlerini yansıtmaktadır. Eğer ki bir yazıda birden fazla varlık hakkında görüş ve düşünceler ifade ediliyorsa, o yazıyı bir bütün olarak pozitif veya negatif olarak sınıflandırmanın hiçbir anlamı olmaz. Aynı şekilde bir yazıda bir tek kişi yerine birden fazla kişinin görüşlerine karışık olarak verilmişse yine bu yazıyı pozitif veya negatif olarak sınıflandırmak manasızdır.

Görüş ifade edilen bir metinde cümlelerin bazılarının polaritesi pozitif, bazılarının ise negatif olabilir. Ayrıca ürünün bir takım özellikleri hakkında pozitif görüşler, bir takım başka özellikleri hakkında ise negatif görüşler yorum içinde karışık olarak yer alabilir. Doküman düzeyindeki duygu analiz çalışmalarında bu düzeydeki detaylar analiz edilmemekte, metin daha ağır basan görüşe göre bütün olarak negatif ya da pozitif olarak sınıflandırılmaktadır.

2.1.2. Cümle düzeyinde duygu analizi

Doküman düzeyinde yapılan duygu analizi çalışması bazı uygulamalar için uygun olmayabilir çünkü doküman düzeyinde yapılan duygu analiz çalışmasında cümlelerin duygu polariteleri tek tek belirlenmeksizin ağır basan duygu polaritesine göre doküman bütün olarak sınıflandırılmaktadır. Bir inceleme yazısı veya ürün yorumunda bazı cümlelerin polaritesi pozitif, bazı cümlelerin polaritesi negatif olabilir. Bazı cümleler ise duygu polaritesi bakımından nötr olabilir. Cümle düzeyinde yapılan duygu analizi çalışmasında metin bütün olarak sınıflandırılmaz. Onun yerine önce her bir cümlenin duygu ifadesi barındırıp barındırmadığı belirlenir, eğer duygu ifadesi varsa daha sonra cümlenin negatif veya pozitif olarak polaritesi belirlenir. Bir cümlenin duygu ifadesi barındırıp barındırmadığına tespit etmeye yönelik yapılan çalışmalara sübjektiflik sınıflandırması denmektedir.

2.1.3. Özellik tabanlı düzeyde duygu analizi

Doküman ve cümle düzeyindeki duygu analizinde duygu ifadelerinin hedefleri belirsiz kalmaktadır. Doküman düzeyinde yapılan duygu analizi çalışmasında bir ürün ile ilgili yapılan yorum pozitif olarak sınıflandırılması, kullanıcının o ürünün tüm yönleri ile ilgili pozitif görüşe sahip olduğu anlamını taşımaz. Örneğin, yorumcu varlığın dört tane özelliği hakkında pozitif, üç tane özelliği hakkında negatif yorum yapmış olduğu varsayılırsa, doküman düzeyinde sınıflandırmada o yorumun polaritesi pozitif olarak belirlenir. Bu durumda bilgi kaybı olmaktadır. Yorumcunun varlığın bazı özellikleri hakkında yapmış olduğu negatif yorumlar ortaya çıkarılamamıştır.

Özellik tabanlı düzeydeki duygu analizi çalışmasında yorumdaki her bir duygu ifadelesinin ürünün hangi özelliği ile ilgili olduğu tespit edilmektedir. Bu şekilde doküman düzeyindeki gibi kaba bir sınıflandırma yerine yorumun detaylarına inilmekte ve daha detaylı bilgi çıkarımı yapılmaktadır. Duygu ifadesinin ürünün hangi özelliği ile ilgili olduğunun ortaya çıkarılmasına yönelik yapılan çalışmalara özellik çıkarımı denmektedir.

Özellik tabanlı duygu analizinde duygu polarite sınıflandırmasında kullanılan yöntemler büyük oranda doküman düzeyi duygu analizinde kullanılan duygu polaritesi tespit etme

yöntemleriyle benzerlik taşımaktadır. Burada önem taşıyan görev duygu ifadesinin kapsam ve hedefinin belirlenmesi ve özellik çıkarımı işleminin doğru şekilde gerçekleştirilmesidir. Çoğu zaman özellik çıkarımı duygu polarite tespitinden daha zorlu bir işlemdir.



3. DUYGU ANALİZİ VE FİKİR MADENCİLİĞİ ALGORİTMALARI

Fikir madenciliği ve duygu analizi çalışmalarında duygu polarite sınıflandırması gerçekleştirilmesi gereken en temel görevdir. Duygu polarite sınıflandırmasına ek olarak özellik tabanlı düzeyde yapılan duygu analizi çalışmalarında özellik çıkarımı da gerçekleştirilmesi gereken bir görevdir. Literatürde yapılan araştırmalarda bu görevler ile ilgili önerilmiş çok sayıda farklı yöntemler olsa da bazı yöntemler ve bu yöntemlerin farklı varyasyonları yaygın olarak kullanılmaktadır. Bu bölümde kullanımı yaygın olan bu yöntemler tanıtılacaktır.

3.1. Duygu Polaritesinin Tespitinde Kullanılan Yöntemler

Duygu polarite sınıflandırması, bütün düzeylerde yapılan duygu analizi çalışmalarında yerine getirilmesi gereken en temel fikir madenciliği görevidir. Duygu polarite sınıflandırmasında kullanılmakta olan yöntemler makine öğrenmesine dayalı yöntemler ve sözcük tabanlı yöntemler olmak üzere iki ana grup altında yer almaktadır.

3.1.1. Makine öğrenmesine dayalı yöntemler

Denetimli makine öğrenmesi ile duygu polarite tespiti, etiketlenmiş eğitim verisi ile sistemin eğitilmesi ve daha sonra eğitilmiş bu sistem ile duygu polaritesinin belirlenmesi ilkesine dayalı olarak yapılır. Burada sınıflandırma yapılırken herhangi bir sınıflandırma algoritması kullanılabilir. Genellikle NB, DVM ve ME gibi sınıflandırıcılar kullanılır. Bu gruptaki yöntemlerin kullanılabilmesi için önceden etiketlenmiş eğitim veri setine gereksinim duyulmaktadır.

Makine öğrenmesine dayalı bir yöntem ile duygu analizi çalışmasının ilk örneği Pang, Lee ve Vaithyanathan tarafından yapılmış olan ve sinema yorumlarını sınıflandıran çalışmadır [8]. Pang ve diğerleri yaptıkları bu çalışmada, özellik vektörü özelliği olarak unigram BoW modelini, sınıflandırma algoritması olarak DVM ve NB'yi kullanmışlardır.

Makine öğrenmesine dayalı yöntemlerin başarılı olabilmeleri için özellik vektöründe etkin özelliklerin kullanılması gerekmektedir. Makine öğrenmesiyle duygu analiz çalışmalarında özellik vektöründe en sık kullanılan özellikler aşağıda anlatılmıştır.

Kelime Torbası (BoW) modeli: Bu özellik, sözcüklerin metinlerde geçip geçmemesine dayanan bir özelliktir. BoW modelinin metin sınıflandırma çalışmalarında kullanımı çok yaygındır. Sözcüklerin metinlerde tek başlarına geçip geçmelerine ilave olarak arka arkaya ikili ya da n’li gruplar şeklinde geçip geçmemelerinin kullanımı da vardır. Kelimelerin metinlerde tek başlarına geçip geçmemelerine dayalı olarak uygulanması unigram modeli, metinlerde arka arkaya ikili gruplar şeklinde geçip geçmemelerine dayalı olarak uygulanması bigram modeli, metinlerde arka arkaya n’li gruplar şeklinde geçip geçmemelerine dayalı olarak uygulanması da n-gram modelidir.

BoW modelinde, sözcüklerin metindeki frekanslarından bağımsız olarak sadece geçip geçmemesine dayalı olarak uygulanması ikili özellik seçimi olarak adlandırılmaktadır. Metinde geçip geçmemesinin yanında frekanslarına dayalı bir şekilde uygulanması da terim frekans ağırlığı özellik seçimi olarak adlandırılmaktadır.

Terim Frekansı - Ters Doküman Frekansı (TF-IDF): TF-IDF, bir kelimenin bir dokümandaki frekansının, diğer tüm dokümanlardaki frekansları toplamına bölünmesiyle hesaplanan bir terim ağırlıklandırma yöntemidir. Bir kelimenin bir dokümanda frekansı yüksek, diğer dokümanlarda frekansı düşükse, yüksek ayırt edici özelliğe sahip bir kelime demektir. Böyle bir sözcük için yüksek bir TF-IDF değeri hesaplanır. Fakat sözcüğün bir dokümanda frekansı yüksek ama diğer dokümanlarda da frekansı yüksekse ayırt edici özelliği daha düşük demektir ve daha düşük bir TF-IDF değeri hesaplanır.

Sözcük türü (POS Tag): Doğal dil işleme ve metin işleme çalışmalarında kullanılan en önemli özelliklerden birisi kelimelerin sözcük türleridir (isim, sıfat, fiil, vs.). Duygu analizi ve fikir madenciliği ile ilgili çalışmalarda da sözcük türü sıkça kullanılmaktadır. Makine öğrenmesinde özellik vektörünü oluşturmada metinlerdeki kelimelerin sözcük türlerinin kullanımı yaygındır.

Nokta Tabanlı Müşterek Bilgi (PMI): PMI iki sözcük arasında ilişkiyi skorlamak için 1989 yılında Church ve Hanks tarafından önerilmiş bir yöntemdir [67]. Doğal dil işleme,

anlamsal çıkarım, makine öğrenmesi ve metin işleme ile ilgili çalışmalarda oldukça yaygın kullanımı oldukça yaygındır. PMI'nin formülasyonu Eş. 3.1'de verilmiştir:

$$PMI(x, y) = \log \frac{p(x, y)}{p(x) * p(y)} \quad (3.1)$$

Burada, $p(x, y)$, x ve y 'nin beraber geçme sayısı, $p(x)$ x 'in geçme sayısı, $p(y)$ ise y 'nin geçme sayısıdır. Metin işleme çalışmalarında terimlerin beraber geçme sayıları çoğunlukla veri setindeki bütün dokümanlar üzerinden hesaplanır. Bazı araştırmalarda bu sayıların web aramalarından elde edildiği durumlar da vardır. PMI skorunun yüksek olması iki terim arasındaki anlamsal yakınlığın fazla olduğu anlamına gelir.

PMI skoru ilk defa fikir madenciliğinde Turney tarafından 2002 yılında yaptığı ve veri seti olarak otomobil ve sinema yorumlarını kullandığı çalışmada kullanılmıştır [7]. PMI'nin makine öğrenmesi özellik vektörü özelliği şeklinde ilk kullanılışı ise Mullen ve Collier tarafından 2004 yılında sinema yorumları üzerinde yapılan duygu analizi çalışmasıdır [68].

3.1.2. Sözcük tabanlı yöntemler

Bu grup altındaki yöntemler doğal dil işleme araç ve yöntemleri ile cümlelerin linguistik analizine dayalı yöntemlerdir. Sözcük tabanlı yöntemlerde denetimli makine öğrenmesinde olduğu gibi eğitim veri setine gereksinim duyulmamaktadır. Doğal dil işleme yöntem ve araçları ile cümlelerin linguistik analizi yapılır. Genellikle duygu terimleri sözlüğü kullanılarak cümlelerdeki duygu terimleri tespit edilir ve cümlelerin linguistik analizi yapılarak anlamsal çıkarımlar yapılır. Cümlelerin linguistik analizi için çoğunlukla üç yöntem ve bunların değişik varyasyonları yaygın olarak kullanılmaktadır:

Koşullu Rastgele Alanlar (CRF) : CRF Lafferty, McCallum ve Pereira tarafından 2001 yılında önerilmiştir. makine öğrenmesi, yapay zeka, örüntü tanıma gibi çalışma alanlarında kullanımı yaygın olan istatistiki bir modelleme yöntemidir [69]. Doğal dil işleme uygulamalarında yaygın şekilde kullanılmaktadır. CRF, doğal dil işleme çalışmalarında cümledeki sözcüklerin ne amaçla kullanıldığını belirlemede, o sözcükten sonra ve önce

gelen sözcüklerin, o sözcüğün cümlesinden sonra ve önce gelen cümlelerin de göz önünde bulundurularak uygulanan istatistik ve olasılık tabanlı yöntemler bütünüdür.

Bağılılık Ağacı: Bağılılık ağacı, cümlelerin linguistik analizinde yaygın bir biçimde kullanılan bir yönlü graf veri yapısıdır. Bağılılık ağacında cümledeki kelimeler (öğeler) düğümlerde tutulur ve hangi ögenin hangi ögeyi nitelediğine göre düğümler arasında kenarlar bulunur. Cümlelerden bağılılık ağacı doğal dil işleme araçları ile oluşturulur. Bağılılık ağacı oluşturulduktan sonra düğümler ve kenarlar arası ilişkilere bakılarak cümlelerin linguistik analizi gerçekleştirilir.

Kural Tabanlı Yaklaşım: Kural tabanlı yaklaşımda cümleler ile ilgili bir takım kurallar tanımlanır ve bu kurallara göre cümleler analiz edilerek çalışma yapılır. Bu kurallar sözcük türleri (sıfat, isim, zarf, fiil), sözcük türü örüntüleri ve bazı diğer cümle özelliklerine dayanan kurallardır. Bu gruptaki yöntemler çoğunlukla CRF ile kombine bir şekilde uygulanır. Kural tabanlı yaklaşımın fikir madenciliği ve duygu analizi çalışmasında ilk kullanımı Turney'in 2002 yılında kullanıcı yorumları veri seti üzerinde yaptığı çalışmada olmuştur [7]. Maharani, Widyanoro ve Khodra da sözcük türü örüntüleri ile ilgili kural tabanlı bir yöntem geliştirmişler ve ürün yorumları veri seti üzerinde duygu analizi çalışması yapmışlardır [55]. Maharani ve diğerleri bu çalışmalarında kuralları tanımlarken Turney'in yaptığı çalışmadaki [7] kurallardan yararlanmışlardır.

3.1.3. Yöntemlerin karşılaştırmalı analizi

Makine öğrenmesi tabanlı ve sözcük tabanlı yöntemlerin birbirlerine karşı üstünlükleri ve zayıf yönleri bulunmaktadır. Makine öğrenmesine dayalı yöntemlerin en zayıf yönü duygu analizi çalışması yapılacak konu ile ilgili etiketlenmiş eğitim veri setine ihtiyaç duyulmasıdır. Üstelik belli bir konudaki etiketlenmiş eğitim veri seti başka bir konuda da kullanılamamaktadır. Mesela diz üstü bilgisayarlar ile ilgili yorumlardan oluşan etiketlenmiş bir eğitim veri seti akıllı telefon yorumları üzerinde yapılacak bir duygu analizi çalışmasında işe yaramamaktadır. Çalışma yapılması gereken bütün alanlar ile ilgili yeteri miktarda eğitim veri seti bulmak da kolay değildir.

Sözcük tabanlı yöntemler, etiketlenmiş eğitim verisine ihtiyaç duymamalarıyla makine öğrenmesine dayalı yöntemlere göre üstünlük sağlarlar ama sözcük tabanlı yöntemler de

bazı zayıf yönleri sahiptir. Sözcük tabanlı yöntemleri uygulayabilmek için doğal dil işleme araçlarına ihtiyaç duyulmaktadır. Ayrıca duygu polarite tespiti için de çoğu zaman duygu terimleri sözlüğü gerekmektedir. Doğal dil işleme araçları ve duygu terimleri sözlükleri her dilde aynı değildir. Kimi dillerde çok sayıda ve güçlü araçlar varken bazı dillerdeki araçların sayısı az ve etkinlikleri çok zayıftır.

Sözcük tabanlı yöntemlerin, duygu terimlerinin kullanılmadığı ve dolaylı olarak görüş ifade edilen durumlarda başarı oranları düşüktür. Makine öğrenmesine dayalı yöntemler bu tür dolaylı ifadelerde daha etkindirler. Mesela “bu telefonun şarjı uzun gidiyor” şeklindeki bir cümle pozitif polariteye sahiptir ama bu sözcük tabanlı bir yöntemde böyle bir cümle polaritesinin pozitif olarak tespit edilmesi oldukça zordur çünkü cümle herhangi bir duygu terimi barındırmamaktadır. Öte yandan yeterli miktar ve çeşitlilikte etiketlenmiş eğitim veri setinin kullanıldığı makine öğrenmesine dayalı bir yöntem böyle bir cümle polaritesini tespit etmede daha başarılıdır.

Literatürdeki çalışmalar, makine öğrenmesine dayalı yöntemlerin duygu polaritesini sınıflandırmadaki başarısının sözcük tabanlı yöntemlere göre daha yüksek olduğunu ortaya koymaktadır. Kennedy ve Inkpen yaptıkları çalışmada [70], Hailong, Wenyan ve Bo yaptıkları karşılaştırmalı analiz çalışmasında [71] testlerde elde edilen sonuçlar denetimli makine öğrenmesi yöntemlerinin sözcük tabanlı yöntemlere üstünlük sağladıklarını göstermiştir. Bu iki grup altındaki yöntemlerin kıyaslaması Çizelge 3.1’de verilmiştir.

Çizelge 3.1. Makine öğrenmesi - Sözcük tabanlı yöntemlerin kıyaslaması

	Avantajlar	Dezavantajlar
Makine öğrenmesine dayalı yöntemler	- Daha başarılıdır - Dolaylı ifadelerde başarı oranı yüksektir.	- Etiketlenmiş eğitim veri setine ihtiyaç vardır. - Bir konudaki etiketlenmiş eğitim verisi başka konuda kullanılamaz.
Sözcük tabanlı yöntemler	- Etiketlenmiş eğitim veri setine ihtiyaç yoktur. - Doğal dil işleme araçları ve duygu terimleri sözlüğüne ihtiyaç olsa da bunlar çalışma alanına özgü değildirler. Elde bu araçlar varsa herhangi bir alandaki çalışmada kullanılabilirler.	- Başarı seviyesi daha düşüktür. - Doğal dil işleme araçları gerekir. - Duygu terimleri sözlüğü gerekir. - Dolaylı ifadelerde başarı oranı daha düşüktür.

3.1.4. Özellik çıkarımında kullanılan yöntemler

Özellik tabanlı düzeyde yapılan duygu analizi çalışmalarında doküman düzeyindeki duygu analizi çalışmalarına ek olarak duygu ifadelerinin hedefleri olan ürün özelliklerinin tespit edilmesi gerekir. Bu görevin özellik çıkarımı olarak adlandırıldığı yukarda belirtilmişti. Literatürdeki çalışmalarda özellik çıkarımı görevi için kullanımı yaygın olan yöntemler şunlardır:

Sık geçen isimler (Frequent names): Sık geçen isimler 2004 yılında Hu ve Liu'nun ürün yorumları veri seti ile yaptıkları özellik tabanlı düzeydeki duygu analizi çalışmasında önerilmiş bir yöntemdir [35]. Bu yöntem şu varsayım üzerine bina edilmiştir: İnsanlar ürünler hakkında yorum yaparken farklı durumlardan ve farklı şeylerden bahsederler ve bunlardan bahsederken de birbirinden çok farklı kelimeler kullanırlar. Ancak, ürün özellikleri hakkında konuşmaya gelindiği zaman, insanların kullandıkları kelimeler çoğunlukla aynı kelimelerdir. Ürün özellikleri de yorumlarda çoğunlukla isim türü kelimelerle ifade edilmektedir. Sonuç olarak veri setindeki tüm yorumlarda frekansı yüksek olan isim türündeki kelimeler genellikle ürünün özelliklerini ifade eden kelimelerdir. Birçok ürün yorumu ve bu konudaki veri setleri incelendiği zaman bu varsayımın doğru olduğu gözlemlenmektedir. Ancak, yorumlarda sık geçen isimlerin hepsi ürün özellikleri ile ilgili olmayabilir. Bazı sık geçen isim türü kelimeler ürün özellikleri haricinde başka kelimeler olabilmektedir. Bu kelimelerin belirlenip ürün özellikleri ile ilgili kelimelerin aralarından çıkarılmaları gerekmektedir. Sık geçen isimler yöntemi 2004 yılında tarafından önerildikten sonra birçok çalışmada değiştirilerek ve geliştirilerek uygulanmıştır.

Popescu ve Etzioni bu yöntemin başarı seviyesi yükseltmek için yaptıkları çalışmada terimler arasındaki anlamsal yakınlığı ölçmekte kullanılan PMI skorlamasını, ürün ve sık geçen isim arasında uygulamışlardır [37].

Goldensohn ve diğerleri bütün cümlelerdeki sık geçen isimleri değil sadece duygu ifadesi barındıran subjektif cümlelerde sık geçen isimleri dâhil ederek bu yöntem üzerinde geliştirme yapmışlardır [72].

Duygu ifadesi - hedef ilişkilerinin linguistik analizi: Bu yöntemde ilk olarak genellikle bir duygu terimleri sözlüğü kullanılarak duygu ifadeleri belirlenir. Belirlenen bu duygu

ifadelerinin hedefleri ya ürünün doğrudan kendisi ya da ürünün bir özelliği/yönüdür. İkinci olarak cümlelerin linguistik analizi yapılarak duygu ifadelerinin hedefleri olan ürünün kendisi ya da ürün özellikleri tespit edilmeye çalışılır. Linguistik analizde sözcük türleri, bağıllık ağacı, CRF gibi doğal dil işleme yöntem ve araçları yoğun şekilde kullanılır.

Zhuang, Jing ve Zhu'nun yaptıkları çalışma [73], Swapna ve Wiebe'nin cümle ayrıştırıcı doğal dil işleme aracı kullanarak yaptıkları çalışma [74], Qiu, Liu, Bu ve Chen'in linguistik analiz için bağıllık ağacı kullandıkları ve "Double Propagation" ismini verdikleri çalışma [44] bu yöntemin kullanıldığı özellik tabanlı düzeydeki duygu analizi çalışmalarına örneklerdir.

Makine öğrenmesine dayalı yöntemler: Denetimli makine öğrenmesi yöntemleri de özellik çıkarımında sıkça kullanılmaktadırlar. Bilgi çıkarımı çalışmalarında kullanılmakta olan denetimli makine öğrenmesine dayalı çok sayıda yöntem vardır. Sıralı öğrenmeye dayalı makine öğrenmesi yöntemleri özellik çıkarımında en yaygın kullanılanlarıdır. CRF [69] ve saklı markov modeli [75] en yaygın şekilde kullanılan sıralı öğrenme metotlarıdır.

Konu Modelleme – Latent Dirichlet Allocation (LDA): LDA 2003 yılında Blei ve diğerleri tarafından büyük metin yığınlarındaki konuları ortaya çıkarmak için önerilmiş bir algoritmadır [2]. Özellik tabanlı düzeyde duygu analizi çalışamalarında ürün özellikleri LDA'daki konulara tekabül etmektedir. LDA algoritması, büyük miktardaki metin yığınlarındaki konuları ortaya çıkarmak için geliştirilmiştir. LDA, fikir madenciliği ve duygu analizi çalışmalarındaki veri setleri olan sosyal medya mesajları, kullanıcı yorumları gibi kısa metinler üzerinde etkin değildir. LDA'yı bu tip kısa metinler üzerinde uygulamak için araştırmalar yapılmaktadır.

Ontoloji tabanlı yöntemler: Ontoloji varlık bilimidir. Semantik web uygulamalarında sıkça kullanılmaktadır. Ontoloji biliminde varlıklar özellikleri ile birlikte modellenir. Bu modellemede varlığın özellikleri ve o özellikler de hiyerarşik olarak alt özellikleri birlikte modellenir. Özellik tabanlı düzeyde duygu analizi çalışmasında metinlerden özellik çıkarırken bu ontoloji modelinden ve modeldeki varlık özelliklerinden faydalanılır. Isidro, Rafael ve Francisco gerçekleştirdikleri özellik tabanlı düzeydeki duygu analizi çalışmasında veri seti olarak sinema yorumlarını kullanmışlar ve özellik çıkarımı görevi

iin IMDB'den alınmıř verilerden oluřturulmuř olan <http://www.movieontology.org> sitesindeki sinema ontolojilerinden yararlanmıřlardır [51].

Bu yntemlerden bařka son yıllarda derin ğrenmeye ve yapay sinir ağılarına dayalı yaklařımlar da duygu analizi alıřmalarında sıklıkla kullanılmaya bařlanmıřtır. Hem duygu polarite tespiti, hem de zellik ıkarımı iin derin ğrenmeye dayalı yntemler kullanılmaktadır.



4. TEZ İLE İLGİLİ ÖN ÇALIŞMALAR VE ÖN BİLGİLER

Bu tezde Türkçe dili üzerinde özellik tabanlı düzeyde duygu analizi çalışması yapılmıştır. Duygu analizi çalışmaları dilden bağımsız düşünülemez. Bu yüzden duygu analizi ve fikir madenciliği çalışmaları doğal dil işleme yöntem ve araçları ile çok yakından ilgilidir. Doğal dil işleme çalışmaları özellikle 1980’li yıllardan sonra, makine öğrenmesi yöntemlerinin yaygınlaşmasından sonra hız kazanmıştır ve bu çalışmalar çoğunlukla İngilizce dili üzerine odaklanmıştır. Aradan geçen yıllar boyunca doğal dil işleme çalışmaları kapsamında İngilizce dili üzerinde çalışan birçok etkili doğal dil işleme araçları geliştirilmiştir.

Türkçe’de duygu analizi çalışması yapmak İngilizce kadar rahat değildir. En başta İngilizce’deki kadar çok sayıda doğal dil işleme araçları mevcut değildir. Mevcut olan araçlarda da hata payı İngilizce’deki araçlara göre daha yüksektir. Hata payının yüksek olmasında araçların yetersizliğinden ziyade Türkçe dilinin sondan eklemeli, üretken ve karmaşık morfolojik yapısı etkilidir. Türkçe’nin sondan eklemeli üretken yapısı, morfolojik belirsizliği çok arttırmaktadır. Yüret ve Türe yaptıkları bir doğal dil işleme çalışmasında, İngilizce’de morfolojik belirsizlik çok nadir iken, üzerinde çalıştıkları veri setinde toplam kelimelerin %42 gibi yüksek bir oranının morfolojik belirsizlik içerdiğini ortaya koymuşlardır [76]. Bu morfolojik belirsizlik doğal dil işleme ve duygu analizi çalışmalarını zorlaştırmaktadır. Oflazer yayınladığı bir makalede de Türkçe’nin bitişken (sondan eklemeli) yapısı, cümle öğelerinin dizilişlerinin çok değişken ve esnek olması gibi unsurların Türkçe’de doğal dil işleme çalışmalarını ne kadar güçleştirdiğini ortaya koymuştur [77].

4.1. Türkçe’deki Doğal Dil İşleme Araçları ve Mevcut Durum

Duygu analizi ve fikir madenciliği çalışmaları, doğal dil işleme yöntem ve araçlarının kullanıldığı metin işleme çalışmaları olduğu için üzerinde çalışılan dil ile çok yakından ilişkilidir. Tezdeki geliştirilen yönteme geçmeden önce doğal dil işleme konusunda Türkçe’deki araçlar, bu araçların yetenekleri ve doğal dil işlemedeki mevcut durum aşağıda özetlenmiştir:

Zemberek: Türkçe’de doğal dil işleme ile ilgili en çok bilinen ve yaygın olarak kullanılan araç Zemberek’tir [78]. Zemberek, kelimeyi kök ve eklerine ayıran, kelimenin sözcük türünü tespit eden bir morfolojik analiz aracıdır. Ayrıca Zemberek yazım yanlışları yapılmış kelimeleri düzeltme işlevi de sunmaktadır. Yanlış yazılmış bir kelimeyi verdiğinizde ona en yakın alternatif kelimeleri sunmaktadır. Zemberek’in iki versiyonu bulunmaktadır. İkinci versiyonunda morfolojik belirsizlik giderici modülü de bulunmaktadır. Java dilinde yazılmış, açık kaynak bir araçtır. Zemberek, masaüstü grafiksel ara yüzü olan bir uygulamadır. Ayrıca da “jar” formatında Java kütüphanesi olarak da sunulmaktadır ve Java uygulamalarında kullanılabilir. Zemberek, masaüstü grafiksel ara yüzü olan bir uygulamadır. Ayrıca da “jar” formatında Java kütüphanesi olarak da sunulmaktadır ve Java uygulamalarında kullanılabilir.

Kemal Oflazer’in Aracı: Kemal Oflazer, Türkçe’de doğal dil işleme üzerine oldukça çok sayıda araştırma yapmış, ürünler ortaya çıkarmış ve makaleler yayınlamıştır. Kemal Oflazer tarafından geliştirilmiş olan, Zemberek’le benzer işlevleri yerine getiren morfolojik analiz aracı mevcuttur [79]. Bu araç, sonlu durum makinası olarak tasarlanmış ve uygulamaya geçirilmiştir ve serbest bir ürün olan Xerox Finite State Tool üzerinde çalışmaktadır.

İTÜ – Doğal Dil İşleme Yazılım Zinciri: İstanbul Teknik Üniversitesi Doğal Dil İşleme grubu tarafından geliştirilmiş olan ve içinde Tokenizer, Morphological Analyzer, Morphological Disambiguator, Dependency Parser, vb. işlevleri sunan, “Türkçe Doğal Dil İşleme Yazılım Zinciri” olarak adlandırdıkları araçlar mevcuttur [80]. Bu araçlar, bir web sayfası aracılığıyla interaktif olarak sunulduğu gibi ticari amaçlar dışında sadece bilimsel araştırma amaçlı olarak “http request - response” mekanizması yoluyla da uygulamalarla entegre edilebilmektedir.

Türkçe Morfolojik Belirsizlik Giderici: Türkçe’de doğal dil işleme çalışmalarında morfolojik belirsizlik çok önemli bir sorundur. Şak, Güngör ve Saraçlar’ın geliştirmiş olduğu bu araç [81] morfolojik belirsizliği gidermekte oldukça etkilidir.

SentiTurkNet: Bu araç İngilizce’deki SentiWordNet [82] aracının Türkçe versiyonudur [83]. Birbiri ile eş anlamlı kelimeler seti (synset) ve bu kelimeler setinin pozitif, negatif ve objektif polarite skorlarını içermektedir. SentiTurkNet’te toplamda 15.000’e yakın eş anlamlı kelimeler seti bulunmaktadır.

4.2. Morfolojik Analiz ve Morfolojik Belirsizlik

Morfolojik analiz, yapısal analiz anlamına gelmektedir. Morfolojik analiz bir kelimenin kök ve eklerine ayrılması, kökünün sözcük türünün (fiil, isim, sıfat, vs.) tespit edilmesi ve her bir ekin türünün tespit edilmesi işlemleridir. Morfolojik analiz bütün doğal dil işleme çalışmalarında yapılması gereken bir görevdir. Doğal dil işleme çalışmalarında kelimelerin sözcük türünün ve köklerinin tespit edilmesi gerekmektedir. Duygu analizi ve fikir madenciliği çalışmalarında da doğal dil işleme araç ve yöntemleri sıkça kullanıldığı için morfolojik analize ihtiyaç duyulmaktadır.

Türkçe sondan eklemeli karmaşık bir morfolojik yapıya sahip bir dil olduğu için morfolojik analizi İngilizce ve diğer birçok dil gibi kolay değildir. Türkçe'nin morfolojik analizinde morfolojik belirsizlik çözülmesi gereken önemli bir problemidir. Morfolojik belirsizlik bir kelimenin birden fazla morfolojik analizinin olması durumudur. Morfolojik belirsizliğin daha iyi anlaşılması için aşağıda birkaç tane örnek verilmiştir.

Mesela ilk örnek olarak “kitabı” kelimesinin Zemberek ile morfolojik analizi yapıldığı zaman aşağıdaki iki farklı analiz ortaya çıkmaktadır:

[(kitap:kitab) (Noun;A3sg+P3sg:1+Nom)]

[(kitap:kitab) (Noun;A3sg+Pnon+Acc:1)]

Her iki analizde de kök isim tipindeki “kitap” kelimesi olarak tespit edilmiştir, ancak birinci analizde sondaki -ı eki “onun kitabı” şeklindeki kullanılan üçüncü tekil şahıs iyelik eki olarak tespit edilmiş, ikinci analizde ise sondaki -ı eki “kitabı okudum” şeklinde kullanılan ismin -i hali olarak tespit edilmiştir. Dolayısıyla “kitabı” kelimesinin morfolojik analizi yapılırken bu iki analizden hangisinin alınması gerektiği belirsizdir.

Başka bir örnek olarak, “geçti” kelimesinin Zemberek ile morfolojik analizi yapıldığı zaman aşağıdaki iki farklı analiz ortaya çıkmaktadır. Her iki analizde de kök “geç” kelimesi olsa da birinci analizde sıfat tipindeki “geç” kelimesi, ikinci analizde ise “geçmek” fiilinin kökü olan fiil tipindeki “geç” kelimesidir.

[(geç:geç) (Adj)(Verb;Past:ti+A3sg)]

[(geçmek:geç) (Verb;Pos+Past:ti+A3sg)]

Daha karmaşık morfolojik belirsizlik içeren bir örnek için “karı” kelimesine bakalım. Zemberek bu kelime için aşağıdaki 6 farklı morfolojik analizi yapmaktadır:

[(kar:kar) (Noun;A3sg+P3sg:ı+Nom)]

[(kar:kar) (Noun;A3sg+Pnon+Acc:ı)]

[(kâr:kar) (Noun;A3sg+P3sg:ı+Nom)]

[(kâr:kar) (Noun;A3sg+Pnon+Acc:ı)]

[(karımak:karı) (Verb;Pos+Imp+A2sg)]

[(karı:karı) (Noun;A3sg+Pnon+Nom)]

Bunlardan ilk ikisinde kök “kar yağdı” şeklinde kullanılan, isim tipindeki “kar” kelimesidir. Üçüncü ve dördüncüde kök “çok kar ettim” şeklinde kullanılan, yine isim tipindeki “kar” kelimesidir. Beşinci de ise ihtiyarlamak anlamında kullanılan “karımak” fiilinin kökü olan fiil tipindeki “karı” kelimesidir. Sonuncuda ise kök “karı koca” şeklinde kullanılan isim tipindeki “karı” kelimesidir.

Örneklerde görüldüğü gibi Türkçe dilinde morfolojik belirsizlik oldukça yaygındır. Yüret ve Türe Türkçe’deki morfolojik belirsizlik üzerine yaptıkları bir çalışmada [76] ellerindeki veri setindeki kelimelerin %42’sinin morfolojik olarak belirsiz olduğunu ortaya koymuşlardır.

Türkçe’de sağlıklı bir şekilde doğal dil işleme çalışması yapabilmek için bu morfolojik belirsizliklerin giderilmesi gerekmektedir. Bu görevi yerine getiren doğal dil işleme araçları mevcuttur. İTÜ Türkçe Doğal Dil İşleme Yazılım Zinciri [80] yazılımları içinde morfolojik belirsizlik giderici modül de mevcuttur. Bu araç online olarak kullanılmakta, http istekleri ile fonksiyonlar çağırılmakta ve sonucu alınmaktadır.

Ayrıca Haşim Sak ve arkadaşları tarafından geliştirilmiş olan morfolojik belirsizlik giderici araç da mevcuttur [81]. Bu perceptron algoritmasını kullanan ve perl ile yazılmış bir araçtır. Perl Unix ve Linux işletim sistemleri üzerinde çalışan bir yazılımdır ama Windows’ta işletim sisteminde de Perl kodlarını çalıştırmak için bir takım araçlar vardır.

Zemberek'in iki farklı sürümü bulunmaktadır. Zemberek'in ikinci sürümü Zemberek-nlp olarak geçmektedir ve bu Zemberek sürümünde morfolojik belirsizlik giderme işlevi de bulunmaktadır.

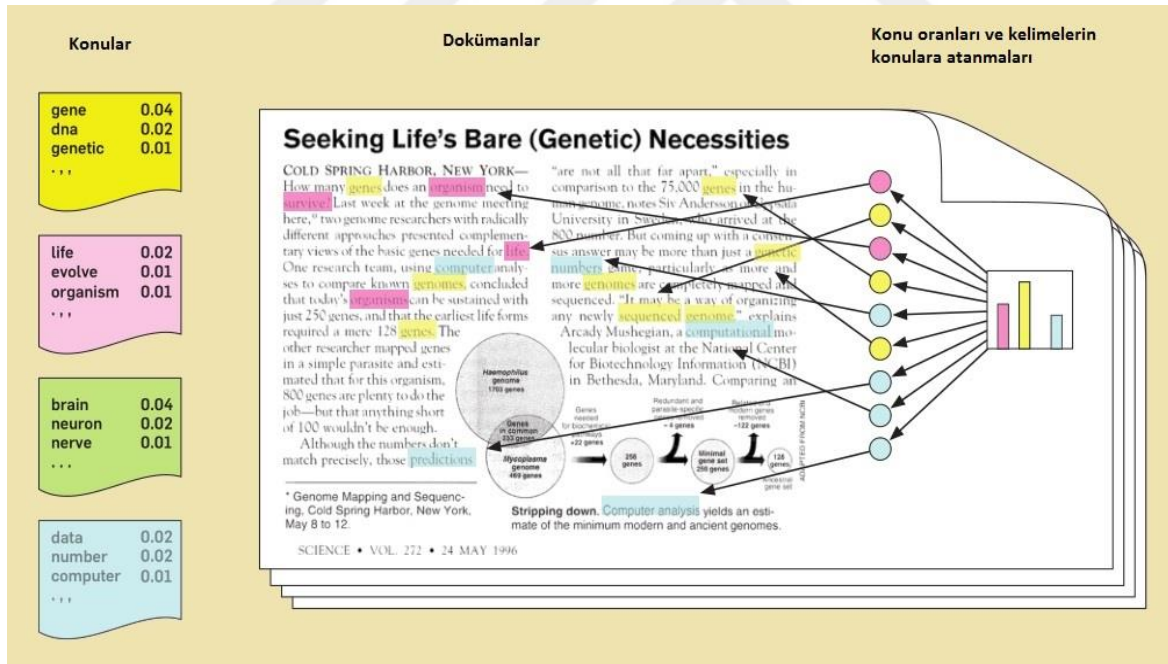
Tez çalışmamız kapsamında bu morfolojik belirsizlik giderici araçlarla ilgili araştırma yapılmıştır. Java ile kullanım kolaylığından ötürü Zemberek-nlp ve İTÜ Doğal Dil İşleme (DDİ) Yazılım Zinciri araçları üzerine durulmuştur. Yapılan bir takım denemelerde İTÜ DDİ Yazılım Zinciri'nin morfolojik belirsizliği gidermede Zemberek-nlp'den biraz daha başarılı olduğu görülmüştür. Bu sebepten ötürü ilk olarak İTÜ DDİ Yazılım Zinciri'nin kullanılması düşünülmüştür. Fakat İTÜ DDİ Yazılım Zinciri bir kütüphane olarak projeye eklenmemekte, çevirim içi olarak http istekleri ile kullanılmaktadır, bundan ötürü çalışma hızı çok düşüktür. Tez çalışmamızdaki toplam yorum ve cümle sayısı hesaba katıldığında bu çalışma süresinin kabul edilebilir seviyenin çok ötesinde olduğu görülmüştür. Bundan ötürü morfolojik belirsizlikleri gidermek için Zemberek-nlp aracının kullanılmasına karar verilmiştir.

Morfolojik belirsizliği gidermek için Zemberek'i kullanırken Zemberek'in bir hatası tespit edilmiştir. Bir cümlede aynı köke sahip iki veya daha fazla kelime varsa, Zemberek bu kelimelerin birden fazla olan morfolojik analizleri arasından seçim yaparken hata yapmakta ve yanlış morfolojik analizi seçmektedir. Zemberek'in kaynak kodu incelenerek bu hatanın nedeni bulunmuştur. Zemberek'in bazı sınıfları "override" edilerek, bu hata giderilmiştir.

4.3. Latent Dirichlet Allocation Algoritması

Bu tezde geliştirilen ve detayları bir sonraki ana başlıkta detaylı şekilde anlatılacak olan özellik tabanlı düzeyde duygu analizi yöntemi, Latent Dirichlet Allocation (LDA) algoritmasının kısa metinler ve ürün yorumları için uyarlanmasına dayalı özgün bir yöntemdir. LDA'nın kısa metinlere uyarlanması literatürde oldukça güncel bir araştırma konusudur. [41, 47, 60-62] bu alanda yapılmış literatürdeki onlarca çalışmalardan sadece birkaç tanesidir. Bu tezde önerilen LDA'ya dayalı yöntemin anlaşılabilmesi için öncelikle LDA algoritmasının anlaşılması gerekmektedir. Bu başlık altında LDA'nın detaylı anlatımı yapılmıştır.

LDA algoritması 2003 yılında Blei, Ng ve Jordan tarafından metin yığınlarındaki saklı olan konuları keşfetmek için önerilmiş bir yöntemdir [2]. LDA algoritması, çok sayıdaki dokümanların ve büyük miktardaki metinlerin içindeki konuları ortaya çıkarmak için kullanılan keşifsel bir yöntemdir. LDA’da, her doküman belli sayıda ve her biri belli orandaki konuların karışımı olarak modellenir. Her bir konu ise o konu ile ilgili kelimeler üzerinde olasılıksal bir dağılım olarak modellenir. Bunlara göre LDA bir dokümanın oluşturulma şeklini şu şekilde modeller: Önce dokümanın hangi konuları hangi oranda içereceğine karar verilir. Sonra dokümanın her bir kelimesinin önce hangi konu ile ilgili olacağına karar verilir ve buna göre o konunun olasılıksal dağılımı altındaki kelimelerden birisi seçilerek dokümana eklenir. Bu sürecin grafiksel gösterimi Şekil 4.4’te gösterilmiştir. Sol tarafta konular ve o konuların kelimeler üzerinde olasılıksal dağılımı gösterilmiştir. En sağdaki bar grafikte dokümanın kapsadığı konular ve oranları gösterilmiştir. Sağdaki bar grafiğin solundaki renkli yuvarlaklarla da dokümandaki kelimelerin konulara atanmaları gösterilmiştir.



Şekil 4.4. Konu Modelleme – LDA grafiksel gösterim

Bu modellemede, dokümanlar ve dokümanlardaki kelimeler gözlenen değişkenlerdir. Konular, konuların kelimeler üzerine olasılıksal dağılımları, dokümanların konular üzerine olasılıksal dağılımları ve dokümanlardaki her bir kelimenin konu atanması ise gizli

değişkenlerdir. Bu gözlenen ve gizli değişkenlerin Şekil 4.4'teki gösterime denk gelen kısımları da işaret edilerek tanımları Çizelge 4.1'de verilmiştir.

Çizelge 4.1. LDA notasyonları

Notasyon	Açıklama
K	Konu sayısı
$\beta_{1:K}$	Konular, her β_k kelimeler üzerine bir dağılımdır (soldaki kelimeler üzerine dağılımlar)
θ_d	d dokümanındaki konu oranları (sağdaki bar grafik)
$\theta_{d,k}$	k konusunun d dokümanındaki oranı (sağdaki bar grafikteki bir sütun)
Z_d	d dokümanın konu atamaları (sağdaki renkli yuvarlakların tümü)
$Z_{d,n}$	d dokümanındaki n 'inci kelimenin konu ataması (sağdaki n 'inci renkli yuvarlak)
W_d	d dokümanının gözlenen kelimeleri
$W_{d,n}$	d dokümanındaki n 'inci kelime

LDA'nın arka planındaki istatistiksel ve matematiksel formülasyonlara ve modele geçmeden önce istatistikteki Dirichlet Dağılımı üzerinde duracağız çünkü LDA'da dokümanların konular üzerine dağılımları ve konuların kelimeler üzerine dağılımları, Dirichlet Dağılımı olarak modellenmektedir.

Dirichlet Dağılımı: Gösterimi $\text{Dir}(\alpha)$ şeklinde olan Dirichlet Dağılımı, parametresi α vektörü olan ve çok değişkenli, sürekli olasılıksal dağılım modelidir. İki değişkenli bir olasılıksal dağılım fonksiyonu olan beta dağılımının çok değişkenli olarak uyarlanmasıdır.

K tane değişkenden ($K \geq 2$) oluşan $X(x_1, x_2 \dots x_k)$ vektörünün, $\alpha(\alpha_1, \alpha_2 \dots \alpha_k)$ parametresi ile Dirichlet Dağılımına göre gösterimi $X \sim \text{Dir}(\alpha)$ şeklindedir ve Dirichlet olasılıksal yoğunluk fonksiyonu Eş. 4.1'deki gibidir:

$$f(X(x_1, x_2 \dots x_k), \alpha(\alpha_1, \alpha_2 \dots \alpha_k)) = \frac{1}{B(\alpha)} \prod_{i=1}^K x_i^{\alpha_i - 1} \quad (4.1)$$

Burada aşağıdaki şartların sağlanması gerekmektedir:

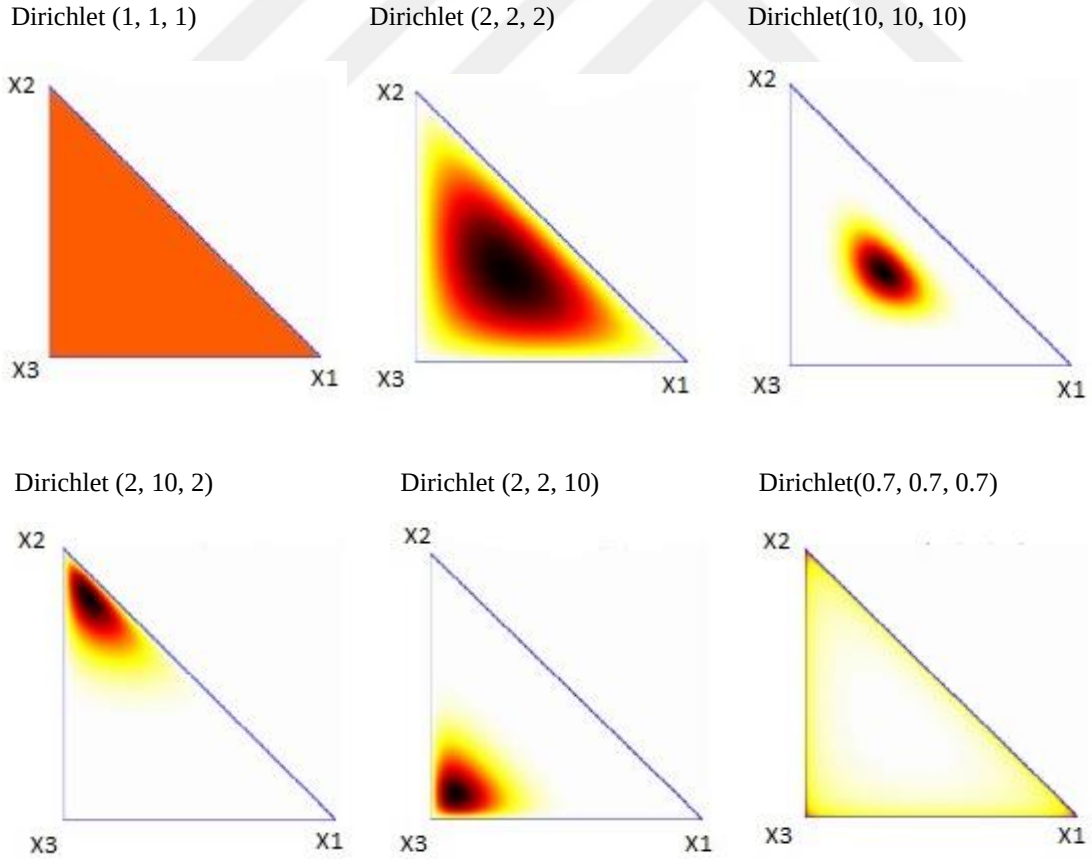
- $x_1, x_2 \dots x_k \geq 0$ (X vektörünün her bir değişkeni 0'dan büyük veya eşittir)
- $x_1 + x_2 \dots x_k = 1$ (X vektörünün değişkenlerinin toplamı 1'dir)
- $\alpha_1, \alpha_2 \dots \alpha_k > 0$ (α vektörünün her bir değişkeni 0'dan büyüktür)

Normalleştirme sabiti olan $B(\alpha)$ fonksiyonunun, gama fonksiyonu türünden tanımı Eş. 4.2'deki gibidir.

$$B(\alpha) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)} \quad \alpha = (\alpha_1, \dots, \alpha_K) \quad (4.2)$$

Dirichlet parametresi olan α vektörünün, yoğunluk dağılımlarını nasıl etkilediğinin daha iyi anlaşılması için üç değişkenli $X(x_1, x_2, x_3)$ vektörünün farklı α vektörü değerlerine göre muhtemel olasılıksal yoğunluk dağılımları Şekil 4.5'teki görsellerde resmedilmiştir.

$\alpha(1, 1, 1)$ Dirichlet parametresi ile x_1, x_2, x_3 değişkenlerinin dağılımlarındaki yoğunluk tamamen homojendir. $\alpha(2, 2, 2)$ Dirichlet parametresi ile dağılım yoğunluğu ortaya doğru kaymıştır yani x_1, x_2, x_3 değişkenlerinin dağılımlardaki ağırlıklarının eşit ve birbirine yakın olduğu durumların sayısı nispeten fazla, bu değişkenlerin tek başlarına veya



Şekil 4.5. Farklı dirichlet parametrelerine göre yoğunluk dağılımları

ikili gruplar halinde ağırlıklarının baskın geldiği durumların dağılımdaki sayısı nispeten azdır (köşe ve kenarlarda yoğunluk az). $\alpha(10, 10, 10)$ Dirichlet parametresi ile yoğunluk iyice merkeze kaymıştır yani x_1, x_2, x_3 değişkenlerinin dağılımlardaki ağırlıklarının eşit ve birbirine yakın olduğu durumların sayısı çok fazladır, neredeyse dağılımın tamamı buraya yığılmıştır (köşe ve kenarlarda yoğunluk çok daha az). $\alpha(2, 10, 2)$ parametresi ile elde edilen dağılımda x_2 'nin ağırlıkta olduğu dağılımların yoğunluğu diğerlerine göre daha fazladır. $\alpha(0.7, 0.7, 0.7)$ parametresi ile seyrek bir dağılım elde edilmektedir. Çoğu durumda bu değişkenlerin ağırlığı çok az veya hiç yoktur. Köşe ve kenar bölgelerde az sayıda dağılım elde edilmektedir.

LDA'da her doküman her konuyu farklı oranlarda içermektedir. Mesela D1 dokümanı K1 konusunu %45, K2 konusunu %25, K3 konusunu %20, K4 konusunu %10 oranında, D2 dokümanı K1 konusunu %35, K2 konusunu %30, K3 konusunu %20, K4 konusunu %15 oranında içerebilir. Bütün dokümanların bu şekilde ayrı ayrı konu oranları vardır. LDA'da dokümanların konular üzerine dağılımları Dirichlet Dağılımı ile modellenmektedir. Aynı şekilde, konuların kelimeler üzerine dağılımları da Dirichlet Dağılımı ile modellenmektedir.

$P(X | Y)$ şeklinde ifade edilen dağılım notasyonu, verilen bir Y değişkenine göre X değişkeninin olasılıksal dağılımı anlamına gelmektedir. Bu notasyona göre LDA'da bir dokümanın üretilmesinin arka planındaki matematiksel modellemenin detaylarına bakalım. LDA'da sırayla önce konuların ve konuların kelimeler üzerine olan dağılımları, ardından da dokümanların konu oranlarının belirlenmesi ve bu konu oranlarına göre, ilgili konulardan kelime seçilerek dokümana eklenmesi süreçleri modellenmektedir.

Buna göre konuların $(\beta_{1:K})$ kelimeler üzerine Dirichlet Dağılımına göre üretim sürecinin matematiksel modeli Eş. 4.3'teki gibidir (notasyonlar Çizelge 4.1'de verilmiştir):

$$\prod_{k=1}^K (P(\beta_k | \alpha)) \quad (4.3)$$

Burada α Dirichlet Parametresi vektördür. Verilen α vektörüne göre her bir konunun (K tane konu) kelimeler üzerine dağılımları (β_k) üretilmektedir (Şekil 5.4'teki soldaki konular).

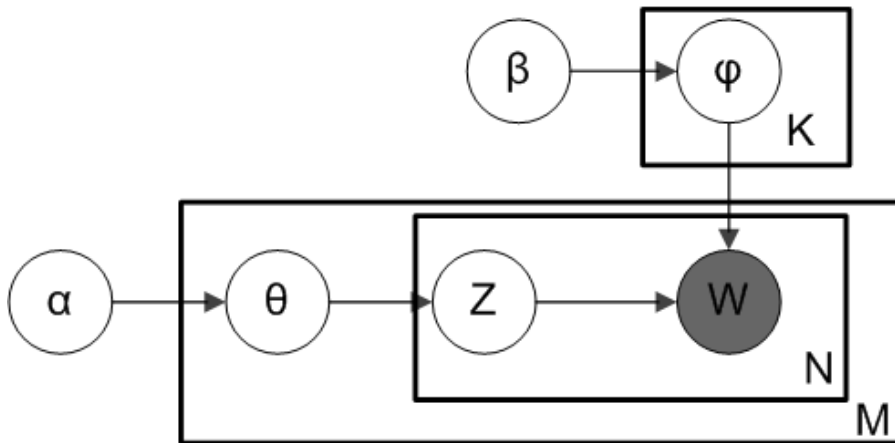
Konular ve konuların kelimeler üzerine dağılımı üretildikten sonra, dokümanların üretilmesinin matematiksel modellemesi ise Eş. 4.4'teki gibidir (notasyonlar Çizelge 4.1'de verilmiştir):

$$\prod_{d=1}^D P(\theta_d | \alpha) \left(\prod_{n=1}^N P(Z_{d,n} | \theta_d) P(W_{d,n} | Z_{d,n}, \beta_{1:k}) \right) \quad (4.4)$$

Burada ilk önce her bir dokümandaki konu oranları (θ_d) α Dirichlet Parametresi ile Dirichlet Dağılımına göre ile belirlenmektedir. Bu Eş. 4.4'teki $\prod_{d=1}^D P(\theta_d | \alpha)$ bölümüdür. α Dirichlet Parametresine göre θ_d değişkenlerinin üretilmesidir.

Dokümanlardaki konu oranları (θ_d) belirlendikten sonra her bir dokümana eklenecek olan her bir kelimenin hangi konu ile ilgili olduğu belirlenmekte yani eklenecek kelimenin konu ataması yapılmakta yani $Z_{d,n}$ değerleri belirlenmektedir. Bu Eş. 4.4'teki $P(Z_{d,n} | \theta_d)$ bölümüdür.

Dokümana eklenecek olan kelimenin konu ataması yapıldıktan sonra yani $Z_{d,n}$ belirlendikten sonra, o konunun kelimeler üzerine olasılıksal dağılımı ($\beta_{1:k}$) da kullanılarak her bir dokümandaki her bir kelimenin ($W_{d,n}$) üretimi gerçekleştirilir. Bu da Eş. 4.4'teki $P(W_{d,n} | Z_{d,n}, \beta_{1:k})$ bölümüdür.



Şekil 4.6. LDA plaka gösterimi

LDA'nın yukardaki formülasyonlar dışında Şekil 4.6'daki gibi plaka şeklinde gösterimi (plate notation) de vardır [2]. Bu gösterimde beyaz yuvarlak düğümler gizli değişkenleri,

gri yuvarlak düğümler gözlenen değişkenleri, oklar bağılılıkları ve dikdörtgen şeklindeki plakalar tekrar eden döngüsel işlemleri göstermektedir. Burada β dirichlet parametresi ile ϕ konusu üretilmektedir. Bu işlem her bir konu için (K tane konu) tekrar ettiğinden ötürü ϕ plaka içine alınmıştır. α dirichlet parametresi ile dokümanın konu dağılımı olan θ üretilmektedir. Bu işlem her bir doküman (M tane doküman) için tekrar ettiği için yani döngüsel olduğu için plaka içine alınmıştır. Dokümanın konu dağılımı olan θ üretildikten sonra dokümandaki her bir kelimenin konu ataması olan Z üretilmektedir. Z de dokümandaki her bir kelime (N tane kelime) için tekrar eden döngüsel bir işlemdir ve plaka içine alınmıştır. Eklenecek olan kelimenin hangi konu ile ilgili olduğu belirlendikten sonra o konunun dağılımı altında bulunan kelimelerden birisi olan W , dokümana eklenir. Bu da tekrar eden döngüsel bir işlem olduğu için plaka içine alınmıştır.

LDA'da dokümanların ve dokümanlardaki kelimelerin üretilmesi bu şekilde modellenmiştir. Fakat bizim elimizde dokümanlar ve dokümanların içindeki kelimeler ($W_{d,n}$) dışında belirli hiçbir şey yoktur. Dokümanlardaki konu oranları (θ_d), dokümanlardaki her bir kelimenin hangi konu ile ilgili olduğu ($Z_{d,n}$), konular ($\beta_{1:K}$) ve konuların kelimeler üzerine dağılımları (β_k) ise bilinmemektedirler yani bunlar gizli değişkenlerdir. LDA algoritmasında yapılan şey tersine gitmeye çalışmak yani dokümanlar ve dokümanlar içindeki kelimelerden yukarda anlatılan doküman üretim modellemesine göre bu gizli değişkenlere gitmektir. Fakat eldeki verilerle yani sadece dokümanlar ve içindeki kelimelerle bunu direk olarak yapmak mümkün değildir, çünkü çok fazla bilinmeyen mevcuttur.

Direk olarak tersine gitmek mümkün değildir fakat bir takım akıllı sezgisel yaklaşımlarla bu dokümanları üretebilecek olan konular, konuların kelimeler üzerine dağılımları ve dokümanlardaki her bir kelimenin bu konulardan hangisi ile ilişkili olduğu gibi gizli değişkenlerde saklı olan değerler yaklaşık olarak çıkartılmaya çalışılır. Bu sezgisel yöntemler içinde en yaygın olanı ise Gibbs örnekleme yöntemidir.

İlk olarak kaç tane konu sayısına (K) karar verilmelidir. Dokümanların genel olarak hangi konular ile ilgili olduğu hakkında ön bilgiye sahipsek, bu konulara göre konu sayısı belirlenebilir. Hiçbir ön bilgiye sahip değilsek, farklı konu sayıları denenerek, kelimelerin konular altına en mantıklı şekilde ayrıştıkları konu sayısı bulunabilir.

Konu sayısı belirlendikten sonra, Gibbs örneklemede ilk olarak her bir dokümandaki her bir kelime rastgele bir konuya atanır. Böyle bir atama yapılmış olmasıyla birlikte gizli değişkenler olan konular, konuların kelime dağılımları ($\beta_{1:k}$), dokümanların konu dağılımları (θ_d) ve dokümandaki her kelimenin konulara atanması (Z_d) değişkenlerine rastgele doğru olmayan değerler (dağılımlar) atanmış olur. Kelimelere bu ilk konu atamaları rastgele yapıldığı için son derece yanlış, birbiri ile tutarsız ve uyumsuz bir model elde etmiş oluruz. Bundan sonra döngüsel bir şekilde her bir dokümanın ve her bir kelimenin üzerinde tekrar tekrar geçerek daha tutarlı bir model elde edilecek doğrultuda kelimelere yeni konular atama işlemi yapılır. Bu döngüsel olarak kelimelere yeni konular belirleme işi tamamen rastgele değil belli bir mantığa göre yapılır. Bir kelime ele alınırken o kelime dışındaki diğer tüm kelimelerin konu atamalarının doğru ama o kelimenin yanlış olduğu varsayılır. Bu varsayıma göre mevcut kelimenin hangi konuya atanırsa daha tutarlı bir model elde edileceği araştırılır. D dokümanındaki, W kelimesine yeni bir konu seçerken, Z konusu için aşağıdaki Eş. 4.5'e göre bir olasılık ağırlığı hesaplanır. Her bir konu için bu hesaplama yapılır.

$$P(Z|W, D) = \left(\frac{D' \text{ de } Z' \text{ ye ait kelimesi sayısı} + \alpha}{D \text{ deki toplam kelime sayısı}} \right) \times \frac{Z \text{ konusunda } W \text{ kelimesi sayısı} + \beta}{Z \text{ deki toplam kelime sayısı} + \beta_{\text{sum}}} \quad (4.5)$$

Bu formül yukarda anlatılan doküman üretiminin matematiksel modelleme ve formülasyonundan gelmektedir. Bu formülün arka planındaki mantık şudur:

1. Bir dokümanda bir konunun ağırlığı ne kadar çoksa, o dokümandaki W kelimesinin o konu ile ilgili olma ihtimali de o derece fazladır. Formüldeki çarpımın sol tarafı D dokümanındaki Z konusunun oranını veren denklemdir.
2. W kelimesinin tüm dokümanlarda Z konusuna atanma oranı nedir? W kelimesinin tüm dokümanlarda Z konusuna atanma oranı ne kadar fazlaysa şu an ele alınan mevcut W kelimesinin de o konuya atanma ihtimali o kadar yüksek olur. Formüldeki çarpımın sağ tarafı W kelimesinin tüm dokümanlarda Z konusuna atanma oranını veren denklemdir.

W kelimesinde her bir konu için bu formüle göre olasılık ağırlıkları hesaplanır ve ağırlıklara göre ağırlıklı rassal şekilde W kelimesine yeni bir konu ataması yapılır. Bu şekilde tüm dokümanların ve tüm kelimelerin üzerinden bir kez geçildiği zaman model bir kademe daha tutarlı hale gelmiş olur. Bu şekilde belli bir iterasyondan (örn: 3000

iterasyon) sonra artık model tutarlı ve stabil hale gelir. Stabil hale geldikten sonra kelimelerin atandığı konular genel olarak değişmemeye başlar.

LDA yöntemi 2003 yılında Blei ve diğerleri tarafından önerilmiş ve daha sonra araştırmacılar tarafından dokümanlardaki konuları ortaya çıkarmak amaçlı birçok çalışmada kullanılmıştır.





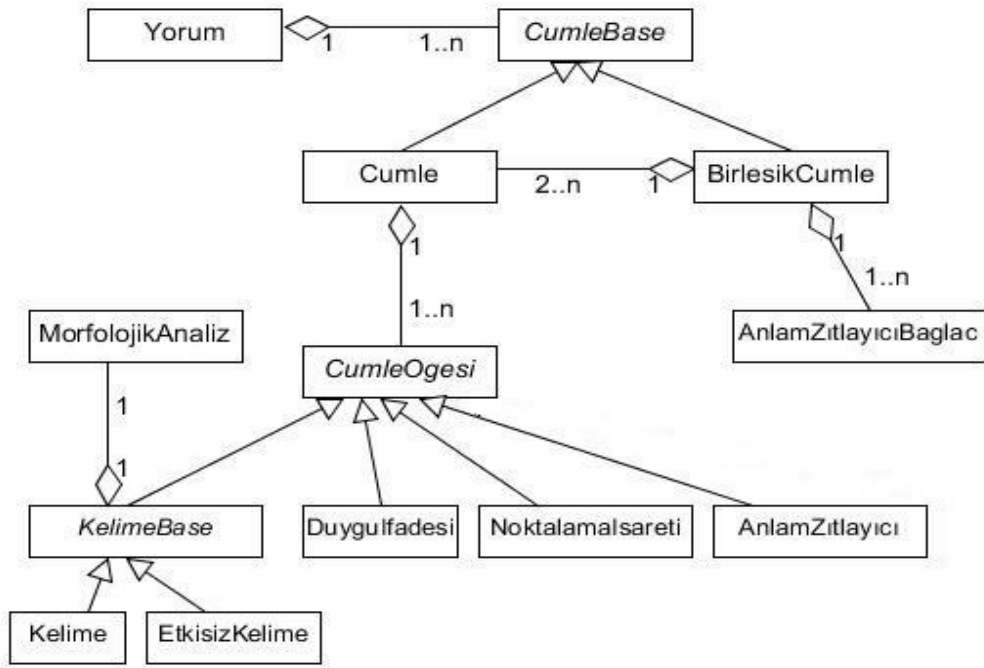
5. GELİŞTİRİLEN YÖNTEM

Bu tezde Türkçe diliyle özellik tabanlı düzeyde duygu analizi çalışması yapılmıştır. Veri seti olarak www.hepsiburada.com e-ticaret sitesinden alınmış 1292 tane akıllı telefon yorumu kullanılmıştır. Çalışmanın Türkçe dili üzerinde yapılmış olması, literatüre önemli bir katkı sağlamıştır. Literatürde İngilizce, Çince gibi yaygın dillerde yapılmış çok sayıda duygu analizi çalışmaları mevcuttur. Bu dillerde her türlü düzeyde ve her türlü yöntemle yapılmış çalışmalar mevcuttur. Fakat Türkçe’de yapılmış duygu analizi çalışmaların sayısı çok azdır. Yapılmış olan çalışmaların da çoğu doküman düzeyinde duygu analizi çalışmalarıdır. Bu çalışmalar yukarda “Duygu Analizi Çalışmaları Literatür Taraması” başlığı altında verilmiştir. Sonuç olarak Türkçe dilinde duygu analizi oldukça bakir bir bilimsel çalışma alanı olarak durmaktadır. Bu tezde bu çalışma alanındaki boşluğun bir doldurulması amaçlanmıştır.

Bu tezde geliştirilen uygulama için programlama dili olarak Java seçilmiştir. Java tercih edilmiş olmasının en önemli nedeni Türkçe’deki en önemli ve etkin doğal dil işleme aracı olan Zemberek’in Java tabanlı bir araç olmasıdır.

Geliştirilen yöntemin detaylarına girmeden önce tasarlanan veri yapısından bahsedilecektir. Ürün yorumlarını, yorumları oluşturan cümleleri, cümleleri oluşturan sözcükleri, sözcüklerin de köklerini ve sonlarına aldıkları ekleri tutabilmek ve bunlar üzerinde etkin şekilde işlemler yapabilmek için Java sınıfları ve veri yapıları oluşturulmuştur. Oluşturulan bu Java sınıflarının UML sınıf diyagramları Şekil 5.1’de verilmiştir.

Bir yorum bir veya daha fazla cümleden oluşmaktadır. Cümleler iki tiptir: birleşik cümle ve normal cümle. Birleşik cümle *ama*, *fakat* gibi anlam zıtlayıcı bağlaç veya bağlaçlar içeren cümlelerdir. Bu tür cümlelerde bağlaçların sol ve sağ tarafındaki alt cümlelerin ayrı ayrı ele alınmaları gerekmektedir. Bu yüzden BirlesikCumle diye sınıfa ihtiyaç duyulmuştur. BirlesikCumle sınıfı en az bir tane anlam zıtlayıcı bağlaç ve en az iki tane alt cümle içerir. BirlesikCumle ile Cumle sınıflarının ortak özellikleri *CumleBase* üst soyut (abstract) sınıfında toplanmıştır.



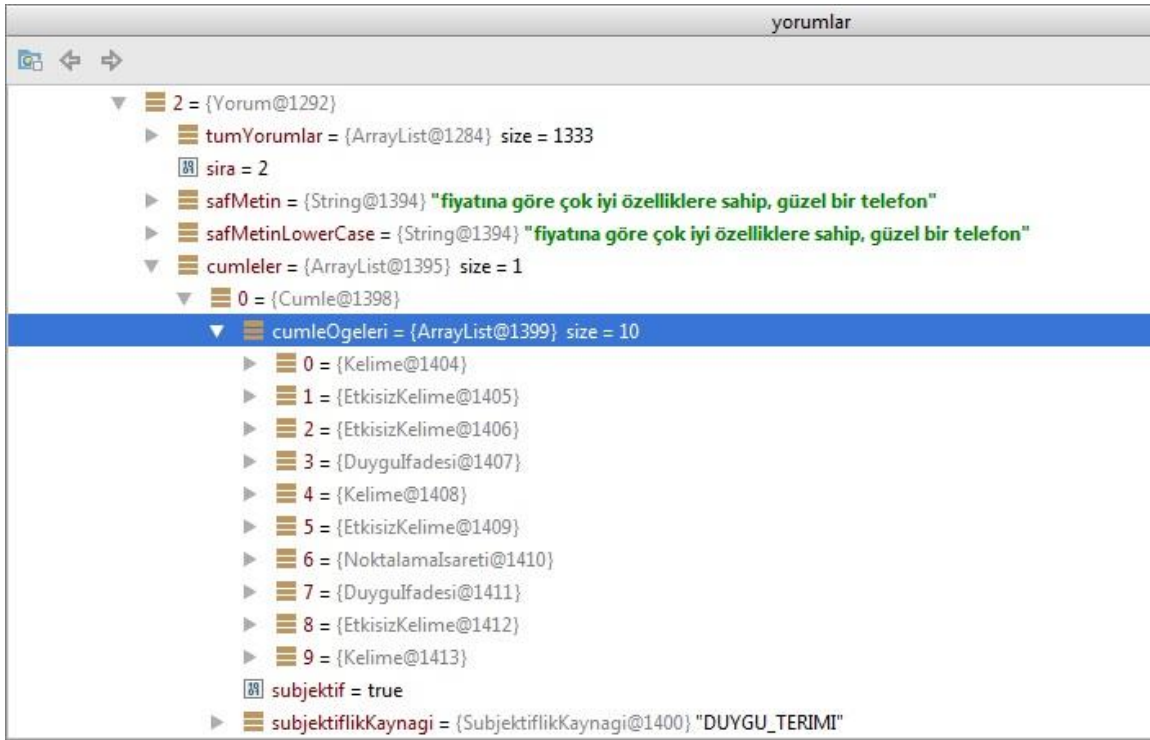
Şekil 5.1. Veri yapısı UML sınıf diyagramı

Cümleyi oluşturan sözcüklerin her bir türü için ayrı sınıf tanımlamaları yapılmıştır. Bunlar DuyguIfadesi, Noktalamasareti, AnlamZitlayici, Kelime ve EtkisizKelime sınıflarıdır. DuyguIfadesi yazarın fikir veya duygusunu ifade eden pozitif veya negatif polariteye sahip sözcükleri tutmak için tanımlanmıştır. Noktalamasareti cümle içindeki virgül, noktalı virgül gibi noktalama işaretlerini tutmak için tanımlanmış sınıftır. AnlamZitlayici cümledeki *fakat*, *ama* gibi anlam zıtlayıcı bağlaçları tutmak için kullanılan sınıftır. EtkisizKelime *daha*, *ve*, *ile*, *vs.* gibi anlamsal bir önemi olmayan etkisiz kelimeleri (stop word) tutmak için tanımlanmış sınıftır. Bu türlerin hiçbirine girmeyen sözcükler de Kelime sınıfı ile temsil edilmektedir. Kelime ve EtkisizKelime türündeki sözcüklerin morfolojik analizine ihtiyaç duyulduğu için bu iki sınıfın ortak özellikleri *KelimeBase* soyut sınıfında toplanmış ve bu iki sınıf bu soyut sınıftan türetilmişlerdir.

Bu veri yapısını birkaç örnek yorum ile açıklayalım. Örnek yorumlar ve bu yorumların veri yapısı ile temsil edilişi aşağıda verilmiştir:

Örnek Yorum 1: “fiyatına göre çok iyi özelliklere sahip, güzel bir telefon”

Örnek Yorum 1'in veri yapısı ile temsil edilmesi Şekil 5.2'de gösterilmiştir. Yorum sınıfının safMetin alanında yorumun metninin yazım hataları ile birlikte saf hali tutulur. Yorumlar üzerinde işlem yapmadan önce ilk yapılan iş yorumdaki tüm karakterleri küçük harfe dönüştürmektir. safMetinLowerCase alanı küçük harfe dönüştürülmüş yorum metninin tutulduğu alandır. cumleler alanı yorumu oluşturan cümlelerin tutulduğu alandır. Bu örnek yorum tek cümleden oluşmaktadır.

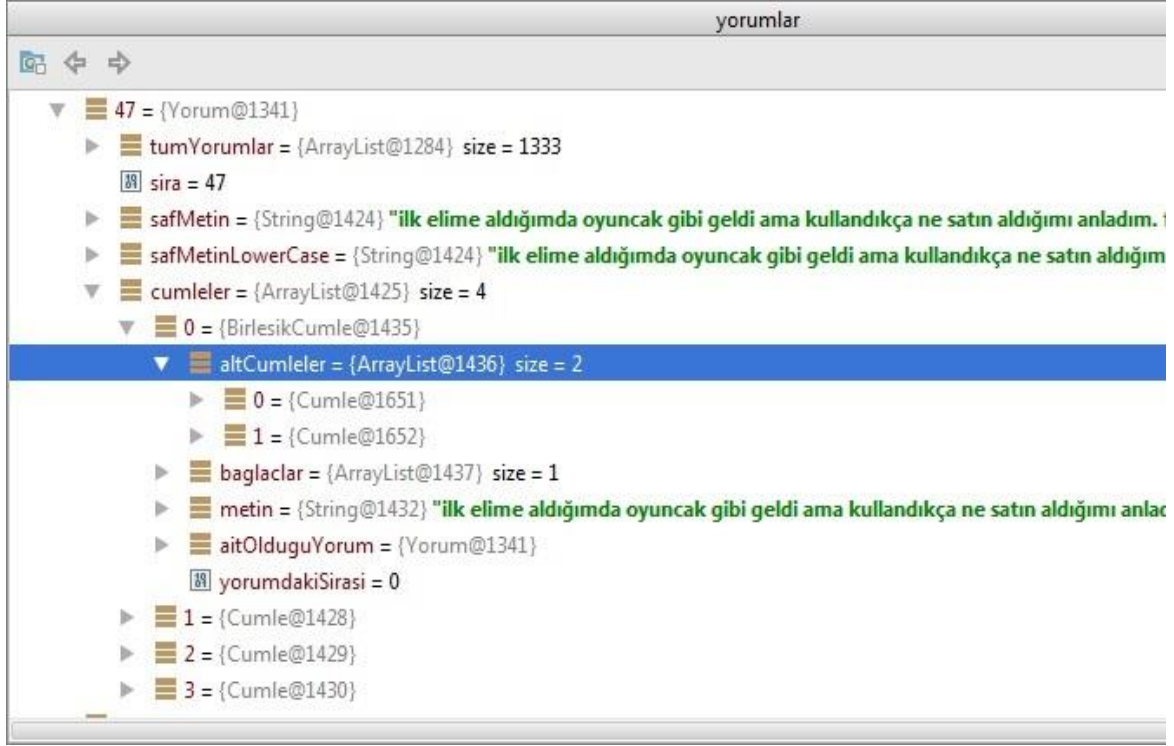


Şekil 5.2. Örnek yorum 1'in sınıf yapısı ile temsil edilmesi

Cumle sınıfı yorumdaki cümleleri tutan alandır. Cumle sınıfı içinde cümledeki her bir sözcüğü tutan cumleOgeleri alanı vardır. Bu yorumdaki cümle öğelerinden *fiyatına* sözcüğü **Kelime** sınıfı ile, *göre* sözcüğü **EtkisizKelime** sınıfı ile, *iyi* sözcüğü **Duygulfadesi** sınıfı ile, cümledeki virgül ise **NoktalamaIsareti** sınıfı ile temsil edilmektedir.

Örnek Yorum 2: *ilk elime aldığımda oyuncak gibi geldi ama kullandıkça ne satın aldığımı anladım. fotoğraf çekimleri harika. bu üründen almayı düşünenler site güvenebilir. taksit imkanı artık biteceği için ben elimi çabuk tuttum.*

Bu yorumdaki ilk cümlede *ama* anlam zıtlayıcı bağlacı bulunmaktadır. Yukarıda da belirtildiği gibi bağlaçlarla bağlı cümleler BirlesikCumle sınıfı temsil edilmektedir. Burada da bu ilk cümle bir bağlaç ve iki alt cümleden oluşan bir BirlesikCumle olarak temsil edilmiştir. BirlesikCumle temsili Şekil 5.3'te gösterilmiştir.



Şekil 5.3. Örnek yorum 2'nin sınıf yapısı ile temsil edilmesi

Hepsiburada sitesinden alınan yorumlar bir XML dosyasında toplanmıştır. XML dosyasından yorumları sırayla okuyan ve her bir yorumu önce cümlelere, sonra her bir cümleyi öğelere ayıran bir ayrıştırıcı (parser) yazılmıştır. Bu ayrıştırıcı, cümleleri öğelere ayırdıktan sonra cümle öğelerinin duygu ifadesi, etkisiz kelime, anlam zıtlayıcı, noktalama işareti ve kelime şeklinde tespitini yapıp bunlara göre gerekli nesneleri oluşturmakta ve veri yapısını oluşturmaktadır.

Ayrıştırıcı tarafından bu tespitlerin yapılabilmesi için Türkçe etkisiz kelimeler listesine ve duygu terimleri sözlüğüne ihtiyaç vardır. İnternette araştırmacıların kullanıma sunulmuş hazır Türkçe etkisiz kelimeler listeleri bulunmaktadır. Bu listelerden birisi alınmıştır. Liste, İngilizce etkisiz kelimeler listesinden de faydalanılarak ve eldeki veri seti incelenerek bizim tarafımızdan genişletilmiştir. Duygu terimleri sözlüğü oluşturulması hakkındaki çalışmalar ise ilerleyen başlıklarda anlatılacaktır.

Özellik çıkarımı yöntemini tasarlarken, özellik tabanlı düzeydeki duygu analizinde göz önünde bulundurulması gereken bütün durumlar incelenmiştir. Göz önünde bulundurulması gereken durumlar madde madde aşağıda sıralanmıştır:

1. Aynı ürün özelliği çok farklı sözcüklerle ve çok farklı şekillerde ifade edilebilmektedir. Sistem aynı ürün özelliğini ifade eden bu farklı sözcükleri ve farklı ifadeleri farklı ürün özellikleri olarak çıkarmamalı, bunları aynı ürün özelliği kategorisi altında gruplandırmalıdır.
2. Bir ürün özelliğinden bahsedilirken her zaman o özelliğin adı söylenmeyebilir. Bazen farklı sözcüklerle dolaylı şekilde o özellikten bahsedilebilmektedir.
3. Bir ürün özelliği her zaman tek kelime ile ifade edilmeyebilir. Bazen iki ya da daha fazla kelimeden oluşan sözcük grupları ile ifade edilebilmektedir.
4. Bir cümle bazı durumlarda bir ürün özelliği ile ilgili, bazı durumlarda da birden fazla ürün özelliği ile ilgili olabilir.
5. Görüş belirtilen bir cümlede, her zaman ürünün bir özelliğinden bahsedilmiyor olabilir. Bazen, kullanıcı o ürün hakkında genel bir beğeni de ifade ediyor olabilir.

Daha iyi anlaşılabilmesi için, bu durumlar aşağıda örnek yorum cümleleri üzerinde gösterilmiştir:

Cümle 1: Telefonun gece fotoğraf çekimi çok iyi.

➔ Telefonun “kamera” özelliğinden bahsedilmektedir. “Kamera” özelliğinden iki ayrı sözcük (altı çizili) ile bahsedilmiştir. Bu sözcükler “kamera” özelliği ile ilişkili kelimeleridir. Ayrıca bu cümlede “kamera” sözcüğü kullanılmadan “kamera” özelliğinden dolaylı şekilde bahsedilmiştir.

2. Bir cümlede birden fazla ürün özelliğinden bahsediliyorsa (Cümle 3 ve 4), sistem cümledeki bu özellikleri tek tek çıkarabilmelidir. Cümlenin tek bir ürün özelliği hakkında olduğu varsayılmamalıdır.
3. Aynı ürün özelliği çok farklı sözcüklerle ifade edildiği zaman, sistem aynı ürün özelliğini ifade eden bu farklı sözcükleri farklı ürün özellikleri olarak çıkarmamalı, bunları aynı ürün özelliği altında gruplayabilmelidir.
4. Kullanıcı herhangi bir ürün özelliğinden bahsetmeden bir görüş ifade ediyorsa (Cümle 5), sistem bunu tespit edebilmeli ve böyle cümlelerden özellik çıkarmaya çalışmamalıdır.

Bütün bu maddeleri göz önüne alarak, geliştirilen yöntem iki safhadan oluşmaktadır: Birinci safha cümlelerin tek bir ürün özelliği içerecek şekilde parçacıklara ayrılması safhasıdır. Buna göre yukardaki örnek cümleler aşağıdaki gibi parçacıklara ayrılmalıdır:

Cümle 1

- Parça 1: “Telefonun gece fotoğraf çekimi çok iyi.”

Cümle 2

- Parça 1: “Düşük ışıkta kameranin odaklanma süresi *tatminkar değil*.”

Cümle 3

- Parça 1: “Dokunmatik ekran hassasiyeti”
- Parça 2: “batarya ömrü”
- Parça 3: “internet hızı çok güzel”

Cümle 4

- Parça 1: “Telefonun şarjı iyi gidiyor ve”
- Parça 2: “bağlantı hızı da süper”

Cümle 5

- Parça 1: “*Hoşuma gitti*.”

Bu safhanın sonunda her biri tek bir ürün özelliği ile ilgili olan cümle parçacıkları elde edilmiş olmaktadır. İkinci safha ise gruplama safhasıdır. Aynı ürün özelliği ile ilgili olan cümle parçacıklarının aynı özellik kategorisi altından gruplanmaları gerekmektedir. Yukardaki örnek cümlelerden elde edilen parçaların Çizelge 5.1'deki gibi gruplanmaları gerekmektedir:

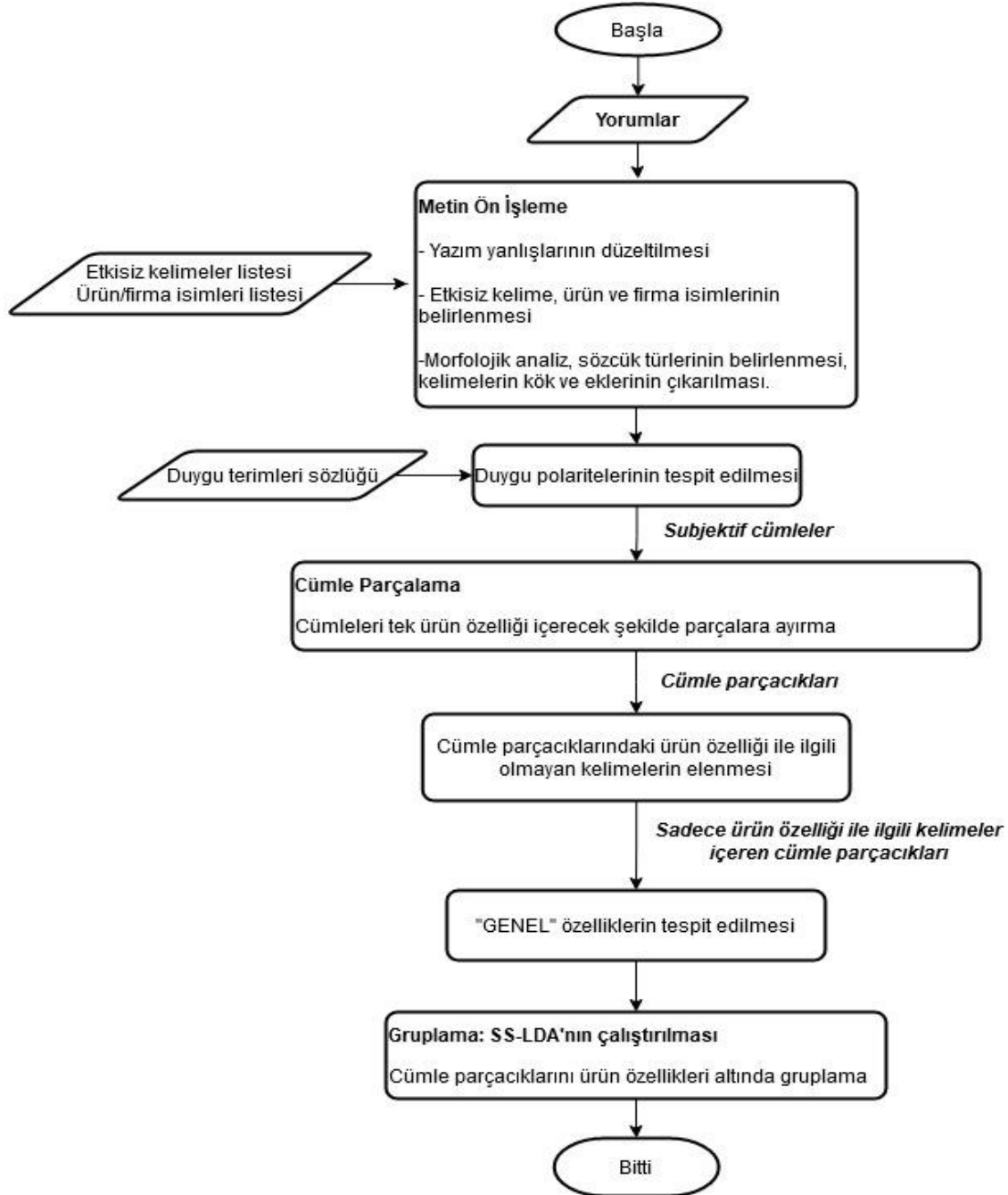
Çizelge 5.1. Cümle parçacıklarının ürün özellikleri altında gruplanması

Ürün özelliği	Ürün özelliği altında gruplanmış cümle parçacıkları
Kamera	- Telefonun gece fotoğraf çekimi çok iyi. (Cümle 1, Parçacık-1) - Düşük ışıktaki kameranın odaklanma süresi tatminkar değil. (Cümle 2, Parçacık-1)
Ekran	- Dokunmatik ekran hassasiyeti (Cümle 3, Parçacık-1)
Batarya	- batarya ömrü (Cümle 3, Parçacık-2) - Telefonun şarjı iyi gidiyor ve (Cümle 4, Parçacık-1)
İnternet	- internet hızı çok güzel (Cümle 3, Parçacık-3) - bağlantı hızı da süper (Cümle 4, Parçacık-2)
Özellik yok	- Hoşuma gitti (Cümle 5, Parçacık-1)

Cümleleri parçacıklara ayırma ve daha sonra bu parçacıkların gruplanması geliştirilen yöntemin iki ana safhasıdır. Ayrıca, tüm sistemde bu iki ana safha dışında bir takım ara görevler de vardır. Cümle parçacıklarının gruplanması, LDA yönteminin kısa metinler için bizim tarafımızdan yapılmış özgün bir uyarlaması ile yapılmıştır. Bu yüzden bu çalışmada önerilen yöntem Sentence Segment LDA (SS-LDA) olarak adlandırılmıştır. Geliştirilen yöntem sözcük tabanlıdır ve eğitim verisi gerektirmemektedir. Tüm sistemin ana hatları Şekil 5.4'de gösterilmiştir. Cümle parçalamanın, parçacıkları gruplamanın ve sistemdeki diğer görevlerin nasıl gerçekleştirildiği izleyen başlıklarda anlatılmaktadır.

5.1. Metin Ön İşleme

Metin ön işleme üç görevden oluşmaktadır: 1. Yazım yanlışlarının düzeltilmesi. 2. Etkisiz kelimelerin, ürün ve firma isimlerinin elenmesi. 3. Morfolojik analiz (kelimelerin kök ve eklerine ayrılması, sözcük türlerinin belirlenmesi).



Şekil 5.4. Sistemin ana hatları

Metin ön işleme çalışmalarındaki en önemli görev yazım yanlışlarının düzeltilmesidir. Çalışmamızda yazım yanlışlarının düzeltilmesinde Zemberek [78] kullanılmıştır. Zemberek'te *denetle* ve *oner* şeklinde fonksiyonlar vardır. *Denetle* fonksiyonu verilen bir kelimenin Türkçe dil kurallarına uygun bir kelime olup olmadığını denetler. *Oner* fonksiyonu ise verilen bir kelimeye en yakın Türkçe dil kurallarına uygun alternatif kelime veya kelimeler önerir.

Zemberek kullanılarak yazım hataları şöyle düzeltilmektedir: Kelime önce *denetle* fonksiyonu ile yazım hatasına karşı denetlenir. Eğer bir yazım hatası varsa bu defa *oner* fonksiyonu ile bu hatalı yazılmış kelimeye benzeyen alternatif kelime veya kelimeler bir dizi yapısı şeklinde alınır. Bu kelimelerden bir tanesi seçilir ve yazımı yanlış olan kelimenin yerine konarak yazım hatası giderilir. Burada önemli bir karar, Zemberek birden fazla kelime önerdiği zaman bu kelimelerden hangisinin seçileceğidir. Önerilen kelimelerden genelde ilk sıradaki tercih edilmektedir çünkü yanlış yazılmış kelimeye en yakın olan kelime ilk sıradakidir. Zemberek ile ilgili kod örneklerinde de kullanımı genelde bu şekildedir. Zaten önerilen kelimelerden hangisinin seçileceği ile ilgili akıllı bir yöntem uygulanmıyorsa, ilk önerilen kelimeyi tercih etmek dışında yapılabilecek çok da bir şey yoktur.

Çalışmamızdaki veri setinden “güzel” yerine yanlışlıkla “guzel” olarak yazılmış bir kelimeyi örnek olarak ele alalım. Zemberek’in *denetle* fonksiyonu “guzel” kelimesi için *false* dönmektedir çünkü Türkçe dil kurallarına uygun bir kelime değildir. False döndüğü için bu defa *öner* fonksiyonu ile Zemberek’ten “guzel” kelimesi için öneriler alınır. Zemberek “guzel” kelimesi için aşağıdaki önerileri sunmaktadır:

guzel → güzel, Uzel, gazel

Görüldüğü gibi Zemberek, 3 tane kelime önerisinde bulunmuştur. Herhangi akıllı bir seçme algoritması uygulanmadığı için ilk öneri olan “güzel” kelimesi alınır ve yazım yanlış başarılı bir şekilde düzeltilir.

Yine çalışmamızdaki veri setinden “sipariş” kelimesinin yanlışlıkla “sıprariş” olarak yazıldığı durumu ele alalım.

sıprariş → sipariş

“Sıprariş” kelimesi için Zemberek sadece “sipariş” kelimesini önermektedir çünkü Türkçe dil kurallarına uygun “sıprariş” kelimesine benzeyen başka kelimeler bulamamıştır. “Sıprariş” yerine “sipariş” konularak yazım yanlış başarılı şekilde düzeltilir.

Bu defa “önce” yerine yanlışlıkla “önece” yazıldığı durumu ele alalım. “Önece” kelimesi için Zemberek aşağıdaki önerilerde bulunmuştur:

önece → nece, önce, öncece, önüce, önlece, öneme, önemce, ötece, önere, önerce, ödece, özece, öznece, ögece, enece, övece, Önence, Önen'e, önele, önelce, ökece, inece, dönece, Dönece

Görüldüğü gibi Zemberek tam 24 tane kelime önerisinde bulunmuştur. Bu önerilerden ilk sıradaki olan “nece” kelimesi alınır. Bu defa ilk sıradaki doğru kelime olmamıştır ve yazım hatası düzeltme görevi başarısız olmuştur. Oysa ilk sıradaki “nece” kelimesi yerine ikinci sıradaki “önce” kelimesi seçilmiş olsaydı başarılı olunacaktı.

Zemberek’in bu son örnekteki gibi çok sayıda öneride bulunduğu ve ilk sıradaki önerilen kelimenin doğru çıkmadığı durumlara sıkça rastlanılmaktadır.

5.1.1. Zemberek ile yazım yanlış düzeltme yöntem önerisi

Bu tez çalışmasında, yazım yanlışlarını düzeltmede Zemberek’in önerdiği kelimelerden seçim yapma üzerine bir çalışma yürütülmüştür. Öncelikle Zemberek’in ilk sırada önerdiği kelimelerin doğruluk oranı tespit edilmeye çalışılmış, daha sonra Zemberek’in önerdiği kelimelerden daha iyi seçim yapabilmeye yönelik bir yöntem önerilmiştir.

Çalışmadaki ürün yorumlarından 470 tanesi bu çalışmaya dahil edilmiştir. Bu 470 yorumda yanlış yazılmış birbirinden farklı kelimelerin sayısı 373 olarak tespit edilmiştir. Yanlış yazılmış birbirinin aynısı kelimeler tek kelime olarak sayılmıştır. Bu 373 yanlış yazılmış kelime için Zemberek’in ilk önerdiği kelimelerden 255 tanesinin doğru, 118 tanesinin ise yanlış olduğu gözlemlenmiştir. Bu, yüzdelerle ifade edilecek olursa %68’i yanlış, %32’si ise doğrudur. Bu sonuçlar toplu olarak Çizelge 5.2’de verilmiştir.

Çizelge 5.2. Zemberek'in yazım yanlışlarını düzeltme istatistikleri

	Sayı	Yüzde(%)
Çalışmaya dahil edilen yorum	470	-
Yanlış yazılmış tekil kelime	373	-
Zemberek ilk sırada doğru kelime öneriyor	255	%68
Zemberek ilk sırada yanlış kelime öneriyor	118	%32

%32 oldukça yüksek bir hata oranıdır. Bu %32'lik, Zemberek'in ilk önerdiği kelimelerin yanlış olduğu durumların birçoğunda doğru kelime Zemberek'in önerdiği tüm kelimelerin arasında bulunmaktadır. Zemberek'in önerdiği kelimeler arasından daha doğru seçimler yapabilmek ve hata oranını düşürmek için şöyle bir yöntem önerilmiştir: Veri seti çok sayıda ürün yorumlarından yani birbiri ile benzer metinlerden oluşmaktadır. Bu metinlerde birbirine benzer konulardan bahsedilmekte, birbirine benzer ifadeler ve kelimeler kullanılmaktadır. Zemberek'in önerdiği kelimelerden birincisini almak yerine bu kelimelerin her birinin tüm yorumlardaki frekansını hesaplayıp en yüksek frekansa sahip olanını almanın daha başarılı olacağı düşünülmüştür.

Yukarıdaki 373 tane yanlış yazılmış kelime üzerinde bu yöntem uygulanmış ve Çizelge 5.3'teki sonuçlar elde edilmiştir:

Çizelge 5.3. Yazım yanlışlarını önerilen yöntemle düzeltme istatistikleri

	Sayı	Yüzde(%)
Çalışmaya dahil edilen yorum	470	-
Yanlış yazılmış tekil kelime	373	-
Önerilen yöntem başarılı	313	%84
Önerilen yöntem başarısız	60	%16
Önerilen yöntem ile zemberek'in aynı kelimeyi üretmesi	304	%82
Önerilen yöntem ile zemberek farklı kelimeler üretmesi	69	%18
Farklı üretilen kelimelerde(69 tane) zemberek doğru	1	%1
Farklı üretilen kelimelerde(69 tane) önerilen yöntem doğru	60	%87
Farklı üretilen kelimelerde(69 tane) ikisi de yanlış	8	%12

373 tane yanlış yazılmış kelimenin 304 tanesinde bu yöntemle önerilen kelime ile Zemberek'in ilk önerdiği kelime aynı olmuş, 69 tanesinde ise bu yöntemle önerilen kelime

farklı olmuştur. Bu 69 farklı kelimenin önerildiği durumların 60 tanesinde bu yöntemin önerdiği kelime, 1 tanesinde ise Zemberek'in ilk önerdiği kelime doğru olmuştur. 8 tane kelimedede ise her iki yöntemin önerdiği kelime de yanlış olmuştur.

Yukarda örneğini verdiğimiz “önce” yerine yanlışlıkla “önece” yazılması durumunu bu yöntemle inceleyelim. “Önece” kelimesi için Zemberek'in ürettiği kelimeler ve bu kelimelerin tüm yorumlardaki frekansları aşağıdaki gibi hesaplanmıştır. Burada ikinci sırada önerilen “önce”nin frekansı en yüksek çıkmış, tercih edilmiş ve başarılı bir seçim gerçekleşmiştir.

önece → nece(0), önce(104), öncece(0), önüce(0), önelece(0), öneme(0), önemce(0), ötece(0), önere(0), önerce(0), ödece(0), özece(0), öznece(0), ögece(0), enece(0), övece(0), ökece(0), önele(0), önelce(0), Önençe(0), Önen'e(0), inece(0), dönece(0)

Bu yöntemin başarılı olamadığı bir örneğe bakalım. “Bugün” yerine yanlışlıkla “bugun” yazılmış bir örnekte “bugun” kelimesi için Zemberek'in önerdiği kelimeler ve frekansları aşağıda verilmiştir:

bugun → bulun(0), busun(0), bumun(0), bunun(20), bugün(18), burun(0), buzun(0), buyun(0), buğun(0), budun(0), bulgun(0), burgun(0), bungun(0)

Burada dördüncü sırada önerilen “bunun” kelimesinin frekansı 20 ile en yüksek çıkmış, tercih edilmiş ama başarısız bir seçim olmuştur. Doğru kelime olan “bugün”ün frekansı ise 18 olarak hesaplanmıştır. Bu örnekte Zemberek'in ilk önerisi de, frekansa göre seçim de başarısız olmuştur.

Bu yöntemle %32 olan hata oranı %16'ya düşürülmüştür. Hata oranı yarıya düşürülerek belirgin bir ilerleme kaydedilmiştir. %16 kabul edilebilir bir hata oranı olsa da yine de çok düşük değildir. Fakat şu durumu göz önünde bulundurmak gerekmektedir: Yapılan çalışma özellik tabanlı bir duygu analizi çalışması olduğu için özellikle duygu ifadesi ve ürün özelliği ifade eden sözcükler önem taşımaktadır. Duygu ifadesi ve ürün özelliği ifade eden kelimelerde yazım yanlışlarının doğru olarak giderilmesi çok önemlidir. Duygu ifadesi ve ürün özelliği ile ilgili olmayan kelimelerde yazım yanlışlarının doğru tespit edilebilmesi nispeten daha az önemlidir. En yüksek frekansa sahip kelimeyi seçme yöntemi özellikle

duygu ifadesi ve ürün özelliği ile ilgili kelimelerde oldukça başarılı çalışmaktadır çünkü bu kelimelerin frekansı yorumlarda oldukça yüksektir. Yapılan test sonuçları da bunu doğrulamaktadır. Frekansa göre kelime seçme yöntemi %84'lük bir başarı oranı ortaya koymuş olsa da duygu ifadesi ve ürün özelliği ile ilgili kelimelerde başarı oranı daha yüksektir. Bu yöntemin başarısız olduğu kelimelerin çoğu duygu ve ürün özelliği ile ilgisi olmayan kelimelerdir.

Sonuç olarak yorumlardaki yazım hatalarının giderilmesinde Zemberek'in önerdiği kelimelerden seçim yaparken bu yöntemin kullanılmasına karar verilmiştir.

Ürün ve firma isimlerinin elenmesi için, ürün ve firma isimlerini içeren bir liste manuel olarak bizim tarafımızdan oluşturulmuştur. Bu liste kullanılarak cümlelerde geçen ürün ve firma isimleri tespit edilmiştir.

5.2. Cümle Polarite Tespiti

Duygu polaritesinin tespiti görevi için, Türkçe duygu terimleri sözlüğü oluşturulmuştur. Duygu terimleri sözlüğü oluşturulurken, Türkçe duygu terimlerini derlemenin yanında, bazı duygu terimleri İngilizce duygu terimleri sözlüklerinden alınmıştır. Oluşturulan duygu terimleri sözlüğü sadece tekil kelimeleri içermemektedir, aynı zaman iki veya üç kelimedenden oluşan kalıpları da içermektedir.

Sözlük, metin dosyası formatındadır ve dosyanın her bir satırında bir sözlük girdisi yer almaktadır. Örnek sözlük girdileri aşağıdaki verilmiştir:

```
güzel / sozcuk_turu:S / polarite:+1
gzl / sozcuk_turu:S / polarite:+1 / kural:KURALSIZ
pişman oldum / sozcuk_turu:K / polarite:-1
şaheser / sozcuk_turu:I / polarite:+1
```

Bir sözlük girdisi önce duygu terimi, sonrasında sırayla / karakteri ile ayrılmış alanlardan oluşmaktadır. Bu alanlar aşağıdaki gibidir:

sozcuk_turu: Duygu teriminin sözcük türünün girildiği alanıdır. S (sıfat), F (fiil), I(isim) veya K(kalıp) seçeneklerinden birisi olabilir ve zorunlu bir alandır.

polarite: Duygu teriminin polaritesinin girildiği alandır. +1 pozitif -1 ise negatif polariteyi ifade etmektedir. İleriye yönelik duygu derecelendirmeyi desteklemesi açısından polarite +1, -1 şeklinde girilmektedir. Zorunlu bir alandır.

kural: Ürün yorumlarında, forum sitelerinde kullanımı yaygın olan ve Türkçe dil kurallarına uygun olmayan bir takım duygu ifadeleri bulunmaktadır. Mesela iyi yerine ii, güzel yerine gzl şeklinde kullanımlar yaygındır. Bu sözcükler üzerinde yazım yanlış düzeltme işlemi yapılmadan oldukları gibi işleme konmaları gerekmektedir. Bu alana bu bilgi girilir. KURALLI veya KURALSIZ şeklinde iki seçenek mevcuttur. Zorunlu bir alan değildir, varsayılan değer KURALLI'dır.

Duygu analizinde duygu polaritesinin tespit edilmesi bütün yönleriyle düşünüldüğünde oldukça karmaşık bir problemdir. Kinaye tespiti, dolaylı şekilde ifade edilen duyguların tespit edilmesi gibi zorlukları olan bir problemdir. Bu çalışmada asıl odaklandığımız görev özellik çıkarımıdır, dolayısıyla duygu polarite tespiti görevi temel seviyede yapılmıştır.

Cümlelerin polaritelerinin tespitinde ilk olarak duygu terimleri sözlüğünden yararlanılarak cümlede duygu terimi olup olmadığına bakılmaktadır. Eğer duygu terimi varsa, cümlede herhangi bir olumsuzluk eki/kelimesi olup olmadığına bakılmaktadır. Olumsuzluk eki/kelimesi yoksa cümlelerin polaritesi duygu teriminin polaritesi ile aynıdır. Olumsuzluk eki/kelimesi varsa, cümlelerin polaritesi duygu teriminin polaritesi ile zıttır. Son olarak, cümlelerin duygu polarite tespiti, anlam zıtlayıcı bağlaçlara (ama, fakat) bakılır. İki ifade birbirine anlam zıtlayıcı bağlaç ile bağlıysa ve bu ifadelerden birinin polaritesi tespit edilmiş durumdaysa, diğer ifadenin polaritesi tersi yönde tespit edilir.

5.3. Cümle Parçalama

Literatürdeki birçok özellik çıkarımı yöntemi, cümlelerdeki ürün özelliği ile ilgili kelimeleri tespit etmeye odaklanmaktadır. Bu çalışmada izlenen yol farklıdır. Ürün özelliği ile ilgili kelimeleri tespit etmeye çalışmak yerine, cümle, her biri tek ürün özelliği ile ilgili olacak şekilde parçalara ayrılmaktadır. Mesela, aşağıdaki örnek cümleyi göz önüne alalım:

“Kameranın fotoğraf çekimi mükemmel”

Burada kullanıcı telefonun “kamera” özelliği ile görüş bildirmektedir. Burada “kamera” sözcüğü özellik ifade eden kelimedir. “Fotoğraf”, “çekim” kelimeleri ise özellik ile ilişkili kelimelerdir. Birçok özellik çıkarımı yöntemi için, burada “kamera” kelimesinin tespit edilmesi yeterlidir. Fakat SS-LDA için, özellik kelimesi olan “kamera” sözcüğü yanında, “fotoğraf”, “çekim” gibi “kamera” özelliği ile ilişkili sözcükler de çok önemlidir. Çünkü bu ürün özelliği ile ilişkili kelimelerin yöntemdeki bir sonraki modül olan Gruplama Modülünün başarısı üzerinde çok büyük bir etkisi vardır.

Cümle parçalama görevi için Hu ve Liu’nun önermiş oldukları “sık geçen isimler” yönteminden [37] ve Agrawal, Imielinski ve Swami’nin önermiş oldukları “birliktelik kuralları” yönteminden [84] faydalanılmıştır. Hu ve Liu yaptıkları özellik tabanlı düzeydeki duygu analizi çalışmasında şu tezi ortaya atmışlardır: “Çok sayıda kullanıcı yorumlarından oluşan bir veri setinde sık geçen isim türündeki sözcükler ürün özellikleri ile ilgili sözcüklerdir”. Bu önermelerini şu şekilde açıklamışlardır: Her kullanıcı, yorumunda farklı şeylerden ve farklı hikayelerden bahseder ve bunları ifade ederken farklı kelimeler kullanır. Fakat ürün yorumlarında kullanıcıların birbirleri ile ortak bahis konuları ürün özellikleridir. Ürün özellikleri hakkında konuşurken de, kullanıcıların kullandıkları isim türü kelimeler genellikle birbirleri ile aynı kelimelerdir. Bu yüzden yorumlarda sık geçen isim türü sözcükler çoğunlukla ürün özellikleridir. Bu yöntem literatürde “sık geçen isimler” olarak adlandırılmış ve birçok duygu analizi çalışmasında kullanılmıştır. Bu yöntemi iyileştirmeye, daha başarılı hale getirmeye yönelik olarak da literatürde çalışmalar mevcuttur. “Sık geçen isimler” yöntemi özellik çıkarımı için önerilmiştir ama biz bu yöntemi burada cümle parçalama görevinde kullanıyoruz.

Birliktelik kuralları Agrawal ve diğerleri tarafından varlıklar arasındaki yakınlığı ölçmek için önerilmiş bir yöntemdir [84]. Birliktelik kural madenciliğinin klasik uygulama alanı market sepeti veri analizidir. Market sepeti analizinde, müşterilerin aldıkları ürünler ve bu ürünler arasındaki ilişkiler ortaya çıkartılır. Bu aşağıdaki gibi bir örnekle açıklanacak olursa:

Peynir → Ekmek [support = 10%, confidence = 75%]

Yukarıdaki birliktelik kuralında, tüm müşterilerin %10'u peynir ve ekmeği birlikte almıştır. Buna support değeri denilmektedir. Peynir alan müşterilerin %75'i aynı zamanda ekmek de almıştır. Buna da confidence değeri denilmektedir. Birliktelik kurallarında X ve Y arasındaki confidence değeri Eş. 5.1'e göre hesaplanır.

$$Confidence(X | Y) = \frac{P(X \cup Y)}{P(X)} \quad (5.1)$$

$P(X \cup Y)$: X and Y'nin birlikte var olma sayısı.

$P(X)$: X'in var olma sayısı.

Not: $Confidence(X | Y)$, $Confidence(Y | X)$ 'e eşit olmayabilir.

Market sepeti ile yukarda açıklanan bu kurallar metin madenciliğinde kelimelerin aynı cümle veya aynı dokümanda birlikte yer almalarına göre uygulanmaktadır. Birliktelik kurallarında kelimelerin sırası göz önüne alınmaz, aynı cümle veya doküman içinde yer almaları yeterlidir.

Cümle parçalama yöntemi örnek bir cümle üzerinde gösterilerek anlatılacaktır. Anlatımda kolaylık olması için burada yazım ve gramer kurallarına uyan bir örnek cümle seçilmiştir. Gerçek hayattaki kullanıcı yorumlarında cümleler genellikle daha kirli, gramer, noktalama ve yazın hatalarıyla doludur. Fakat yöntemin çalışma şekli ve prensibi çok değişmemektedir.

Örnek cümle:

Kameranın gece fotoğraf çekimi, android işletim sistemi ve dokunmatik ekran hassasiyeti çok iyi.

“Kamera” özelliği

“işletim sistemi” özelliği

“ekran” özelliği

Bu cümledeki “kamera” özelliği ile ilgili olarak, “kamera” ürün özelliği kelimesi, “fotoğraf” ve “çekim” sözcükleri de özellik ile ilişkili kelimelerdir. “Ekran” özelliği ile ilgili olarak da “ekran” sözcüğü, ürün özelliği kelimesi, “dokunmatik” ve “hassasiyet” kelimeleri de özellik ile ilişkili kelimelerdir.

“Kamera”, “işletim sistemi” ve “ekran” özellikleri ile ilgili kelimeler cümlede sıralanmıştır. Cümlelerin doğru bir şekilde parçalanması için hangi noktalardan parçalara

ayırmamız gerektiğine nasıl karar verilecek? Örnek cümlelerin ilk parçasını “android” kelimesinden önce, ikinci parçasını “dokunmatik” kelimesinden önce bölmemiz gerektiğini nereden bileceğiz?

Yapılması gereken ilk şey ürün özelliği ve özellik ile ilişkili kelimeleri ortaya çıkarmaktır. Hu ve Liu’nun önermesi [37] burada da geçerlidir. Kullanıcı yorumlarından oluşan bir veri setinde ürün özelliği ve özellik ilişkili kelimelerin toplam frekansları yüksektir. Hu ve Liu önerdikleri yöntemde sadece isim türündeki sözcükleri dahil etmişlerdir. Fakat, veri seti üzerinde yapılan inceleme ve analizler, ürün özelliği ve özellik ile ilişkili kelimelerin çoğunlukla isimler (*kamera, ekran*) ve bazen de belli çekim tipine sahip fiiller (*telefon ısınıyor*) olduğunu ortaya koymuştur. Dolayısıyla bu çalışmada isim türü sözcüklere ek olarak belli çekim tipine sahip fiiller de dahil edilmiştir. Bu çekim tipi şimdiki zaman üçüncü tekil şahıs çekimidir (*telefon ısınıyor*). Morfolojik analizde sözcüklerin türleri tespit edilmekte ve sözcükler kök ve eklerine ayrılmaktadır. Fiil çekimleri de eklerin analiz edilmesi ile tespit edilmektedir. Hu ve Liu, “sık geçen isimler” ismini kullanmışlardır çünkü yöntemlerinde sadece isim türü sözcükleri dahil etmişlerdir. Çalışmamızda isim türü dışında belli bir çekime sahip fiilleri de dahil edildiği için “sık geçen sözcükler” ismini kullanılmaktadır.

İçinde kullanıcının pozitif veya negatif fikir ve görüşlerini belirttiği cümlelere “subjektif cümle” denilmektedir. İçinde kullanıcının fikir veya görüş belirtmeyip objektif gerçekleri ifade ettiği cümlelere ise “objektif cümle” denilmektedir. Bunlara birer örnek aşağıda verilmiştir:

Subjektif cümle: “Telefonun dokunmatik hassasiyeti çok iyi”

Objektif cümle : “Uzun araştırmalar sonunda bu telefonu almaya karar verdim”

Ürün özelliği ve özellik ilişkili sözcükler çoğunlukla subjektif cümlelerde bulunmakta, objektif cümlelerde çok daha az bulunmaktadır. Objektif cümleler çoğunlukla ürün özelliği ve özellik ilişkili sözcükler dışında başka sözcükleri içermektedir. Yaptığımız testlerde “sık geçen sözcükleri” tespit etmek için kelime frekanslarını sadece subjektif cümlelerde hesaplamamızın, tüm cümlelerde hesaplamaya göre ürün özelliği ve özellik ile ilişkili kelimeleri keşfetmede daha etkin olduğunu gözlemlenmiştir. Dolayısıyla kelime frekansları sadece subjektif cümlelerde hesaplanmaktadır. Yukardaki örnek cümledeki

“kamera”, “fotoğraf”, “çekim”, “android”, “işletim”, “sistem”, “dokunmatik”, “ekran”, “hassasiyet” sözcükleri bizim veri setimize göre “sık geçen sözcükler”dir ve bunlar sistem tarafından ürün özelliği ve özellik ile ilişkili sözcükler olarak keşfedilmektedir.

Ürün özelliği ve özellik ile ilişkili sözcükler keşfedildikten sonra, birliktelik kuralları kullanılarak cümleyi parçalama işlemine başlanmaktadır. Aynı ürün özelliği ile ilişkili sözcüklerinin ikililer şeklinde birbirleri arasındaki confidence değerleri oldukça yüksektir, çünkü bu sözcükler yorumlarda sık sık birlikte geçmektedirler. Öte yandan, farklı ürün özellikleri ile ilişkili kelimeleri arasındaki confidence değerleri oldukça düşüktür çünkü bu sözcükler ürün yorumlarında birlikte geçme sayıları düşüktür. Mesela “kamera-fotoğraf”, “kamera-çekim”, “ekran-dokunmatik” gibi kelime ikilileri arasındaki anlamsal yakınlık güçlüdür, dolayısıyla bunların arasındaki confidence değerleri yüksektir. Öte yandan, örnek olarak “kamera-android”, “kamera-işletim”, “sistem-dokunmatik”, “sistem-ekran” gibi farklı ürün özelliklerine ait kelime ikilileri arasındaki anlamsal yakınlık zayıftır, dolayısıyla aralarındaki confidence değerleri düşüktür.

Geliştirilen yöntemde confidence değerlerini hesaplanırken, iki kelimenin birlikte geçiyor sayılma kuralı şudur: İki kelime aynı cümlede geçiyorsa ve aralarındaki mesafe (aralarındaki kelime sayısı) belli bir değerin altında ise, bu iki kelime birlikte geçiyor sayılır, değilse sayılmaz. Bizim testlerimizde bu değer 5 olarak alınmıştır. Dolayısıyla iki kelime aynı cümlede geçiyorlarsa ve aralarındaki mesafe 5 veya daha az ise “birlikte geçiyor” sayılmaktadırlar.

Çizelge 5.4. Veri setinden elde edilmiş örnek confidence değerleri

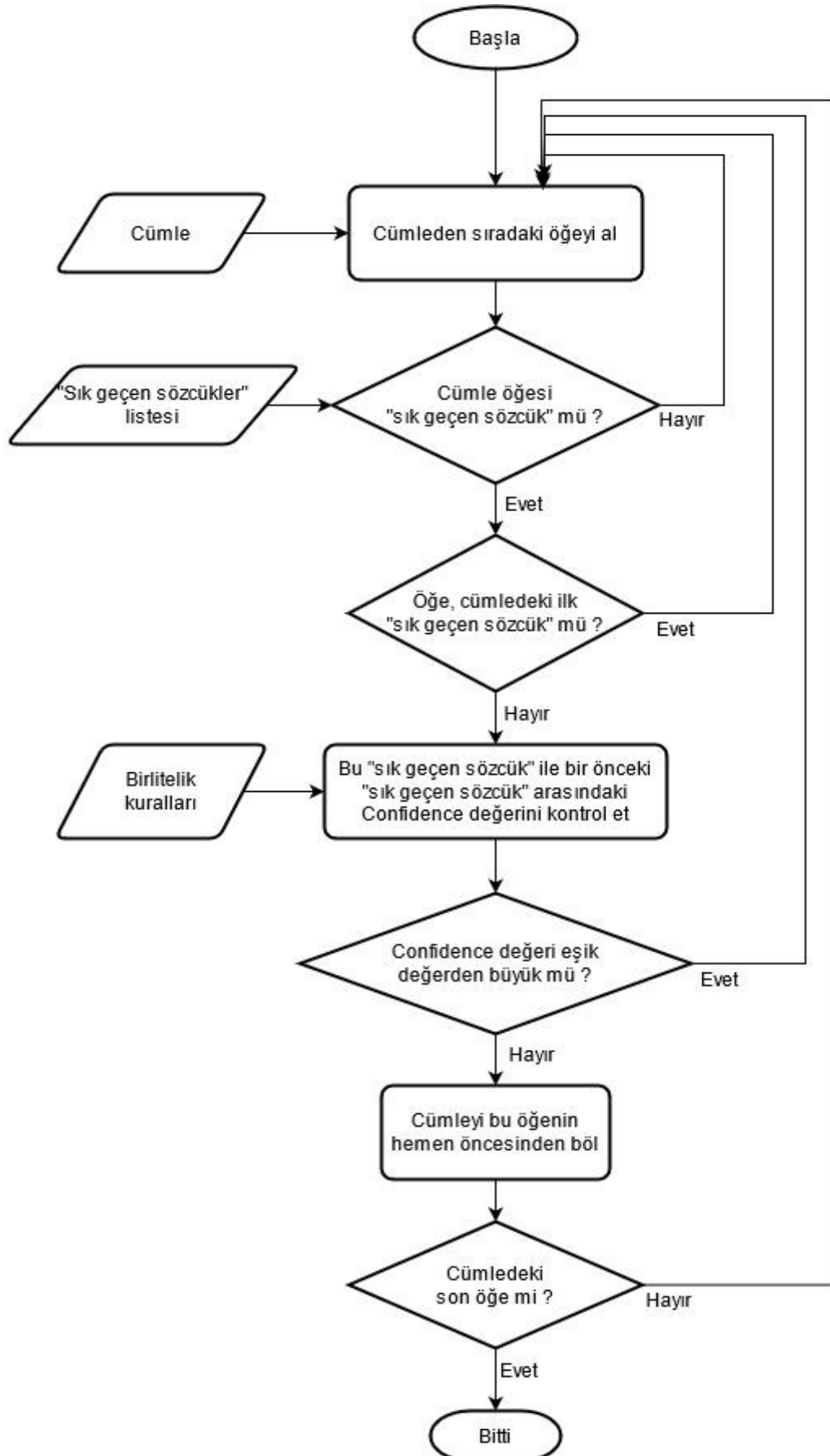
Yüksek confidence değerleri	Düşük confidence değerleri
<i>Confidence (sistem işletim): 0.82609</i>	<i>Confidence (android çekim): 0.00000</i>
<i>Confidence (işletim sistem): 0.95000</i>	<i>Confidence (çekim android): 0.00000</i>
<i>Confidence (fotoğraf çekim): 0.34127</i>	<i>Confidence (android hassasiyet): 0.01794</i>
<i>Confidence (çekim fotoğraf): 0.36364</i>	<i>Confidence (hassasiyet android): 0.01653</i>
<i>Confidence (fotoğraf kamera): 0.36208</i>	<i>Confidence (ekran sistem): 0.00889</i>
<i>Confidence (kamera fotoğraf): 0.29715</i>	<i>Confidence (sistem ekran): 0.02174</i>
<i>Confidence (dokunmatik ekran): 0.26860</i>	<i>Confidence (işlemci kamera): 0.00889</i>
<i>Confidence (ekran dokunmatik): 0.28889</i>	<i>Confidence (kamera işlemci): 0.02174</i>
<i>Confidence (sd kart): 1.00000</i>	<i>Confidence (internet ekran): 0.04592</i>
<i>Confidence (kart sd): 0.23810</i>	<i>Confidence (ekran internet): 0.04000</i>
<i>Confidence (dokunmatik hassasiyet): 0.0578</i>	
<i>Confidence (hassasiyet dokunmatik): 0.7000</i>	

Veri setinden elde edilmiş örnek confidence değerleri Çizelge 5.4’te verilmiştir. Çizelgede görüldüğü gibi aynı ürün özelliği ile ilişkili kelimeler arasındaki confidence değerleri yüksek, farklı ürün özellikleri ile ilgili kelimeler arasındaki confidence değerleri düşüktür.

Ürün özellikleri ile ilişkili kelimeler keşfedildikten ve sözcükler arası confidence değerleri hesaplandıktan sonraki görev cümlelerin parçalanmasıdır. Yöntemimizde kullanıcının herhangi bir görüş belirtmediği objektif cümleler herhangi bir işlem yapılmadan geçilmektedir. Eğer cümle subjektif ise, cümlenin ögeleri üzerinden tek tek geçilir ve “sık geçen sözcük” olup olmadığına bakılır. Eğer “sık geçen sözcük” ise:

- Eğer cümledeki ilk “sık geçen sözcük” ise hiçbir şey yapılmadan bir sonraki cümle ögesine geçilir.
- Eğer cümledeki ilk “sık geçen sözcük” değilse, mevcut “sık geçen sözcük” ile bir önceki “sık geçen sözcük” arasındaki confidence değerine bakılır. Confidence değerine bakarken hem Confidence(bir önceki sözcük | mevcut sözcük), hem de Confidence(mevcut sözcük | bir önceki sözcük) değerine bakılır ve büyük olan alınır.
 - Eğer confidence değeri eşik değerinden büyükse, sözcükler arasındaki anlamsal yakınlık güçlü demektir. Bu, mevcut “sık geçen sözcük” ile bir önceki “sık geçen sözcük”ün aynı ürün özelliği ile ilişkili kelimeler olduğu anlamına gelmektedir. Dolayısıyla bu noktada herhangi bir bölme işlemi yapılmamaktadır.
 - Eğer confidence değeri eşik değerinden küçükse, sözcükler arasındaki anlamsal yakınlık zayıf demektir. Bu, mevcut “sık geçen sözcük” ile bir önceki “sık geçen sözcük”ün farklı ürün özelliği ile ilişkili kelimeler olduğu anlamına gelmektedir. Dolayısıyla cümlede mevcut “sık geçen sözcük”ten hemen önce bölme gerçekleştirilir. Bölme gerçekleştirilirken, isim tamlaması, sıfat-isim, sıfat-fiil, zarf-fiil gibi cümle yapıları göz önüne alınır ve bölme noktası gerekirse bir kelime geri alınır.

Bu algoritmanın akış diyagramı şeklinde gösterimi Şekil 5.5’de, pseudo-code şeklinde resmi gösterimi ise Algoritma 5.1’de yapılmıştır.



Şekil 5.5. Cümle parçalama algoritması akış diyagramı

Algoritma 5.1. Cümle Parçalama algoritması

Input: 1) Sentence 2) Frequent Words List 3) Association Rules 4) Confidence Threshold

Output: Sentence segments

Process:

```

firstFrequentWordPassed := false
previousFrequentWord := null
for each element e in Sentence
    if e is in Frequent Words List then
        if firstFrequentWordPassed then
            if confidence(e, previousFrequentWord) > Confidence Threshold or
                confidence(previousFrequentWord, e) > Confidence Threshold then
                splitSentenceBeforeElement(e);
            endif
        else
            firstFrequentWordPassed := true
        endif
        previousFrequentWord := e
    endif
endfor

```

Bu algoritmanın uygulanışı iki örnek cümle üzerinde gösterilmiştir. Bu örnek cümlelerden birisi kısa ve tek ürün özelliğinden bahsedilen bir cümledir. Diğer örnek cümle ise üç tane ürün özelliğinden bahsedilen nispeten daha uzun bir cümledir.

Örnek cümle 1: *İşlemci hızı harika.*

Cümlelerin öğeleri üzerinden tek tek geçilir ve her öğede yapılan işlemler aşağıdaki gibidir:

İşlemci: “Sık geçen sözcük” ama cümledeki ilk sık geçen sözcük. Herhangi bir işlem yapılmaz

hızı: “Sık geçen sözcük”. Bir önceki sık geçen sözcük olan “İşlemci” ile arasındaki *confidence* değeri yüksek. Bu iki sözcük arasındaki anlamsal yakınlık yüksektir ve dolayısıyla aynı ürün özelliği ile ilişkili sözcüklerdir. Bölme yapılmaz.

harika: Duygu terimi.

Sonuç olarak bu örnek cümlede herhangi bir parçalama işlemi yapılmamıştır. Cümle parçalama algoritması bu cümleyi tek bir parça olarak dönmektedir.

Örnek cümle 2

Kameranın gece fotoğraf çekimi, android işletim sistemi ve dokunmatik ekran hassasiyeti çok iyi.

Cümlelerin öğeleri üzerinden tek tek geçilir ve her öğede yapılan işlemler aşağıdaki gibidir:

Kamera: “Sık geçen sözcük” ama ilk sık geçen sözcük. Herhangi bir işlem yapılmaz.

gece: “Sık geçen sözcük” değil. Herhangi bir işlem yapılmaz.

fotoğraf: “Sık geçen sözcük”. Bir önceki sık geçen sözcük olan “kamera” sözcüğü ile confidence değeri yüksek. Bu iki sözcük arasındaki anlamsal yakınlık yüksektir, dolayısıyla aynı ürün özelliği ile ilişkili sözcüklerdir. Bölme yapılmaz.

çekim: “Sık geçen sözcük”. Bir önceki sık geçen sözcük olan “fotoğraf” sözcüğü ile confidence değeri yüksek. Bu iki sözcük arasındaki anlamsal yakınlık yüksektir, dolayısıyla aynı ürün özelliği ile ilişkili sözcüklerdir. Bölme yapılmaz.

android: “Sık geçen sözcük”. Bir önceki sık geçen sözcük olan “çekim” sözcüğü ile confidence değeri düşük. Bu iki sözcük arasındaki anlamsal yakınlık zayıftır ve dolayısıyla aynı ürün özelliği ile ilişkili sözcükler değildir. Bu sözcükten hemen önce bölme yapılır.

işletim: “Sık geçen sözcük”. Bir önceki sık geçen sözcük olan “android” sözcüğü ile confidence değeri yüksek. Bu iki sözcük arasındaki anlamsal yakınlık yüksektir, dolayısıyla aynı ürün özelliği ile ilişkili sözcüklerdir. Bölme yapılmaz.

sistem: “Sık geçen sözcük”. Bir önceki sık geçen sözcük olan “işletim” sözcüğü ile confidence değeri yüksek. Bu iki sözcük arasındaki anlamsal yakınlık yüksektir, dolayısıyla aynı ürün özelliği ile ilişkili sözcüklerdir. Bölme yapılmaz.

dokunmatik: “Sık geçen sözcük”. Bir önceki sık geçen sözcük olan “sistem” sözcüğü ile confidence değeri düşük. Bu iki sözcük arasındaki anlamsal yakınlık zayıftır ve dolayısıyla aynı ürün özelliği ile ilişkili sözcükler değildir. Bu sözcükten hemen önce bölme yapılır.

ve: etkisiz sözcük

ekran: “Sık geçen sözcük”. Bir önceki sık geçen sözcük olan “dokunmatik” sözcüğü ile confidence değeri yüksek. Bu iki sözcük arasındaki anlamsal yakınlık yüksektir, dolayısıyla aynı ürün özelliği ile ilişkili sözcüklerdir. Bölme yapılmaz.

hassasiyet: “Sık geçen sözcük”. Bir önceki sık geçen sözcük olan “ekran” sözcüğü ile confidence değeri yüksek. Bu iki sözcük arasındaki anlamsal yakınlık yüksektir, dolayısıyla aynı ürün özelliği ile ilişkili sözcüklerdir. Bölme yapılmaz.

çok: etkisiz kelime

iyi: duygu terimi

Sonuç olarak bu örnek cümleden aşağıdaki 3 cümle parçacığı elde edilir (etkisiz kelimeler elenmiş):

- Kameranın gece fotoğraf çekimi,
- android işletim sistemi
- dokunmatik ekran hassasiyeti iyi.

Eğer cümle subjektif ve birden fazla duygu terimi içeriyorsa, öncelikle cümle duygu terimlerinden hemen sonra parçalanır. Bu şekilde, tek duygu terimi içeren alt cümleler elde edilmiş olur. Tek duygu terimi içeren alt cümleler ise yukardaki algoritmaya göre parçalara ayrılır. Eğer cümlede hiçbir “sık geçen sözcük” yoksa o zaman yukardaki algoritmaya göre herhangi bir parçalama yapılmaz ve cümlenin tamamı tek parça olarak dönülür.

Cümleleri parçalara ayırdıktan sonraki adım cümle parçacıklarındaki ürün özellikleri ile ilişkili olmayan kelimelerin elenmesi işlemidir. Kendi veri setimiz ve başka ürün yorumları üzerinde yapılan incelemeler sonucunda ürün özellikleri ile ilişkili sözcüklerin aşağıdaki maddelerdeki tanımlara kapsamında olduğu gözlemlenmiştir.

- İsim türündeki sözcükler
- Birinci tekil şahıs şimdiki zaman çekimine sahip fiiller
- Bazı duygu terimleri (zarif, şık, vs.). Bütün duygu terimleri, ürün özellikleri ile ilişkili değildir. Mesela “iyi”, “kötü” gibi duygu terimleri geneli niteleyen duygu

terimleridir, ürün özellikleri ile ilişkili değildir. Hangi duygu terimleri ürün özellikleri ile ilişkili, hangileri değil bilemediğimiz için bütün duygu terimleri dahil edilmiştir.

Bu maddelerdeki tanımlara uyan sözcükler cümle parçalarında bırakılır ve diğer sözcükler elenir. Buradaki amaç bir sonraki modül olan Gruplama modülüne sadece ürün özellikleri ile ilişkili sözcükleri geçirmektir. Ürün özellikleri ile ilişkili olmayan sözcükleri elemek Gruplama modülünün başarısını yükseltmektedir.

Sıradaki görev “Genel” özellikleri belirlemektir. Bazı cümle parçacıklarında kullanıcı ürün hakkında her hangi bir ürün özelliğini belirtmeden pozitif veya negatif bir görüş belirtebilir. Mesela “beğendim” , “Herkes tavsiye ederim” şeklindeki yorumlar bunlara örnek gösterilebilir. Geliştirilen yöntemde bu tip yorumlar “Genel” isimli özel bir kategori altına alınmaktadır. Bu şekilde herhangi bir ürün özelliğine işaret etmeyen ifadeleri “Genel” adlı özel bir kategori altında toplamak ilk defa Liu [1] tarafından yapılmıştır. Yöntemimizde bir cümle parçacığının “Genel” özellik olup olmadığının tespit edilmesi için “sık geçen sözcük” içerip içermediğine bakılır. Eğer cümle parçacığında herhangi bir “sık geçen sözcük” yoksa o cümle parçacığının “Genel” ile ilgili olduğuna karar verilir. “Genel” olarak tespit edilmiş cümle parçacıkları Gruplama modülüne geçirilmemektedir.

5.4. Gruplama – Sentence Segment LDA (SS-LDA)

Cümleleri tek ürün özelliği içerecek şekilde parçalara ayırdıktan ve ürün özellikleri ile ilişkili olmayan kelimeleri eledikten sonra elimizde her biri tek ürün özelliği hakkında olan ve bir veya birkaç kelimeden oluşan cümle parçacıkları vardır. Bir sonraki adım bu cümle parçacıklarını aynı ürün özellikleri ile ilgili olanlar aynı grubun altında olacak şekilde gruplamaktır.

Gruplama görevi için, LDA’nın en uygun yöntem olduğuna karar verildi. LDA, denetimsiz öğrenmeye dayalı olması, eğitim verisi gerektirmemesi nedeniyle oldukça etkili bir yöntemdir. Fakat konu modelleme yöntemi ve LDA algoritması büyük miktardaki metin kitlelerindeki konuları keşfetmek üzere geliştirilmiş ve büyük miktarda veri üzerinde etkili olan bir yöntemdir. Duygu analizi çalışmalarının çalışma sahası olan sosyal medya

mesajları, ürün yorumları gibi metinler ise oldukça kısa metinlerdir. Kısa dokümanlardan oluşan metin yığınlarında LDA algoritması etkili olamamaktadır.

LDA algoritmasının ürün yorumlarında özellik çıkarımında etkili olamamasının iki önemli nedeni vardır: Birincisi, LDA algoritmasında bir kelimenin bir dokümanda geçme sayısı, LDA'nın arka planında bulunan istatistiksel ve olasılıksal dağılım gereği önem taşımaktadır. Kısa metinlerde bir kelime çoğu zaman sadece bir defa geçmektedir, dolayısıyla o dokümanda LDA için o kelimenin istatistiksel dağılımı bakımından anlamlı bir sonuç elde edilememektedir. İkincisi ürün yorumlarında bulunan konular sadece ürün özellikleri değildir. Bu yüzden LDA algoritması ürün özelliklerinin yanında ürün özellikleri ile ilgisi olmayan konuları da ortaya çıkarmaktadır.

5.4.1. LDA'nın kısa metinlerde uygulanması

Konu modelleme yönteminin ve LDA algoritmasının ürün yorumları, Twitter mesajları gibi kısa metinlerde kullanılacak şekilde uyarlanması son yıllardaki sıcak araştırma alanlarından birisidir. Bu alanda son yıllarda çok sayıda çalışma yapılmış ve yapılmaktadır. Yapılan çalışmalarda farklı yöntemler uygulanmış olsa da belli sayıda çalışma tarafından uygulanmış bazı yaygın yöntemler mevcuttur. Bu yöntemler sırayla aşağıda anlatılmıştır:

1. Yöntem: Bu yöntemde kısa dokümanlardaki metin seyrekliği sorununu aşabilmek için kısa metinler birleştirilerek sahte uzun dokümanlar elde edilir ve LDA algoritması bu sahte uzun metinler üzerinde uygulanır. Hong ve Davison bu yöntemi Twitter mesajları üzerinde uygulamışlardır [85]. Twitter mesajlarını birleştirmeyi hem kullanıcı bazlı, hem de hashtag bazlı olarak gerçekleştirmişlerdir. Mehrotra, Sanner, Buntine ve Xie [86], Melis ve Saveski [87] yaptıkları çalışmalarda Twitter üzerinde benzer şekilde mesajları birleştirerek sahte uzun dokümanlar elde edip, daha sonra da LDA algoritmasını uygulamışlardır. Bu yöntem kullanılarak yapılan çalışmalar arasındaki farklar genellikle kısa metinleri birleştirmede kullanılan yöntemlerdeki farklılıklardır.

2. Yöntem: Bu yöntem Yin ve Wang tarafından Google News mesajlarında ve Twitter mesajları üzerinde [88], Jo ve Oh tarafından ürün yorumları üzerinde [89] uygulanmıştır. Bu yöntemde kısa metnin Dirichlet Dağılımına göre birden fazla konunun olasılıksal dağılımından oluştuğu değil sadece bir konudan oluştuğu varsayılır. Mesela bir tweet'in

sadece bir konu hakkında olduğu, ürün yorumlarında ise yorumdaki her cümlenin ürünün sadece bir özelliği hakkında olduğu varsayılmıştır.

Bu varsayıma göre uygun olarak da LDA modelinin arka planındaki olasılıksal dağılımlar ve bu dağılımların matematiksel modellemeleri güncellenmiştir. Yin ve Wang önerdikleri bu yeni kısa metinler için konu modelleme yöntemini Dirichlet Multinomial Mixture Model [88] olarak isimlendirmişlerdir. Yohan Jo ve Alice Oh ise ürün yorumlarında bir cümlede bir ürün özelliğinden bahsedildiğini varsaydıkları için, önerdikleri yöntemi Sentence-LDA [89] olarak isimlendirmişlerdir.

3. Yöntem: Bu yöntem Titov ve McDonald tarafından restoran yorumlarında [90], Alam ve Lee tarafından otel yorumlarında [91], Xianghua, Guo, Yanyan ve Zhiqiang tarafından Çince sosyal medya mesajlarında [92] uygulanmıştır. LDA algoritması ürün yorumlarında uygulandığı zaman konu olarak ürün özellikleri dışında yorumlardaki ürün özelliği olmayan konuları da ortaya çıkarmaktadır. Bu yöntemde ürün özelliği olan konulara lokal konular, ürün özelliği olmayan konulara da global konular denilmektedir. Global konuları keşfetmek için LDA algoritması tüm yorumları kapsayacak şekilde uygulanmaktadır. Lokal konuları yani ürün özelliklerini keşfetmek için ise yorumlar üzerinde Kayan Pencere yöntemi uygulanmaktadır. Kayan Pencere belli bir sayıdaki cümleyi kapsayacak şekilde tanımlanır ve tüm yorumlar boyunca her seferinde bir cümle ilerleyecek şekilde kaydırılarak ilerlenir. Her kaydırmada pencerenin içinde kalan metin bir doküman olarak kabul edilerek LDA algoritması uygulanır.

5.4.2. Sentence Segment LDA (SS-LDA)

Sentence Segment LDA (SS-LDA) bu tez çalışması kapsamında geliştirilen kısa metinler için özgün bir LDA uyarlaması yöntemidir.

Geleneksel LDA yöntemi dokümanları kelime torbası (BoW) modeli olarak ele alır. Kelime torbası modeli ile dokümandaki cümlelerdeki kelimelerin dizilişi ve cümle öğeleri arasındaki ilişkiler göz önüne alınmaz. Dokümandaki her bir kelimenin konu atanması önceki ve sonraki kelimelerin konu atanmalarından tamamen bağımsızdır. SS-LDA ise dokümanı kelime torbası modeli olarak ele almaz. Kelimelerin konu atamaları önceki ve sonraki kelimelerden bağımsız değildir.

Aslında SS-LDA’da konu atamaları kelimelere değil cümle parçacıklarına yapılır. Bir cümle parçacığına bir konu atandığı zaman, o cümle parçacığındaki bütün kelimeler o konuya atanmış olur.

SS-LDA’da bir dokümanın (yorumun) üretim sürecinin modellenmesi aşağıdaki gibidir:

1. Her bir z konusu (ürün özelliği) için

1.1. O konunun kelime dağılımını oluştur: $\varphi_z \sim \text{Dirichlet}(\beta)$

2. Her bir d dokümanı (yorumu) için:

2.1. d dokümanının konu (ürün özelliği) dağılımını oluştur : $\vartheta_d \sim \text{Dirichlet}(\alpha)$.

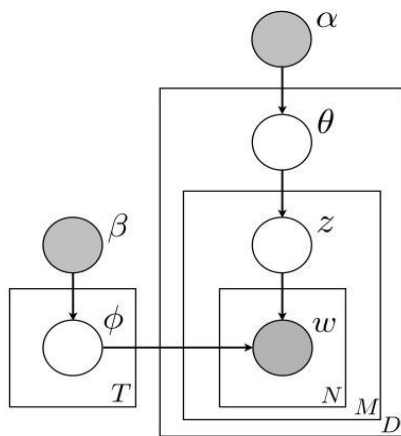
2.2. Her bir cümle parçacığı (Sentence Segment) SS için ,

2.2.1. SS için bir konu (ürün özelliği) z seçilir

2.2.2. SS ’deki her bir kelime için

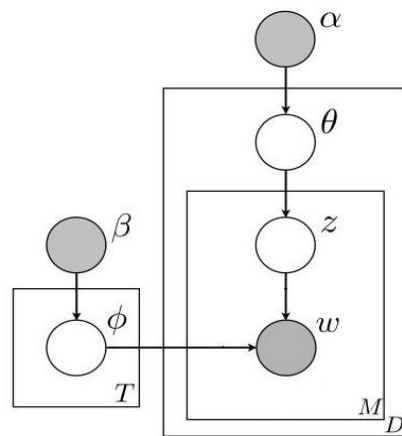
2.2.2.1. z ’nin altındaki kelimelerden birisi seçilerek SS ’e eklenir.

SS-LDA’nın plaka gösterimi Şekil 5.6’da gösterilmiştir. Kıyaslama yapılabilmesi ve aradaki farkın görülmesi için geleneksel LDA’nın da plaka notasyonu yanda gösterilmiştir. Görüldüğü gibi SS-LDA’da LDA’dan farklı olarak cümle parçacığı (Sentence Segment) üretim süreci vardır.



SS-LDA

D : doküman sayısı
 M : D dokümanındaki Sentence Segment sayısı
 N : Sentence Segment’teki kelime sayısı
 T : Konu (ürün özelliği) sayısı



LDA

D : doküman sayısı
 M : D dokümanındaki kelime sayısı
 T : Konu sayısı

Şekil 5.6. SS-LDA ve LDA’nın plaka notasyonları.

SS-LDA'nın Gibbs örnekleme iterasyonlarındaki konu atama formülü de buna göre uyarlanmalıdır. Fakat daha önce geleneksel LDA'da Gibbs örnekleme iterasyonlarındaki konu atama formülünü hatırlayalım. 4.3 nolu başlıkta verildiği gibi geleneksel LDA'nın Gibbs örnekleme iterasyonlarında, D dokümanındaki W kelimesi için yeni bir konu atanırken, Z konusunu atanma ihtimali Eş. 4.5'te verilmişti. Hatırlatma için Eş. 5.2'de tekrar verilmiştir.

$$P(Z|W, D) = \left(\frac{D'de Z'ye ait kelimesi sayısı + \alpha}{D'deki toplam kelime sayısı} \right) \times \frac{Z konusunda W kelimesi sayısı + \beta}{Z'deki toplam kelime sayısı + \beta sum} \quad (5.2)$$

Bu formülün arka planındaki mantık da 4.3 nolu başlıkta anlatılmıştı. Tekrar edilecek olursa W kelimesi için yeni bir konu seçilirken iki kritere bakılır:

- Z konusunun bu dokümandaki ağırlığı ne kadardır (çarpmanın sol tarafı): Eğer bu dokümanda Z konusunun ağırlığı fazla ise mevcut W kelimesinin Z konusuna atanma ihtimali de fazla olur.
- W kelimesi tüm dokümanlarda Z konusuna ne kadar atanmıştır (çarpmanın sağ tarafı): Eğer W kelimesi tüm dokümanlarda Z konusunda sık sık geçiyorsa, o zaman mevcut W kelimesinin de Z konusuna atanma ihtimali yüksektir.

SS-LDA'nın Gibbs örnekleme iterasyonlarında, cümle parçacıklarının (Sentence Segment) üzerinden geçilmektedir (LDA'daki gibi kelimelerin üzerinden değil) ve konu atamaları kelimelere değil bu cümle parçacıklarına yapılmaktadır. Geleneksel LDA'ya ait Eş. 5.2'deki formül ve arka planındaki mantık, SS-LDA'nın doküman üretim süreç modeline göre ve kullanıcı yorumlarının kısa metin olduğu göz önüne alınarak modifiye edilmiştir. Bu modifikasyonlar aşağıdaki gibidir:

1. “Z konusunun (ürün özelliği) bu dokümandaki ağırlığı ne kadardır” modifiye edilmesi: Bu kriter ürün yorumlarında anlamsızdır çünkü sadece birkaç cümleden oluşan bir kullanıcı yorumunda bir ürün özelliği çoğunlukla sadece bir defa geçer. Bunun yerine “Z konusu (ürün özelliği) tüm yorumlardaki ağırlığı ne kadardır” kriterine bakılır. Eğer Z

konusundan (ürün özelliği) diğer tüm kullanıcılar sık sık bahsediyorlarsa, o zaman bu cümle parçacığının da Z konusu (ürün özelliği) hakkında olma ihtimali yüksektir.

2. “W kelimesi tüm dokümanlarda Z konusuna ne kadar atanmıştır” modifiye edilmesi: Burada elimizde sadece bir tane kelime yok, birkaç kelimeden oluşan bir cümle parçacığı var. “W kelimesi tüm dokümanlarda Z konusuna ne kadar atanmıştır” kriteri cümle parçacığındaki her bir kelime için tek tek göz önünde bulundurulur. Dolayısıyla, Eş. 5.2’in sağ tarafındaki ihtimal cümle parçacığındaki her bir kelime için hesaplanır ve bu ihtimaller birbirleri ile çarpılır. Bu şekilde, cümle parçacığındaki her bir kelime hesaplama dâhil edilmiş olur.

Sonuç olarak SS-LDA’nın Gibbs örnekleme iterasyonları için aşağıdaki Eş. 5.3’teki formül elde edilmiş olur.

$$P(Z|SS, D) = \left(\frac{\text{tüm dokümanlarda } Z' \text{ ye atanmış SS sayısı}}{\alpha} + \alpha \right) * \prod_{w=1}^{N_{SSW}} \frac{w \text{ kelimesinin } Z' \text{ ye atanma sayısı} + \beta}{Z' \text{ ye atanmış toplam kelime sayısı} + \beta_{sum}} \quad (5.3)$$

Bu formülün formal gösterimi Eş. 5.4’te verilmiştir ve notasyonlar Çizelge 5.5’te açıklanmıştır.

$$P(Z | SS, D) = (N_z^{SS} + \alpha) * \prod_{w=1}^{SS_{nw}} \left(\frac{N_{zw}^W + \beta}{N_z^W + \beta_{sum}} \right) \quad (5.4)$$

Çizelge 5.5. SS-LDA notasyonları

Notasyon	Açıklama
SS	Cümle parçacığı (Sentence Segment)
N_z^{SS}	Tüm dokümanlarda Z konusuna (ürün özelliği) atanmış cümle parçacığı sayısı
N_{zw}^W	W kelimesinin Z konusuna (ürün özelliği) atanma sayısı
N_z^W	Z konusuna (ürün özelliği) atanmış toplam kelime sayısı
D	Doküman (yorum) sayısı
N_{SSW}	Cümle parçacığındaki (SS) kelime sayısı
α	Yorumun konular (ürün özellikleri) üzerine dağılımı için Dirichlet Parametresi
β	Konuların (ürün özellikleri) kelimeler üzerine dağılımı için Dirichlet Parametresi
β_{sum}	β * Kelime Hazinesi (vocabulary) boyutu

SS-LDA’da konu atamalarının tekil kelimeler yerine cümle parçacıklarına yapılması gruptaki başarıyı belirgin şekilde yükseltmektedir. Geleneksel LDA’da, kelime torbası modeli ile kelimeler arası anlamsal yakınlık ilişkileri ve kelimelerin birlikte yer alma örüntüleri göz önüne alınmamaktadır. Öte yandan, SS-LDA’da kelimelerin birliktelik örüntüleri ve aralarındaki anlamsal yakınlık ilişkileri göz önüne alınmaktadır. Mesela “ekran” ve “dokunmatik” kelimelerine bakalım. “Dokunmatik” kelimesi sık sık “ekran” kelimesi ile aynı cümle parçacığı içinde yer almaktadır ve aynı cümle parçacığı içinde oldukları zaman bu kelimeler aynı konuya atanmaktadırlar. “Dokunmatik” kelimesi “ekran” kelimesi olmadan tek başına geçtiği zaman SS-LDA’nın onu “ekran” ile aynı konuya atama ihtimali çok yüksektir çünkü “dokunmatik” kelimesi “ekran” ile birlikte yer aldığı cümle parçacıklarında hep “ekran” ile hep aynı konuya atanmıştır. Bu, SS-LDA’da kelimelerin birliktelik örüntülerinin göz önüne alınmasının gruptaki başarıyı nasıl arttırdığını göstermektedir.

LDA’da ve dolayısıyla SS-LDA’da önemli parametrelerden birisi de konu (ürün özelliği) sayısıdır. Akıllı telefonlar için aşağıda Çizelge 5.6’da verilen dokuz tane ürün özelliği belirlenmiştir.

Çizelge 5.6. Akıllı telefon özellikleri

1. Ekran	6. Tasarım-Kasa
2. Kamera-Çoklu Ortam	7. Batarya
3. Hız-İnternet Hızı	8. Fiyat
4. İşletim Sistemi-Yazılım	9. Kargo-Sipariş
5. Hafıza-Bellek	

5.5. Özellik birleştirme

Yöntemimizde cümlelerin tek ürün özelliği içerecek şekilde parçalara ayrılması, hatasız gerçekleştirilen bir işlem değildir. Bu işlem esnasında, cümlelerin parçalanmaması gerekirken parçalanması, parçalanması gerekirken parçalanmaması gibi hatalar olabilmektedir. Cümlelerin parçalanmaması gerekirken parçalanması cümlede fazladan özellik çıkarımı yapılması anlamına gelmektedir. Özellik birleştirme işlemi, işte bu şekilde fazladan hatalı olarak gerçekleşmiş olan özellik çıkarımlarını düzeltmeyi hedeflemektedir.

Cümle parçalamadan sonra SS-LDA ile cümle parçacıklarının ürün özellikleri (konu) altında gruplanmaları gerçekleştirilmektedir. Bir cümle içerisinde arka arkaya aynı ürün özelliğine atanmış cümle parçacıkları varsa yani aynı cümle içinde arka arkaya aynı ürün özelliğinden bahsediliyorsa bunlar birleştirilerek tek bir ürün özelliği haline getirilmektedir. Testlerimizde gerçekleşmiş örnek bir özellik birleştirme işlemi aşağıdadır:

Örnek Cümle: *Estetik olması, ince olması da avantaj.*

Bu cümle, cümle parçalama modülü tarafından "*estetik olması*" ve "*ince olması da avantaj*" şeklinde ikiye parçalanmış ve her iki parça da SS-LDA tarafından "*tasarım-kasa*" özelliği altında gruplanmıştır. "*Tasarım*" hakkında pozitif görüş şeklinde iki tane özellik çıkarımı yapılmıştır. Aslında bu cümlenin tamamı "*tasarım*" hakkında tek bir pozitif görüş şeklinde değerlendirilmesi daha doğrudur. Özellik Birleştirme modülü burada aynı cümle içinde arka arkaya gelen aynı ürün özelliği altında gruplanmış olan cümle parçacıklarını birleştirmiş ve "*tasarım*" hakkında bir tane özellik çıkarımı şekline dönüştürmüştür.

6. TESTLER VE DENEYSEL SONUÇLAR

Bu çalışmadaki veri seti akıllı telefonlar hakkında 1292 adet kullanıcı yorumundan oluşmaktadır. Yorumlar www.hepsiburada.com e-ticaret sitesinden alınmıştır. Hepsiburada'dan alınan her bir yorum, tek tek tarafımızca Çizelge 5.6'daki ürün özellikleri ve bu özelliğe göre duygu polariteleri şeklinde etiketlenmiştir. Etiketlenmiş yorumlar XML formatında bir dosyada saklanmaktadır. Örnek bir yorum ve bu yorumun örnek bir etiketlenmesi Şekil 6.1'de gösterilmiştir.

```
<review id="74">
  <text>Ekran boyutu ve renk kalitesi olarak piyasada ki eş değer telefonlardan daha kaliteli
  diyebilirim. bir çok özelliği mevcut ancak bunların bir çoğu günlük yaşantımızda kullanmayacağımız
  özellikler. bu özelliklerin dezavantajı şarjı çabuk bitirmesi. bunun dışında telefon çok güzel.
  </text>
  <aspects>
    <aspect category="ekran" polarity="positive"/>
  </aspects>
  <aspects>
    <aspect category="şarj" polarity="negative"/>
  </aspects>
  <aspects>
    <aspect category="genel" polarity="positive"/>
  </aspects>
</review>
```

Şekil 6.1. Veri setinden etiketlenmiş örnek bir yorum

Veri setimizdeki yorumlarda “genel” özellik dahil etiketlenen ürün özelliklerinin sayıları Çizelge 6.1’de verilmiştir.

Çizelge 6.1. Etiketlenen ürün özellikleri sayıları

Ürün özelliği	Etiketlenme Sayısı
Kargo-Sipariş	122
Ekran	304
Kamera-Çoklu Ortam	323
Hız-İnternet Hızı	448
İşletim Sistemi	156
Hafıza-Bellek	89
Tasarım-Kasa	428
Batarya	182
Fiyat	493
“Genel”	1469

Ayrıca kendi veri setimiz dışında, literatür kıyaslaması için yöntemimiz SemEval-2016 [93] Türkçe restoran yorumları üzerinde de test edilmiştir.

Rakamsal başarı ölçüm sonuçlarına geçmeden önce SS-LDA'nın yukardaki 9 adet ürün özelliği altında hangi sözcükleri grupladığı Çizelge 6.2'de verilmiştir. Çizelge 6.2'deki veriler incelendiğinde SS-LDA'nın yorumlardaki kelimeleri, ürün özellikleri altında grupta oldukça başarılı olduğu gözlemlenmektedir. Aynı ürün özelliği ile ilgili farklı kelimelerin doğru bir şekilde gruplandıkları görülmektedir. Mesela “ekran”, “dokunmatik”, “hassas”, “çözünürlük” kelimeleri başarılı bir şekilde “Ekran” özelliği altında gruplanmıştır. “Kamera”, “kalite”, “fotoğraf”, “ses”, “çekim”, “video” gibi kelimeler de “Kamera-Çoklu Ortam” özelliği altında gruplanmışlardır. Diğer özelliklerin altında gruplanan kelimeler de incelendiğinde de grupta son derece başarılı olduğu görülmektedir.

Çizelge 6.2. Kelimelerin ürün özellikleri altında gruplanması.

Özellik	Çıkarılan Kelimeler
Ekran	Ekran, dokunmatik, hassas, çözünürlük, boyut, hassasiyet, parlaklık, algılama büyük, hızlı
Kamera-Çoklu Ortam	Kamera, kalite, görüntü, fotoğraf, ses, çekim, güzel, video, resim, harika, çek, mp, flaş, makine, net, yeterli, ışık, zoom, hd
Hız-İnternet Hızı	Hızlı, hız, İnternet, işlemci, uygulama, wifi, bağlantı, flash, çalışıyor, program, menü, video.
İşletim Sistemi	Android, sistem, işletim, uygulama, market, kullanışlı, güncelleme, ad, ücret, yük, program.
Hafıza-Bellek	Hafıza, ram, oyun, dahili, kart, düşük, mb, gb, yetersiz, uygulama, program, sd
Tasarım-Kasa	Şık, kullanım, tasarım, hafif, zarif, renk, menü, görünüm, kibar, görünüş, rahat, dur, kullanma, görsel, harika, estetik, kullanım, hoş, yorum, beyaz, kılıf, duruş, tür, kasa, pratik, dizayn, boyut.
Batarya	Şarj, gün, git, süre, batarya, pil, gitme, kullanım, bit, ömür, ısınma, durum, kullan, dayan.
Fiyat	Fiyat, uygun, performans, para, piyasa, pahalı, ucuz, android, makul, cazip, kullanışlı, marka, garanti, ideal, bütçe, hesaplı.
Kargo-Sipariş	hızlı, gün, kargo, site, sipariş, hediye, ulaş, gel, gönderi, geç, teslimat, süre, çık, yıl, teslim, güvenilir, kullanışlı, alış, hizmet

Başarı ölçümlerinde üç başarı düzeyi kriterine göre ölçümler yapılmıştır. Bunlar Precision, Recall ve F-Score başarı ölçüm kriterleridir. Precision, Recall ve F-Score ölçüm metriklerinin formüllerini vermeden önce True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) kavramlarının ne demek olduğu önce resmi formüllerine göre, daha sonra da sözlü olarak açıklanacaktır.

Elimizde belli sayıda herhangi bir varlık olsun. Bunlar Positive/Negative şeklinde iki sınıftan birine ait olsunlar. Burada ikili sınıflandırma Positive/Negative olarak söylene de herhangi bir ikili sınıflandırma olabilir. Geliştirilen bir sistemin bu varlıkları Positive/Negative şeklinde sınıflandırmaya çalıştığını varsayalım. Sistemin yaptığı bu sınıflandırmada:

Doğru Pozitif (TP): Sınıfı Positive olan ve sistem tarafından da doğru şekilde Positive olarak sınıflandırılanları ifade eder.

Doğru Negatif (TN): Sınıfı Negative olan ve sistem tarafından da doğru şekilde Negative olarak sınıflandırılanları ifade eder.

Yanlış Pozitif (FP): Sınıfı Negative olan ancak sistem tarafından yanlış şekilde Positive olarak sınıflandırılanları ifade eder.

Yanlış Negatif (FN): Sınıfı Positive olan ancak sistem tarafından yanlış şekilde Negative olarak sınıflandırılanları ifade eder.

Bu tanımlar ışığında precision ve recall başarı ölçüm metrikleri Eş. 6.1 ve Eş. 6.2’de verilmiştir.

$$Precision = \frac{TP}{TP+FP} \quad (6.1)$$

$$Recall = \frac{TP}{TP+FN} \quad (6.2)$$

Precision, sistemin çıkarım yaptığı değerlerin doğruluk oranıdır. Precision’da sistemin yaptığı çıkarımların ne kadarı doğru ona bakılır. Sistemin bulamadıkları ve çıkarım

yapamadıklarının sayısına bakılmaz. Recall ise, sistemin çıkarması gereken verilerin ne kadarını çıkarabildiğinin oranıdır. Recall’da sistem çıkarması gerekenleri çıkarırken doğruların yanında ne kadar yanlış çıkardığına bakılmaz, ona bakan kriter precision’dır.

Bir örnek ile precision ve recall’u açıklayalım. Elimizde 20 tane varlık olsun ve bunların içinde 10 tanesi Positive olsun. Bir sistem bu 20 varlığın içindeki Positive’leri bulmaya çalışsın.

Sistemin 3 tane seçim yaptığını, 3 tane varlığı Positive olarak tespit ettiğini ve bu 3 varlığın da gerçekten Positive olduğunu varsayalım. Bu durumda:

- Precision çok yüksektir hatta %100’dür. Sistem 3 tane tespit yapmıştır ve 3 tespit de doğrudur. Sadece precision gözüyle bakarsak sistem çok çok başarılıdır.
- Recall çok düşüktür. Varlıkların içindeki 10 tane Positive’den sadece 3 tanesi bulunabilmiş, diğer 7’si kalmıştır. Sadece recall olarak bakarsak çok başarısızdır.

Bir de sistemin 16 tane seçim yaptığını, bu 16 varlığı Positive olarak sınıflandırdığını ama bunlardan 9 tanesinin gerçekten Positive olduğunu varsayalım. Bu durumda:

- Precision oldukça düşüktür, %56’dır. Sistem 16 varlığa Positive demiştir ama bunların 9 tanesi gerçekten Positive’dir. Sadece precision gözüyle bakarsak sistem oldukça başarısızdır.
- Recall ise çok yüksektir, %90’dır. Varlıkların içindeki 10 tane Positive’den 9 tanesi sistem tarafından bulunabilmiştir. Sadece recall gözüyle bakarsak çok başarılıdır.

Bu yüzden tek başına precision veya tek başına recall başarı ölçümü için yeterli değildir. Her ikisinin de ölçülmesi gereklidir. Burada F-Score devreye girmektedir. F-Score, hem precision ve hem de recall’u göz önüne alan ölçüm metriğidir ve Eş. 6.3’te verilmiştir:

$$F - Score = 2 \times \frac{Precision * Recall}{Precision + Recall} \quad (6.3)$$

Veri setimizdeki ürün yorumlarının kendimiz tarafından etiketlendiği yukarda belirtilmişti. 1292 adet akıllı telefon yorumunda yukarda belirtilen dokuz ürün özelliği ve “Genel”

özelliğine göre toplan 4014 ürün özelliği etiketlenmiştir. Çizelge 6.3’de yöntemimizin yorumlarda özellik çıkarımı konusundaki başarı ölçüm sonuçları verilmiştir.

Çizelge 6.3. Önerilen yöntemin ürün özelliklerini çıkarmadaki başarı ölçüm sonuçları

Ölçüm kistası	Değer
Yorum sayısı	1292
Etiketlenmiş ürün özelliği sayısı	4014
SS-LDA tarafından çıkarılmış ürün özelliği sayısı	4116
SS-LDA tarafından doğru şekilde çıkarılmış ürün özelliği sayısı	3349
Precision	%81.36
Recall	%83.43
F-Score	%82.39

LDA ve dolayısıyla SS-LDA metinlerdeki kelimelerin istatistiksel dağılımına dayalı yöntemlerdir. Dolayısıyla veri setinin toplam boyutu bu yöntemlerin başarısındaki önemli faktörlerden birisidir. Bu yüzden başarı ölçümünde yöntemin sadece tüm veri seti üzerinde başarı seviyesi ölçülmemiş, aynı zamanda farklı içeriğe sahip ve farklı sayıdaki ürün yorumları üzerinde de başarı ölçümleri yapılmış, yöntemin değişen veri seti boyutlarına göre başarısı test edilmiştir. Tüm veri setinden farklı boyutlarda ve farklı içeriğe sahip test veri setleri elde etmek için K-Fold Cross Validation yöntemine benzer bir yöntem uygulanmıştır.

K-Fold Cross Validation, eğitim verisi gerektiren denetimli makine öğrenmesine dayalı yöntemlerde veri setini, eğitim verisi – test verisi şeklinde bölmek için kullanılan bir yöntemdir. Mesela elimizdeki veri setinden 5-Fold Cross Validation yöntemine göre farklı içeriklere sahip eğitim verisi – test verisi elde etmek için tüm veri seti 5’e bölünür ve her seferinde bu 5 parçadan birisi test verisi, kalan 4 parça ise eğitim verisi olarak belirlenir ve buna göre testler yapılır. Bu şekilde farklı içeriğe sahip eğitim verileri ve test verileri ile sistem test edilmiş olur. Bu yöntemin grafiksel gösterimi Şekil 6.2.’deki gibidir.

Test	Eğitim	Eğitim	Eğitim	Eğitim	Test 1
Eğitim	Test	Eğitim	Eğitim	Eğitim	Test 2
Eğitim	Eğitim	Test	Eğitim	Eğitim	Test 3
Eğitim	Eğitim	Eğitim	Test	Eğitim	Test 4
Eğitim	Eğitim	Eğitim	Eğitim	Test	Test 5

Şekil 6.2. K-Fold Cross Validation

SS-LDA denetimli makine öğrenmesine dayalı bir yöntem değildir, dolayısıyla eğitim verisine ihtiyaç yoktur. Bizim yapmak istediğimiz farklı içeriklere sahip veri setleri elde etmektir. Bunu yapmak için veri setimiz K-Fold Cross Validation yönteminde olduğu gibi 4 eşit parçaya bölünmektedir. Her bir parça 323 adet yorumdan oluşmaktadır. Her seferinde 1 parçayı dışarda tutularak, $1292 - 323 = 969$ yorumdan oluşan farklı içeriklere sahip 4 farklı veri setleri elde edilmektedir. Bu süreç Şekil 6.3’de grafiksel olarak gösterilmiştir.

Hariç	Dahil	Dahil	Dahil	Test 1
Dahil	Hariç	Dahil	Dahil	Test 2
Dahil	Dahil	Hariç	Dahil	Test 3
Dahil	Dahil	Dahil	Hariç	Test 4

Şekil 6.3. Test veri setleri elde etme yöntemi

Özellik çıkarım yöntemimizde ürün özelliklerini gruplamak için kullandığımız SS-LDA yöntemi konu modellemeye dayalı olduğu için az miktardaki metinler üzerinde etkili olmayan bir yöntem. Bundan ötürü yöntemimizin küçülen veri seti ile başarısının nasıl etkilendiğini gözlemlemek için 500 ve 300 yorumdan oluşan veri setleri üzerinde de testler

yapılarak başarı ölçümleri yapıldı. Ayrıca SS-LDA yöntemi rastgelelik içeren stokastik bir yöntemdir. O yüzden her bir çalıştırma aynı sonuçlar elde edilemeyebilir. Bundan ötürü başarı ölçümü için yapılan her bir test defalarca tekrarlanmış ve en iyi – en kötü başarı düzeyleri ölçülmüştür.

Buna göre tüm veri seti olan 1292 yorum üzerinde, K-Fold Cross Validation yöntemine benzer bir yöntem ile elde edilmiş farklı içeriklere sahip 969 yorumdan oluşan farklı veri setleri üzerinde, 500 yorum üzerinde ve 300 yorum üzerinde elde edilen en iyi ve en kötü başarı düzeyleri Çizelge 6.4’de verilmiştir. Sonuçlar incelendiğinde SS-LDA’nın tüm veri seti olan 1292 adet yorumdaki başarı seviyesi ile farklı içeriklere sahip 969 adet yorumdaki başarı seviyesi hemen hemen aynıdır. Yani veri setini 1292 yorumdan 969 yoruma düşürmek başarı seviyesinde belirgin bir düşmeye neden olmamıştır. Veri seti 500 yoruma düşürüldüğü zaman ise başarı seviyesinde belirgin düşme gerçekleşmiştir. Veri setini 300 yoruma düşürüp yapılan ölçümlerde başarı seviyesi daha da düşmüştür. Sonuç olarak veri setinin büyüklüğü SS-LDA’nın başarısında önemli bir etkindir ama belli bir noktaya etkilidir. Belli bir büyüklükten sonra veri setinin daha da büyük olması yöntemin başarısı üzerinde önemli bir iyileşme sağlamamaktadır.

Çizelge 6.4. Önerilen yöntemin özellik çıkarımında başarı ölçüm sonuçları

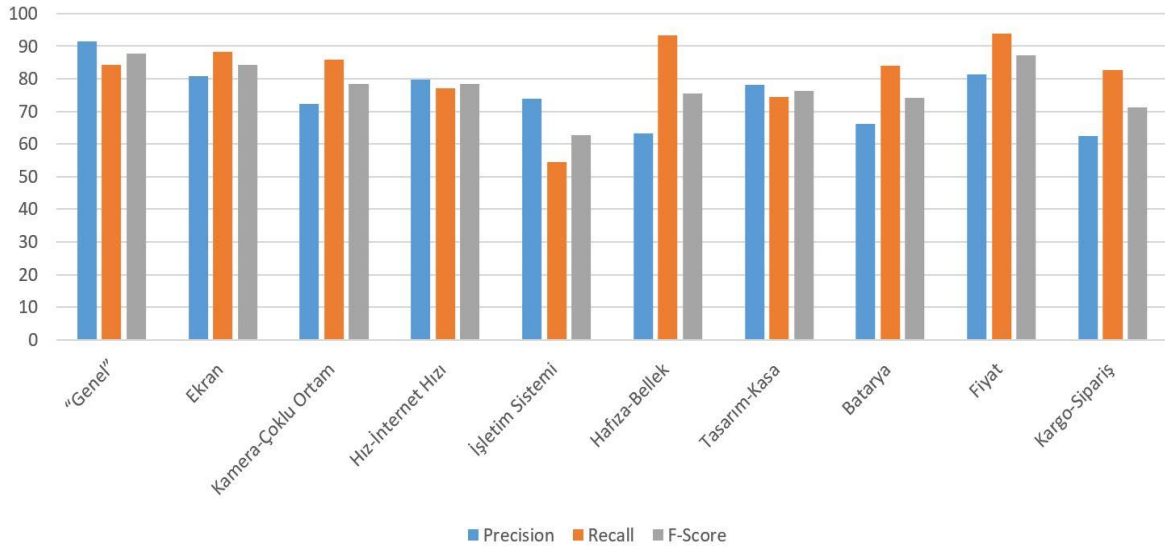
Veri Seti		En iyi sonuç			En kötü sonuç		
		<i>Precision</i>	<i>Recall</i>	<i>F-Score</i>	<i>Precision</i>	<i>Recall</i>	<i>F-Score</i>
Tüm Yorumlar		%81,36	%83,43	%82,39	%77,1	%81,4	%79,19
969 Yorum	1. 323 dışında	%80,3	%84,9	%82,54	%75,7	%80,6	%78,07
	2. 323 dışında	%79,2	%83,5	%81,29	%76,1	%80,3	%78,14
	3. 323 dışında	%81,7	%84,6	%83,12	%76,4	%80,5	%78,40
	4. 323 dışında	%80,8	%84,8	%82,75	%75,9	%79,7	%77,75
500 Yorum		%77,4	%80,3	%78,82	%72,6	%75,5	%74,02
300 Yorum		%74,1	%76,9	%75,47	%69,2	%72,3	%70,72

SS-LDA’nın ayrı ayrı her bir ürün özelliği için de başarı ölçümü yapılmıştır. “Genel” özellik dahil her bir özellik için yapılan başarı ölçümleri Çizelge 6.5’te ve Şekil 6.4’de verilmiştir. Her ürün özelliğinde elde edilen başarı seviyeleri aynı değildir. Bu, o özelliğin bahsedildiği cümlelerdeki ifadelerin linguistik karmaşıklığı, kelimelerin birliktelik örüntüleri ve yorumlarda o özelliğin toplam frekansı ile bağlantılıdır. Mesela “fiyat”

özelliğinde elde edilen başarı seviyesi oldukça iyidir çünkü “fiyat” özelliğinden bahsedilirken dolaylı ve karmaşık anlatımların kullanımı oldukça azdır, çoğunlukla direk olarak “fiyat” kelimesi kullanılmaktadır. Ayrıca veri setimizde fiyat özelliği kullanıcıların en çok bahsettikleri ürün özelliğidir. Fiyat kelimesi ise en yüksek frekansa sahip kelimedir. Yorumlarda fiyat özelliğinden çok bahsedilmiş olması bu özellik ile ilgili daha sağlıklı istatistiksel dağılımlar ve daha sağlıklı birliktelik örüntüleri elde edilmesini sağlamaktadır.

Çizelge 6.5. Önerilen yöntemin her bir ürün özelliği için başarı ölçümleri

Ürün özelliği	Precision	Recall	F-Score
“Genel”	%91,49	%84,15	%87,67
Ekran	%80,72	%88,16	%84,28
Kamera-Çoklu Ortam	%72,32	%85,76	%78,47
Hız-İnternet Hızı	%79,68	%77,18	%78,41
İşletim Sistemi	%73,91	%54,48	%62,73
Hafıza-Bellek	%63,35	%93,25	%75,45
Tasarım-Kasa	%78,19	%74,53	%76,31
Batarya	%66,23	%84,07	%74,09
Fiyat	%81,37	%93,91	%87,19
Kargo-Sipariş	%62,35	%82,79	%71,13



Şekil 6.4. Önerilen yöntemin her bir ürün özelliği için başarı ölçümleri

Öte yandan hafıza-bellek özelliğinde precision ve dolayısıyla F-Score nispeten düşük çıkmıştır. Bunun iki nedeni vardır: Yorumlarda kullanıcılar hafıza-bellek hakkında konuşurken genellikle direk olarak “hafıza”, “bellek” gibi kelimeleri kullanmayıp “mb”, “gb”, “sd”, “kart” gibi kelimeler kullanarak dolaylı anlatım yapmaktadırlar. İkincisi

“hafıza-bellik” özelliğinin tüm yorumlarda ağırlığı oldukça azdır, çoğu yorumda bu özellikten hiç bahsedilmemektedir. SS-LDA istatistiksel dağılımlara bağlı çalıştığı için bu ürün özellikte başarı oranı nispeten daha düşük gerçekleşmiştir.

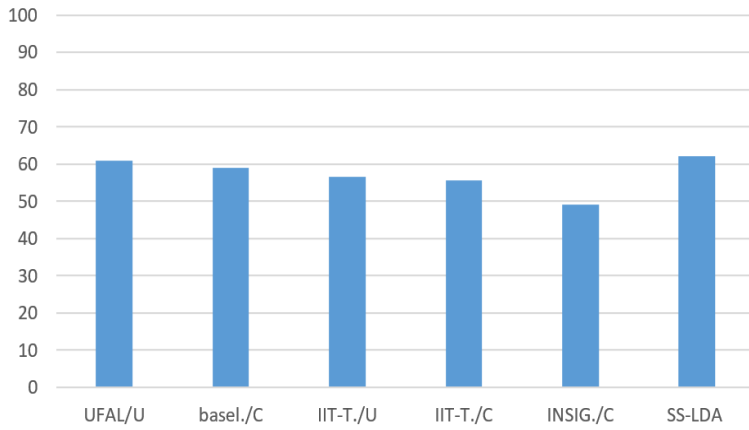
Özellik tabanlı düzeyde duygu analizinde duygu polaritesinin tespitindeki başarı oranının ölçümü biraz karışık bir konudur. Yorumlar özellik#polarite ikilileri şeklinde etiketlenmiştir. Polaritenin tespit edilebilmesi için öncelikle ilişkili olduğu ürün özelliğinin çıkarılması gerekmektedir. Dolayısıyla ürün özelliği doğru şekilde çıkarılamadığı zaman, o özellik ile ilgili polarite de tespit edilememektedir. Bundan ötürü özellik tabanlı duygu analizinde duygu polarite tespitinin precision, recall ve F-Score'un nasıl ölçüleceği konusu belirsizdir. Bu yüzden polarite tespiti başarı ölçümü, çıkarılmış uzun özelliklerindeki polarite tespitinin başarıları ölçülerek yapılmıştır. Bu şekilde ürün özellikleri üzerinde yapılan polarite tespiti başarı ölçüm sonucunda %96,78 oranında polaritelerin doğru tespit edildiği görülmüştür. Fakat çıkarılamamış olan özellik#polarite ikililerindeki polariteler tespit edilememiş olarak durmaktadır. O yüzden "duygu polarite tespiti başarı seviyesi %96,78" demek gerçeği tam olarak yansıtmamaktadır.

Bu çalışmada literatür kıyaslaması iki farklı yolla yapılmıştır. Birincisi SemEval-2016 [93] duygu analizi çalıştayındaki Türkçe dilindeki çalışmalarla kıyaslanmıştır. SemEval-2016 çalıştayında duygu analizi ile ilgili 3 tane görev vardır ve Task-5 olarak isimlendirilen görev yorumlardan ürün özellik çıkarımıdır. Farklı dillerde veri setleri mevcuttur ve bu diller ile ilgili çalışmalar kendi aralarında birbirleri ile kıyaslanmaktadır. SemEval-2016 çalıştayında Türkçe’de restoran yorumları ve telekomünikasyon yorumları şeklinde iki tane veri seti mevcuttur. Türkçe restoran yorumları ile ilgili SemEval-2016’da dört tane çalışma mevcuttur. Bunların haricinde de SemEval-2016 tarafından sağlanan varsayılan (baseline) bir uygulama mevcuttur. Çalışmalar F-Score değerine göre kıyaslanmaktadır.

SS-LDA SemEval-2016 Türkçe restoran yorumları veri seti üzerinde çalıştırılmış ve başarı ölçümü yapılmıştır. Precision olarak %65.84, recall olarak %59.04 ve F-Score olarak %62.25 değerleri elde edilmiştir. Aynı veri seti üzerindeki dört çalışma, varsayılan uygulama ve SS-LDA’nın sonuçları Çizelge 6.6. ve Şekil 6.5.’te gösterilmiştir. SS-LDA’nın %62.25’lik F-Score sonucu diğer çalışmalardaki en yüksek skor olan %61.029’dan biraz daha iyidir.

Çizelge 6.6. SS-LDA ve SemEval-2016 çalışmaları başarı ölçümü

Metot/Submission	F-Score
UFAL/U	%61,029
basel./C	%58,896
IIT-T./U	%56,627
IIT-T./C	%55,728
INSIG./C	%49,123
SS-LDA	%62,250



Şekil 6.5. SS-LDA ve SemEval-2016 çalışmaları başarı ölçümü

Sonuçları kıyaslarken şu gerçeğin göz önüne alınması gerekmektedir: SemEval-2016’da eğitim veri setleri ve test veri setleri sunulmaktadır. Eğitim verisi sağlandığı için gönderilen çalışmaların tamamında denetimli öğrenmeye dayalı yöntemler kullanılmaktadır. Fakat, SS-LDA’da eğitim verisi gerektirmeyen denetimsiz öğrenmeye dayalı bir yöntem kullanılmaktadır. Denetimli öğrenmeye dayalı yöntemler eğitim verisinden faydalanmak gibi çok önemli bir avantaja sahiptirler. Kennedy ve Inkpen [70], Hailong, Wenyan ve Bo [71] yaptıkları çalışmalarda denetimli makine öğrenmesine dayalı yöntemlerin başarı oranının, sözcük tabanlı yöntemlere göre daha yüksek olduğunu ortaya koymuşlardır. Öte yandan denetimli öğrenmeye dayalı yöntemlerin eğitim verisi gerektirmeleri gibi çok önemli bir dezavantajları bulunmaktadır çünkü çalışma yapılan her alan ile ilgili etiketlenmiş eğitim verisi bulmak her zaman mümkün değildir. Bu gerçekler göz önüne alındığında SS-LDA’nın SemEval-2016 Türkçe veri setinde ortaya koyduğu başarının oldukça iyi olduğu söylenebilir.

İkinci olarak, SS-LDA geleneksel LDA ve literatürdeki iki tane kısa metinler için LDA adaptasyonu çalışması ile kıyaslanmıştır. Bunlar Jo ve Oh tarafından önerilen Sentence-

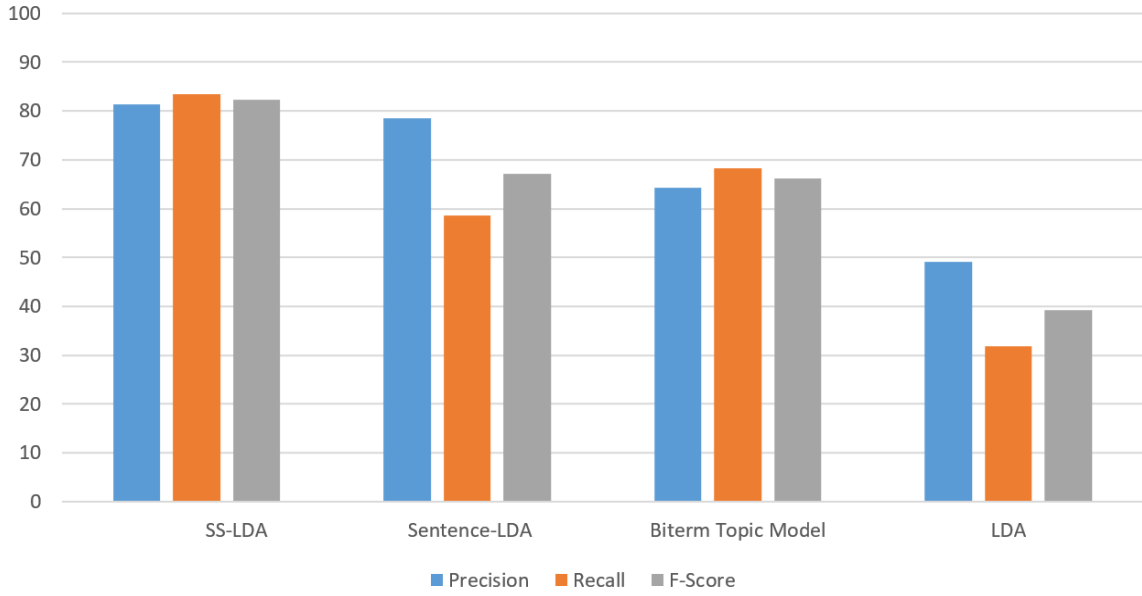
LDA [89] ve Yan, Guo, Lan ve Cheng tarafından önerilen Biterm Topic Model [94] çalışmalarıdır. Biterm Topic Model yönteminin kaynak kodu makalenin yazarı tarafından github üzerinde paylaşılmış durumdadır. Github'dan kaynak kodu alındı ve uygulamanın nasıl bir girdi dosyası istediğini incelenip kendi veri setimiz o girdi formatına dönüştürülmüştür. Daha sonra Biterm Topic Model uygulaması kendi veri setimiz üzerinde başarıyla çalıştırılmış ve başarı ölçümleri yapılmıştır.

Sentence-LDA ise bizim yöntemimiz olan SS-LDA'ya oldukça benzemektedir. Sentence-LDA'da "Subjektif her cümle sadece bir ürün özelliği hakkındadır" şeklinde bir varsayım yapılmış ve LDA'nın kısa metinlere adaptasyonu bu varsayıma göre yapılmıştır. Bu yöntem bizim yöntemimiz olan SS-LDA'ya benzemektedir. O yüzden SS-LDA'nın kodu üzerinde gerekli değişiklikler tarafımızdan yapılarak Sentence-LDA yöntemi gerçekleştirilmiş, veri setimiz üzerinde başarı ile çalıştırılmış ve başarı ölçümleri yapılmıştır.

Sonuç olarak SS-LDA, Sentence-LDA, Biterm Topic Model ve geleneksel LDA kendi veri setimiz üzerinde çalıştırılmış ve başarı ölçümleri yapılmıştır. Sonuçlar Çizelge 6.7 ve Şekil 6.6'da gösterilmiştir.

Çizelge 6.7. SS-LDA, Sentence-LDA, Biterm Topic Model ve LDA başarı ölçüm sonuçları

Metrik	SS-LDA	Sentence-LDA	Biterm Topic Model	LDA
Çıkarılan özellik sayısı	4116	2353	4266	2603
Doğru çıkarılmış özellik sayısı	3349	1846	2740	1279
Precision	%81,36	%78,48	%64,23	%49,16
Recall	%83,43	%58,62	%68,26	%31,88
F-Score	%82,39	%67,11	%66,18	%39,16



Şekil 6.6. SS-LDA, Sentence-LDA, Biterm Topic Model ve LDA başarı ölçüm sonuçları

Sentence-LDA'da subjektif cümleler tek bir ürün özelliği ile ilgilidir şeklinde bir varsayıma dayanmaktadır. Bu varsayım precision'u olumsuz etkilemese de recall'un oldukça düşük çıkmasına neden olmuştur. Mesela dört özellikten bahsedilen bir cümleyi ele alalım. Sentence-LDA cümlelerin tek özellik hakkında olduğunu varsayarak bu dört özellikten üçünü kaçırmaktadır. Üç özelliğin çıkarılamamış olması recall'un düşük çıkmasına neden olmaktadır. O cümleden çıkardığı bir tane ürün özelliğini ise çoğunlukla doğru çıkardığı için precision olumsuz etkilenmemektedir.

Biterm Topic Model'de kelimeler bigram modeli gibi ikililer şeklinde ele alınmaktadır. Bu ürün özellikleri ile ilgili cümlelerdeki kelimelerin birliktelik örüntülerinin kısmen hesaplamaya dahil edilmesini sağlamıştır ama bu birliktelik örüntüleri hesaba katılırken kelimelerin farklı ürün özelliklerine ait kelimeler olması durumu göz önüne alınmamaktadır. Bu durum precision'ın düşük çıkmasına neden olmaktadır. SS-LDA'da kelimelerin birliktelik örüntüleri göz önüne alınırken uyguladığımız cümle parçalama ile kelimelerin farklı ürün özelliklerine ait olma durumu göz önüne alınmaktadır.

SS-LDA bu yöntemler içinde en iyi sonuçları ortaya koymuştur. Hem precision hem de recall bakımından en iyi performansı ortaya koymuştur. Geleneksel LDA'nın kısa metinlerde etkili olamayacağını tahmin ediyorduk. Elde ettiğimiz test sonuçları bunu

doğrular nitelik çıkmıştır. Karşılaştırmada en düşük başarı seviyesi geleneksel LDA ile elde edilmiştir.

Bunların dışında yöntemimizin García-Pablosa, Cuadrosa ve Rigau tarafından önerilmiş LDA kısa metin uyarlaması olan W2VLDA [61] ile de kıyaslanması için çalışma yapılmıştır. W2VLDA, LDA hesaplamalarına Word Embeddings yöntemini dahil eden ve word vektör bilgilerini içeren Word2Vec dosyalarından faydalanan bir yöntemdir. W2VLDA'nın kaynak kodu GitHub üzerinde paylaşılmıştır. W2VLDA'yı çalıştırmak için kaynak kodunu indirmek yeterli değildir. Hangi dilde çalıştırılacaksa kapsamlı ve etkin bir Word2Vec dosyası gerekmektedir. En kapsamlı Türkçe Word2Vec dosyası olarak Abdullah Köksal tarafından hazırlanmış [96] ve GitHub üzerinde paylaşılmış olan Türkçe Word2Vec dosyasını kullanarak W2VLDA yöntemi bu tez çalışmasında kullanılan veri kümesi üzerinde çalıştırılmıştır. Yapılan deneysel çalışmalarda %29 civarında başarı elde edilmiştir. García-Pablosa ve diğerleri makalelerinde W2VLDA'nın farklı veri setlerindeki başarı ölçümlerinde %71 ile %83 arası değişen başarı oranları elde ettiklerini belirtmişlerdir. Ancak, Word2vec dosyası kullanmadan yaptıkları başarı ölçümünde %31 civarında başarı oranı elde etmişlerdir. Sonuç olarak iyi hazırlanmış bir Word2Vec dosyası olmadan W2VLDA yöntemi çok başarısız olmaktadır. Kendi çalışmamızda kullanılan veri kümesinde W2VLDA yöntemi ile başarı oranı %29 olarak ölçülmüştür. Kullanılan Türkçe Word2Vec dosyasının yeterli düzeyde etkin olmadığı değerlendirilmiştir ve W2VLDA ile elde edilen %29 oranındaki başarı seviyesinden ötürü bu yöntem ile deneysel karşılaştırma yapılmamıştır.



7. SONUÇ VE ÖNERİLER

Bu tez çalışmasında özellik tabanlı düzeyde duygu analizinde özellik çıkarımı için özgün bir yöntem önerisi yapılmıştır. Önerilen yöntem en önemli konu modelleme algoritması olan Latent Dirichlet Allocation (LDA) algoritmasının kısa metinlere göre yapılmış özgün bir uyarlamasıdır. Önerilen yöntem Sentence Segment LDA (SS-LDA) olarak isimlendirilmiştir. SS-LDA denetimsiz öğrenmeye dayalı ve etiketlenmiş eğitim verisi gerektirmeyen bir yöntemdir. SS-LDA'da ürün özellikleri ile ilgili kelimelerin birliktelik örüntülerinin LDA'daki hesaplamalara dahil edilerek LDA'nın kısa metinler ve ürün yorumları için uyarlaması yapılmıştır. Bu birliktelik örüntülerinin hesaplamalara dahil edilmesi başarıyı belirgin şekilde yükseltmiştir. Çalışmada veri seti olarak www.hepsiburada.com e-ticaret sitesinden alınmış Türkçe 1292 adet akıllı telefon yorumu kullanılmıştır.

Literatürde İngilizce, Çince gibi dillerle yapılmış çok sayıda özellik tabanlı düzeyde çalışma mevcuttur. Bu dillerde özellikle de İngilizce'de çok sayıda ve çok etkin doğal dil işleme, duygu analizi araç ve kütüphaneleri mevcuttur. Fakat Türkçe'de İngilizce'deki kadar çok sayıda ve etkinlikte doğal dil işleme araç ve kütüphaneleri mevcut değildir. Bir de Türkçe sondan eklemeli yapıya sahip çok esnek ve üretken bir dildir. Dilin bu özelliği doğal dil işleme ve dolayısıyla duygu analizi çalışmalarını oldukça zor hale getirmektedir. İngilizce ile yapılmış birçok duygu analizi çalışmasının veri setleri de paylaşılmış durumdadır. Bu sayede literatür kıyaslaması yapmak da zor değildir. Fakat Türkçe diliyle yapılmış özellik tabanlı düzeydeki duygu analizi çalışmalarının sayısı oldukça azdır ve yabancı bir çalıştay olan SemEval-2016'daki Türkçe veri seti dışında paylaşılmış bir veri seti mevcut değildir. Bu durum literatür kıyaslaması yapmayı da oldukça zor hale getirmektedir. Sonuç olarak Türkçe diliyle özellik tabanlı düzeyde duygu analizi alanında literatürde önemli bir boşluk bulunmaktadır. Bu tezin bu alanda yapılmış bir çalışma olması bu boşluğu doldurması bakımından önem taşımaktadır.

SS-LDA akıllı Türkçe telefon yorumları üzerinde test edilmiştir ve deneysel sonuçlar SS-LDA'nın özellik çıkarımında oldukça başarılı olduğunu ortaya koymuştur. Test sonuçları incelendiğinde SS-LDA'nın özellikle aynı ürün özelliğini ifade eden çok farklı kelimeleri

ve çok farklı ifadeleri farklı ürün özellikleri olarak sınıflandırmayıp bunları aynı ürün özelliği altında sınıflandırmada çok başarılı olduğu gözlemlenmiştir.

SS-LDA kelimelerin istatistiksel dağılımına ve birliktelik örüntülerine dayalı bir yöntemdir. Bu tür istatistiksel dağılımlara dayalı yöntemlerin başarılı olabilmeleri için sağlıklı dağılım değerleri elde edilebilecek büyüklükte bir veri setinin olması gerekmektedir. Dolayısıyla SS_LDA'da veri setinin boyutu önem taşımaktadır. SS-LDA tüm veri seti dışında veri setinin alt kümeleri olan daha küçük veri setleri ile de çalıştırılmış ve küçülen veri setlerinin başarıyı nasıl etkilediği test edilmiştir. Bu konuda yapılan testlerde veri seti küçüldükçe belli bir noktadan sonra başarı oranının da düştüğü gözlemlenmiştir.

Literatür kıyaslaması için SS-LDA SemEval-2016 Türkçe restoran yorumları veri seti üzerinde de test edilmiş ve başarı ölçümü yapılmıştır. SemEval Farklı dillerde veri setlerinin olduğu ve aynı dillerdeki çalışmaların birbirleri ile kıyaslandıkları bir duygu analizi, fikir madenciliği çalıştadır. 2016 yılında yapılan SemEval-2016 çalıştayında Türkçe restoran yorumları veri seti de bulunmaktadır. Bu veri seti ile ilgili 4 tane çalışma gönderilmiştir ve sonuçları yayınlanmıştır. SS-LDA da SemEval-2016 restoran yorumları üzerinde test edilmiş ve aynı bu veri seti üzerindeki diğer çalışmalarla kıyaslanmıştır. Kıyaslamalar F-Score başarı ölçüm kriteri üzerinden yapılmıştır. Yapılan kıyaslamada SS-LDA'nın en yüksek başarıyı ortaya koyan çalışmadan az farkla da olsa daha iyi bir F-Score değeri ortaya koyduğu gözlemlenmiştir.

İkinci olarak SS-LDA geleneksel LDA ve başka iki kısa metin için LDA uyarlamaları olan Sentence-LDA ve Biterm Topic Model ile karşılaştırılmıştır. Yapılan karşılaştırmada Sentence-LDA oldukça iyi bir precision değeri ortaya koymuştur ama recall değeri düşük çıkmıştır. Biterm Topic Model ise hem precision hem de recall bakımından ortalama denilebilecek bir performans göstermiştir. SS-LDA yöntemi ise precision ve recall bakımından hem Sentence-LDA hem de Biterm Topic Model'den daha başarılı sonuçlar ortaya koymuştur.

Bu çalışmada SS-LDA Türkçe sadece veri setleri üzerinde test edilmiştir. Geliştirilen yöntemde dile bağımlı kısımlar vardır ama bu bağımlılık düşük düzeydedir. SS-LDA bazı küçük uyarlamalarla başka dillerde de kullanılabilecek bir yöntemdir. Gelecek çalışması

olarak SS-LDA'nın İngilizce gibi zengin veri setlerinin olduđu bir dilde çalışacak hale getirilmesi ve bu veri setleri üzerinde test edilmesine yönelik çalışmalar yapılması planlanmaktadır.





KAYNAKLAR

1. Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. ABD: Morgan & Claypool Publishers, 18.
2. Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3(4-5), 993-1022.
3. Tetsuya, N. and Jeonghee, Y. (2003, October). *Sentiment analysis: Capturing favorability using natural language processing*. Paper presented 2nd International Conference on KnowledgeCapture, Sanibel Island, FL, ABD.
4. Kushal, D., Steve, L. and Pennock, D.M. (2003, May). *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. Paper presented at 12th International Conference on World Wide Web, Budapest Congress Centre, Macaristan, 519-528.
5. Vasileios, H. and Janyce, M.W. (2000, July). *Effects of adjective orientation and gradability on sentence subjectivity*. Paper presenyed at 18th International Conference on Computational Linguistics, Saarbrücken, Almanya, 299-305.
6. Tong, R.M. (2001, September). *An operational system for detecting and tracking opinions in on-line discussion*. Paper presented at SIGIR 2001 Workshop on Operational Text Classification, New Orleans, Louisiana, ABD, 1-6.
7. Turney P.D. (2002, July). *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews*. Paper presented at 40th Annual Meeting of the Association for Computational Linguistics, Pennsylvania, ABD, 417-424.
8. Pang, B., Lee, L. and Vaithyanathan, S. (2002, July). *Thumbs up?: sentiment classification using machine learning techniques*. Paper presented at 2002 Conference on Empirical Methods in Natural Language Processing, University of Pennsylvania, ABD, 79-86.
9. Mcdonald, R., Hannan, K., Neylon, T., Wells, M. and Reynar, J. (2007, June). *Structured Models for Fine-to-Coarse Sentiment Analysis*. Paper presented at 45th Annual Meeting of the Association of Computational Linguistics, Prag, Çek Cumhuriyeti, 432-439.
10. Nakagawa, T., Inui, K. and Kurohashi, S. (2010, June). *Dependency tree-based sentiment classification using CRFs with hidden variables*. Paper presented at 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, ABD, 786-794.

11. Davidov, D., Tsur, O. and Rappoport, A. (2010, August). *Enhanced Sentiment Learning Using Twitter Hashtags and Smileys*. Paper presented at 23rd International Conference on Computational Linguistics, Pekin, Çin, 241-249.
12. Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267-307.
13. Kang, H., Yoo, S.J. and Han, D. (2012). Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39(5), 6000-6010.
14. Mejova, Y. and Srinivasan, P. (2011, July). *Exploring Feature Definition and Selection for Sentiment Classifiers*, Paper presented at 5th International AAAI Conference on Weblogs and Social Media, Barselona, İspanya, 546-549.
15. Bespalov, D., Bai, B., Qi, Y. and Shokoufandeh, A. (2011, October). *Sentiment Classification Based on Supervised Latent n-gram Analysis*. Paper presented at 20th ACM International Conference on Information and Knowledge Management, Glasgow, İngiltere, 375-382.
16. Balahur, A., Hermida, J.M. and Montoyo, A. (2012). Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53(4), 742-753.
17. İnternet: ISEAR Databank. Cipeç.URL: <http://emotion-research.net/toolbox/toolboxdatabase.2006-10-13.2581092615>, Son Erişim Tarihi: 12.04.2021.
18. Zhang, L., Hua, K., Wang, H., Qian, G. and Zhang, L. (2014). Sentiment Analysis on Reviews of Mobile Users. *Procedia Computer Science*, 34, 458-465.
19. Zhang, K., Xie, Y., Yang, Y., Sun, A., Liu, H. and Choudhary, A. (2014). Incorporating conditional random fields and active learning to improve sentiment identification. *Neural Networks*, 58, 60-67.
20. Choi, Y. and Cardie, C. (2008, October). *Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis*. Paper presented at Proceedings of EMNLP'08, Conference on Empirical Methods in Natural Language Processing, Waikiki, Hawaii, 793-801.
21. Habernal, I., Ptacek, T. and Steinberger, J. (2015). Supervised sentiment analysis in Czech social media. *Information Processing & Management*, 50(5), 693-707.
22. Park, S., Lee, W. and Moon, I.C. (2015). Efficient extraction of domain specific sentiment lexicon with active learning. *Pattern Recognition Letters*, 56, 38-44.

23. Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E. and González-Castaño, F.J. (2016). Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications*, 58, 57-75.
24. Khan, F.H., Qamar, U. and Bashir, S. (2016). eSAP - A decision support framework for enhanced sentiment analysis and polarity classification. *Information Sciences*, 367, 862-873.
25. Chen, T., Xu, R., He, Y. and Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*, 72, 221-230.
26. Eroğul, U. (2009). *Sentiment Analysis In Turkish*, Yayınlanmış Yüksek Lisans Tezi, Orta Doğu Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, Ankara.
27. Kaya, M., Fidan, G. and Toroslu, I.H. (2012, December). *Sentiment Analysis of Turkish Political News*. Paper presented International Joint Conferences on Web Intelligence and Intelligent Agent Technology, Macau, Çin, 174-180.
28. Şimşek, M.U. and Ozdemir, S. (2012, December). *Analysis of the Relation between Turkish Twitter Messages and Stock Market Index*. Paper presented at 6th Conference on Application of Information and Communication Technologies, Tiflis, Gürcistan, 1-4.
29. Çetin, M. and Amasyalı, M.F. (2013, April). *Supervised and traditional term weighting methods for sentiment analysis*. Paper presented at 21st Signal Processing and Communications Applications Conference (SIU), Haspolat, KKTC, 1-4.
30. Martineau, J. and Finin, T. (2009, May). *Delta TFIDF: An Improved Feature Space for Sentiment Analysis*. Paper presented at 3rd International Conference on Weblogs and Social Media, San Jose, California, ABD, 258-261.
31. Aytekin, Ç. (2013). An Opinion Mining Task in Turkish Language A Model for Assigning Opinions in Turkish Blogs to the Polarities. *Journalism and Mass Communication*, 3(3), 179-198.
32. Sevindi, B.İ. (2013). *Türkçe Metinlerde Denetimli Ve Sözlük Tabanlı Duygu Analizi Yaklaşımlarının Karşılaştırılması*, Yüksek Lisans Tezi, Gazi Üniversitesi, Bilgisayar Mühendisliği Bölümü, Ankara.
33. Meral, M. and Diri, B. (2014, April). *Sentiment Analysis on Twitter*. Paper presented at IEEE 22nd Signal Processing and Communications Applications Conference (SIU), Trabzon, Türkiye, 690-693.

34. Akba, F., Uçan, A., Sezer, E.A. and Sever, H. (2014, July). *Assessment of feature selection metrics for sentiment analyses: Turkish movie reviews*. Paper presented at 8th European Conference on Data Mining, Lizbon, Portekiz, 180-184.
35. Hu, M. and Liu, B. (2004, August). *Mining and summarizing customer reviews*. Paper presented at Proceedings of KDD'04 International Conference on Knowledge Discovery and Data Mining, Seattle, ABD, 168-177.
36. Ding, X., Liu, B. and Yu, P.S. (2008, February). *A holistic lexicon-based approach to opinion mining*. Paper presented at Proceedings of WSDM-2008, Conference on Web Search and Web Data Mining, Stanford ABD, 231-240.
37. Popescu, A.M. and Etzioni, O. (2005, October). *Extracting product features and opinions from reviews*. Paper presented at Proceedings of EMNLP-2005, Conference on Empirical Methods in Natural Language Processing, Vancouver, Kanada, 339-346.
38. Wu, Y., Zhang, Q., Huang, X. and Wu, L. (2009, August). *Phrase Dependency Parsing for Opinion Mining*. Paper presented at Proceedings of EMNLP'09, Conference on Empirical Methods in Natural Language Processing, Singapur, 1533-1541.
39. Wei, W. and Gulla, J.A. (2010, July). *Sentiment learning on product reviews via sentiment ontology tree*. Paper presented at 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, İsviçre, 404-413.
40. Cesa-Bianchi, N. and Gentile, C., Zaniboni, L. (2006). Incremental Algorithms for Hierarchical Classification. *Journal of Machine Learning Research*, 7, 31-54.
41. Brody, S. and Elhadad, N. (2010, June). *An unsupervised aspect-sentiment model for online reviews*. Paper presented at Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, LA, ABD, 804-812.
42. Moghaddam, S. and Ester, M. (2010, October). *Opinion digger: an unsupervised opinion miner from unstructured product reviews*. Paper presented at 19th ACM international conference on Information and knowledge management. Toronto, Kanada, 1825-1828.
43. Jiang, L., Yu, M., Zhou, M., Liu, X. and Zhao, T. (2011, June). *Target dependent twitter sentiment classification*. Paper presented at 49th Annual Meeting of the Association for Computational Linguistics, Portland, ABD, 151-160.
44. Qiu, G., Liu, B., Bu, J. and Chen, C. (2011). Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics*, 37(1), 9-27.

45. Eirinaki, M., Pisal, S. and Singh, J. (2012). Feature-based opinion mining and ranking. *Journal of Computer and System Sciences*, 78(4), 1175-1184.
46. Kontopoulos, E., Berberidis, C., Dergiades, T. and Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, 40(10), 4065-4074.
47. Xianghua, F., Guo, L., Yanyan, G. and Zhiqiang, W. (2013). Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. *Knowledge-Based Systems*, 37, 186-195.
48. Bagheria, A., Saraee, M. and Jong, F. (2013). Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems*, 52, 201-203.
49. Nakagawa, H. and Mori, T. (2003). Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2), 201-219.
50. Kang, D. and Park, Y. (2014). Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Systems with Applications*, 41(4), 1041-1050.
51. Isidro, P.M., Rafael, V.G. and Francisco, G.S. (2011, June). *Ontology-Guided Approach to Feature-Based Opinion Mining*. Paper presented at 16th international conference on Natural language processing and information systems, Alicante, Ispanya, 193-200.
52. Quan, C. and Ren, F. (2014). Unsupervised product feature extraction for feature-oriented opinion determination. *Information Sciences*, 272, 16-28.
53. Kansal, H. and Toshniwal, D. (2014). Aspect based Summarization of Context Dependent Opinion Words. *Procedia Computer Science*, 35, 166-175.
54. Bhadane, C., Dalal, H. and Doshi, H. (2015). Sentiment Analysis: Measuring Opinions. *Procedia Computer Science*. 45, 808-814.
55. Maharani, W., Widyanoro, D.H. and Khodra M.L. (2015). Aspect Extraction in Customer Reviews Using Syntactic Pattern. *Procedia Computer Science*, 59, 244-253.
56. Li, S., Zhou, L. and Li, Y. (2015). Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures. *Information Processing and Management*, 51(1), 58-67.
57. Liao, C., Fengn, C., Yang, S. and Huang, H. (2016). Topic-related Chinese message sentiment analysis. *Neurocomputing*, 210, 237-246.

58. Liu, Q., Gao, Z., Liu, B. and Zhang, Y. (2016). Automated rule selection for opinion target extraction. *Knowledge-Based Systems*, 104, 74-88.
59. Akhtar, M.S., Gupta, D., Ekbal, A. and Bhattacharyya, P. (2017). Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. *Knowledge-Based Systems*, 125, 116-135.
60. Xiong S., Wang K., Ji D. and Wang B. (2018). A short text sentiment-topic model for product reviews. *Neurocomputing*, 297, 94-102.
61. García-Pablosa A., Cuadros M. and Rigau G. (2018). W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis. *Expert Systems with Applications*, 91, 127-137.
62. Akbaş, E. (2012). *Aspect Based Opinion Mining on Turkish Tweets*, Yüksek Lisans Tezi, Bilkent Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Ankara.
63. Dehkharghani, R., Yanikoglu, B., Saygin, Y. and Oflazer, K. (2016). Sentiment analysis in Turkish at different granularity levels. *Natural Language Engineering*, 23(4), 1-25.
64. Dalkılıç, F.E., Gelişli, S. and Diri B. (2010, April). *Named Entity Recognition from Turkish texts*, Paper presented at IEEE 18. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, Diyarbakır, Türkiye, 918-920.
65. Akın, G. and Eryiğit, G. (2012, January). Initial Explorations on using CRFs for Turkish Named Entity Recognition. *Paper presented at 24th International Conference on Computational Linguistics*, IIT, Bombay, Hindistan, 2459-2474.
66. Küçük, D. and Yazıcı, A. (2009, October). *Named Entity Recognition Experiments on Turkish Texts*. Paper presented at 8th International Conference on Flexible Query Answering Systems, Roskilde, Danimarka, 524-535.
67. Church, K.W. and Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, 16, 22-29.
68. Mullen, T. and Collier, N. (2004, July). *Sentiment Analysis using Support Vector Machines with Diverse Information Sources*. Paper presented at Conference on Empirical Methods in Natural Language Processing, Barcelona, İspanya, 412-418.
69. Lafferty, J., McCallum, A. and Pereira F. (2001, June). *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. Paper presented at 18th International Conference on Machine Learning, Williams College, Williamstown, MA, ABD, 282-289.

70. Kennedy, A. and Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters, *Computational Intelligence*, 22(2), 110-125.
71. Hailong, Z., Wenyan, G. and Bo, J. (2014, September). *Machine Learning and Lexicon Based Methods for Sentiment Classification: A Survey*. Paper presented at 11th Web Information System and Application Conference, Tianjin, Çin, 262-265.
72. Goldensohn, S.B., Hannan, K., McDonald, R., Neylon, T., Reis, G.A. and Reynar, J. (2008, January). *Building a sentiment summarizer for local service reviews*. Paper presented at Proceedings of WWW'08 Workshop: NLP in the Information Explosion Era, Pekin, Çin.
73. Zhuang, L., Jing, F. and Zhu, X. (2006, November). *Movie review mining and summarization*. Paper presented at 15th ACM International Conference on Information and Knowledge Management, Indianapolis, ABD, 43-50.
74. Swapna, S. and Wiebe, J. (2009, August). *Recognizing stances in online debates*. Paper presented at Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapur, 226-234.
75. Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, 77(2), 257-286.
76. Yüret D. and Türe F. (2006, June). *Learning morphological disambiguation rules for Turkish*. Paper presented at Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, New York, USA, 328-334.
77. Oflazer K. (2012). Türkçe ve Doğal Dil İşleme. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 5(2).
78. İnternet: Zemberek NLP. *Cipac*.URL: <http://zembereknlp.blogspot.com.tr>, Son Erişim Tarihi: 13.04.2021
79. Oflazer K. (1994). Two-level description of Turkish morphology. *Literary and linguistic computing*, 9(2), 137-148.
80. İnternet: ITU Türkçe Doğal Dil İşleme Yazılım Zinciri. *Cipac*.URL: <http://tools.nlp.itu.edu.tr/>, Son Erişim Tarihi: 13.04.2021.
81. Sak H., Güngör T. and Saraçlar M. (2007, May). *Morphological disambiguation of Turkish text with perceptron algorithm*. Paper presented at 8th International Conference on Computational Linguistics and Intelligent Text Processing, Berlin, Almanya, 107-118.

82. İnternet: SentiWordNet. *Cippec*.URL: <http://sentiwordnet.isti.cnr.it/>, Son Erişim Tarihi: 13.04.2021.
83. Dehkharghani R., Saygın Y., Yanikoglu B. and Oflazer K. (2016). SentiTurkNet: a Turkish polarity lexicon for sentiment analysis, *Language Resources and Evaluation*, 50, 667-685.
84. Agrawal R., Imielinski T. and Swami A. (1993, June). *Mining Association Rules between Sets of Items in Large Databases*. Paper presented at Proceedings Of the ACM SIGMOD International Conference on Management of Data, Washington, ABD, 207-216.
85. Hong L. and Davison B. (2010, July). *Empirical study of topic modeling in twitter*. Paper presented at Proceedings of the First Workshop on Social Media Analytics. Washington DC, ABD, 80-88.
86. Mehrotra R., Sanner S., Buntine W. and Xie L. (2013, July). *Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling*. Paper presented at 36th international ACM SIGIR conference on Research and development in information retrieval, Dublin, İrlanda, 889-892.
87. Alvarez-Melis D. and Saveski M. (2016, May). *Topic Modeling in Twitter: Aggregating Tweets by Conversations*. Paper presented at 10th International AAAI Conference on Web and Social Media, Cologne, Almanya, 519-522.
88. Yin J. and Wang J. (2014, August). *A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering*. Paper presented at 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, ABD, 233-242.
89. Jo Y. and Oh A. (2011, February). *Aspect and sentiment unification model for online review analysis*. Paper presented at 4th ACM international conference on Web search and data mining, Hong Kong, Çin, 815-824.
90. Titov I. and McDonald R. (2008, April). *Modeling Online Reviews with Multi-grain Topic Models*. Paper presented at 17th international conference on World Wide Web, Pekin, Çin, 111-120.
91. Alam M. H. and Lee S.K. (2012, December). *Semantic Aspect Discovery for Online Reviews*. Paper presented at 12th International Conference on Data Mining, Brüksel, Belçika.
92. Xianghua F., Guo L., Yanyan G. and Zhiqiang W. (2013). Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon, *Knowledge-Based Systems*, 37, 186-195.

93. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar S., AL-Smadi M., Al-Ayyoub M., Zhao Y., Qin B., Clercq O. D., Hoste V., Apidianaki M., Tannier X., Loukachevitch N., Kotelnikov E., Bel N., Jiménez-Zafra S. M. and Eryigit G. (2016, June). *Semeval-2016 Task 5: Aspect Based Sentiment Analysis*. Paper presented at 10th International Workshop on Semantic Evaluation (Semeval-2016), 19-30.
94. Yan X., Guo J., Lan Y. and Cheng X. (2013, May). *A biterm topic model for short texts*. Paper presented at 22nd international conference on World Wide Web Pages, Rio de Janeiro, Brezilya, 1445-1456.
95. Bajpai, R., Hazarika, D., Singh, K., Gorantla, S., Cambria, E. and Zimmerman, R. (2019). Aspect-Sentiment Embeddings for Company Profiling and Employee Opinion Mining, *arXiv preprint, arXiv: 1902.08342*.
96. İnternet: Akoksal, Türkçe Word Embeddings Dosyası. *Cipec*.URL: <https://github.com/akoksal/Turkish-Word2Vec>. Son Erişim Tarihi: 13.04.2021.
97. Ozyurt B. and Akcayol M. (2021). A new topic modeling based approach for aspect extraction in Aspect Based Sentiment Analysis: SS-LDA. *Expert Systems With Applications*, 168.
98. Ozyurt B. and Akcayol M. (2018). Fikir madenciliği ve duygu analizi, yaklaşımlar, yöntemler üzerine bir araştırma. *Selçuk Üniversitesi Mühendislik Bilim ve Teknoloji Dergisi*, 6(4), 668-693.

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : ÖZYURT, Barış
 Uyruğu : T.C.
 Doğum tarihi ve yeri :
 Medeni hali :
 Telefon :
 Faks :
 e-mail :

Eğitim

Derece	Eğitim Birimi	Mezuniyet Tarihi
Doktora	Gazi Ü. /Bilişim Ens./Bilgisayar Bilimleri	Devam ediyor
Yüksek lisans	ODTÜ/Bilgisayar Mühendisliği	2003
Lisans	ODTÜ/Bilgisayar Mühendisliği	2000
Lise	Antalya Anadolu Lisesi	1996

İş Deneyimi

Yıl	Yer	Görev
2020-Halen	Etstur Ersoy Turistik Servisleri	Yazılım Mühendisi
2017-2020	OTI Holding	Yazılım Mühendisi
2003-2017	Tarnet A.Ş.	Yazılım Mühendisi
2001-2003	Meteksan Sistem A.Ş.	Yazılım Mühendisi
2000-2001	Cybersoft A.Ş.	Yazılım Mühendisi

Yabancı Dil

İngilizce

Yayınlar

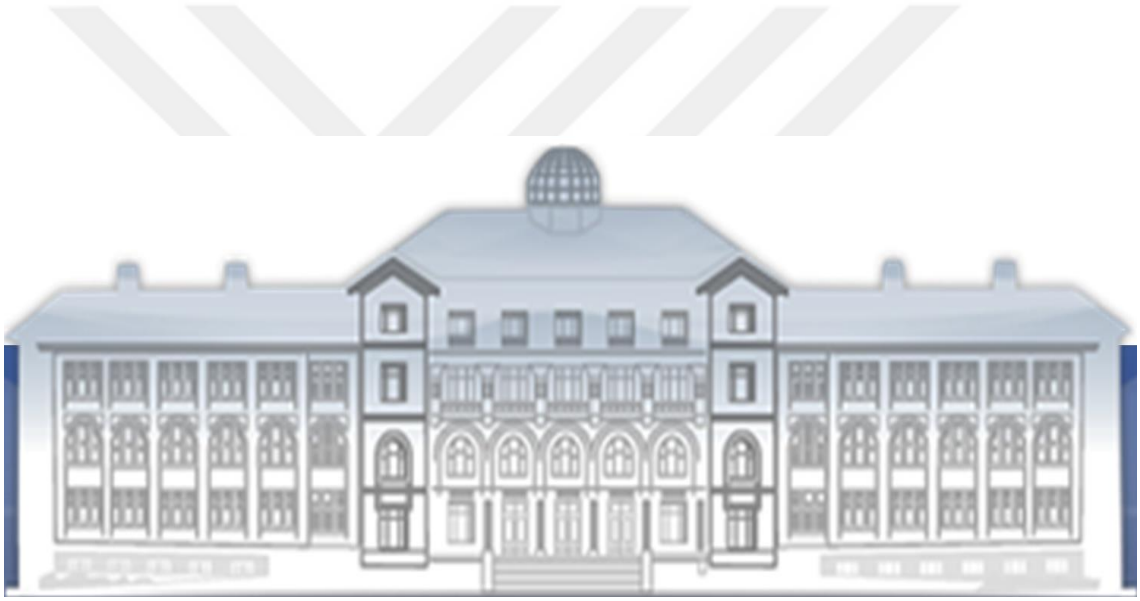
1. Ozyurt B., Akcayol M. (2021). A new topic modeling based approach for aspect extraction in Aspect Based Sentiment Analysis: SS-LDA. *Expert Systems With Applications*, 168. DOI: 10.1016/j.eswa.2020.114231.

2. Ozyurt B., Akcayol M. (2018). Fikir madenciliđi ve duygu analizi, yaklaşımlar, yöntemler üzerine bir araştırma. *Selçuk Üniversitesi Mühendislik Bilim ve Teknoloji Dergisi*, 6(4), 668 - 693. DOI: 10.15317/Scitech.2018.16.
3. Ozyurt B., Alper I.D., Akcayol M. (2017). Analyzing the Route Cache Timeout Parameter of DSR Protocol in Mobile Ad Hoc Networks. *International Journal of Computer and Communication Engineering*, 6(1), 40 - 48. DOI: 10.17706/ijcce.2017.6.1.40-48

Hobiler

Bisiklet, yüzmek, turistik geziler.





GAZİLİ OLMAK AYRICALIKTIR.