

KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

YÜKSEK LİSANS TEZİ

ÜRÜN ÖZELLİKLERİNİN TÜRKÇE KULLANICI YORUMLARINDAN
ÖZELLİK TABANLI SENTİMENT ANALİZİ İLE KEŞFEDİLMESİ

HAZAL TÜRKMEN

KOCAELİ 2016

KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI

YÜKSEK LİSANS TEZİ

ÜRÜN ÖZELLİKLERİNİN TÜRKÇE KULLANICI
YORUMLARINDAN ÖZELLİK TABANLI SENTİMENT
ANALİZİ İLE KEŞFEDİLMESİ

HAZAL TÜRKMEN

Doç.Dr. Sevinç İLHAN OMURCA

Danışman, Kocaeli Üniv.

Yrd.Doç.Dr. Alev MUTLU

Jüri Üyesi, Kocaeli Üniv.

Yrd.Doç.Dr. Tuğba YILDIZ

Jüri Üyesi, İstanbul Bilgi Üniv.



Tezin Savunulduğu Tarih: 15.06.2016

ÖNSÖZ VE TEŞEKKÜR

İnternet kullanımının yaygınlaşması ve bunun sonucunda insanların çok farklı alanlarda kişisel fikirlerini internet üzerinden paylaşımları ile sosyal medya, kişisel bloglar, e-ticaret siteleri önemli bilgi kaynakları haline gelmiştir. Gün geçtikçe insanların hayatlarında internet üzerinden sosyal paylaşımlara daha çok yer vermesi, firmaların da sosyal ağ ortamlarında daha fazla yer almasına neden olmuştur. Bu açıdan sosyal ağlar çok önemli veri kaynaklarıdır ve buradan elde edilen bilgi birçok sektöre yön vermektedir. İnsanların bir ürün ya da hizmet hakkındaki olumlu ve olumsuz görüşleri, internet ağı aracılığı ile bütün dünyaya yayılmaktadır. Sonuç olarak hem popülerlik hem de ticari açıdan, ürün ve ya hizmetlerin zayıf ve güçlü özelliklerinin kullanıcı yorumları yardımıyla çıkarılması önemli bir araştırma konusu haline gelmiştir.

Tezimin konusunun belirlenmesinde, araştırma aşamasında, yön tayininde ve tamamlanmasında desteğini esirgemeyen değerli hocam Doç. Dr. Sevinç İlhan Omurca'ya, birlikte çalışmaktan memnunluk duyduğum Arş. Gör. Ekin Ekinci'ye, TÜBİTAK 114E422 nolu proje kapsamındaki tezim sırasında beni maddi açıdan destekleyen Tübitak Araştırma Destek Programları Başkanlığı'na ve bugünlere gelmemde büyük pay sahibi olan aileme teşekkürlerimi sunarım.

Mayıs - 2016

Hazal TÜRKMEN

İÇİNDEKİLER

ÖNSÖZ VE TEŞEKKÜR	i
İÇİNDEKİLER	ii
ŞEKİLLER DİZİNİ.....	iii
TABLolar DİZİNİ	iv
SİMGELER VE KISALTMALAR DİZİNİ	v
ÖZET.....	vi
ABSTRACT	vii
GİRİŞ	1
1. GENEL BİLGİLER	3
1.1. Tez Çalışmasının Amacı ve Başlatılma Sebepleri	3
1.2. Tez Çalışmasının Katkıları.....	4
1.3. Literatür Taraması	4
1.4. Tezin Yapısı	12
2. GELİŞTİRİLEN SİSTEM ve TEORİK ALTYAPISI	13
2.1. Sentiment analizi nedir?	13
2.2. Sentiment Analizi Teknikleri	15
2.2.1. Makine öğrenmesi ile sentiment analizi	15
2.2.2. Sözlüksel benzeşim ile sentiment analizi.....	16
2.3. Sentiment analizi seviyeleri.....	17
2.4. Özellik Tabanlı Sentiment Analizi	17
2.4.1. Problem tanımı.....	18
3. GELİŞTİRİLEN SİSTEM.....	21
3.1. Sistem Mimarisi	21
3.1.1. Veri toplama ve önışleme	22
3.1.2. Sık geçen özelliklerin çıkarılması.....	25
3.1.3. Sık geçmeyen özelliklerin çıkartılması.....	31
3.1.4. Özellik-sentiment çiftlerinin çıkartılması	32
3.1.5. Özellik tabanlı skorlama	37
4. DENEYSEL ÇALIŞMALAR	42
4.1. Kullanılan Teknolojiler	42
4.2. Kullanılan Veri Kümleri.....	42
4.3. Yazılım Mimarisi Tasarımı ve Uygulama Adımları	43
4.3.1. Veri önışleme.....	46
4.3.2. Sık geçen tekli özelliklerin çıkarılması.....	46
4.3.3. Sık geçen çok terimli özelliklerin çıkarılması	47
4.3.4. Sık geçmeyen özelliklerin çıkarılması	48
4.3.5. Dokümanın etiketlenmesi	49
4.3.6. Özellik-sentiment çiftlerinin çıkarılması	49
4.3.7. Özellik tabanlı skorlama	56
5. SONUÇLAR VE ÖNERİLER	60
KAYNAKLAR	62
KİŞİSEL YAYIN VE ESERLER	68
ÖZGEÇMİŞ	69

ŞEKİLLER DİZİNİ

Şekil 2.1. Sentiment analizinin Google Trends grafiği	13
Şekil 2.2. Sentiment analizinin kavramsal modeli	14
Şekil 3.1. Geliştirilen sistem mimarisi	22
Şekil 3.2. Bir web arama sisteminin temel adımları	23
Şekil 3.3. Çok terimli özellik çıkarımının akış şeması.....	28
Şekil 3.4. Algoritma adımları.....	32
Şekil 3.5. Özellik tabanlı skorlama adımları.....	40
Şekil 4.1. Sistem UML diyagramı.....	45
Şekil 4.2. Özellik tabanlı skorlama sonucu.....	57



TABLÖLAR DİZİNİ

Tablo 3.1. Yedi durum için oluşturulan PDA'lar.....	34
Tablo 4.1. Yorum setinin detaylı bilgisi.....	43
Tablo 4.2. Sık geçen tekli özellikler ve PMI skoru.....	47
Tablo 4.3. Sık geçen çok terimli özellikler	47
Tablo 4.4. Sık geçmeyen tekli özellikler.....	48
Tablo 4.5. Özellik-sentiment çiftleri	50
Tablo 4.6. Gruplanan özellikler	57



SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar

DFA	: Deterministic Finite Automata (Deterministik Sonlu Otamata)
ME	: Maximum Entrophy
NB	: Naïve Bayes
ÖSÇ	: Özellik Sentiment Çiftleri
PDA	: Push Down Automata (Yığıtlı Özdevinirler)
SVM	: Support Vector Machine

ÜRÜN ÖZELLİKLERİNİN TÜRKÇE KULLANICI YORUMLARINDAN ÖZELLİK TABANLI SENTİMENT ANALİZİ İLE KEŞFEDİLMESİ

ÖZET

İnternetteki sınırsız bilgi ve yorum akışı çok farklı alanlardaki kullanıcı yorumlarının analiz edilebilmesine olanak vermiştir. Günümüzde firmalar kendi markalarının ürünlerini değerlendirirken ya da geliştirirken müşterilerden alınan geri bildirimlerden faydalanma yolunu seçmektedirler. Bu geri bildirimler, firmaların ürettikleri ürünlerin niteliklerinin ve firmaların rekabetçi güçlerinin artmasına yardımcı olan en önemli unsurlar olarak kabul görmektedir. Artık müşteriler ve firmalar kullanılan ürün ile ilgili çok sayıda kullanıcı yorumuna internet üzerinden kolayca erişebilmektedir. Büyük bir doküman kümesi içinden herhangi bir ürün hakkında yapılan olumlu/olumsuz yorumların çıkarılması; bu yorumların ürüne ait hangi özellikler için yapıldığının keşfi, el ile gerçekleştirilemeyecek bir işlemdir. Özetle ürünlerin zayıf ve güçlü özelliklerinin kullanıcı yorumlarından otomatik olarak keşfedilmesi önemli bir araştırma ve uygulama alanı haline gelmiştir. Sentiment analizi (duygu analizi) alanında yapılan çalışmaların büyük çoğunluğu metnin anlamının olumlu ve ya olumsuz olduğu üzerine odaklanmıştır. Günümüzde ise kullanıcı yorumlarının daha derinlemesine analizi için özellik tabanlı sentiment analizi öne çıkan çalışma alanı haline gelmiştir. Tamamlanan tez kapsamında internetteki Türkçe kullanıcı yorumları üzerinde olumlu/olumsuz sınıflama yapabilen özgün bir özellik tabanlı sentiment analizi yöntemi geliştirilmesi hedeflenmiştir. Literatürdeki çalışmalarda, Türkçe metinlerin özellik tabanlı sentiment analizine özgü bir çalışmaya ve yazılıma rastlanmamış olması bu alanda yapılacak olan tez çalışmasının değerini artıracaktır. Geliştirilen sistem ile bir ürün hakkında genel kullanıcı fikrinin olumlu ya da olumsuz olduğu keşfedilebilecek aynı zamanda ürünün güçlü ve zayıf özellikleri belirlenebilecektir. Tez kapsamında internetten toplanan otel yorumlarının üzerinden analiz yapılmış, bu analiz sonucunda uzman kişinin bulduğu özellikler dışında daha detaylı özellikler yakalandığı gözlenmiştir ve bunların kullanıcı yorumlarına göre otomatik puanlaması elde edilmiştir.

Anahtar Kelimeler: Özellik Çıkarımı, Özellik Tabanlı Sentiment Analizi, Sentiment Skorlaması, Türkçe Sentiment Analizi.

DISCOVERING PRODUCT FEATURES FROM TURKISH REVIEWS BY USING ASPECT BASED SENTIMENT ANALYSIS

ABSTRACT

Unlimited stream of information-review in Internet enables analyzing of customer reviews from many different domains. Nowadays, companies utilize customer feedbacks for evaluating or developing their own brand products. These feedbacks are regarded as the most important factors which assist to increase the quality and competitive power of products which are produced by companies. From now on, customers and companies can easily reach the large number of customer reviews about products via the Internet. Within such a large amount of document set, processes of extraction of positive and negative customer reviews about any product, discovering for which aspects of the product these reviews are made, cannot be performed manually. In brief, auto-discovery of aspect weakness of the products based on customer reviews has become important. The majority of the studies on sentiment analysis focus on whether or not the meaning of the text is positive or negative. Nowadays, aspect based sentiment analysis has become a prominent field of study for in-depth analysis of customer reviews. Auto-detection of aspects of products is a very important step for orientation (positive/negative) detection of customer reviews on these products. In this project, development of an original aspect based sentiment analysis method which is capable of classifying Turkish customer reviews on the Internet as positive/negative is aimed. In the literature, a study and a software specific to aspect based sentiment analysis of Turkish texts has not been detected so this would increase the value of the project in this field. In this framework, it can be discovered that the general customer opinion about a product is positive or negative and at the same time both strong and weak aspects of this product can be determined. Hotel and electronic reviews which are commented by users are analyzed by developed system and more aspects are found expect from aspects which are found by human expert and sentiment score for aspects is achieved according to user reviews.

Keywords: Aspect Extraction, Aspect Based Sentiment Analysis, Sentiment Scoring, Turkish Sentiment Analysis.

GİRİŞ

Sentiment analizi; kişilerin ürünler, servisler, firmalar, bireyler, görevler, olaylar, başlıklar ve bunların özellikleri üzerine fikirleri, duyguları, değerlendirmeleri, değer biçmeleri, tutumları ve hislerini analiz etme işidir. Dilbilim (Linguistic) ve doğal dil işleme üzerine çok eski yıllardan beri çalışmalar yapılmasına rağmen 2000’li yıllara kadar üzerine çok fazla çalışma bulunmayan sentiment analizi, 2000’lerden itibaren oldukça aktif bir alan haline gelmiştir. Bu alanın bu kadar önemli bir hale gelmesinin nedenlerinin başında, sosyal medyada bulunan büyük miktardaki duygu içeren dokümanlar gösterilebilir. Günümüzde sentiment analizinin sosyal medya üzerine yapılan çalışmaların merkezinde yer aldığı da bilinmektedir. Sentiment analizi çalışmalarının popüler olmasının diğer nedenleri arasında kişilerin görüşlerinden etkilenen; işletme, siyaset bilimi, ekonomi ve sosyal bilimler üzerine de uygulama geliştirilebilir olması, daha önceden üzerine çalışılmamış pek çok zorlu araştırma problemine çözüm önermesi ve ticari uygulamaların yaygınlaşması gösterilebilir.

Kişi görüşleri neredeyse tüm insan aktivitelerinin merkezinde yer almaktadır. Bunun nedeni ise kişilerin görüşlerinin, insan davranışlarını belirleyen en önemli etkileyici olmalarıdır. Tüketiciler bir ürünü almadan önce o ürünü kullanan diğer tüketicilerin fikirlerini ve ya oy verecek kişilerse adaylar ile ilgili olarak diğer oy vereceklerin fikirlerini alarak karar verme ihtiyacı hissetmektedirler. İş hayatında ise firmalar, kendi ürünleri ile ilgili olarak tüketicinin ve ya halkın fikrini dikkate almak istemektedirler. Yalnız bu tür bilgilerin elde edilmesinde insan gücünün kullanılması oldukça zorludur ve yavaş bir şekilde işlemektedir. Proje bu açıdan otomatik bir yöntem sunmakta, maliyet ve zaman yönündeki olumsuz etkilenmeyi gidermektedir.

Sosyal paylaşımların internet üzerinden yapılması eğilimdeki önlenemez artış ile karar verme konusunda internet üzerindeki veri içeriklerinden yararlanılmaya başlanmıştır. Mesela bir tüketici herhangi bir ürünü almadan önce bu ürünü kullanan arkadaşına, ailesine sormak yerine bu ürün ile ilgili yapılan kullanıcı yorumlarını dikkate almaktadır. Çünkü kendisine aydınlatacak pek çok fikre bu yorumlar sayesinde sahip olabilmektedir. Aynı şekilde bir firmayı ele alırsak anketler,

kamuoyu yoklaması ya da hedef kitle ile uğraşmak yerine halkın ürün ile ilgili yorumlarını sosyal medya üzerinden fazlasıyla elde edebilmektedir. Peki, neden otomatik sentiment analizine ihtiyaç duyulmaktadır? Bu sorunun cevabı şu şekilde verilebilir. Bir ürün ile ilgili kullanıcı yorumlarına bakmak istenildiğinde sadece bir web sitesinde olmadığını pek çok sitede aynı ürün ile ilgili farklı kullanıcıların yapmış olduğu çok miktarda yorumun olduğu görülmektedir. Bu kadar fazla verinin olduğu bir ortamda istenilen bilginin yer aldığı web sitelerinin bulunması ve veriler üzerinden özetleme yapılması ortalama bir insan için oldukça zor bir iştir. Bundan dolayı, bir ürün ile ilgili otomatik olarak görüş ve ya duygu keşfetmek, analiz etmek ve özetlemek için özel sistemlere ihtiyaç vardır. Görüş madenciliği (opinion mining) olarak da bilinen sentiment analizi, bu ihtiyaçlardan doğar. Özellik tabanlı sentiment analizi ise sentiment analizinin ürünün parçası olarak adlandırılan özelliklerin seviyesinde yapılan analiz çeşididir. Örneğin; otel ile ilgili olan bir yorumda otelin bir özelliği olan oda ile ilgili olumlu/olumsuz değerlendirme yapılması daha derinlemesine bir analiz gerçekleştirilmesini sağlar.

Bu tez çalışmasında, bir ürün ile ilgili Türkçe kullanıcı yorumları toplanarak, bu yorumlar üzerinde olumlu ve ya olumsuz sınıflama yapan özellik tabanlı sentiment analizi yöntemi modellenmiştir. Tez çalışması dört temel aşamadan oluşmaktadır. Birinci aşamada veri kümesinin oluşturulması için ürünlere ait Türkçe kullanıcı yorumları internetten toplanmıştır. Tezin ikinci aşamasında tekli ve öbek şeklindeki isimlerden özellik olarak kullanılanların saptanması için genelleştirilmiş bir yöntem uygulanmıştır. Üçüncü aşamada hangi sentimentin hangi özelliği nitelediği bulunmuş ve son aşamada, oluşturulmuş bir sentiment sözlüğü kullanılarak özelliklerin duygu skorlaması gerçekleştirilmiştir.

1. GENEL BİLGİLER

Bu bölümde, tez çalışmasının amacı ve başlatılma sebebi açıklanmış; literatürde sentiment analizi alanında gerçekleştirilen çalışmalar özetlenmiş ve son olarak da tezin katkısı belirtilmiştir.

1.1. Tez Çalışmasının Amacı ve Başlatılma Sebepleri

Web'in günden güne büyümesiyle birlikte e-ticaret, bloglar, haber portalları ve sosyal medya siteleri tarafından çok büyük ölçüde verilerin depolandığı bir kaynak haline gelmiştir. Bunun sonucu olarak Web, veri analizi için vazgeçilmez bir kaynak haline gelmiştir. Günümüzde özellikle e-ticaret siteleri ürünler hakkında birçok kullanıcının yorumlarını analiz ederek satış politikalarını belirlemektedir. Ve ya kullanıcılar bir ürün almadan önce bu ürünü alan kişilerin yorumlarını okuyarak ürünü satın aldığı anda neyle karşılaşacağına dair ipuçları almak istemektedirler. Ancak verinin bu derece büyük olması insan gücü ile raporlanmasını ve analizini olanaksız kılmaktadır. Bu durumda bu verilerin bilgisayarlar tarafından işlenerek istatistiksel ve algoritmik çözümler ile otomatik olarak analiz yapılması kaçınılmaz olmuştur. Örneğin, bu otomatik analiz işlemi için yaklaşık 250 milyon kullanıcıya ve 2 milyar ürüne sahip Amazon gibi dev e-pazarlama siteleri kendi makine öğrenmesi modellerini geliştirmişlerdir.

Günümüzde araştırmacılar ilk olarak büyük dokümanlardan başlayarak, kısa metinlerin cümlelerine hatta bu cümlede geçen bir varlığa ait özelliklerin duygularını belirlemek için çalışmaktadırlar. Dokümanlarda geçen özellikler, duyguların ifade edildiği başlıklardır. Bir sözcük ve ya söz öbeklerinden oluşabilirler. Analiz işleminin özellik seviyesinde bu kadar derinlemesine yapılması verimli ve verimsiz olan verinin daha iyi ayrıştırılarak daha gerçekçi duygu analizinin yapılmasını sağlamaktadır.

Bu çalışmada belirli bir ürüne ait kullanıcı yorumları analiz edilerek, bu ürüne ait zayıf ve güçlü özelliklerin keşfedilmesi amaçlanmaktadır. Örneğin; bir otel ile ilgili yorumdan otele ait oda, yemek, havuz, personel vb. gibi özelliklerin olumlu/olumsuz

nasıl yorumlandığının çıkarılması otel ile ilgili yapılan analizin daha gerçekçi olmasını sağlayacaktır. Önerilen sistemin gerçekleştirilmesinde doğal dil ile yazılmış yorumlar internetten toplanarak Türkçe Dil Kuralları'na bağlı birçok yöntem denenmiş; yapılan test ve sonuçlara göre uygun olan yöntem bulunarak sonuçlar elde edilmiştir.

1.2. Tez Çalışmasının Katkıları

Bu tezde aşağıdaki katkıların sağlanması amaçlanmıştır.

Tez kapsamında Türkçe metinler üzerinde yapılan özellik tabanlı sentiment analizi için özgün yöntemler önerilmiştir ve bu yöntemler yardımı ile ürünlerin zayıf ve ya güçlü taraflarının belirlendiği bir uygulama geliştirilmiştir. Geliştirilen uygulama kolayca başka alanlara da adapte edilebilecek yapıda tasarlanmıştır. Böylece ülkemizdeki farklı markalar ürünlerinin kullanıcılar tarafından beğenilen ya da beğenilmeyen özelliklerini değerlendirebileceklerdir. Gerçekleştirilen tez çalışması bu yönü ile sanayi ile akademik dünya arasında köprü kurabilecek kapasitededir.

Tez çalışmasında, özellik sentiment çiftlerinin bulunması adımının gerçekleşmesi için Push Down Automata (PDA) çözüm olarak önerilmiştir. Daha önce Türkçe doğal dil işleme alanında özellik sentiment çiftlerinin bulunmasına yönelik böyle bir çalışma gerçekleştirilmemiştir.

Türkçe metinler için metinlerden özellik çıkarımı gerçekleştiren bir çalışma bulunmamaktadır. Gerçekleştirilmesi hedeflenen sistemde özelliklerin tümü otomatik olarak keşfedilmiştir. Bu açıdan tez çalışmasının her aşamasındaki elde edilen sonuçlar Türkçe için literatüre önemli bir katkı sağlamıştır ve bu analiz yönteminin geliştirilmesi için gelecek çalışmalara yardımcı olacak niteliktedir.

1.3. Literatür Taraması

Literatürde özellik tabanlı sentiment analizinde danışmanlı ve danışmansız yöntemler kullanılmaktadır. Danışmanlı yöntemler daha önceden etiketlenmiş eğitim verilerine ihtiyaç duymaktadır. Daha önceden etiketlenmiş eğitim verisinin oluşturulması, çok yüksek insan eforu isteyen zor bir adımdır. Bu nedenle danışmansız veri kümesi ile

çalışmak daha geçerli bir yoldur. İnternette farklı ürünler için yapılmış kullanıcı yorumları ile gerçekleştirilen sentiment analizi uygulamalarında, danışmansız ve alan bağımsız şekilde tasarlanan yöntemlerin daha etkin olduğu kabul gören bir gerçektir. Hu ve Liu [1] sentiment analizinde temel olarak kabul edilen çalışmalarında, ürünler üzerine yapılan yorumları kullanarak özet çıkarmayı amaçlamışlardır. Buradaki özet çıkarma bir yorum içerisinde temsil edici cümleler seçilmesi şeklinde tanımlanan klasik özet çıkarmadan farklıdır. Bu çalışma kapsamında tanımlanan özet çıkarma, bir ürün için ürünün parçaları ve özelliklerinin eldeki yorumlardan çıkartılması; çıkartılan özellikler ile ilişkili onları niteleyen sentimentlerin çıkartılması ve sonuç olarak ilgili özellik için pozitif ve negatif yorumların listelenmesidir. Yani kullanıcının ürün ile ilgili neyi sevip neyi sevmediğinin tespit edilmesi amaçlanmaktadır. Sistem iki adımda gerçekleştirilmiştir. Birinci adımda kullanıcıların duygularını belirttiği özelliklerin çıkartılması ve bu özelliklerin frekanslarına göre derecelendirilmesi; ikinci adım ise özellik ile ilgili kaç kullanıcının pozitif kaç kullanıcının da negatif duygu belirttiğinin tespitidir ki bu adım özetleme adımdır. Çalışmada çıkartılan özellikler sık geçen ve sık geçmeyen özelliklerdir ve asıl önemli olan ise sık geçen özelliklerin çıkartılmasıdır. Sık geçen özellikleri (isim/isim öbekleri) çıkarmak için birliktelik kuralları kullanılmıştır. Ancak çıkartılan her isim/isim öbeği ürün ile ilgili özellik olmayabilir bu nedenle de bir filtreleme işlemi gerçekleştirilmiştir. Bu çalışmada tekli isimler için “gereksizlerin budanması” (Redundancy Pruning), isim öbekleri için de “kompaktlık budama” (Compactness Pruning) olmak üzere 2 çeşit filtreleme önerilmiştir. Sentimentleri çıkarmak için ise sentimentlerin özelliklerin yakınında bulunma isteği göstermeleri yaklaşımı kullanılmıştır. Bu yaklaşımdan yola çıkarak eğer bir sıfat sık geçen bir özelliği niteliyorsa, bu sıfat sentiment olarak belirlenmektedir. Sık geçmeyen özellikler, özellik kümesi için sadece tamamlayıcı görevindedir ve sık geçen özellikler kadar önemli değildirler. Bu özelliklerin çıkartılması için kullanılan yaklaşım ise farklı özellikleri tanımlamak için aynı sentimentlerin kullanılmasıdır. Çıkartılan sentimentlerin nitelediği ve sık geçen özellik olarak işaretlenmemiş isim/isim öbekleri sık geçmeyen özellik olarak belirlenmiştir. Nasıl ki sentimentleri çıkartırken sık geçen özellikler kullanılmışsa sık geçmeyen özellikleri bulmak için de sentimentler kullanılmıştır. Sentimentin pozitif ya da negatif duygu verip vermediğinin tespiti de WordNet üzerinden iteratif bir yöntem ile yapılmıştır.

Hu ve Liu'nun [2] aynı yılda yaptıkları ikinci çalışma da ilk çalışma ile aynı olup ürünler üzerine yapılan yorumlardan özet çıkarmayı amaçlamaktadır. Buradaki sistem 3 temel adımdan oluşmaktadır. İlk adım ürün özelliklerinin çıkartılması adımıdır. Burada da birliktelik kuralları madenciliği kullanılmış, çıkartılan özellikler aday özellik olarak kabul edilip gereksizlerin budanması ve kompaktlık budama ile filtreleme gerçekleştirilmiştir. Bir sonraki adımda, ilk adımda elde edilen ürün özelliklerinin bir ve ya daha fazlasının geçtiği kişilerin görüşlerini içeren cümlelerin bulunması amaçlanmıştır. Cümlelerin pozitif ya da negatif duygu içerip içermediğini belirlemek için ise üç alt adım gerçekleşmiştir. Bu alt adımlar; sıfatların sık geçen özelliklerin kullanımı ile çıkartılması ve çıkartılan sıfatlara bağlı olarak da sık geçmeyen özelliklerin çıkartılması; iteratif yöntem ile her sentimentin pozitif ya da negatif duygu vermesinin tespit edilmesi ve cümlenin polaritesinin bulunmasıdır. Temel adımların sonuncusu ise özetlemedir. Bu özetlemede ilk çalışmadaki gibi bir özetleme olup yapısal özetleme olarak isimlendirilmiştir.

Popescu ve Etzioni [3] tarafından gerçekleştirilen çalışmada, kullanıcı yorumlarından belirgin ürün özelliklerini çıkarmak amacıyla OPINE danışmansız bilgi çıkarım sistemi tasarlanmıştır. OPINE girdi olarak aldığı temel varlık ve kullanıcı yorumlarına göre aşağıdaki 4 işlevi gerçekleştirmektedir. Aday özellikler PMI (Pointwise Mutual Information) kullanılarak filtrelenmiştir. Bu filtreleme işi (PMI hesabı) temel varlık ve alt parçaları (meronym) arasındaki ilişkiye dayalı olarak gerçekleşmiştir. Ürün özelliklerine dayalı olarak sentimentlerin çıkartılması işleminde 10 adet kural belirlenmiş ve eğer cümlede belirgin ürün özeliği varsa bu kurallar cümleye uygulanarak sentimentler belirlenmiştir. Sentimentin polaritesini belirlemede “relaxation labeling” adında özel bir yöntem kullanılmıştır.

Brody ve Elhadad [5] önerdikleri sistem ile yorumlar üzerinden özellik çıkaran ve sentiment belirleyen danışmansız bir yöntem gerçekleştirmişlerdir. Özellikleri çıkarmak için yerel LDA (Local Latent Discriminant Analysis) yöntemi kullanılmıştır. Cümle seviyesinde çalışan ve özellikleri otomatik olarak çıkarmak için az sayıda “konu” ihtiyacı duyan bu yöntem “yerel konu modeli” olarak isimlendirilmektedir. Çıkartılan özellikler için temsil edici sözcükleri bulmada ise “paylaşılan bilgi” (Mutual Information) kullanılmıştır. Mesela “yemek” özelliğinin temsil edici sözcüklerine bakıldığında “menü, taze, suşi, balık, şef, kuzine” nin elde

edildiği gözlenmiştir. Sentimentleri çıkarırken ise her özellik ile ilişkili sentimentler olumsuzlama ve bağlaçlar dikkate alınarak belirlenmiştir; daha sonra da bu sentimentlerden bir çekirdek küme (polariteleri ile birlikte) otomatik olarak oluşturulmuştur. Geriye kalan sentimentlerin polaritelerinin belirlenmesi için de bir bağlantı grafindan (Conjunction Graph) yararlanılmıştır. Çekirdek küme ve bağlantı grafi kullanılarak polaritesi bilinmeyen sentimentler üzerine yayılım yapılarak onların da polaritesi belirlenmiştir. Sentiment polaritesinin belirlenmesinde özelliklerin etkisi de göz önünde bulundurulmuştur.

Wei ve arkadaşları [5] ürün özelliklerini semantik tabanlı çıkartan bir sistemin geliştirilmesinden bahsetmişlerdir. Buradaki sistem Hu ve Liu'nun [2] geliştirdikleri sistem üzerine kurulu olmakla birlikte mevcut sistemin eksik yönlerini gidermeyi amaçlamaktadır. Ürün özelliklerini çıkarmada birliktelik kurallarını kullanılarak çıkartılan aday özellikler üzerinde, gereksizlerin budanması ve kompaktlık budama filtreleme yöntemlerini uygulayan sistem, sık geçen olarak belirlenen ürün özelliklerini sentimentleri kullanarak iyileştirmeyi amaçlamaktadır. Bu adım "semantik tabanlı iyileştirme" (Semantic Based Refinement) olarak tanımlanmaktadır. Asıl hedef ise üç başlık altında; ürün ile ilgili olmayan özellikleri elemek, kişilerin görüş bildirmediği olmayan özellikleri elemek ve sık geçmeyen özellikleri etkin bir şekilde çıkarmak şeklinde belirlenmiştir.

Thet ve arkadaşları [6] sinema filmleri ile ilgili yapılan yorumlardaki özellikleri çıkartarak bu özellikler üzerinden sentiment polaritesinin ve sentiment etkisinin başarılı bir şekilde belirlenmesi için yöntem önermişlerdir. Ayrıca bu çalışmada cümlecik/cümlelerin sentiment puanı belirlenmekte ve bu puan ile cümlecik/cümleler pozitif/negatif/tarafsız olarak sınıflandırılmaktadır. Sentiment puanı belirlemede ise sentiment etkisi devreye girmektedir. "Harika" ve "güzel" karşılaştırıldığında "harika" sentimentinin verdiği duygunun "güzel" sentimentinin verdiği duygudan daha baskın olduğu görülmektedir. Sentiment puanının kullanımı ile en pozitif ve negatif cümlecik/cümleler de belirlenmektedir. Bu çalışmada önerilen yöntem, özelliklerin sentiment etkisini ve polaritesini belirlemek için dilbilimsel bir yaklaşım ile cümlecik seviyesinde sentiment analizini gerçekleştirmektedir. Cümledeki her kelimenin sentiment puanını belirlemek için hem alan tabanlı hem de genel sentiment sözlüklerinden yararlanılmıştır. Yorumları oluşturan cümlelerde yer alan kelimeler

arasındaki bağımlılığa dayanarak bir bağımlılık ağacı (dependency tree) üretilmiştir. Oluşturulan ağaç alt ağaçlardan oluşmaktadır. Alt ağaçlar cümleciklere karşılık gelmektedir ve her cümlecikte bir özellik bulunmaktadır. Her cümlecik elde edildikten sonra o cümlecğin içeriksel sentiment puanı aynı cümlecikteki sinema filmi özelliği üzerinden hesaplanmaktadır.

Alexander ve Patrick [7] danışmalı yöntem izleyerek twitter verilerinin sentiment analizini yapmışlardır. 300.000 metin yazısını twitter sayfalarından otomatik olarak toplayarak Part of Speech (POS) etiketi ve n-gram yöntemi ile özellik çıkarımını gerçekleştirmişlerdir. Literatürde Pang ve diğerleri (2002) bigram yönteminin film yorumlarının sınıflandırılmasında daha iyi sonuç verdiğini söylerken, Dave ve arkadaşları [8] bigram ve trigram yöntemlerinin birlikte ürün yorumlarının analizinde daha iyi başarımlar verdiğini raporlamıştır. Bu çalışmada da çeşitli yöntemlerin karşılaştırılması yapılmış ve en yüksek başarımlar bigram yönteminde elde edilmiştir. Sınıflandırmada Naïve Bayes ve SVM sınıflayıcılarını kullanmışlardır.

G. Qiu ve arkadaşları [9] özellik ve duygu sözcüklerini çıkarmada başlatma (bootstrapping) algoritması tabanlı bir yöntem önermişlerdir. Kullandıkları bu yöntemde bağımsal dilbilgisi tanımlamalarını belirleyerek özellik ve düşünce belirten sözcükler arasında ilişkiler belirlemişler ve bu tanımlamaları özellik ve düşünce belirten sözcükleri bulmak için kullanmışlardır. Algoritma başlangıç olarak fikir içeren sözcüklerden başlayarak belirlenen söz dizimsel ilişkiler doğrultusunda yeni özellikleri bulmaya çalışmaktadır. Yeni özellik ve ya düşünce içeren bulunmadığında da durmaktadır.

Zhang ve arkadaşları [10] tarafından Çince üzerine yapılan bir çalışmada, ünlü bir kozmetik firmasının ürünlerinin zayıf özelliklerinin, kullanıcı yorumlardan özellik tabanlı sentiment analizi ile tespit edilmesi amaçlanmıştır. Burada sunulan ve 5 adımdan oluşan sistemin en önemli özelliği ise zorlu ancak önemli bir iş olan örtülü özelliklerin de eldeki yorumlardan çıkartılabilmesidir. Sistem aşağıdaki şekilde özetlenmektedir. 1) Elde edilen kullanıcı yorumları bir ön işleme tabi tutulmuş ve aday ürün özellikleri yorumlardan elde edilmiştir. Aday özelliklerin çıkartılmasında frekans tabanlı bir yol izlenmiştir. Aday özellikler üzerinden hem belirgin özellikler hem de örtülü özellikler çıkartılmıştır. Belirgin özelliklerin çıkartılması için semantik

tabanlı anlambirim (Morpheme) yöntemi ve Hownetten (sözlük) yararlanılmıştır. Anlambirim ile sık geçen özellikler tespit edilirken Hownet ile sık geçmeyen özellikler tespit edilmiştir. 2) İkinci adımda elde edilen belirgin ve örtülü özellikler ilgili özellikler altına semantik yöntemlerin kullanılması ile gruplandırılmıştır. 3) Her cümle için o cümlede bulunan özelliğin sentiment polaritesi belirlenmiştir: Cümle seviyesinde sentiment analizi ile özelliklerin polaritesi belirlenmiştir. Bu aşamada pozitif ve negatif sentimentlerin, oldukça önemli olan zarflar ve dereceleri, olumsuzlamalar ve ama bağlacı göz önünde bulundurularak analizi gerçekleştirilmiştir. 4) Belli bir ürün için o ürüne ait yorumlardan zayıflıklar tespit edilmiştir: Her özellik için pozitif, negatif ve tarafsız cümleler dikkate alınmıştır. 5) Belli bir ürünü farklı markaların ürünleri ile karşılaştırarak o ürün için zayıflıkları tespit edilmiştir.

Ayoub Bagheri ve arkadaşları [11] sentiment analizi için sık geçen ve örtülü olan özelliklerin çıkarımının çok önemli bir adım olduğunu belirtmiş ve bu aşama için danışmansız ve alandan bağımsız bir yöntem önermiştir. Çalışmanın bir önemli özelliği de birden fazla sözcük içeren ve çok terimli özellik (multi-word aspect) olarak tanımlanan özelliklerin bulunmasıdır. Model kullanıcı yorumlarını giriş olarak özellik listesini çıkış olarak vermektedir. İlk adım olarak sezgisel kurallar işletilerek aday özellikler bulunur. Listeyi genişletmek için bootstrapping algoritmasını kullanmışlardır. Başlangıç çekirdek özellikler içinde önerdikleri skorumla yöntemi ile en yüksek puanlı aday özellikleri seçerek algoritmaya sokmuşlardır ve daha sonra algoritmanın işletilmesi ile aday listeyi genişletmişlerdir. Genişletilen listenin içindeki gereksizlerin elenmesi için budama yönetmelerini uygulayarak nihai listeyi elde etmişlerdir. Çok terimli özelliklerin (multi-word aspect) bulunması için de genelleştirmiş FLR (Frequencies and Left and Right of the current word) yöntemini kullanarak birden fazla sözcük içeren çok terimli özellikleri bulmuşlardır.

V.K. Singh ve arkadaşları [12] Hintçe yazılmış film yorumlarından doküman ve özellik seviyesinde sentiment analizinde kullanılacak sentiwordnet tabanlı bir model geliştirmişlerdir. Detaylı araştırmalar sonucunda duygu içeren sözcüklerin ya sıfat olarak ya da sıfat+zarf olarak bulunduğunu keşfetmişler ve bunların skorlamasının sentiwordnet üzerinden yapıldığı bir yapı tasarlamışlardır. Kurulan bu modeli ilk

doküman seviyesinde uygulamışlar, daha sonra özellik seviyesinde kullanmışlardır. Özellikler el ile geniş çapta bir araştırma sonucunda çıkarılmıştır. Özellikler çıkarıldıktan sonra yorumlar cümle bazında bölünerek özelliğin yeri cümlede keşfedilmiştir. Bulunan özelliğin 5-gram öncesine ve sonra bakılarak belirlenen model çerçevesinde sıfatlar ve ya sıfat+zarf birliktelikleri bulunmuş daha sonra sentiwordnet üzerinde polarite hesaplanması gerçekleştirilmiştir. Bulunan sonuçlar Alchemy API yazılımı ile bulunan sonuçlar ile karşılaştırılarak raporlanmıştır.

Wanxiang Che ve arkadaşları [13] özellik tabanlı sentiment analizinin daha iyi sonuç vermesi için analiz aşamasından önce cümleden gereksiz bilgileri silen “Sent_Comp” adında cümle özetleme metodu geliştirmişlerdir. Kullandıkları veri kümesi Çince yazılmış ürün kullanıcı yorumlarıdır ve sistem alan bağımsız olarak tasarlanmıştır. Bu çalışmaya göre sentiment analizi probleminde sözdizimsel özelliklerin çözülmesi büyük sorun yaratmaktadır ve bu cümle özetleme metodu bu sorundan kaynaklı geliştirilmiştir. Klasik otomatik cümle özetleme çalışmaları en kısa dilbilgisel cümleye erişmeye çalışırken; bu çalışmadaki özetleme methodu duygu içeren sözcükler gibi sentiment bilgilerinin cümlede kalmasını sağlar. Özetleme methodu için “discriminative conditional random field” modeli uygulanmıştır. Çalışmanın doğruluğunu test etmek için [9] çalışmasındaki özellik ve polarite sözcüklerini çıkarmak amacıyla kullanılan “double propagation” algoritmasını kullanmışlardır ve %75 doğruluğa ulaşmışlardır.

Salud M. Jime'nez-Zafra ve arkadaşları[14] verilen bir alana bağlı özelliklerin ve sentimentlerin çıkarılmasını hedefleyen bir danışmasız yöntem önermişlerdir. Özelliklerin bulunması için 12 milyondan fazla varlığı ilişkisel model ile kaydeden herkesin erişimine açık veritabanı Freebase [15] yazılımını kullanmışlardır. Polaritelerinin belirlenmesi içinde sözlük tabanlı yaklaşım kullanmışlar ve en çok bilinen üç sözlük üzerinde çalışmışlardır.

Türkçe üzerinde çok sayıda sentiment analizi çalışması yer almadığından hazır derlem ve ya veri kümeleri bulunmamaktadır ve daha çok makine öğrenmesi yöntemleri ile çalışılmıştır.

İlk çalışma Eroğlu ve arkadaşları [16] tarafından film yorumları üzerinde yapılmıştır. Bu çalışmada makine öğrenmesi algoritmalarından olan SVM, doğal dil işleme

teknikleri ve N-gram yöntemi ile modelleme yapılmıştır. Yorumlar olumlu ve ya olumsuz şekilde etiketlenerek %85 sistem başarımı elde edilmiştir.

Boynukalın [17] İngilizce veri kümesini manuel olarak Türkçeye çevirip ve verileri etiketleyerek makine öğrenmesi yöntemlerini uygulamış ortalama %80 başarımla elde etmiştir.

Vural ve arkadaşları [18], yaptıkları çalışmada sosyal medya siteleri üzerinden elde ettikleri sinema yorumları için danışmansız sentiment analizi gerçekleştiren bir sistem tasarlamışlardır.

Meral ve Diri çalışmalarında [19] yaklaşık 8.500 adet twitter metnini manuel olarak etiketleyip makine öğrenmesi yöntemlerinden Rastgele Orman (Random Forest), SVM ve NB (Naïve Bayes) sınıflayıcılarıyla duygu Analizi yapmışlardır. Korelasyon tabanlı öznelik seçimi yaparak %90 civarında başarı elde etmişlerdir.

Kaya ve arkadaşları [20] yaptıkları tez çalışmasında İngilizce ve Türkçe dilinde topladıkları haber yorumları üzerinde makine öğrenmesi algoritmaları ile transfer öğrenme yöntemini uygulayarak duygu analizi yapmıştır ve ortalama %77 başarımla sağlamıştır.

Akba ve arkadaşları [21] Türkçe film yorumları üzerinde öznelik seçme yöntemleri ile makine öğrenmesi metodlarından SVM ve NB algoritmalarını uygulayarak duygu çıkarımı yapmışlardır. Olumlu ve ya olumsuz olarak etiketlenen film yorumları ile yaklaşık %84 başarımla elde etmişlerdir.

Balahur ve arkadaşları [22] İngilizce tweetlerden oluşan veri kümesini makine çevirisi başka dillere çevirerek; hazır makine öğrenmesi ve öznelik seçme yöntemlerini uygulayarak analiz gerçekleştirmişlerdir. Türkçe duygu analizi sonucunda elde edilen başarı %60'a yakındır.

Akbaş [23] yaptığı çalışmada twitter metinleri üzerinde çeşitli ön işleme metodlarından sonra ürünler hakkındaki alt konular çıkarmışlar ve bu konulara göre veriyi gruplamışlardır. Elle oluşturulan duygusal kelime listesi kullanılarak öznelik vektörü oluşturulmuştur. Çalışmada makine öğrenmesi yöntemiyle pozitif ve ya negatif olarak sınıflandıran model yaklaşık %85 başarımla elde etmiştir.

Özsert ve Özgür [24] yaptıkları sözcüklerin polaritesinin belirlenmesinde çok dilli bir yaklaşım önermişlerdir. Wordnet'teki sözcük ilişkilerini kullanarak bir graf yapısı oluşturmuşlardır. İngilizce çekirdek kelimelerden yarı otomatik bir yöntem ile Türkçedeki negatif ve pozitif duygu içeren çekirdek sözcükleri tespit etmişlerdir. Tespit ederken elde ettikleri başarıyı yaklaşık %90 civarındadır.

Çetin ve Amasyalı [25] iki farklı Twitter verisi üzerinden sentiment analizini gerçekleştirirken terim ağırlıklandırmada danışmalı yöntemlerin klasik danışmasız yöntemlere oranla başarılı olduğunu göstermeyi amaçlamıştır.

1.4. Tezin Yapısı

Tez çalışması beş bölümden oluşmaktadır. Tezin girişi niteliğinde olan birinci bölümde gerçekleştirilen çalışma tanıtılmış, amacı ve öneminden bahsedilmiştir. Konu ile ilgili yapılan çalışmalara dair literatür taraması giriş bölümünün son kısmında verilmiştir. Bölüm 2'de duygu analizine dair kısa bir açıklama verilmiş ve özellik tabanlı duygu analizine geçiş yapılmıştır. Bölüm 3'de ise tez çalışmasının temelini oluşturan ürünlerin zayıf ve güçlü özelliklerinin kullanıcı yorumlarından keşfedilmesinden bahsedilmiştir. Çalışmada önerilen yöntemler detaylandırılarak bölüm 3 sonlandırılmıştır. Bölüm 4'de uygulanan sistemden ve deneysel sonuçlar açıklanmıştır. Bu bölümde çalışma alt başlıklar şeklinde detaylandırılmış ve her aşama anlatıldıktan sonra sonuçları verilmiştir. Son bölüm olan bölüm 5'de tez çalışması değerlendirilmiş ve ileriye dönük yapılabilecek olan çalışmalardan bahsedilmiştir.

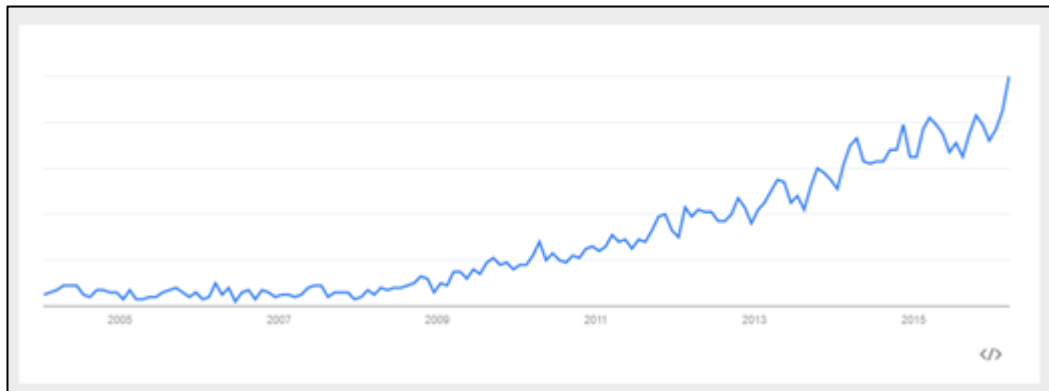
2. GELİŞTİRİLEN SİSTEM ve TEORİK ALTYAPISI

Bu bölümde tez kapsamında geliştirilen sistem ve teorik altyapısına dair detaylar verilmektedir.

2.1. Sentiment analizi nedir?

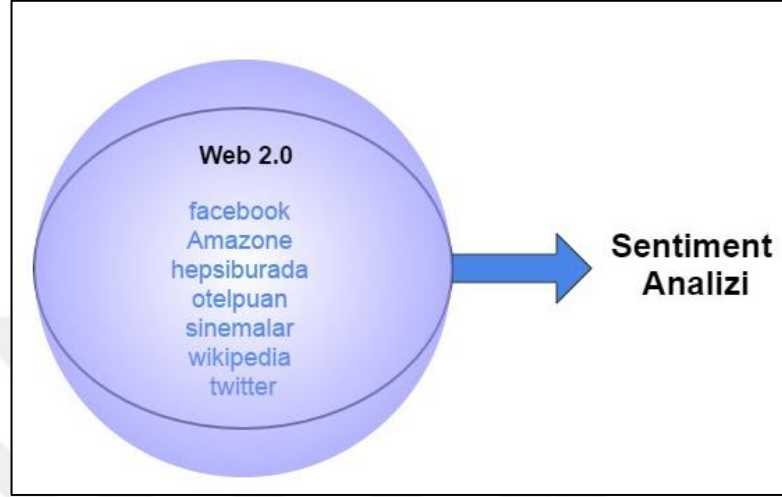
19. yüzyılın sonlarına doğru doküman sınıflandırma yöntemleri sayesinde belge, metin gibi benzer içerikli dokümanlar alt sınıflara ve ya kategorilere algoritmik olarak ayrılabilir hale gelmiştir. Dokümanların sınıflandırılması çalışmaları yaygınlık kazandıkça, dokümanların analizinde daha derinlemesine analiz gerçekleştirebilen paragraf, cümle ve hatta cümlelerin bileşenleri seviyesinde analize olanak veren pek çok farklı yöntem ortaya çıkmıştır. Her geçen gün yeni yöntemlerin çıkmasıyla birlikte bilgiye erişimde en çok kullanılan yöntemlerden olan metin sınıflandırma yöntemleri daha da olayı indirgeyerek bir paragrafın ya da bir cümlelerin ana fikrini çıkarabilecek kadar gelişmiştir.

2000’li yıllarda ise araştırmacılar dokümanların sadece fikir içermediğini duygu da içerebileceğini düşünerek dokümanların özelliklerini de araştırmaya başlamışlardır. Dokümanların genel duygusunu keşfettikten sonra cümlelerin, daha da derinleşerek ifadelerin duygularını belirleme yoluna gitmişlerdir [26]. Sentiment (duygu) analizi bir belgenin, cümlelerin ve ya ifadenin duygusunun çıkartılıp, yorumlanması işidir. Büyüyen Web ile birlikte oldukça popüler bir araştırma alanı haline gelmiştir.



Şekil 2.1. Sentiment analizinin Google Trends grafiği [68]

Bu büyümeye paralel olarak, insanlar duygularını Web 2.0 tarafından desteklenen internet forumları, bloglar, sosyal paylaşım siteleri gibi alanlarda kullanıcı tarafından yaratılan içerikler paylaşarak, uygulamalar için çok büyük veri yığınları oluşturmuşlardır [1].



Şekil 2.2. Sentiment analizinin kavramsal modeli [27]

Bu içerikler sayesinde ürün satın almak isteyen bir kişi, başka bir insana sormak yerine, bu bilgileri kullanarak ürün hakkında yorum sahibi olabilmektedir. Ve ya çoğu şirket piyasaya sürdüğü bir ürün hakkında anket yapmak yerine, bu içerikleri toplayarak gerekli bilgi çıkarımını yapabilmektedir. Ancak bu kadar büyük veri yığnında duygu içeren kaynakları bulmak ve bu verileri Web’ de taramak oldukça zorlu bir süreçtir. Birçok durumda, duygular kullanıcılar tarafından üretilen devasa boyutlu ve doğal dilde yazılmış içeriklerde gizli durumdadır. Çok yüksek boyutlardaki veri kaynaklarından gerekli olan bilginin çıkarılması, duygu ile alakalı olan cümlelerin yakalanması, okunup özetlenmesi gibi işlemler insan gücü ile yapılması çok zor olan işlemlerdir. Sonuç olarak bunların hepsini otomatik olarak gerçekleştiren bir sistem modellemek zorunludur. Bu ihtiyaçların ve zorunlulukların sonucunda, temelinde doğal dil işleme ve metin madenciliği süreçlerinin var olduğu sentiment analizi bir çalışma alanı olarak doğmuştur ve gittikçe önem kazanmıştır [28].

2.2. Sentiment Analizi Teknikleri

Sentiment analizi alanında yapılan güncel bilimsel çalışmalarda iki temel yaklaşım yer almaktadır. Bunlar makine öğrenmesi ile sentiment analizi ve sözlüksel benzeşim ile sentiment analizi yaklaşımlarıdır.

Metin sınıflandırma tekniklerini kullanan makine öğrenmesi yöntemlerinde, makine öğrenmesi algoritmaları ve dilbilimsel özellikler kullanılmaktadır. Makine öğrenmesi yaklaşımı da kendi içinde danışmalı ve danışmasız öğrenme yöntemleri olmak üzere iki kategoride incelenmektedir. Danışmalı öğrenme yöntemi önceden gözlemlenmiş ve sonuçları bilinen (etiketlenmiş) verileri kullanarak, bu verileri ve sonuçları içeren fonksiyonu modellemeyi amaçlayan öğrenme metotlarıdır. Danışmasız öğrenme, danışmalı öğrenmeden farklı olarak etiketlenmemiş veriden gizli kalmış örüntülerin keşfedilmesine dayanan metotları içermektedir.

Sözlüksel benzeşim ile sentiment analizinde duygusal anlam taşıyan sözcüklerin bulunarak, bu sözcüklerle metnin analiz yapılması söz konusudur. Bu yaklaşım iki yöntem içermektedir. Sözlük tabanlı yaklaşım, önceden belirlenen çekirdek duygu içeren sözcüklerin, belirli bir sözlükte eş anlamlarının ve zıt anlamlarının bulunmasına dayalı yöntemleri kapsamaktadır. Derlem (corpus) tabanlı yaklaşım ise, çekirdek duygu içeren sözcüklerin yardımıyla büyük bir derlemde diğer duygu içeren sözcüklerin ve bu alana bağlı polaritelerinin bulunmasına dayalı yöntemleri kapsamaktadır [29].

Yukarıda bahsedilen bu yaklaşımlar takip eden bölümlerde detaylandırılmıştır.

2.2.1. Makine öğrenmesi ile sentiment analizi

Makine öğrenmesi yaklaşımında, metin sınıflandırmada da sıkça kullanılan NB ve SVM sınıflandırıcı gibi danışmanlı öğrenme yöntemleri kullanılmaktadır. Bu yaklaşımda temel problem, elimizde kayıtlı olan doküman kümesindeki $D = \{ d_1, d_2, \dots, d_n \}$ dokümanların önceden belirlenmiş $S = \{ s_1, s_2, \dots, s_n \}$ kümesindeki sınıflara ait olup olmadığının belirlenmesidir. Makine öğrenmesi yöntemleri ile bir model kurularak j . dokümanın i . sınıfa ait ise doğru, değil ise yanlış değerini üreten fonksiyon tasarlanmaktadır.

Makine öğrenmesi yaklaşımlarından danışmalı öğrenme tekniğiyle yapılan birçok çalışma literatürde mevcuttur. Bunlardan sentiment analizi alanındaki ilk çalışmalardan biri olan Pang ve arkadaşları [30] tarafından kurulan modelde, internetten toplanan film yorumları üzerinde NB, ME (Maximum Entropy) ve SVM metotları kullanılmıştır; en yüksek başarımlar unigram ile elde edilen özellikler üzerinde SVM uygulanması ile elde edilmiştir. Zang ve arkadaşları [31] NB ve SVM sınıflandırıcılarını kullanarak kurdukları modelde yaklaşık %96 başarımlar elde etmişlerdir. Veri kümesi olarak restoranlar hakkında yapılmış 1500 pozitif, 1500 negatif kullanıcı yorumu kullanılmıştır. Dang ve arkadaşları [32] SVM sınıflandırıcı ile farklı özellik eleme metotları kullanmışlardır. Film yorumları üzerinde Bilgi Kazancı (Information Gain) yöntemi uygulanarak, özellik uzayındaki gereksiz olanların elenmesi sağlanmıştır ve sonuç olarak kurdukları modelde ortalama %85 başarımlar elde etmişlerdir.

Literatürde danışmasız öğrenme tekniği ile yapılan çalışmalar danışmalı öğrenme tekniği kadar sık bulunmasa da uygulamaları mevcuttur. Turney tarafından yapılan çalışmada [33] POS etiketleri kullanılarak cümlenin sıfat ve zarf olan sözcükleri çıkartılmış daha sonra bu sözcüklerin sentiment polariteleri PMI yöntemi ile belirlenmiştir. Telefon yorumları üzerinde ortalama %84 başarımlar elde edilirken, film yorumları üzerinde ortalama %66 başarımlar elde edilmiştir.

2.2.2. Sözlüksel benzeşim ile sentiment analizi

Duygu içeren sözcüklerin bulunması için uygulanan manuel yöntemler yüksek maliyet ve zaman kaybına sebep olmalarından kaynaklı önerilmemektedir. Bundan dolayı bu işlemin otomatik bir şekilde modellenmesi zorunludur. Sözlüksel benzeşim yöntemlerinde iki farklı yöntem yer almaktadır. Bunlar duygu sözlüğü ve duygu derlemi yaklaşımlarıdır. Literatürde genelde bu iki yöntem birlikte kullanılarak sentiment analizi gerçekleştirilmektedir.

Duygu sözlüğü tabanlı yaklaşımda duygu sözlüğü, manuel olarak kurulan çekirdek sentimentlerin, eş ve zıt anlamları bir sözlükte bulunarak genişletilmektedir. WordNet [34] sözlüğü bu tür uygulamalarda en çok kullanılan sözlük veritabanıdır.

Duygu derlemi yaklaşımında duygu derlemi, sentiment analizi için gerekli olan duygu içeren kelime gruplarından oluşan veri yığınlarıdır. Bu derlemi oluştururken her sınıfa ait çekirdek kelimeler, her bağlam için yeniden oluşturulmaktadır. Duygu derlemi tabanlı yöntemler ile sadece bir bağlam için duygu içeren sözcük listesi oluşturulmaktadır [26]. Bu yöntemlerde çalışılan dile ait dilbilimsel özelliklerden yararlanılmaktadır. Hatzivassiloglou ve McKeown [35] çalışmalarında çekirdek olarak belirlenen sıfatları kullanarak, dilbilimsel kısıtlamalar ile duygu derlemini genişletmişlerdir. Çalışmadaki kısıtlamalar bağlaçlar olarak belirlenmiştir.

2.3. Sentiment analizi seviyeleri

Sentiment analizi doküman, cümle ve özellik olmak üzere üç seviyede gerçekleştirilmektedir. Doküman seviyesinde yapılan sentiment analizinde tüm doküman genel olarak incelenip, baskın olan polarite doküman polaritesi olarak belirlenmektedir [36, 37, 38]. Cümle seviyesinde yapılan sentiment analizi, doküman seviyesine göre daha detaylı bir analiz olup, literatürde birçok çalışması mevcuttur [39, 40, 41]. Bu analizde objektif olan cümle bulunduğundan sonra cümlenin polaritesi belirlenmektedir.

Cümlelerde birer ufak doküman olarak kabul edilebileceğinden, doküman ve cümle seviyesinde yapılan sentiment analizi arasında temel bir fark bulunmamaktadır. Doküman ve ya cümle seviyesinde yapılan sentiment analizinde, varlıkların özellikleri için kullanılan duygular dikkate alınmamaktadır. Bundan dolayı daha ayrıntılı bir analiz için özellik seviyesinde sentiment analizi gerçekleştirilmelidir.

2.4. Özellik Tabanlı Sentiment Analizi

Dokümanın ya da cümlenin tamamının negatif olarak belirlenmesi, sözü geçen ürünün her özelliğinin zayıf olduğu anlamına gelmemektedir. Daha doğru ve geçerli sentiment analizi için ürünün hangi özelliklerinin negatif ve ya pozitif olduğu belirlenmelidir. Diğer bir deyişle cümledeki duygunun etkin şekilde kullanılması ancak cümledeki olumlu/olumsuz duygu içeren özelliklerin belirlenmesi ile mümkündür. Örneğin “otelin odaları gayet temizdi ancak gündüzleri girdiğimiz barın önündeki havuz iyi değildi” kullanıcı yorumu “oda” özelliği için olumlu, “havuz” özelliği için ise olumsuz duygu taşımaktadır. Burada cümlenin tamamının duygusunu

çıkarmak yerine özellikler için duyguları çıkarmak daha kullanışlı bir sistem tasarlanmasını sağlamaktadır.

2.4.1. Problem tanımı

Duygu (sentiment) analizi ve ya düşünce madenciliği (opinion mining) bir cümlenin, dokümanın ya da bir varlığa ait özelliğin taşıdığı duygunun tespit edilip, analiz edilmesi işlemine verilen addır. Özellik tabanlı sentiment analizi ise sentiment analizinin özellik seviyesinde yapılan analizi olup, dokümanda geçen özelliklerin yansıttığı duyguyu (olumlu/olumsuz) yakalama işlemidir. Özellik tabanlı sentiment analizinin problem tanımı aşağıda detaylı bir şekilde verilmiştir.

Bir otel yorumu örneği ile özellik tabanlı sentiment analizi problemi ele alınırsa [42];

Örnek:

(1) merhaba geçen sene bu otele gittik. (2) genel olarak otel güzeldi. (3) otelin odaları gayet temizdi. (4) ancak gündüzleri girdiğimiz barın önündeki havuz iyi değildi.

Bu örnekte sorulması gereken asıl soru: yorumdan hangi bilgi çıkarılmak istenmektedir? Öncelikle yorumda birçok duygu içeren sözcük bulunmaktadır. (2) ve (3) numaralı cümleler olumlu duygu, (4) numaralı cümle ise olumsuz duygu içermektedir. (2) numaralı cümlede bütün otelden olumlu bahsedilirken (3) numaralı cümlede otelin odasından olumlu bahsedilmektedir. (4) numaralı cümlede ise otelin havuzundan olumsuz bahsedilmektedir. Sonuç olarak bu problem, duyguyu içeren ve bu duygunun nitelediği sözcüklerin bulunmasına odaklıdır. Problemin daha iyi anlaşılması için aşağıda bazı önemli tanımlara yer verilmektedir.

Varlık (entity): Varlık (v) bir ürün, servis, kişi, olay, vb gibi herhangi başlık olarak tanımlanabilmektedir. Varlığı v: (T, A) çiftleri ile temsil edersek; bu çiftteki T, bir bileşenin ve bu bileşene bağlı alt bileşenleri gösteren hiyerarşiyi, A ise bu varlığa ait özellik kümesini ifade etmektedir. Bu tanıma bağlı kalarak, varlık bir ağaç yapısı ve ya hiyerarşik bir düzen şeklinde temsil edilebilmektedir. Bu yapıda her düğüm, varlığa ait bir bileşen ve ya alt bileşendir ve her bir düğüm de özellik kümesiyle

ilişkilidir. Yukarıdaki örnekte; otel temel düğümü ifade ederken, odalar ise bu düğüme bağlı bir özelliği ifade etmektedir.

Özellik (aspect): Özellikler duygunun nitelediği başlıklardır ve varlığın bir alt bileşenidir [11]. Bir ve ya birden fazla kelime şeklinde temsil edilebilmektedirler. Genellikle de isim şeklinde kullanılmaktadırlar. “otelin odaları gayet temizdi” cümlesindeki “oda” sözcüğü otel temel varlığı için özelliği temsil etmektedir. Türkçe dil bilgisi kurallarına göre de cümledeki sıfatın nitelediği isimler özellik olarak temsil edilmektedir. Özellikler iki biçimde tanımlanmaktadır [42]: eğer cümlede belirli durumda iseler belirgin (expilicit), gizli ise örtülü (implicit) olarak adlandırılmaktadırlar [10]. Tez kapsamında sentiment çıkarımı sadece belirgin özellikler temel alınarak gerçekleştirilmiştir.

Duygu taşıyıcı (sentiment holder) : Duyguyu belirten kişi ve ya kurumlar duygu taşıyıcı olarak temsil edilmektedir. Duygunun kaynağı olarak da adlandırılan duygu taşıyıcı, kişilerin özel düşüncelerini belirttiği haber kaynaklarında önem taşımaktadır. Örneğin; “Ali yapılan anlaşmanın güvenilir olduğunu belirtti” cümlesinde “Ali” kişisi duygu taşıyıcı görevindedir.

Duygu (Sentiment): Cümlede duygu taşıyıcı tarafından belirtilen, bir özelliğin duygusunu içeren öğelerdir [43]. Duygu direkt ve karşılaştırmalı olarak iki biçimde tanımlanmaktadır. Tez kapsamında direkt duygular ile çalışılmıştır. Direkt duygu $(v_j, \bar{o}_{jk}, sp_{ijk}, h_i, t_l)$ beşli bir fonksiyon şeklinde temsil edilir. Bu beşlide v_j temel varlığı, $\bar{o}_{jk}$ bu temel varlığa ait özelliği, $[sp]_{ijk}$ sentiment polaritesini, h_i duygu taşıyıcısını, t_l duygu taşıyıcı tarafından duyguyu belirtme zamanını ifade etmektedir. Yukarıdaki örneğe göre duyguyu gösterirsek, “genel olarak otel güzeldi” cümlesindeki güzel sözcüğü duyguyu nitelemektedir. Türkçe dilbilgisinde sıfatlar duyguyu temsil etmektedir.

Duygu yönelimi (opinion orientation): Duygu yönelimi, diğer bir deyişle sentiment polaritesi özelliğin taşıdığı duygunun pozitif ve ya negatif yöneliminde olup olmadığını belirtmektedir.

Yukarıdaki tanımları kullanarak özellik tabanlı sentiment analizi problemini tanımlarsak;

Varlık (v), özelliklerden oluşan $A=\{a_1, a_2, \dots, a_n\}$ kümesi şeklinde temsil edilebilir. Bir dokümanda varlık kümesini $\{v_1, v_2, \dots, v_q\}$ ve duygu taşıyıcı kümesinden $\{h_1, h_2, \dots, h_q\}$ kümesini içerdüğinden dolayı, özellik tabanlı sentiment analizi problemi dokümanda ki bütün duygu beşlilerinin $(v_j, \bar{o}_{jk}, sp_{ijk} h_i, t_l)$ bulunmasına dayanmaktadır.



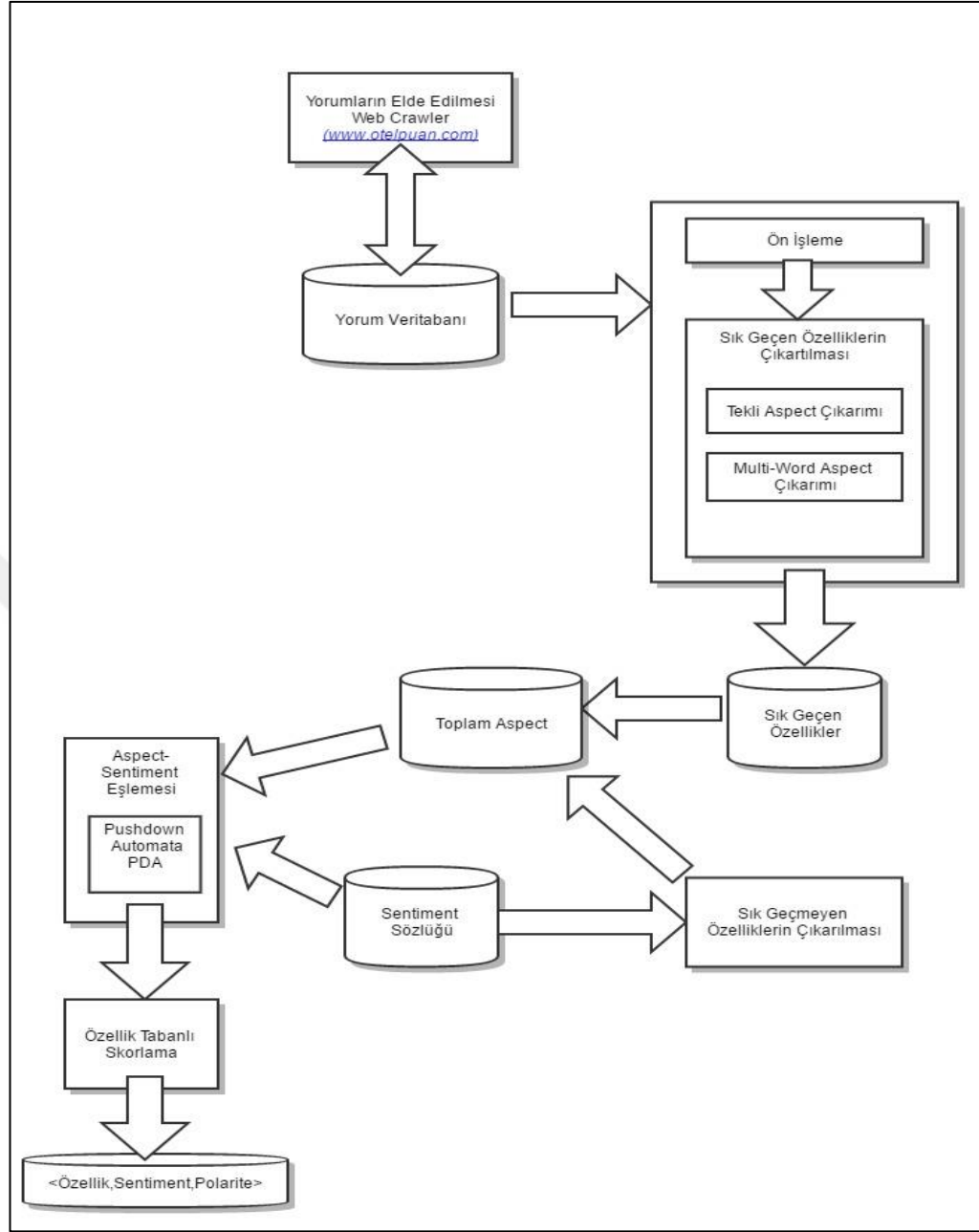
3. GELİŞTİRİLEN SİSTEM

Tez kapsamında geliştirilen sistem bu bölümde detaylandırılmıştır. Sistem mimarisi verilerek, her birimde kullanılan yöntemler açıklamasıyla verilmiştir. Sistemi oluşturan birimler birer problem olarak ele alınmıştır ve literatürde yapılan temel çalışmalara da yer verilerek detaylandırılmıştır.

3.1. Sistem Mimarisi

Geliştirilen sistem 6 temel birimden oluşmaktadır. İlk adım, veri kümesinin oluşturulmasıdır. Herkesin erişebileceği Türkçe kullanıcı yorumlarından oluşan veri kümeleri bulunmadığından dolayı sistemde kullanılacak veri kümesi otomatik bir şekilde oluşturulmuştur. Daha sonra veri kümesi üzerinde etkisiz sözcüklerin temizlenmesi, kök bulma gibi doğal dil işlemedeki ön işleme teknikleri kullanılmıştır. 2. ve 3. adımda metinden sık geçen özelliklerin çıkarımı gerçekleştirilmiştir. Sık geçmeyen özellikler 4. adımda keşfedilerek metinde geçen toplam özellikler çıkartılmıştır. 5. adımda özellik-sentiment çiftleri bulunmuş son adımda bu çiftler üzerinden özellik tabanlı skorelama gerçekleştirilmiştir.

Bu birimlerde gerçekleştirilen adımlar alt başlıklar halinde aşağıda ayrıntısıyla anlatılmaktadır.



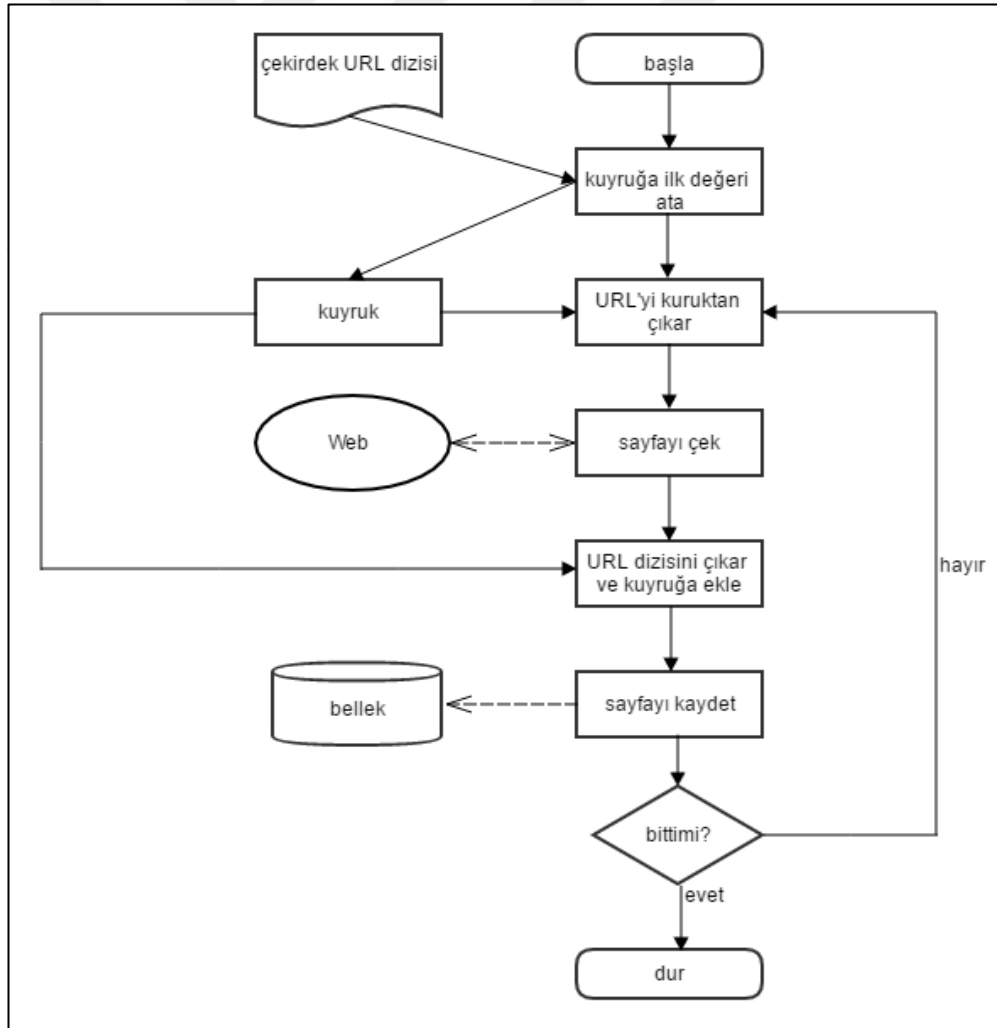
Şekil 3.1. Geliştirilen sistem mimarisi

3.1.1. Veri toplama ve ön işleme

Özellik tabanlı sentiment analizi, film yorumlar [6], dijital kamera [2], bilgisayar [4] gibi elektronik ürün yorumları, ve restoran yorumları [44] gibi birçok farklı alanda uygulanmıştır. Bu tez çalışmasında ise bu projeye özel veri kümesi otomatik olarak oluşturulmuştur.

Veri kümesinin oluşturulması için www.otelpuan.com sitesindeki yorumlar kullanılmış ve bu veriler için internetten otomatik link toplayan bir web arama

sistemi (crawler) program kodlanmıştır. Web arama sistemleri, diğer bilinen adıyla “örümcek”, “robot”, “ajan” sistemleri, arama motorlarının kullanımına yönelik Web sayfalarına erişen sistemlerdir [45]. Bu sistemler genel olarak daha önceden tanımlanmış başlangıç URL (Uniform Resource Locator) girişleri ile başlar ve bu sayfaları indirir. Daha sonra her bir sayfa için o sayfanın URL bağlantılarını URL listesine ekler. Bu listeyi de kuyruk (frontier) adı verilen bir hafızada tutar. URL listesindeki bu bağlantılar sırasıyla işlenir ve sayfanın içeriği ortaya çıkarılır. Sistem, kullanıcı tarafından belirlenen sayfa sayısı kadar sayfayı gezdikten sonra sonlanabilir. Şekil 3.2’de yukarıda anlatılan bu adımlar özetlenmiştir. Web’in dinamik olarak büyümesinden dolayı, internetten bilgi toplamak amacıyla web arama sistemi tasarlamak iyi bir çözümdür. En bilinen web arama sistemi ise “Googlebot”tur.



Şekil 3.2. Bir web arama sisteminin temel adımları [45]

Tek iş parçacığı ile çalışan bir web arama sistemi, uzaktaki bilgisayar ile yapılan istek ve cevap işleminde çok vakit harcayabilir [46]. Bundan dolayı çoklu iş parçacıklı ve mümkün olduğunca yeniden düzenlenebilir bir sistem şeklinde tasarlanması gerekmektedir.

Web, düğümleri sayfalardan oluşan, düğümleri birbirine bağlayan kenarları da linklerinden oluşan bir graf gibi düşünülebilir. Bir web arama sistemi, bir ve ya birden fazla düğümden başlayarak, başka düğümlere ulaşmak için kenarları takip etmektedir.

Web'den toplanan veriler belirli işlemlerden geçmediği sürece uygulamalar tarafından kullanılamaz ve anlam ifade etmez. Bu veri türlerine yapısal olmayan veriler denir [47]. Web'den toplanan veriler yapısal olmayan veri tipindedir ve ön işleme adımından geçmediği sürece herhangi bir yöntem uygulanıp bilgi çıkarımı yapılamaz durumdadır. Bu verileri ön işleme adımlarıyla organize hala getirmek mümkündür. Belirli bir ürüne ait kullanıcı yorumların metin halinde elde edilmesinden sonra ön işleme adımlarından geçirilip önerilen yöntemlerin daha az maliyetli ve daha efektif gerçekleştirilmesi sağlanmıştır. Ön işleme adımları aşağıda detaylandırılmıştır.

İlk adımda, Web' den toplanan dokümanlardaki metin dışı işaret ve etiketlerin temizliği gerçekleştirilmektedir. Özellikle internet sayfaları diğer doküman sayfalarından farklı olarak reklam, menü sistemi gibi birçok bölümden oluşan içeriklere de sahip olduğu için bu işlem çok önemlidir. Sayfadan bu bölümlerin doğru bir şekilde tespit edilip ayrıştırılması gerekmektedir. Aksi takdirde yapılacak olan analiz işlemlerinde hatalar olacaktır. Buna ek olarak bütün metin küçük harfe çevrildikten sonra kısaltmalar ve sayılar da bu ayrıştırmanın içindedir. Bu adımda yapılan bir diğer işlemde etkisiz sözcüklerin (stopwords) metinden ayıklanmasıdır.

Gereksiz bilgilerin metinden ayrıştırılmasından sonra, türemiş ve ya ek gelerek değişmiş bir kelimenin orijinal haline (köküne) indirgeme işlemi (stemming) gerçekleştirilmektedir [48]. Kök bulma işleminden sonra dilin morfolojik yapısına göre kelimelerin etiketleme işlemi yapılmaktadır. Metinlerdeki bilgiyi organize etmek ve kolay erişilmesini sağlamak için bu adım önem taşımaktadır.

3.1.2. Sık geçen özelliklerin çıkarılması

Özellikler, bir varlığı tanımlayan, varlığı oluşturan her şeydir. Bu çalışma için özellik içermeyen cümleler sentiment analizi için hiçbir anlam ifade etmemektedir. Bundan dolayı sık geçen özelliklerin çıkarılması önemli bir adımdır.

İnternette bir ürün ile ilgili kullanıcı yorumları incelendiğinde, ürünün kendisi ile birlikte ürüne ait pek çok özelliğinde de değerlendirildiği görülmektedir. Ürünün genel olarak pozitif olarak değerlendirilmesi bu ürünün özelliklerinin de pozitif olarak değerlendirilmiş olması anlamına gelmemektedir. Ancak detaylı olarak yorumlar incelendiğinde, ürün özelliklerinin kullanıcılar tarafından farklı değerlendirildiği tespit edilebilmektedir. Örneğin bir kullanıcı yorumu “tesis güzeldi ama yemekler temiz değildi” gibi bir cümle içerebilir. Büyük bir doküman uzayından böyle bir cümlenin bulunarak bilgi çıkarılması oldukça zor bir iştir. Bundan dolayı bu bilgi çıkarımının otomatik olarak gerçekleştirilmesi için özellik tabanlı sentiment analizi bize yardımcı olacaktır. Özellik, üzerine yorum yapılan temel varlık, bu varlığın özellikleri, alt parçaları ve alt parçalarının özellikleri şeklinde genel bir anlama karşılık gelmektedir. Bunu bir ağaç yapısı şeklinde düşünecek olursak kök düğüm temel varlık yaprak düğümler de alt parçalar ve özellikler olarak karşımıza çıkmaktadır [42]. Özellikler tek bir söz öbeğinden oluşabileceği gibi birden fazla söz öbeğinden de oluşabilmektedir. Bu açıdan bu problem iki aşamalı incelenmelidir. Bu aşamalar tek sözcükten oluşan tekli özelliklerin çıkarılması ve birden fazla sözcükten oluşan çok terimli özelliklerin (multi-word aspect) çıkarılmasıdır.

Bu adımda en sık uygulanan yöntem, isim olan sözcüklerin POS (part of speech) etiketleriyle belirlenmesinden sonra doküman içerisindeki frekanslarına göre önem derecesinin belirlenmesidir [1]. Kullanıcılar önem verdikleri özelliklerden sıkça bahsetmektedirler. Bu da söz konusu frekans tabanlı yöntemin etkili olmasını sağlamaktadır. Çalışmada ki algoritmanın doğruluğunun ispatlanması için başka bir yöntem geliştirilmiştir[49]. Yapılan çalışmada bulunan özelliklerin ne derece temel varlıkla ilişkili olduğuna dayalı bir skor hesaplaması yapan Pointwise Mutual Information (PMI) yöntemiyle, gereksiz olan özelliklerin silinmesi gerçekleştirilmektedir. Blair-Goldensohn ve arkadaşları [50] tarafından sık geçen isimlerin bulunması için el ile gerçekleştirilen bir hesaplama kullanılmıştır.

Hesaplama isimlerin sentimentlerle birlikte geçmesine dair bir ağırlık hesabı gerçekleştirilmiştir. Zhu ve arkadaşları [51] ise çok terimli özellikleri çıkartmak için “C değeri” hesabına dayalı bir yöntem önermişlerdir.

Tez kapsamında, kullanıcı yorumlarından özelliklerin çıkarılmasına dair önerilen ve uygulanan yöntemler aşağıdaki iki başlık altında verilmiştir.

Sık geçen tekli özelliklerin çıkarılması: Tekli özelliklerin kullanıcı yorumlarından çıkarılması işleminin ilk adımı olarak frekans tabanlı yöntem kullanılmıştır. Ürün ile yapılmış yorum kümesi $R=\{R_1, R_2, \dots, R_n\}$ şeklinde gösterilsin. Bu R kümesi içerisinde herhangi bir R_i yorumu, ürüne ait özellikler ve bu özellikler için olumlu ya da olumsuz sentimentler içeren cümlelerden oluşsun. Özellik tabanlı sentiment analizinin yapılabilmesi için ürün ile ilgili özelliklerin bu cümlelerden çıkarılması gerekmektedir. POS etiketleriyle tutulan yorumlardan isim ile etiketlenen sözcükler uygulama esnasında belirlenen bir frekans değerine göre elenerek, frekansın üstünde kalan sözcükler $A=\{A_1, A_2, \dots, A_m\}$ aday özellik kümesine konulmuştur. Bu adımdan sonra bu kümedeki özelliklerin ne derece seçilen temel varlık (otel, cep telefonu gibi) ile ilişkili olduğunun hesaplanması için PMI kullanılmıştır. PMI doğal dil işlemede sözcüklerin temel varlıkla ilişkisinin ölçülmesinde sıkça kullanılan bir ölçüttür [51]. PMI gözlemlenen iki sözcüğün birlikte geçme olasılığı ile ayrı geçme olasılıklarının karşılaştırılmasına dair bir hesaplama gerçekleştirmektedir [11]. Bu hesaplama sonucunda aday özelliğin düşük PMI skoruna sahip olması temel varlıkla ilişkisinin düşük olduğu anlamına gelmektedir. Aday özellik kümesindeki özelliklerin PMI skorlarının ortalaması alınarak bir frekans değeri belirlenmiş ve bu değerin altında kalan özellikler aday özellik kümesinden silinmiştir. Çalışmada otel temel varlığı d ile aday özellik kümesindeki herhangi bir özellik ise A_i ile gösterilsin. Bu durumda A_i ve d arasındaki ilişki PMI yöntemi ile aşağıdaki şekilde belirlenmektedir.

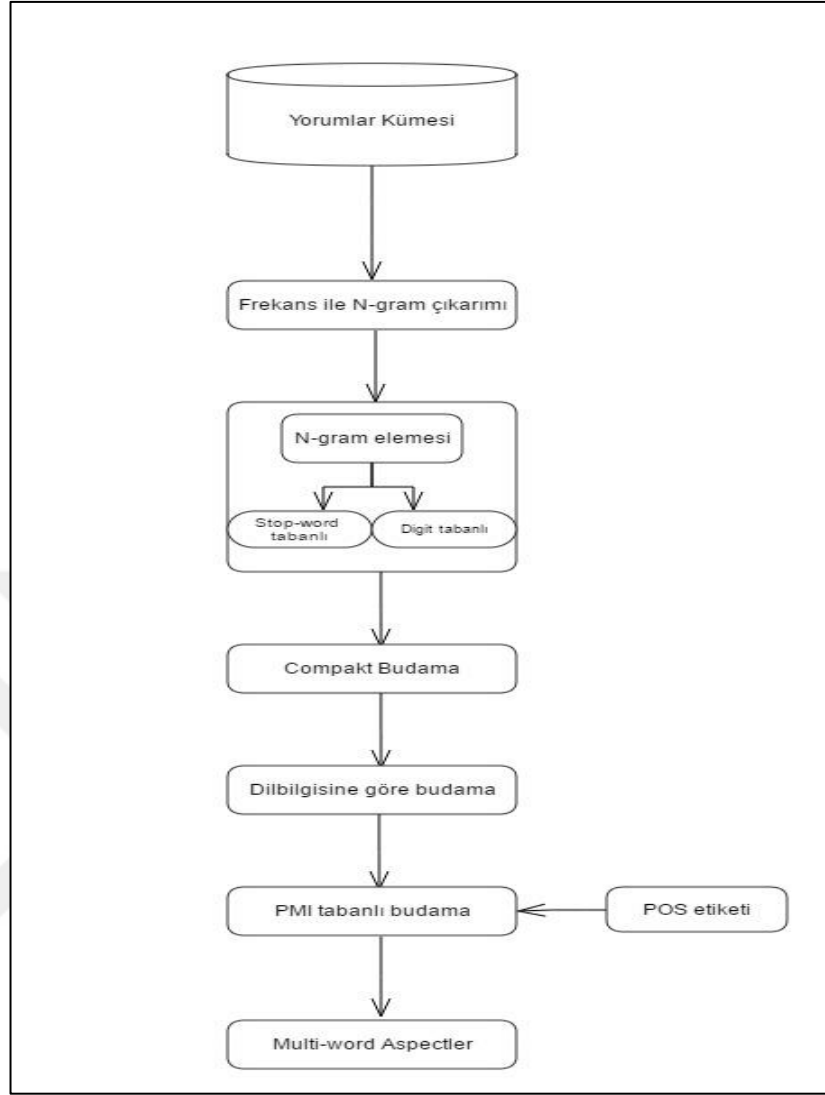
$$PMI(A_i, d) = \log \frac{\text{hits}(A_i \wedge d)}{\text{hits}(A_i) \text{hits}(d)} \quad (3.1)$$

Yukarıdaki Denklem (1.1)’de $\text{hits}(A_i \wedge d)$, A_i özelliği ile d temel varlığının Web’de birlikte geçme sayısını verirken, $\text{hits}(A_i)$ A_i özelliğinin, $\text{hits}(d)$ ise d temel varlığının

tek başına Web’de geçme sayısını vermektedir. PMI skoruna göre aday özelliklerin filtrelmesi sonucunda, kullanıcı yorumlarından tekli özellikler elde edilmiştir.

Sık geçen çok terimli özelliklerin çıkarılması: Bazen varlıklara ait özelliklerin, belirtimi birden fazla kelimenin oluşturduğu kelime öbekleri ile yapılmaktadır. Bu kelime öbeklerine Türkçe’ de tamlamalar denmektedir. Kullanıcılar bir ürün ile ilgili özelliklerden bahsederken bu tamlamalara sıklıkla başvurumaktadırlar. Örneğin; otel ile ilgili yorum kümesinde “otelin konumu”, “köy kahvaltısı” gibi tamlama şeklinde kurulan özellikler yer alabilmektedir. Bu tür özellikler çok terimli özellik (multi-word aspect) olarak isimlendirilmektedir. Kullanıcı yorumlarından ürün ile ilgili özelliklerin tam olarak çıkarılabilmesi için tamlama şeklinde kurulan çok terimli özelliklerin belirlenmesi gerekmektedir.

Tez kapsamında çok terimli özelliklerin çıkarılması için izlenen akış şeması Şekil 3.3’de verilmektedir.



Şekil 3.3. Çok terimli özellik çıkarımının akış şeması

Ön işlemeden geçirilmiş kullanıcı yorumlardan, kelime n-gramların frekansları ile birlikte çıkartılması gerçekleştirilmiştir. Çıkartılan n-gramlar arasındaki etkisiz kelimeleri ve rakam tabanlı olanlar filtrelendikten sonra kompaktlık budaması adımı uygulanmaktadır. Kompaktlık budaması adımı sonucunda büyük ölçüde indirgenmiş olan çok terimli özellikler sezgisel kuralların uygulanması ile yeni bir budamadan geçirilmekte olup; son adımda temel alan ile ilişkili olan çok terimli özelliklerin seçilmesi için PMI yöntemi uygulanmaktadır. Aşağıda bu adımların detaylı açıklamaları verilmiştir.

N-gram tabanlı yöntemler, kolay uygulanmaları ve etkili olmaları nedeniyle doğal dil işleme problemlerinde sıklıkla tercih edilmektedir. Yöntem dilden bağımsız olması

ve bu bağlamda çok fazla dilbilgisi kurallarına bağlı kalmamasından dolayı başarılı sonuçlar vermektedir.

N-gramlar, dokümanda birlikte geçen N elemanlı grupların dizisi olarak adlandırılmaktadır. Bu elemanlar çalışmanın amacına göre seçilen; dokümanı oluşturan karakterler (harf, noktalama, boşluk vs.), kelimeler, POS etiketleri, ya da birbiri ardına bulunabilecek diğer özellikler olabilirler. Literatürde kelime n-gramlar, birden fazla sözcük içeren kelime gruplarının bulunmasında kullanılmaktadır [52,53].

N-gramları bir örnek cümle üzerinden $N=2$ olacak şekilde açıklayacak olursak:

Cümle: “Odalar çok temizdi”

Karakter N-gramlar: “Od”, “da”, “al”, “la”, “ar”, “n_”, “_ç”, ..., “zd”, “di”

Kelime N-gramlar: “odalar çok”, “çok temizdi”

Doğal dil işleme uygulamalarında sıklıkla başvuru alan n-gram tabanlı yöntemler, oluşturulan n-gramların doküman içerisindeki frekansına bağlı bir model sunmaktadır. Bu çalışmada kelime n-gramlar kullanılmıştır. Yöntemin kullanılmasındaki amaç ise belli bir yorum kümesindeki çok terimli özelliklerin elde edilmesidir. Türkçedeki çok terimli özellikler incelendiğinde genellikle iki kelimeden oluştukları gözlenmiştir. Bu nedenle çalışmada n değeri 2 olarak seçilmiştir. Tüm yorumlar içerisindeki ikili birliktelikler frekansları ile birlikte çıkartılmıştır.

N-gram yöntemiyle frekansları yakalanan ikili birlikteliklerin içerisindeki gereksiz olan birlikteliklerin elenmesi için; bir sonraki adımda kompakt budama yöntemi uygulanmıştır [1, 2]. Bu yaklaşıma göre n-gram ile belirlenen aday çok terimli özellikler cümlelerde peş peşe geçme sayısı, belirlenen eşik frekansın altında kalıyorsa; bu birliktelikler listeden silinerek budama işlemi yapılmaktadır. Deneysel çalışmalar esnasında farklı temel alanlardaki test sonuçlarına göre söz konusu olan eşik frekansı üç olarak belirlenmiştir. Bu frekansın üstünde kalan aday çok terimli özellikler bir sonraki adıma giriş olarak uygulanmaktadır.

Kompaktlık budama işleminden çıkan liste, belirlenen sezgisel kurallar ile tekrardan budama işleminden geçirilerek aday çok terimli özellikler listesinin karmaşıklılığı büyük ölçüde azaltılmıştır. Sezgisel kurallar, problem uzayının çok büyük olduğu zaman kesin bir biçimde sınırlamak ve ya sadeleştirmek için doğal dil işleme uygulamalarında sıkça kullanılmaktadır. Türkçe dilbilgisine göre söz öbekleri belirli kurallar çerçevesinde oluşmaktadır ve bu dilbilgisi kuralları sezgisel kurallar olarak tanımlanıp aday çok terimli özellik listesine uygulamak süratiyle; büyük boyuttaki liste daha küçük boyutlu bir listeye indirgenebilmektedir. Sezgisel kuralları belirlemek için Türkçe'deki bu söz öbeklerinin oluşumunda kullanılan dilbilgisi kuralları incelenmiştir. Türkçe dilbilgisi kurallarına göre söz öbeklerinde vurgulanmak istenen sözcük en sonda bulunurken; sözcüğü vurgulayan sözcük ve ya sözcükler önce yer almaktadır. Bu yaklaşıma göre, bu söz öbekleri tamlamalar olarak karşımıza çıkmaktadır. Tamlamalarda vurgulanmak istenen varlık sonda yer alıp tamlanan ismini alırken, tamlananı vurgulayan sözcük/ler tamlayan olarak adlandırılmaktadır. Tamlamalar tamlayanın tipine göre ad ya da sıfat tamlaması olarak ikiye ayrılmaktayken, ad tamlamaları da tamlananın aldığı eke göre dörde ayrılmaktadır [54].

1. İsim Tamlamaları, aralarında anlamca ilgi bulunan, en az iki farklı isimden oluşan kelime gruplarıdır. Aşağıdaki gibi 4 çeşitten oluşmaktadır.

- Belirtili İsim Tamlaması: Belirtili isim tamlamalarında tamlayan ve tamamlanan isimdir. Birinci isim tamlayan ekini (ın, in, un, ün), ikinci isim de tamlanan ekini (ı, i, u, ü) alması gerekir.

Örnek: “otelin konumu”, “hizmetin kalitesi”

- Belirtisiz İsim Tamlaması: Belirtisiz isim tamlamasında tamlayan ve tamlanan isimdir. Tamlayanın ek almadığı, tamlananın ise -i (-ı, -u, -ü) ekini alması gerekir.

Örnek: ”köy kahvaltısı”

- Takısız İsim Tamlaması: Takısız isim tamlamalarında tamlayan da tamlanan da isimdir. Bu çeşit isim tamlamalarında hem tamlayanın, hem de tamlanan ek almaz.

Örnek: “demir parmaklık”

- Zincirleme İsim Tamlaması: En az üç ismin, birbirini anlam bakımından tamamlamalarıyla oluşan sözcük gruplarına zincirleme isim tamlaması denir. Çalışma kapsamında iki sözcükten oluşan tamlama grupları ele alındığından zincirleme isim tamlaması kullanılmamıştır [bunlar için referans vermek gerek].

Örnek: “oda kapısının kolu”

2. Sıfat Tamlaması, Türkçe’deki sıfat tamlamaları tamlayanı sıfat, tamlananı isim olan kelime gruplarından oluşur. Sıfat tamlamaları o ismin işaret ettiği varlığın adedi, biçimi, rengi, durumu, konumu vs. hakkında bilgi verir.

Örnek: ”temiz tesis”

Sıfat fiiller, fiillerin cümlede sıfat görevinde kullanılan halidir. Fiilerin sonuna sıfat-fiil ekleri olan -en, -esi, -mez, -ar, -di(k), -di(ği), -ecek ve -miş getirilerek oluşturulmaktadırlar. Sıfat fiiller isimden önce gelerek cümlede sıfat tamlamasını oluşturmaktadırlar.

Yukarıda belirtilen sezgisel kurallar aday çok terimli özellik listesine uygulandıktan sonraki adım, kalan özelliklerin belirlenen domain ile ne derece bağlantılı olduğunun skorunu veren PMI yönteminin uygulanmasıdır.

Bütün adımların sonucunda kullanıcı yorumlarından ürün ile ilgili özelliklerin tümü çıkarılmış olunur.

3.1.3. Sık geçmeyen özelliklerin çıkartılması

Sık geçen özellikler kullanıcıların bir ürüne ait özelliklerden bahsederken en çok kullandıkları özellikler kümesinden oluşmaktadır. Ancak bazı durumlarda az kullanılan ama potansiyel kullanıcılar için önemli olan özelliklerde yorum kümesinde bulunabilir. Bu özelliklerin çıkarımının sık geçen özelliklerin çıkarımı kadar önemli olmasa bile, ürüne ait özelliklerin çıkarımı aşamasının tamamlanması için gerekli olan bir adımdır.

Her sentimentin nitelemek istediği bir hedef varlık (özellik) vardır. Bu ilişki sentiment olan sözcük biliniyorsa özellik çıkarımında kullanabilir. Sık geçmeyenler olarak tanımlanan bu yöntem özelliklerin çıkarılmasındaki son adımdır [1]. Bir

sentiment birden fazla birbirinden farklı olan özelliği niteleyebilmektedir. Bu durumda, eğer bir cümle sık geçen bir belirgin özellik içermiyor ancak sentiment ve bu sentimentin yakınında isim içeriyorsa; bu isimler aday sık geçmeyen belirgin özellikler olarak belirlenmektedirler [1, 2]. Çalışma kapsamında sık geçmeyen belirgin özelliklerin çıkartılması amacıyla izlenen adımlar Şekil 3.4’de verilmektedir.

```
Sık geçmeyen belirgin özelliklerin çıkarılmasında kullanılacak algoritma adımları:  
1:for Yorumlar içerisindeki tüm Cümleler  $Sen_i$  için do  
2:   if  $Sen_i$  Sık geçen özellik içermiyorsa ama Sentiment içeriyorsa then  
3:     Bu sentimente en yakın olan isim/isim öbekleri aday sık geçmeyen özellik olarak kaydedilir  
4:   end if  
5:end for
```

Şekil 3.4. Algoritma adımları

Önerilen bu algoritma sonucunda bulunan aday özellikler ilk önce frekanslarına göre elendikten sonra bu özellik kümesindeki ürün ile ilgisi olmayan aday özellikler NGD (Normalized Google Distance) ölçütü ile filtrelenerek sık geçmeyen özellik kümesi çıkartılmaktadır.

3.1.4. Özellik-sentiment çiftlerinin çıkartılması

Bu adımda özellik ve sentiment eşleşmesinin doğru olduğu çiftlerin çıkartılması amaçlanmaktadır. Bunun gerçekleştirilmesi için özellik ve sentimentler kullanıcı yorumları üzerinde etiketlenmiş ve bu etiketler bir JSON veri yapısı ile temsil edilmiştir [55]. Önerilen sistem, oluşturulan JSON veri yapısını giriş olarak almakta ve çıkışında geçerli bir Özellik-Sentiment Çifti (ÖŞÇ) üretmektedir. Türkçe dilbilgisi kurallarına göre ÖŞÇ’ler belirli dizilimlere göre cümlelerde bulunmaktadır. Dilin kurallarına uyan dizilimlerin bulunmasında daha önce ÖŞÇ çıkarımında kullanılmamış olan Pushdown Automata (PDA) çözüm olarak önerilmiştir. Sistem iki aşamadan oluşmaktadır. Birinci aşamada geçerli ÖŞÇ içeren dizilimlerin PDA lar ile bulunması sağlanmaktadır. İkinci aşamada PDA’nın tanıdığı dizilimlerde geçen özellik ve sentimentlerin eşleştirilmesi gerçekleştirilmektedir. Sistem, her adımın çıkışı diğer adımın girişi olacak şekilde tasarlanmıştır.

Otomata notasyonların ve tanımların yazımında [56] çalışması kullanılmıştır.

Bir uç durumla tanıyan PDA Denklem (3.2)’de gösterildiği gibi 7 değişkenden oluşmaktadır. Denklemden sırasıyla, Q sonlu sayıda durum kümesi, Σ sonlu sayıda

giriş simgesinden oluşan giriş alfabesi, Γ sonlu sayıda simge içeren yığın alfabesi, q_0 başlangıç durumu, Z_0 başlangıç yığın sembolü, F kabul edilen durumlar kümesi ($F \subset Q$), δ geçiş ya da hareket işlevini ifade etmektedir. Deterministik PDA modelinde geçiş işlevi, $[(Q \times (\Sigma \cup \{ \lambda \}) \times \Gamma)]$ ‘dan $[Q \times \Gamma^*]$ ‘a bir eşleme oluşturulurken, deterministik olmayan modelde ise geçiş işlevi, $[Q \times (\Sigma \cup \{ \lambda \}) \times \Gamma]$ ‘dan $[Q \times \Gamma^*]$ ‘ın sonlu alt kümelerine bir eşleme oluşturulmaktadır.

$$M = \langle Q, \Sigma, \Gamma, \delta, q_0, Z_0, F \rangle \quad (3.2)$$

M ‘nin anlık tanımının (instantaneous description) gösteriminde $c = (q, x, \omega) \in Q \times \Sigma^* \times \Gamma^*$ sırasıyla, q mevcut durum, x giriş dizgisinin henüz işlenmemiş kesimi, ω yığının mevcut içeriğini ifade etmektedir ve geçiş işlevi δ bu M ‘ye ait anlık tanımlar arasında hareket ilişkisini $\vdash_M$ tanımlar. M tarafından tanınan dizgiler kümesi aşağıdaki Denklem (3.3) ‘deki gibi tanımlanmaktadır.

$$L(M) = \{ x \in \Sigma^* \mid (q_0, x, Z_0) \vdash_M^* (p, \varepsilon, \omega), p \in F \} \quad (3.3)$$

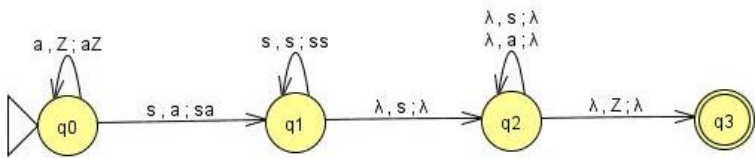
Türkçe’de isimlerin önüne, sonrasına gelerek onları niteleyen ve ya belirten sözcüklere sıfat denir. Türkçe Dilbilgisi kurallarına göre sıfatlar cümlede varlıklarla ortaya çıkarlar ve bu nedenden dolayı tek başına kullanılamazlar. “Temiz havuza bayıldım” cümlesinde “temiz” sözcüğü “havuz” isminden önce kullanılmış ve “havuz” ismini nitelemektedir. Ancak sıfatın mutlaka isimden önce gelmesi gerekmemektedir. Bazen bir ismin niteliğini bildirmesine rağmen cümlede isimden önce yer almadığı durumlar da olabilir. “havuz temizdi” cümlesinde “temiz” sıfatı isimden sonra kullanılmış ve bir önceki örnekte olduğu gibi “havuz” ismini nitelemektedir. Cümlelerde bir sıfat bir ismi niteleyebileceği gibi birden fazla ismi de niteleyebilmektedir. Örneğin virgöl noktalama işareti (,) eş görevli kelime ve kelime gruplarını birbiri ardınca sıralamak için kullanılabilir [57]. “otelin konumu, havuz güzeldi” cümlesinde “güzel” sıfatı “otelin konumu” ve “havuz” isimlerini nitelemektedir. Başka bir dilbilgisi kuralına göre de, Türkçede “ve ” bağlacı bir cümlede aynı duyguyu veren sıfatları birbirine bağlamak için kullanılabilir. Bu kuralın uygulanışını anlayabilmek için “servis hızlı ve iyiydi” cümlesi ele alalım; “hızlı” ve “iyi” sıfatlarının ikisi de “servis” ismini nitelemektedir. “yemekler güzeldi ve servis hızlıydı” cümlesindeki “ve” bağlacı ise anlam yönünde ilişkili ve sıralı

cümleleri bağlamaktadır ve bu durumda “güzel” sıfatı “yemek” ismini, “hızlı” sıfatı ise “servis” ismini nitelemektedir. Bu kurala benzer Türkçe dilbilgisinde isimlerin ve sıfatların sıralı bir şekilde birlikte kullanımına dair başka kurallarda bulunmaktadır. Bu bilgiler ışığında sıfat ve isimler sıralı bir örüntü şeklinde cümlelerde bulunmaktadır. Veri madenciliği uygulamalarında örüntüleri keşfetmek için kullanılan kısıtlamalar, sıfat ve isimlerden oluşan bu sınırlı dizilimlerin belirlenmesi için iyi bir çözüm sunmaktadır.

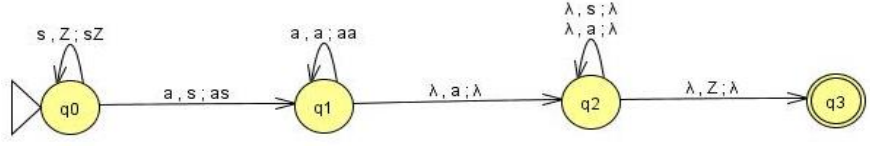
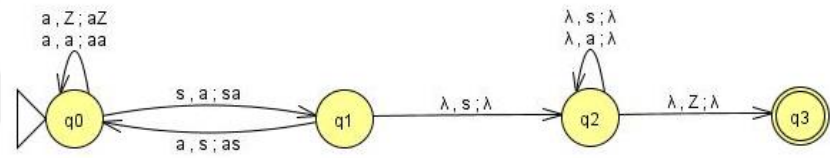
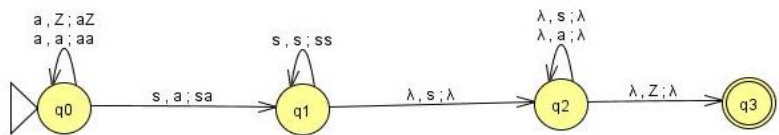
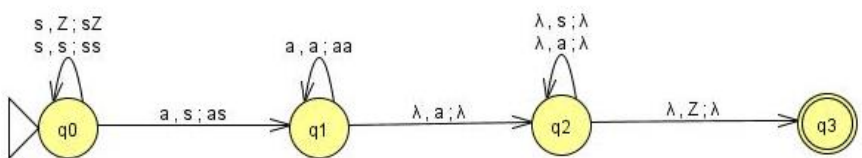
Sıralı örüntü madenciliğinde kullanılan kısıtlamalar deterministik sonlu otomata (DFA) ve ya PDA’lar ile düzenli ve ya içerikten bağımsız dillerin üretilmesi ile temsil edilebilir [58]. PDA ile cümlelerdeki sıralı dizilimler tespit edildikten sonra, bir sonraki aşama olarak cümlelerden hangi sıfatın hangi ismi nitelemek için kullanıldığını çıkarılabilir.

Tez kapsamında, cümledeki isimler özellikleri, sıfatlar ise sentimentleri temsil etmektedir. Bu durumda Türkçe’ de ÖŞÇ’lerin bulunması için Türkçe dilbilgisi kurallarına uygun isim-sıfat sıralı örüntülerinin bulunması gerekmektedir. Bu örüntülerin keşfedilmesi için kullanılan içerik bağımsız gramer kısıtları (context-free grammar constraints) ve PDA’lar açıklamasıyla aşağıdaki Tablo 3.1’de halinde verilmiştir.

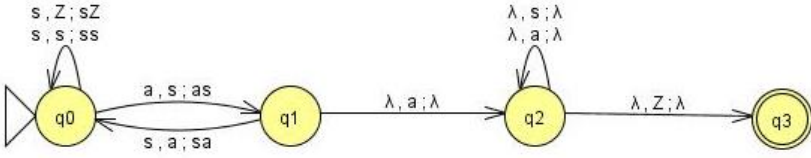
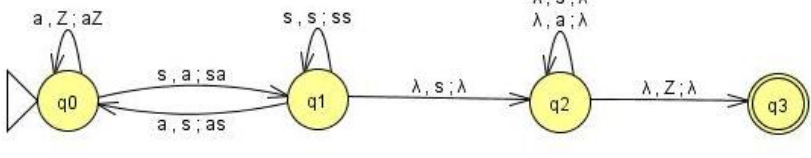
Tablo 3.1. Yedi durum için oluşturulan PDA’lar

Durum	İçerikten bağımsız gramer kısıtları ve PDA’ lar
1	$\{ a^1 s^m \mid m \in \mathbb{N}, m \geq 1 \}$ Birden fazla sentiment önlerindeki bir özelliği niteler. 

Tablo 3.1. (Devam) Yedi durum için oluşturulan PDA'lar

2	$\{ s^1 a^m \mid m \in \mathbb{N}, m \geq 1 \}$ Bir sentiment sonrasında gelen birden fazla özelliği niteler. 
3	$\{ as \}^*$ Sırası ile her sentiment kendisinden önce gelen özellik niteler. 
4	$\{ a^n s^m \mid m, n \in \mathbb{N}, m, n \geq 1 \}$ Sentimentlerin tamamı önünde gelen özelliklerin tamamını niteler. 
5	$\{ s^n a^m \mid m, n \in \mathbb{N}, m, n \geq 1 \}$ Sentimentlerin tamamı sonrasında gelen özelliklerin tamamını niteler. 

Tablo 3.1. (Devam) Yedi durum için oluşturulan PDA'lar

6	$\{ sa \}^*$ Sırası ile her sentiment kendisinden sonra gelen özelliği niteler. 
7	$\{ a^1s^* \}^*$ Sentimentlerin tamamı kendisinden önce gelen bir özelliği niteler. 

Yukarıda görüldüğü üzere tez çalışmasında ÖŞÇ'lerin çıkarılmasında Türkçe dilinde isim-sıfat tanımlaması kurallarını tanıyan 7 farklı PDA modeli kullanılmıştır. Uç durum ile tanıyan PDA otomatlarının girişi, özellik (aspect) ve ya sentiment etiketlerinden oluşan string dizgisinden oluşmaktadır. ($\Sigma : \{a,s\}$). PDA'lar tarafından tanınan cümleler bizim için geçerli olan özellik-sentiment dizilimine sahiptir ve bu cümlelerden ÖŞÇ'ler çıkarılmaktadır. Kurulan otomatlar ile Türkçe bir cümlede olası tüm özellik-sentiment dizgileri yakalanmıştır. Yapı yeni kuralların kolayca eklenebilmesi ve mevcut kuralların değiştirilebilmesi yönünden oldukça esnek bir çözüm sunmaktadır. Kurulan PDA'lar türkçe dili için yapılan sentiment analizi çalışmalarının tümünde özellik-sentiment eşlemede kullanılabilecek niteliktedir.

PDA'ların geçerli olan isim-sıfat çiftlerinin çıkarılmasındaki kullanımını bir örnekte anlatacak olursak;

Ham yorum: Yemekler güzeldi ve servis hızlıydı.

Ham yoruma ait Json veri yapısı:

```
[{"cümle":[{"kelime":"yemek","usage":"aspect","tip":"ISIM","pol":0},{kelime":"g  
üzel","usage":"seedsentiment","tip":"SIFAT","pol":1},{kelime":"ve","usage":"bagl  
ac","tip":"baglac","pol":0},{kelime":"servis","usage":"aspect","tip":"ISIM","pol":0  
},{kelime":"hızlı","usage":"yok","tip":"SIFAT","pol":0}]]}]
```

Yukarıdaki örnekte olduğu gibi json veri yapısı yorumu oluşturan sözcükler ve bu sözcüğe ait etiketlerden oluşmaktadır. Bu yapıya göre özellikleri ‘a’ etiketi, sentimentleri ise ‘s’ etiketi ile temsil edersek; otomatanın girişine ‘ass’ string dizgisi verilecektir. ‘ass’ string dizgisi Türkçe dil kurallarına göre geçerli bir dizilime sahiptir ve bu kuralı otomata tanır. Bu kuralı PDA’nın tanınmasından sonra verilen ham yorumdan aşağıdaki ÖSÇ’ler çıkarılmıştır.

<yemek, güzel> ve <servis, hızlı>

3.1.5. Özellik tabanlı skortlama

Başlangıçta sözlük tabanlı yaklaşım ile oluşturulan olan sentiment çekirdekleri bu adımda özellik tabanlı bir şekilde genişletilmiştir. Sentimentler incelendiğinde alan-bağımlı ve alan-bağımsız olabilecekleri görülmektedir. Alan-bağımsız olan sentimentler sözlük tabanlı yaklaşım ile belirlenmekteyken, alan-bağımsız sentimentlerin belirlenmesinde daha karmaşık bir durum ile karşı karşıya kalınmaktadır. Burada belirtilen alan, üzerinde çalışılan temel varlığa karşılık gelmektedir. Alan-bağımlı sentimentler temel varlık (otel) ve ya temel varlık ile alakalı olan özellik değişirse farklı polariteye sahip olabilmektedirler [2]. Bunu bir örnek ile açıklarsak; “otel” ve “telefon” temel varlıkları için “kısa” sentimentini ele alalım. “otelin yemek servisi süresi oldukça kısa.” ve “Telefonun batarya ömrü çok kısa.” cümlelerinde kısa sentimentini farklı polarite göstermektedir. Servis süresinin kısa olması pozitif bir duygu vermekten batarya ömrünün kısa olması negatif bir duygu vermektedir. Bu durum aynı temel varlığa ait özellikler arasında da görülebilmektedir. “otelin elemanları müşterilere karşı soğuktu” ve “bu otelde her zaman soğuk içecek bulabilirsiniz” cümlelerinde “otel” temel varlığına ait farklı özellikler için kullanılan “soğuk” sentimentini ilk cümlede negatif polariteye sahipken, diğer cümlede pozitif polariteye sahiptir. Bu açıdan değişen özelliğe göre sentimentin polaritesinin belirlenmesi oldukça önemlidir.

Özellik tabanlı yöntemle oluşturulacak olan sözlük özellik, sentiment, polarite şeklindeki ifadelerden oluşacaktır. Keşfedilen her sentiment, bu sentimentin tanımladığı özellik ve polarite ile sözlüğe ayrı kayıt satırları olarak eklenecektir. Benzer şekilde önceden belirlenen sentiment çekirdekleri ve ilgili özellikleri de bu sözlükte polariteleri ile yer alacaktır. Özelliklere göre değişen sentiment polaritelerini hesaplayabilmek için Türkçe dilbilgisi kurallarına dayalı çıkartılan ”sezgisel kurallar” uygulanmıştır. Bu kurallar 4 sınıfta ele alınmıştır.

- Karşıtlık Bağlaçları: Karşıtlık bağlaçları, iki farklı duyguyu veren sentimentleri birbirine bağlamaktadır ve “ama”, “fakat”, “lakin”, “yalnız”, “ancak”, “zira”, “ne var ki”, “ne yazık ki” olarak belirlenmişlerdir. Bağlaçların yeri cümle içerisinde tespit edildikten sonra, bu cümle içerisindeki polaritesi bilinen bir sentimentten polaritesi bilinmeyen sentimentin polaritesi bulunabilmektedir. Örneğin “Yemek az ama nefis.” cümlesinde “nefis” pozitif sentiment iken kendisine ama bağlacı ile bağlanan “az” negatif sentiment olarak belirlenmekte ve yemek, az, negatif olarak sözlüğe eklenmektedir. Karşıtlık bağlacı cümledeki duygunun kırıldığı noktadır. Bu bağlaç bulundurmayan tüm cümlelerde duygu devam etmektedir. Örneğin; “yemek güzel ve lezzetliydi.” Cümlesinde güzel pozitif sentiment olduğundan lezzetli sentimentinde polaritesi pozitif olarak yemek, lezzetli, pozitif olarak sözlüğe eklenmektedir.

Yukarıdaki kural dışında kalan durumlar için literatürde, olumlu bir cümleden sonra ya da önce gelen cümlelerin olumlu anlam taşıdığı; olumsuz cümleden sonra ya da önce gelen cümlelerin ise olumsuz anlam taşıdığı varsayılmaktadır [59, 60]. Bu varsayımı özellik tabanlı sentiment analizi için de uyarlayabiliriz. “Pizza çok lezzetliydi. İçinde bol malzeme vardı.” şeklinde bir yorumdaki gibi peş peşe kullanılan iki cümle elimizde olsun. Birinci cümlede pizza özelliği ile ilgili pozitif bir sentiment var bu durumda ikinci cümlede malzeme özelliğini niteleyen bol sentimentinin de pozitif polariteye sahip olduğu anlaşılmaktadır:

- Olumsuzlama kuralı: Olumsuzluk duygusu veren kelimeler ve ekler cümledeki sentimentin polaritesini pozitif ise negatife, negatif ise de pozitif çevirmektedir. Türkçe de olumsuzlama yapan sözcük “değil” olarak belirlenmiştir. Olumsuzlama kuralı ve karşıtlık bağlacı kuralı ile cümledeki polaritesi bilinmeyen sentimentin polaritesi belirlenebilmektedir. Örneğin “Yemek güzel değildi ama servis hızlıydı.” cümlesinde “Yemek” ve “servis” belirgin özellikleri ve “değil” olumsuzluk ifadesi

bulunmaktadır. Bu ifade polaritesi pozitif olan “güzel” sentimentini polaritesini negatife dönüştürmüştür. Bu durumda “güzel” sentimentinin geçtiği cümleye ama bağlacı ile bağlı olan ikinci cümledeki “hızlı” sentimentinin polaritesinin pozitif olduğu bulunmuştur.

- Pekiştirme zarfları: Türkçede cümledeki duyguyu (pozitif/negatif) pekiştirmek amacıyla miktar zarfları (“çok”, “hiç”, “kısmen”, “biraz”, “fazla” vb.) kullanılmaktadır. Bu zarfların anlamı ne oranda güçlendirdiği farklılık göstermektedir. Örneğin “yemek çok kötüydü” cümlesindeki olumsuz duygu “yemek biraz kötüydü” cümlesindeki olumsuz duygudan daha güçlüdür. Uygulamada, bu zarfların ağırlık dereceleri manuel olarak atanmıştır.
- Düşünce içeren fiiller: Dilbilimsel açıdan fiiller sentimentin belirlenmesinde önemli bir role sahiptirler. “hoşlanmak” ,”beğenmek” gibi bazı fiiller sentiment polaritesine sahiptirler ve bu tarz fiillere düşünce içeren fiiller (opinion verbs) adı verilmektedir [61]. Örneğin; “havuzu beğenmedim” cümlesinde beğenmek fiili kullanılmış ve Türkçede olumlu fiili olumsuz biçime dönüştürenme ekiyle beğenmemek fiili olumsuz anlam kazanarak havuz özelliğini negatif olarak etkilemiştir.

Yukarıda tanımlanan sezgisel kurallar çerçevesinde gerçekleştirilen örnek bir polarite hesabı aşağıdaki cümle için gösterilmiştir.

“odalar biraz kötüydü $[-1 \times 0,5]$ ama havuz çok temizdi $[+1 \times 2]$.”

Bu örnek cümlede, biraz zarfı, -1 olan kötü sentimentinin skorunu 0,5 katsayısı oranında güçlendirmiştir. Çok zarfı ise temiz sentimentinin pozitif anlamını 2 kat güçlendirmiştir.

Kullanıcı yorumlarında yer alan farklı cümlelere göre cümlelerde yer alan özelliklerin sentiment polaritelerini hesaplayabilmek için yukarıda belirtilmiş olan Türkçe dilbilgisi kurallarına dayalı olarak gerçekleştirilen sentiment skorum işlemi için önerilen algoritma Şekil 3.5’de verilmektedir.


```

1:for Cumle icerisindeki tum Cumlecikler icin do
2:   $P(w_{|clause|}) \leftarrow 0$ 
3:end for
4:for Cumle icerisindeki tum kelimeler  $w_j$  icin do
5:  if  $w_j$  Sentiment List (SL) icinde bulunuyorsa then
6:     $P(w_j) \leftarrow (positive:1 \text{ or } negative:-1)$ 
7:  else
8:     $P(w_j) \leftarrow 0$ 
9:  end if
10:end for
11:for Cumle icerisindeki tum kelimeler  $w_j$  icin do
12:  if  $w_j$  Negation List (NL) icinde bulunuyorsa then
13:    for  $w_j$  ile ayni Cumlecikte bulunan  $w_i$  ler icin do
14:       $P(w_i) \leftarrow -1 \times P(w_i)$ 
15:    end for
16:    if  $w_j$  ile ayni Cumlecikte bulunan  $w_i$  lerin hicbiri SL icerisinde bulunmuyorsa then
17:       $P(w_{|clause|}) \leftarrow P(w_{|clause|}) \times (-1)$ 
18:    end if
19:  end if
20:  if  $w_j$  Adverb List (AdvL) icinde bulunuyorsa then
21:    for  $w_j$  ile ayni Cumlecikte bulunan  $w_i$  ler icin do
22:       $P(w_i) \leftarrow degree(w_j) \times P(w_i)$ 
23:    end for
24:    if  $w_j$  ile ayni Cumlecikte bulunan  $w_i$  lerin hicbiri SL icerisinde bulunmuyorsa then
25:       $P(w_{|clause|}) \leftarrow degree(w_j)$ 
26:    end if
27:  end if
28:end for
29:for Cumle icerisindeki tum kelimeler  $w_j$  icin do
30:  for tum Cumlecikler icin do
31:    if  $w_j \in w_{|clause|}$  then
32:       $P(w_{|clause|}) \leftarrow P(w_{|clause|}) + P(w_i)$ 
33:    end if
34:  end for
35:end for
36:for Cumle icerisindeki tum kelimeler  $w_j$  icin do
37:  if  $w_j$  Transition List (TL) icinde bulunuyorsa then
38:    for  $w_j$  ile ayni Cumlecikte bulunan  $w_i$  lerin hicbiri SL icerisinde bulunmuyorsa then
39:       $P(w_{|clause|}) \leftarrow |P(w_{|clause|})| \times value(w_j) \times (diger\ cumlecik\ positive:1 \text{ or } negative:-1)$ 
40:    end for
41:  end if
42:end for

```

Şekil 3.5. Özellik tabanlı skoreleme adımları

Algoritmadaki cümle değişkeni farklı alt cümlelerden oluşan bir kullanıcı yorumuna karşılık gelmektedir. $w_{|clause|}$: alt cümleyi ifade etmektedir. $P(w_{|clause|})$: özellik-sentiment çiftinin sentiment skorunu ifade etmektedir. $P(w_j)$: eğer w_j mevcut sentiment listede ise değerini (pozitif:1, negatif:-1) vermektedir. NL (Negation List): olumsuzluk duygusu veren kelimeler ve ekleri belirtmektedir. AdvL: sentiment analizinde ele alınan pekiştirme zarflarını ve ağırlıklarını belirtmektedir. Transition List (TL): karşıtlık bağlaçlarını belirtmektedir. SL (Sentiment List): başlangıçta bizim tarafımızdan belirlenmiş olan sentiment çekirdekleri belirtmektedir.

Yorumdaki cümleler nokta “.” ile birbirinden ayrılmıştır. Bu tür cümleler alt cümleciklere bölünür, her alt cümleciğin sentiment puanı hesaplanarak özelliğe ait sentiment polaritesi belirlenir. Cümleler birbirinden “.” ile ayrılmış ise; cümlelerden herhangi biri sentiment içermiyorsa ve yukarıda belirtilmiş olan bağlaçlardan biri ile bağlı değilse “.” işareti “ve” bağlacı olarak kabul edilir. Bu kabulün sebebi literatürde peş peşe gelen cümlelerin aynı duyguyu verdiklerinin varsayılmasıdır.

Bu adımlar sonunda tez çalışmasında kullanılacak sentiment sözlüğü kurulmuş olacaktır ve sözlüğe başka kelime eklenmeyecektir.



4. DENEYSEL ÇALIŞMALAR

Bu bölümde tez kapsamında geliştirilen sistemin uygulamasında kullanılan teknolojilerden bahsedilip ve elde edilen sonuçlar incelenecektir.

4.1. Kullanılan Teknolojiler

Sistemin uygulama kısmı Eclipse IDE platformunda, Java dili kullanılarak kodlanmıştır. Java, açık kaynak kodlu, nesneye yönelik, zeminden bağımsız bir programlama dilidir. Java dili veri madenciliği uygulamalarında sıkça kullanılmaktadır. Türkçe doğal dil işleme kütüphanesi zemberek [62] java platformunda düzgün bir biçimde çalışması uygulamanın java dilinde yazılmasını avantajlı kılmıştır. Doğal dil işleme adımlarından olan sözcüklerin kökünün ve bu sözcüklerin Türkçe dilbilgisine göre sözcük tipilerinin bulunması adımları bu kütüphane ile gerçekleştirilmiştir. Kütüphanenin kök bulma fonksiyonu türemiş kelimelerde doğru çalışmadığından dolayı sözcüğün gövdesini bulmak için ek fonksiyonlar yazılmak zorunda kalınmıştır.

Yukarıda belirtilen teknolojiler dışında projenin her adımında problemlerin çözülmesi için ek başka teknolojilerde kullanılmıştır. Bunlar uygulamanın adımları detaylandırılırken anlatılmıştır.

4.2. Kullanılan Veri Kümleri

Veri kümesinin oluşturulması için tasarlanan uygulama iki modülden oluşmaktadır. Bu aşamalardan birincisi link toplayan web crawler modülü, ikincisi de web dokümanından gerekli bilginin çekilmesini sağlayan web scraper modülüdür. Web'i otomatik olarak dolaşarak link toplayan bir web crawler modülünü tasarlamak için crawler4j kütüphanesi kullanılmıştır. Crawler4j java dilinde yazılmış ve çoklu iş parçacıklı yapısında çalışan açık kaynak kodlu bir kütüphanedir [63]. Web crawler'ı hedef siteye yönelik çalışması için gerekli olan parametreler girildikten sonra çoklu iş parçacıklı yapısında gerekli olan linkleri toplaması için çalıştırılabilir. Bu parametreler; hedef sitenin domain adresi, gezilecek sayfa sayısı, iş parçacık sayısı

ve gezinirken gereksiz linkleri filtrelemesi için dosya uzantılarıdır. Bu projede kullanıcı yorumları için birçok site taranmıştır. Ancak Türkçe’de kullanıcı yorumlarından özellik tabanlı sentiment analizi ile ürünlerin zayıf özelliklerinin çıkarılması uygulaması için uygun yorum kümelerini içeren iyi bir site bulunmamaktadır. Testler sonucunda en iyi sonuç otel alanı için www.otelpuan.com sitesindeki yorumların kullanılması ile elde edilmiştir.

Web scraper modülü için web uygulamalarının testi için kullanılan açık kaynak kodlu selenium kütüphanesi kullanılmıştır [64]. Bu kütüphane ile html etiketleri arasındaki kullanıcı yorumlarını içeren metinler web sayfasından çekilmiştir. Elde edilen yorum kümesi Tablo 4.1’de detaylı bir şekilde verilmiştir.

Tablo 4.1. Yorum setinin detaylı bilgisi

Veri kümesi	Yorum sayısı	Cümle sayısı
Otel	1518	6852

4.3. Yazılım Mimarisi Tasarımı ve Uygulama Adımları

Tez çalışmasında geliştirilmiş olan uygulamanın nesne uml diyagramı Şekil 4.1’de verilmektedir. Şekilde görüldüğü gibi gerçekleştirilen proje modeli sekiz sınıftan oluşmaktadır. Proje, her alt problem bir sınıf nesnesi olacak şekilde tasarlanmıştır.

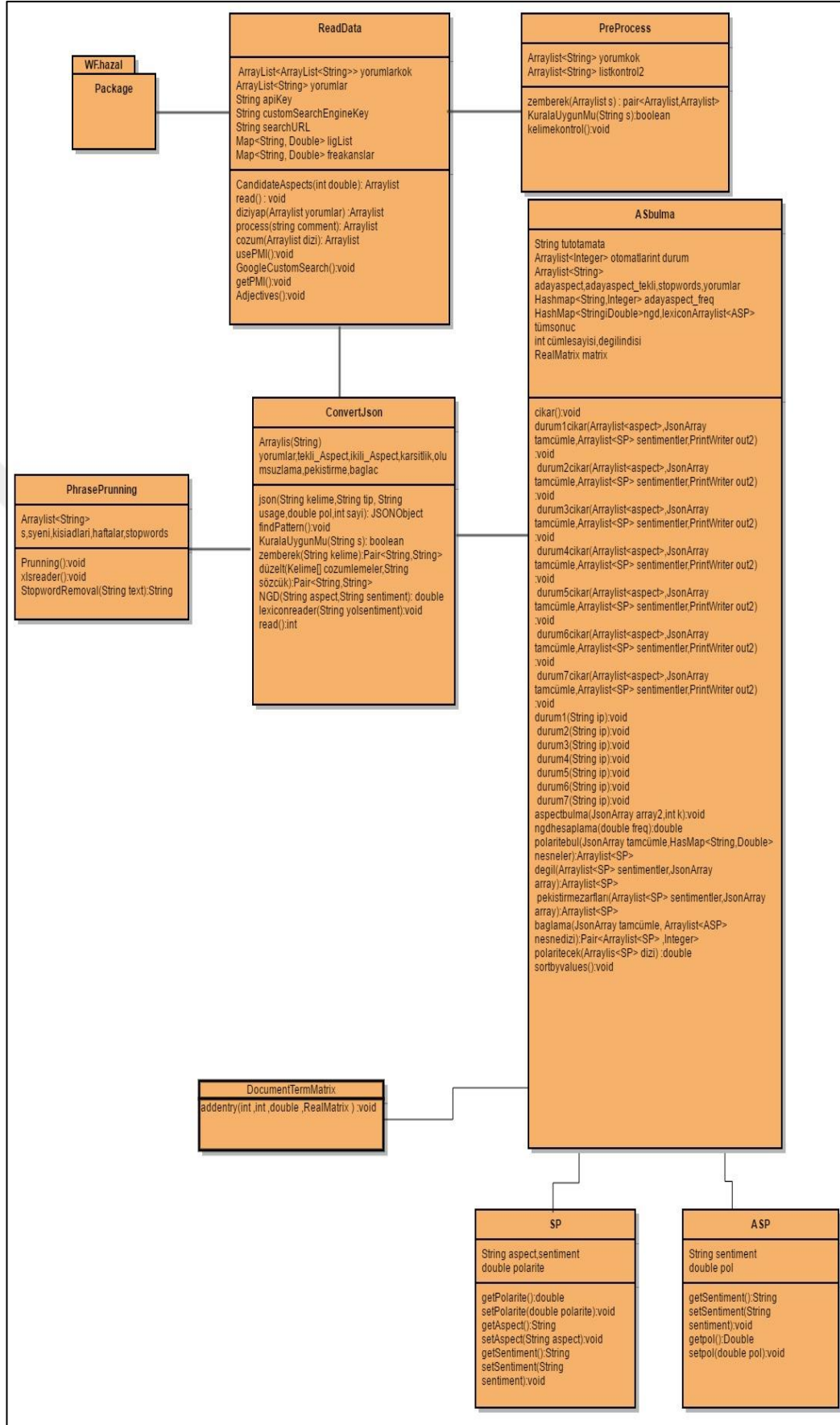
Bu bölümde bu sekiz sınıfın açıklaması, sonuçları ve kullanılan teknoloji anlatılmaktadır.

Projenin adımları ve uygulamadaki nesneler ile olan ilişkisi aşağıdaki maddeler ile verilmektedir;

- ReadData sınıfı: yorumların “.txt” dosyasından okunması, veri ön işleme adımları, tekli özelliklerin bulunması
- Preprocess sınıfı: dokümanın dilin morfolojik yapısına göre etiketlenmesi
- Phrasepruning sınıfı: çok terimli özelliklerin bulunması
- ConvertJson sınıfı: tekli özellikler, çok terimli özellikler, sentimentler ve dilin morfolojik yapısına göre json veri yapısında dokümanın etiketlenmesi
- ASbulma sınıfı: oluşturulan json veri yapısından PDA’lar ile ÖSÇ’lerin bulunması, ÖSÇ’lerin polaritelerinin hesaplanması
- DocumentTermMatrix sınıfı: doküman terim matrisinin oluşturulması

- AS sınıfı: PDA’lardan çıkan ÖSÇ’lerin yapısının temsili
- ASP sınıfı: tüm proje sonucunda oluşan <özellik,sentiment,polarite> yapısının temsili





Şekil 4.1. Sistem UML diyagramı

4.3.1. Veri ön işleme

Metinlerin analizi işleminde ilk adım olan veri ön işlemede, “.txt” dosyasında kaydedilmiş yorum kümesinden okuma yapılmakta ve gereksiz bilgiler ayrıştırılmaktadır. Veri ön işleminin son adımı olan sözcüklerin köküne ve ya gövdesine indirgeme (stemming) işlemi zemberek kütüphanesi ile gerçekleştirilerek, isim tipindeki sözcükler sık geçen tekli özelliklerin bulunması için fonksiyona gönderilmektedir. Aşağıda zemberek kütüphanesinin çalışma yapısına dair örnek verilmiştir.

“odalar” sözcüğünün zemberek çıkışı: [Kok: oda, ISIM] Ekler: ISIM_COGUL_LER

4.3.2. Sık geçen tekli özelliklerin çıkarılması

Bu adımda ön işleme adımından çıkan isim tipindeki sözcüklerden sık geçen tekli özellikler çıkarılmıştır. ReadData sınıfında ait CandidateAspect() fonksiyonu bu işlemi gerçekleştirmek için kodlanmıştır. İsim tipindeki sözcüklerden bütün doküman setinde geçme sayılarına göre eleme yapılarak aday özellik listesi elde edilmiştir. Frekans sayısı uygulama aşamasında yapılan test sonuçlarına göre 27 olarak belirlenmiştir. Bu aday özellik listesi PMI skorlarının hesaplanması için usePMI() fonksiyonu kullanılarak, özelliklerin “otel” temel varlığı ile ilişkisini gösteren skorları hesaplanmıştır. Projede önerilen skor hesaplanmasında özelliklerin Google’da geçme sayısı kullanılmaktadır. Bu sayı Google Custom Search Api’nin kullanılması ile elde edilmiştir. Google Custom Search Api Google’daki arama sonuçlarına erişmek için program geliştiren kullanıcılara Google tarafından sunulan bir teknolojidir [65]. Google bu uygulama ara yüzünün kullanımında kullanıcılara kotada kısıtlama sunduğundan dolayı ara yüzün kullanımını düzgün sağlamak için Google Developer Console sayfasından kota satın alımı yapılmıştır.

PMI skorlarına göre filtreleme işleminde frekans, bütün skorlar toplanıp ortalamaları alınarak belirlenmiştir. PMI skorunun küçük olması aday özelliğin temel varlık “otel” ile ilişkisinin az olduğunu göstermektedir. Bu işlemden sonra elde edilen nihai tekli özellik listesi PMI skorlarıyla birlikte Tablo 4.1’de verilmektedir.

Tablo 4.2. Sık geçen tekli özellikler ve PMI skoru

ID	Özellik	PMI skoru
1	Servis	-28,73643338490531
2	Deniz	-28,71411046898712
3	Olanak	-28,414967402784878
4	Animasyon	-28,11980292432237
5	Konum	-27,677682218889817
6	İçecek	-27,481107888996185
7	Kalite	-27,352348263286455
8	Eğlence	-27,217047853170865
9	Personel	-27,040287641122173
10	Aktivite	-26,2310605897874
11	Tatil	-26,017652087649243
12	Otel	-25,910290528330556
13	Havuz	-25,100847057181436

Tablodaki verilere göre “havuz” özelliği “otel” temel varlığı ile ilişkisi en yüksek olan, “ilişki” özelliği ise en düşük olan özelliktir.

4.3.3. Sık geçen çok terimli özelliklerin çıkarılması

Sık geçen çok terimli özelliklerin çıkarımında n-gram tabanlı bir yöntem izlenmiştir. Kelime n-gramlardan oluşan liste ilk olarak kompaktlık budama yöntemi ile filtrelenmiştir. Bu yöntemdeki frekans test sonuçlarına göre 3 olarak belirlenmiştir. Kompaktlık budama sonucunda elde edilen liste dilbilgisi kurallarına göre tekrardan filtrelenmiştir. Dilbilgisi kurallarına göre eleme yapılırken zemberek kütüphanesi kullanılmıştır. PhrasePruning sınıfındaki pruning() metodu bu budama işlemini gerçekleştirmektedir. Zemberek kütüphanesi ile sözcükler hecelerine ayrıştırılarak eklerine ulaşılmış, bu ekler yardımıyla sözcüklerin tamlama oluşturup oluşturmadıkları belirlenmiştir. Özelliklerin temel varlıkla ilişkisinin skorunu hesaplamak için kullanılan PMI yöntemi ile son kez budamadan geçirilen multi-word özelliklerin son hali Tablo 4.3’da verilmektedir.

Tablo 4.3. Sık geçen çok terimli özellikler

ID	Çok terimli özellik	ID	Çok terimli özellik
1	Hizmet anlayışı	20	Tesis yemekleri
2	İçme imkânları	21	İçecek kalitesi
3	Yemeklerin kalitesi	22	Yemeklerin lezzeti
4	Personelin hizmeti	23	Misafir ilişkileri
5	Bahçe odası	24	Havuz imkânları
6	Hizmet kalitesi	25	Personelin ilgisi

Tablo 4.3. (Devam) Sık geçen çok terimli özellikler

7	Köy kahvaltısı	26	Tesis konumu
8	Otelin konumu	27	Açık büfe
9	Çalışan personel	28	Yemek kalitesi
10	İçme olanakları	29	Oda kalitesi
11	Yiyecek içecek	30	Tesis olanakları
12	Tesis personel	31	Otel personelinin
13	Tesis hizmet	32	Müşteri ilişkileri
14	Yemek çeşitliliği	33	Müşteri hizmetleri
15	İlgi alaka	34	Personellerin hizmeti
16	Personelin kalitesi	35	Servis kalitesi
17	Fiyat kalite	36	Personel yaklaşımı
18	Yüzme olanakları	37	Ala carte
19	Otel personel	38	Kaliteli hizmet

4.3.4. Sık geçmeyen özelliklerin çıkarılması

Sık geçmeyen özelliklerin çıkarılması sık geçenler kadar önemli olmasa bile özellik çıkarımının kullanıcı yorumlarından çıkarımının tamamlanması için gereken bir aşamadır. Çalışmada önerilen yöntem ASbulma sınıfında aspectbulma() fonksiyonu tarafından gerçekleştirilmektedir. Aspectbulma() methodu ile sentiment içeren ancak özellik içermeyen cümlelerdeki isimler özellik olarak etiketlenmiştir. Bu aday sık geçmeyen özelliklerin budanması ngdhesaplama() methodu ile sağlanmıştır. Bu fonksiyon içerisinde her aday özelliğe ait NGD skorunun ve dokümanda geçme frekansının belirlenen bir ağırlık ile çarpılarak, bir skor elde edilmiştir. Bu skora bağlı budama işlemi gerçekleştirilmiştir. Bu ağırlık uygulamada 0,5 olarak kabul edilmiştir. Budama sonunda elde edilen sık geçmeyen özellik listesi Tablo 4.4’de verilmektedir.

Tablo 4.4. Sık geçmeyen tekli özellikler

ID	Özellik
1	Restoran
2	Ortam
3	Garson
4	Yiyecek
5	Alan
7	Oda
8	club
9	Hizmet
10	Yemek
11	Tesis

4.3.5. Dokümanın etiketlenmesi

Metinlerdeki bilgiyi organize etmek ve kolay erişilmesini sağlamak için dokümandaki sözcüklerin etiketlenmesi önem taşımaktadır. Dilin morfolojik yapısına göre sözcüklerin tipleri, çıkarılan tüm özellikler ve sentiment sözlüğündeki tüm sentimentler json veri yapısında etiketlenmiştir.

Json, hafif, dilden bağımsız veri değişim formatıdır. Json strings, numbers, Booleans ve null olmak üzere dört temel tipi, object ve arrays olmak üzere iki yapısal tipi temsil edebilir [66]. json veri formatı bu çalışmadaki uygulama için birçok kolaylık sağlamaktadır. Dokümanın etiketler ile temsili uygulamaların hızı ve karmaşıklığı açısından önem taşımaktadır. Json veri formatı bellekte az yer kaplar ve diğer bir veri değişim tipi olan XML'e göre çok büyük hızda çözümlemesi yapılır [67]. Json modern tarayıcılar ve mobil cihazlar tarafından XML'e göre daha çok tercih edilen bir formattır. Aşağıda kullanıcı yorumlarında elde edilen bir cümle için json veri yapısıyla oluşturulmuş şekli verilmiştir.

```
[{"cümle":[{"kelime":"yemek","usage":"aspect","tip":"ISIM","pol":0},
{"kelime":"güzel","usage":"seedsentiment","tip":"SIFAT","pol":1},
{"kelime":"ve","usage":"baglac","tip":"baglac","pol":0},
{"kelime":"servis","usage":"aspect","tip":"ISIM","pol":0},
{"kelime":"hızlı","usage":"yok","tip":"SIFAT","pol":0}]]]
```

Yukarıda görüldüğü gibi bir cümle JSONarray nesnesi olarak, bu cümle içindeki sözcüklerde JSONobject nesnesi olarak temsil edilmiştir. Bir JSONobject nesnesi 5 adet etiketten oluşmaktadır. Bunlar “kelime”, “usage”, “tip” ve “pol” etiketleridir. Bu aşama için uygulamada ConvertJson sınıfı oluşturulmuştur.

4.3.6. Özellik-sentiment çiftlerinin çıkarılması

Oluşturulan json veri yapısı üzerinden gerçekleştirilen özellik-sentiment çıkarımı için ASbulma sınıfına ait cıkar() fonksiyonu kodlanmıştır. Bu fonksiyon önerilen yöntemdeki yedi adet otomata fonksiyonunu içermektedir. Bu fonksiyonlardan birine

uyan cümle, kurala uyan fonksiyona özellik-sentiment eşlemesi için girdi olarak verilir. Bu fonksiyonun çıktısı da AS nesnelerinden oluşan dizidir. Tablo 4.5’de çıkarılan özellik-sentiment çiftleri gösterilmektedir.

Tablo 4.5. Özellik-sentiment çiftleri

açık büfe güzel	animasyon keyifli
aktivite güzel	animasyon muhteşem
aktivite hoş	animasyon mükemmel
aktivite iyi	animasyon rahatsız
aktivite keyifli	animasyon tatlı
aktivite türlü	animasyon temiz
ala carte başarılı	animasyon yetersiz
ala carte çeşitli	club alakalı
ala carte görevli	club başarılı
ala carte güzel	club bazı
ala carte harika	club fena
ala carte ilgili	club güzel
ala carte iyi	club ilgili
ala carte kaliteli	club iyi
ala carte kolay	club küçük
ala carte lezzetli	club rahat
ala carte muhteşem	club sıkıntılı
ala carte mükemmel	club uygun
ala carte süper	club yeşil
ala carte tatlı	çalışan personel alakalı
ala carte usta	çalışan personel düzgün
ala carte yeni	çalışan personel güleryüzlü
ala carte yeşil	çalışan personel ilgili
ala carte zengin	çalışan personel iyi
ala carte zor	çalışan personel sıcak
alan başarılı	deniz amaçlı
alan bol	deniz belli
alan büyük	deniz berbat
alan dağınık	deniz büyük
alan faydalı	deniz çakıllı
alan geniş	deniz dalgalı
alan güzel	deniz eski
alan hijyenik	deniz fena
alan hoş	deniz güleryüzlü
alan iyi	deniz güzel
alan rahat	deniz harika
alan temiz	deniz iyi
alan uzak	deniz kararsız
alan yeşil	deniz keyifli
animasyon başarılı	deniz kirli
animasyon boş	deniz kötü
animasyon eğlenceli	deniz muhteşem
animasyon etkin	deniz pis
animasyon güzel	deniz rahat

Tablo 4.5. (Devam) Özellik-sentiment çiftleri

animasyon iyi	deniz sıcak
deniz sıkıntılı	havuz keyifli
deniz tam	havuz küçük
deniz taşlı	havuz muhteşem
deniz temiz	havuz mükemmel
deniz ufak	havuz sakın
deniz uygun	havuz temiz
deniz uzak	havuz uzak
deniz yakın	havuz yaşlı
deniz yeşil	havuz yeterli
deniz yeterli	hizmet anlayışı başarılı
deniz yosunlu	hizmet anlayışı güleryüzlü
eğlence başarılı	hizmet anlayışı güzel
eğlence etkin	hizmet anlayışı harika
eğlence güzel	hizmet anlayışı iyi
eğlence iyi	hizmet anlayışı kaliteli
eğlence keyifli	hizmet anlayışı mükemmel
eğlence lezzetli	hizmet anlayışı rahat
eğlence rahat	hizmet artı
eğlence süper	hizmet başarılı
eğlence terli	hizmet başka
eğlence uygun	hizmet çocuklu
eğlence yeterli	hizmet dengeli
eğlence zayıf	hizmet eksik
fiyat kalite güzel	hizmet geniş
fiyat kalite iyi	hizmet görevli
garson alakalı	hizmet güleryüzlü
garson çalışkan	hizmet güzel
garson eksik	hizmet harika
garson güleryüzlü	hizmet hızlı
garson güzel	hizmet hijyenik
garson ilgili	hizmet ilgili
garson iyi	hizmet iyi
garson mükemmel	hizmet kaliteli
garson temiz	hizmet kalitesi fena
garson zayıf	hizmet kalitesi güzel
havuz başarılı	hizmet kalitesi iyi
havuz büyük	hizmet kalitesi mükemmel
havuz derin	hizmet kalitesi rahat
havuz eski	hizmet kalitesi standart
havuz genç	hizmet kalitesi yüksek
havuz güzel	hizmet kalitesi zor
havuz hoş	hizmet muhteşem
havuz imkanları güzel	hizmet mutlu
havuz iyi	hizmet mükemmel
havuz kapalı	hizmet nasıl
hizmet niyetli	kalite rahat
hizmet odaklı	kalite uygun
hizmet önemli	kalite yeterli
hizmet temiz	kalite yüksek

Tablo 4.5. (Devam) Özellik-sentiment çiftleri

hizmet türlü	kaliteli hizmet kaliteli
hizmet uygun	konum güzel
hizmet yakın	konum harika
hizmet yavaş	konum hoş
hizmet yüksek	konum iyi
hizmet zayıf	konum kolay
içecek alkollü	konum mükemmel
içecek alkolsüz	konum temiz
içecek çeşitli	konum tertemiz
içecek düzen	konum uygun
içecek güzel	konum uzak
içecek iyi	konum yeşil
içecek kaliteli	köy kahvaltısı harika
içecek kalitesi güzel	misafir ilişkileri görevli
içecek kalitesi iyi	misafir ilişkileri güleryüzlü
içecek kapalı	misafir ilişkileri güzel
içecek kutulu	misafir ilişkileri ilgili
içecek markalı	misafir ilişkileri iyi
içecek muhteşem	müşteri hizmetleri iyi
içecek sorunsuz	müşteri ilişkileri başarılı
içecek temiz	müşteri ilişkileri güzel
içecek uygun	müşteri ilişkileri harika
içecek yerli	müşteri ilişkileri ilgili
içme imkanları güzel	müşteri ilişkileri iyi
içme imkanları iyi	müşteri ilişkileri kuvvetli
içme imkanları mükemmel	müşteri ilişkileri sıcak
içme olanakları güzel	müşteri memnuniyeti güleryüzlü
içme olanakları iyi	müşteri memnuniyeti güzel
içme olanakları kaliteli	müşteri memnuniyeti hassas
içme olanakları temiz	müşteri memnuniyeti ilgili
ilgi alaka başarılı	müşteri memnuniyeti iyi
ilgi alaka güzel	oda ateşli
ilgi alaka hoş	oda bakımsız
ilgi alaka iyi	oda basit
ilgi alaka yenilenmiş	oda başarısız
kalite başarılı	oda bazı
kalite güzel	oda büyük
kalite iyi	oda dar
kalite kaliteli	oda derin
kalite orantılı	oda eksik
kalite önemli	oda eski
oda farklı	ortam ferah
oda faydalı	ortam güzel
oda fena	ortam hoş
oda ferah	ortam huzurlu
oda geniş	ortam iyi
oda güzel	ortam kaliteli
oda harika	ortam mavi
oda ilgili	ortam mükemmel
oda iyi	ortam sakin

Tablo 4.5. (Devam) Özellik-sentiment çiftleri

oda kalitesi eski	ortam yeşil
oda kalitesi iyi	otel alakalı
oda keyifli	otel ali
oda kirli	otel ateşli
oda konforlu	otel başarılı
oda kötü	otel başka
oda kumlu	otel bayağı
oda küçük	otel bazı
oda modern	otel büyük
oda rahat	otel devamlı
oda rahatsız	otel doğru
oda sıkıntılı	otel eğlenceli
oda tamam	otel eksik
oda tedirgin	otel eski
oda temiz	otel ferah
oda tertemiz	otel geniş
oda ufak	otel güler yüzlü
oda uygun	otel güvenli
oda uzak	otel güzel
oda yeni	otel hijyenik
oda yenilenmiş	otel hoş
oda yetersiz	otel huzurlu
oda yüksek	otel iddialı
oda zayıf	otel ilgili
olanak ateşli	otel iyi
olanak başarılı	otel kaliteli
olanak dalgalı	otel kararsız
olanak eksik	otel kibar
olanak eski	otel kolay
olanak geniş	otel kötü
olanak güzel	otel kusursuz
olanak iyi	otel kuvvetli
olanak muhteşem	otel mutlu
olanak mükemmel	otel mükemmel
olanak yeterli	otel nadir
olanak yüksek	otel nasıl
otel nezih	personel kibar
otel önemli	personel muhteşem
otel özel	personel mükemmel
otel personel güler yüzlü	personel müthiş
otel personel güzel	personel odaklı
otel personelinin iyi	personel önemli
otel rahat	personel özverili
otel sakin	personel rahat
otel sıcak	personel saygılı
otel süper	personel sıcak
otel tam	personel standart
otel temiz	personel şanslı
otel tertemiz	personel tamam
otel türlü	personel temiz

Tablo 4.5. (Devam) Özellik-sentiment çiftleri

otel uygun	personel terbiyeli
otel yanlış	personel uzun
otel yeni	personel yaklaşımı güzel
otel yenilenmiş	personel yaklaşımı iyi
otel yeşil	personel yetersiz
otel yetkili	personel yoğun
otel yüksek	personelin hizmeti güzel
otel zevkli	personelin hizmeti iyi
otelin konumu güzel	personelin ilgisi güzel
otelin konumu iyi	personelin ilgisi iyi
otelin konumu yeşil	personellerin hizmeti hızlı
personel alakalı	personellerin hizmeti iyi
personel başarılı	restoran başarılı
personel boş	restoran büyük
personel çalışkan	restoran çeşitli
personel devamlı	restoran ender
personel düzgün	restoran geniş
personel edepi	restoran güleryüzlü
personel eksik	restoran güzel
personel geçerli	restoran harika
personel güleryüzlü	restoran hijyenik
personel güzel	restoran ilgili
personel harika	restoran iyi
personel hoş	restoran kaliteli
personel ilgili	restoran kolay
personel ilgisiz	restoran kötü
personel iyi	restoran lezzetsiz
personel kaliteli	restoran muhteşem
personel kalitesi güzel	restoran mükemmel
personel kanlı	restoran rahat
personel kaygılı	restoran rahatsız
restoran sıcak	tesis bakım
restoran süper	tesis büyük
restoran tatlı	tesis canlı
restoran temiz	tesis çocuklu
restoran uzak	tesis düzen
restoran yeni	tesis düzgün
restoran zayıf	tesis eğlenceli
restoran zor	tesis eksik
servis ateşli	tesis eski
servis ender	tesis geniş
servis güleryüzlü	tesis güleryüzlü
servis güzel	tesis güvenli
servis hızlı	tesis güzel
servis ilgili	tesis hijyenik
servis iyi	tesis hizmet kaliteli
servis kaliteli	tesis hoş
servis kalitesi güzel	tesis huzurlu
servis kalitesi iyi	tesis ilgili
servis kibar	tesis iyi

Tablo 4.5. (Devam) Özellik-sentiment çiftleri

servis mükemmel	tesis kaliteli
servis temiz	tesis kararsız
servis yüksek	tesis keyifli
tatil büyük	tesis kısıtlı
tatil ender	tesis kibar
tatil güzel	tesis kolay
tatil harika	tesis konforlu
tatil huzurlu	tesis kusursuz
tatil iyi	tesis küçük
tatil kaliteli	tesis misafirperver
tatil kapalı	tesis muhteşem
tatil kararsız	tesis mutlu
tatil keyifli	tesis mutsuz
tatil kötü	tesis mükemmel
tatil kusursuz	tesis özel
tatil muhteşem	tesis pahalı
tatil mükemmel	tesis personel güleryüzlü
tatil pozitif	tesis personel ilgili
tatil rahat	tesis rahat
tatil sakin	tesis sakin
tatil şahane	tesis sessiz
tatil uzak	tesis sıcak
tatil ücretsiz	tesis sorunsuz
tatil üstün	tesis tamam
tatil yegane	tesis temiz
tesis alakalı	tesis türlü
tesis uygun	yemek muhteşem
tesis yakın	yemek mükemmel
tesis yemekleri güzel	yemek nezih
tesis yeni	yemek önemli
tesis yenilenmiş	yemek özensiz
tesis yeşil	yemek rahatsız
tesis yeterli	yemek resmen
tesis yoğun	yemek sınırlı
tesis yorgun	yemek standart
tesis yüksek	yemek süper
yemek ağır	yemek şahane
yemek ağırlık	yemek tatlı
yemek ateşli	yemek taze
yemek başarılı	yemek temiz
yemek belli	yemek uygun
yemek boş	yemek yeşil
yemek büyük	yemek yeterli
yemek çeşitli	yemek yüksek
yemek çeşitliliği güzel	yemek zayıf
yemek çeşitliliği iyi	yemeklerin lezzeti uygun
yemek çeşitsiz	yemeklerin lezzeti yeterli
yemek düzenli	yiyecek belirli
yemek güleryüzlü	yiyecek bol
yemek güzel	yiyecek çeşitli

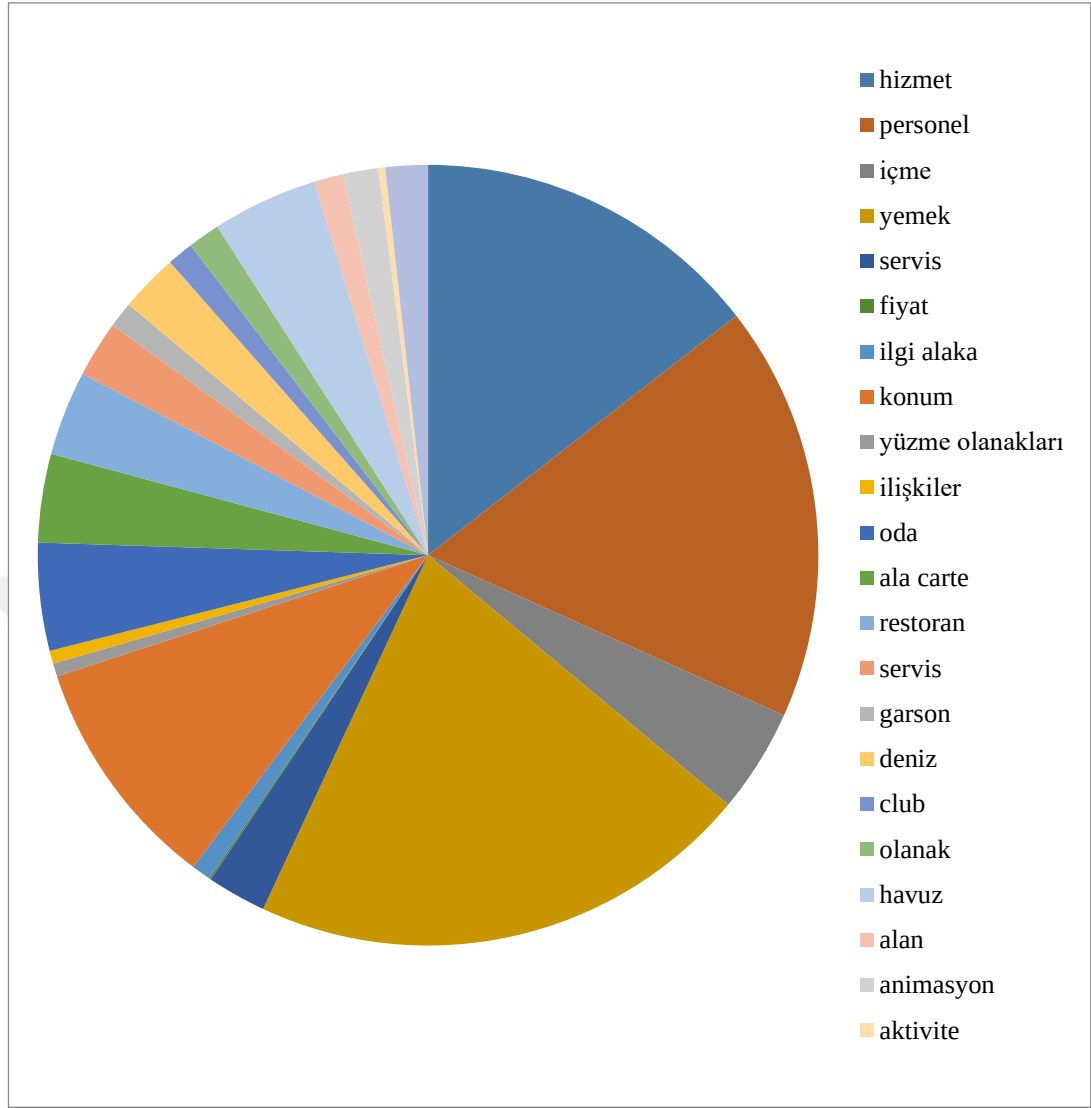
Tablo 4.5. (Devam) Özellik-sentiment çiftleri

yemek hafif	yiyecek düzen
yemek harika	yiyecek güzel
yemek harikulade	yiyecek içecek bol
yemek ilgili	yiyecek içecek güzel
yemek iyi	yiyecek içecek iyi
yemek kaliteli	yiyecek içecek kaliteli
yemek kalitesi artı	yiyecek içecek olağan
yemek kalitesi duyarlı	yiyecek içecek yüksek
yemek kalitesi güzel	yiyecek iyi
yemek kalitesi harika	yiyecek kaliteli
yemek kalitesi iyi	yiyecek kapalı
yemek kalitesi muhteşem	yiyecek lezzetli
yemek kalitesi süper	yiyecek lezzetsiz
yemek kalitesi yüksek	yiyecek markalı
yemek keyifli	yiyecek sorunsuz
yemek kırmızı	yiyecek temiz
yemek kötü	yiyecek uygun
yemek küçük	yiyecek yeşil
yemek lezzetli	yüzme olanakları güzel
yemek lezzetsiz	yüzme olanakları iyi

4.3.7. Özellik tabanlı skorlama

Özellik tabanlı skorlama adımı ÖSÇ'lerin çıkarılmasıyla iç içe çalışmaktadır. Sistem cümle içerisindeki ÖSÇ'lerin yakalanması ve sonrasında bu çiftlerdeki polarite hesabının yapılması şeklinde tasarlanmıştır. Bu hesabın gerçekleştirilmesi için ASbulma sınıfındaki polaritebul() fonksiyonu kodlanmıştır. Özellik tabanlı skorlama için kullanılan olumsuzlama ve pekiştirme zarfları kurallarına ait fonksiyonlar yazılmıştır.

Oluşturulan {özellik, sentiment, polarite} ifadeler manuel şekilde gruplandırılarak skorları toplanmış ve aşağıdaki Şekil 4.2'de gösterilmiştir. Bu grafiğe göre 502 puanla en çok yemek özelliği olumlu şekilde kullanıcılar tarafından konuşulmuştur. Fiyat özelliği ise 2 puanla en düşük skora sahip özellikler olarak belirlenmiştir. Yapılan manuel gruptama ise Tablo 4.5'de verilmiştir.



Şekil 4.2. Özellik tabanlı skorelama sonucu

Tablo 4.6. Gruplanan özellikler

Özellik	Skor	TESİS
tesis hizmet	1,0	
kaliteli hizmet	1,0	
müşteri hizmetleri	4,0	
personellerin hizmeti	7,0	
hizmet anlayışı	10,0	
hizmet kalitesi	82,0	
hizmet	243,0	
toplam	348,0	
		PERSONEL
tesis personel	2,0	
personel kalitesi	2,0	
otel personelinin	2,0	
otel personel	3,0	

Tablo 4.6. (Devam) Gruplanan özellikler

çalışan personel	5,0	PERSONEL
personel yaklaşımı	6,0	
Personellerin hizmeti	7,0	
personelin ilgisi	8,0	
ilgi alaka	19,0	
personel	360,0	
toplam	414,0	
içecek kalitesi	5,0	İÇME
içme olanakları	6,0	
içme imkanları	10,0	
yiyecek içecek	15,0	
içecek	67,5	
toplam	103,5	
yemeklerin kalitesi	0,0	YEMEK
yemeklerin lezzeti	1,0	
köy kahvaltısı	1,0	
tesis yemekleri	1,0	
açık büfe	4,0	
yemek çeşitliliği	4,0	
yemek kalitesi	13,0	
yiyecek içecek	15,0	
yiyecek	72,5	
yemek	390,5	
toplam	502,0	
servis kalitesi	3,0	SERVİS
servis	56,0	
toplam	59,0	
otelin konumu	35,0	KONUM
konum	198,0	
toplam	233,0	
misafir ilişkileri	-1,0	İLİŞKİ
müşteri ilişkileri	14,0	
toplam	13,0	
yüzme olanakları	21,0	YÜZME OLANAKLARI
toplam	21,0	
tesis olanakları	0,0	TESİS
tesis hizmet	1,0	
tesis yemekleri	1,0	
tesis	221,0	
toplam	223,0	
ala carte	88,0	ALA CARTE

Tablo 4.6. (Devam) Gruplanan özellikler

toplam	88,0	ALA CARTE
otel	205,5	OTEL
toplam	205,5	
restoran	84,5	RESTORAN
toplam	84,5	
animasyon	35,0	ANİMASYON
toplam	35,0	
aktivite	7,0	AKTİVİTE
toplam	7,0	
eğlence	42,0	EĞLENCE
toplam	42,0	
servis	56,0	SERVİS
toplam	56,0	
garson	25,5	GARSON
toplam	25,5	
deniz	58,0	DENİZ
toplam	58,0	
club	26,0	CLUB
toplam	26,0	
havuz imkanları	2,0	HAVUZ
havuz	103,0	
toplam	105,0	
olanak	32,0	OLANAK
toplam	32,0	
alan	29,0	ALAN
toplam	29,0	

Sistemin çıkışında elde edilen 61 tane özellik gruplanarak 22 tane özelliğe indirgenmiştir. 22 tane özelliğe ait toplam puanlamada yukarıdaki tabloda raporlanmıştır. Hedef internet kaynağında otel temel alana ilişki beş özelliğe göre oylama yapılırken, çalışmadaki sistem toplamda 61 tane özelliğe göre oylama yapmıştır. Tabloya göre bir kullanıcı yemek ile ilgili özelliğe bakarken, yemeklerin lezzeti, tesisin yemekleri, köy kahvaltısı, açık büfe gibi 10 alt özelliğin puanlamasına da bakıp, daha detaylı bir değerlendirme yapabilmektedir.

5. SONUÇLAR VE ÖNERİLER

Bu bölümde tez çalışmasının genel bir değerlendirilmesi yapılarak elde edilen önemli sonuçlar verilecek, önerilen ve uygulanan sistemin üstünlükleri değerlendirilecek ve sonraki çalışmalar için bazı öneriler tartışılacaktır.

İnternetin hızla yaygınlaşmasıyla beraber insanlar belli konular hakkındaki görüşlerini forumlar, bloglar ve sosyal medya platformlarında paylaşmaya başlamışlardır. Bu alanda hızla biriken veri ve bu verilere ulaşımın hızlı olması araştırma ve analiz için oda noktası haline gelmiştir. Günümüzde insanlar bir ürün, olay ve ya hizmet hakkında bilgi sahibi olmak istediklerinde, bir başkasına sormak yerine; ilgili kullanıcı yorumlarını okuyarak karar vermektedir. Çok büyük veri yığınları halindeki kullanıcı yorumları toplanarak; toplumun, bir ürün ve ya özellikleri hakkındaki genel yönelimlerini, fikirleri özellik tabanlı sentiment analizi ile elde etmek mümkündür. Bu tez çalışmasında bir ürünün zayıf ve güçlü özelliklerinin kullanıcı yorumlarından özellik tabanlı sentiment analizi ile keşfedilmesi amaçlanmıştır.

Sistemin ilk adımı olan, ürüne ait kullanıcı yorumlarının elde edilmesi için birçok internet veri kaynağı araştırılmıştır. Bu araştırmalar sonucunda www.otelpuan.com adresinde yer alan oteller hakkındaki kullanıcı yorumlarının, Türkçe dil kurallarına uygun yazılmış, daha nitelikli metinler içermesi sebebiyle deneysel çalışmalar için uygun olduğuna karar verilmiştir. Belirli bir otel için kullanıcı yorumlarının kaydedilmesinden sonra, özelliklerin çıkarımı adımı gerçekleştirilmiştir. Bu adımda özellikler tekli ve çok terimli olmak üzere Türkçe dilbilgisi kuralları doğrultusunda iki yapıda incelenmiş; çıkarılan özellikler ayırt 4.4.2 ve 4.4.3’de raporlanmıştır. Sonraki adım olan özellik sentiment çiftlerinin çıkarımında Türkçe metinlerde bu alanda daha önce uygulanmamış olan PDA yapıları kullanılarak yeni bir yöntem önerilmiş ve uygulanmıştır; elde edilen özellik-sentiment çiftleri ayırt 4.4.6’da listelenmiştir. Gerçeklenen sistemin son adımı olarak da çıkartılan özellik-sentiment çiftlerinin önceden belirlenmiş ve sonrasında otel temel alanı için genişletilmiş bir sentiment sözlüğü ile skorlanması gerçekleştirilmiştir.

Tez çalışması süresince özellik tabanlı sentiment analizinde karşılaşılan bir takım zorluklar yaşanmıştır. Bunlardan en önemlisi, sistemin en önemli adımı olan özellik çıkarımında kullanıcı yorumlarının dilbilgisi açısından bozuk olmasından dolayı kullanıcı yorumları üzerinde manuel düzeltmeler gerçekleştirilmiştir. Yine bu zorluk nedeniyle yalnızca Otel alanı için sonuçlar raporlanmıştır.

Çalışma Türkçe kullanıcı yorumları üzerinde yapılan ilk özellik tabanlı sentiment analizi çalışması olmasıyla önem taşımaktadır. Tezin katkıları aşağıda maddeler halinde verilmektedir:

- Türkçe’de sentiment analizi çalışmalarında kullanılmak üzere genel erişime açık kullanıcı yorumlarından oluşan veri kümeleri bulunmamaktadır. Tez kapsamında otel temel alanında kullanıcı yorumları veri kümeleri oluşturulmuştur.
- Otel puanda Kaynak sitelerde dört özelliğe (hizmet, yüzme, oda, yemek) ait puanlama yapılırken, geliştirilen sistem dört özellik ile birlikte başka özellikler de keşfederek daha detaylı puanlama gerçekleştirmiştir.
- Sistem farklı bir temel alana ait kullanıcı yorumlarına kolayca adapte olacak şekilde tasarlanmıştır. Bu özelliği ile alandan bağımsız ve oldukça esnek bir özellik tabanlı sentiment analizi uygulaması gerçekleştirilmiştir.
- Tezin temel adımları olarak belirlenen farklı uygulama süreçlerinde gerçekleşen yöntemler Türkçe için ilk olarak uygulanan yöntemlerdir. Bu açıdan tez çalışması ileride gerçekleştirilecek Türkçe sentiment analizi çalışmaları için yol gösterebilecek niteliktedir.

Bu çalışmanın devamı niteliğinde gerçekleştirilebilecek bazı çalışmalar aşağıda sunulmaktadır:

- Geliştirilen sistem mobil platforma adapte edilerek, kullanıcılara uygulama şeklinde sunulabilir. Kullanıcılar satın alacakları ürün ve ya hizmet ile ilgili özelliklerin puanlamasını uygulama ile cep telefon vb. cihazlardan görüntüleyebilir.
- Sistemin çıkışında verilen puanlamaya göre bir öneri sistemi geliştirilebilir. Örneğin; otel alanı için en yüksek puanı alan özelliğe sahip oteller gruplanarak kullanıcıya otel önerisi sunulabilir.

KAYNAKLAR

- [1] Hu M., Liu B., Mining and Summarizing Customer Reviews, *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, DOI:10.1145/1014052.1014073.
- [2] Hu M., Liu B., Mining Opinion Features in Customer Reviews, *AAAI*, 2004, 4(4), 755-760.
- [3] Popescu A. M., Nguyen B., Etzioni O., OPINE: Extracting Product Features and Opinions from Reviews, *HLT/EMNLP on Interactive Demonstrations*, DOI: 10.3115/1225733.1225750.
- [4] Brody S., Elhadad N., An Unsupervised Aspect-Sentiment Model for Online Reviews, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, USA, 2-4 June 2010.
- [5] Wei C., Chen Y. M., Yang C. S., Yang C. C., Understanding What Concerns Consumers: A Semantic Approach to Product Feature Extraction from Consumer Reviews, *Information Systems and E-Business Management*, 2010, 8(2), 149-167.
- [6] Thet T. T., Na J. C., Khoo C. S., Aspect-Based Sentiment Analysis of Movie Reviews on Discussion Boards, *Journal of Information Science*, DOI:0165551510388123.
- [7] Pak A., Paroubek P., Twitter as a Corpus for Sentiment Analysis and Opinion Mining, *Language Resources and Evaluation Conferences*, Valletta, Malta, 19-21 May 2010.
- [8] Dave K., Lawrence S., Pennock D. M., Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Review, *In WWW-03: Proceedings of the 12th International Conference on World Wide Web*, DOI:10.1145/775152.775226
- [9] Qiu G., Liu B., Bu J., Chen C., Opinion Word Expansion and Target Extraction Through Double Propagation, *Computational Linguistics*, 2011, 37(1), 9-27.
- [10] Zhang W., Xu H., Wan W., Weakness Finder: Find Product Weakness from Chinese Reviews by Using Aspects Based Sentiment Analysis, *Expert Systems with Applications*, 2012, 39(11), 10283-10291.
- [11] Bagheri A., Saraee M., De Jong F., Care more about Customers: Unsupervised Domain-Independent Aspect Detection for Sentiment Analysis of Customer Reviews, *Knowledge-Based Systems*, 2013, 52(11), 201-213.

- [12] Singh V. K., Piryani R., Uddin A., Waila P., Sentiment Analysis of Movie Reviews: A New Feature-Based Heuristic for Aspect-Level Sentiment Classification, *In Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, 2013 International Multi-Conference on, IEEE, Kerela, India, 22-23 March 2013, 712-717.
- [13] Che W., Zhao Y., Guo H., Su Z., Liu T., Sentence Compression for Aspect-Based Sentiment Analysis, *Audio, Speech and Language Processing, IEEE/ACM Transactions on*, 2015, **23**(12), 2111-2124.
- [14] Jiménez-Zafra S. M., Martín-Valdivia M. T., Martínez-Cámara E., Ureña-López L. A., Combining Resources to Improve Unsupervised Sentiment Analysis at Aspect-Level, *Journal of Information Science*, DOI:0165551515593686.
- [15] Bollacker K., Evans C., Paritosh P., Sturge T., Taylor J., Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge, *In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, San Francisco, USA, 10-12 June 2008.
- [16] Eroğul U., Sentiment Analysis in Turkish, Master Thesis, Middle Earth Technical University, Institute of Science, Ankara, 2009, 238615.
- [17] Boynukalın Z., Makine Öğrenimi Teknikleriyle Türkçe Metinlerde Duygu Analizi, Yüksek Lisans Tezi, Orta Doğu Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 2012, 318919.
- [18] Vural A. G., Cambazoglu B. B., Senkul P., Tokgoz Z. O., A Framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish, *Computer and Information Sciences III*, DOI: 10.1007/978-1-4471-4594-3_45.
- [19] Meral M., Diri B., Twitter Üzerinde Duygu Analizi, 24. *IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU)*, Zonguldak, Türkiye, 16-19 Mayıs 2014.
- [20] Kaya M., Sentiment Analysis of Turkish Political Columns with Transfer Learning, Master Thesis, Middle Earth Technical University, Institute of Science, Ankara, 2013, 34112.
- [21] Akba F., Uçan A., Akçapınar Sezer E., Sever H., Assessment of Feature Selection Metrics for Sentiment Analyses: Turkish Movie Reviews, *8th European Conference on Data Mining 2014*, Lisbon, Portugal, 15-17 July 2014.
- [22] Balahur A., Turchi M., Steinberger R., Perea-Ortega J. M., Jacquet G. D., Zavarella V., El Ghali A., Resource Creation and Evaluation for Multilingual Sentiment Analysis in Social Media Texts, *The 9th Edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, 28-30 May 2014.

- [23] Akbas E., Aspect Based Opinion Mining on Turkish Tweets, Master Thesis, Bilkent University, Institute of Science, Ankara, 2012, 309443.
- [24] Özsert C., Özgür A., Word Polarity Detection Using a Multilingual Approach, *Computational Linguistics and Intelligent Text Processing*, 8th Ed., Springer, Berlin, 75-82, 2013.
- [25] Çetin M., Amasyalı F., Active Learning for Turkish Sentiment Analysis, *Innovations in Intelligent Systems and Applications (INISTA)*, 2013 International Symposium, Albena, Bulgaria, 19-21 June 2013.
- [26] Uçan A., Otomatik Duygu Sözlüğü Çevirimi ve Duygu Analizinde Kullanımı, Yüksek Lisans Tezi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 2014, 379634.
- [27] Kumar A., Teeja M. S., Sentiment Analysis: A Perspective on Its Past, Present and Future, *International Journal of Intelligent Systems and Applications*, 2012, **4**(10), 1.
- [28] Wang J., Encyclopedia of Data Warehousing and Mining, *Information Science Reference*, 2005, **2**(3).
- [29] Medhat W., Hassan A., Korash H., Sentiment Analysis Algorithms and Applications: A Survey, *Ain Shams Engineering Journal*, 2014, **5**(4), 1093-1113.
- [30] Pang B., Lee L., Vaithyanathan S., Thumbs up?: Sentiment Classification Using Machine Learning Techniques, *ACL-02 Conference on Empirical Methods in Natural Language Processing*, 2002, **10**, 79-86.
- [31] Zhang Z., Ye Q., Zhang Z., Li Y., Sentiment Classification of Internet Restaurant Reviews Written in Cantonese, *Expert Systems with Applications*, 2011, **38**(6), 7674-7682.
- [32] Dang Y., Zhang Y., Chen H., A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews, *Intelligent Systems IEEE*, 2010, **25**(4), 46-53.
- [33] Turney P. D., Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews, *40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, USA, 6-7 July 2002.
- [34] Miller G. A., WordNet: A Lexical Database for English, *Communications of the ACM*, 1995, **38**(11), 39-41.
- [35] Hatzivassiloglou V., McKeown K. R., Predicting the Semantic Orientation of Adjectives. *35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, PA, USA, 10-15 May 1997, 174-181.

- [36] Moreo A., Romero M., Castro J. L., Zurita J. M., Lexicon-Based Comments-Oriented News Sentiment Analyzer System, *Expert Systems with Applications*, 2012, **39**(10), 9166-9180.
- [37] Bosco C., Patti V., Bolioli A., Developing Corpora for Sentiment Analysis: the Case of Irony and Senti-Tut, *IEEE Intelligent Systems*, 2013, **28**(2), 55-63.
- [38] Cho H., Kim S., Lee J., Lee J. S., Data-Driven Integration of Multiple Sentiment Dictionaries for Lexicon-Based Sentiment Classification of Product Reviews, *Knowledge-Based Systems*, DOI:10.1016/j.knosys.2014.06.001.
- [39] Peñalver-Martinez I., Garcia-Sanchez F., Valencia-Garcia R., Rodríguez-García M. Á., Moreno V., Fraga A., Sánchez-Cervantes J. L., Feature-Based Opinion Mining Through Ontologies, *Expert Systems with Applications*, 2014, **41**(13), 5995-6008.
- [40] Qiu L., Rui H., Whinston A., Social Network-Embedded Prediction Markets: The Effects of Information Acquisition and Communication on Predictions, *Decision Support Systems*, 2013, **55**(4), 978-987.
- [41] Weichselbraun A., Gindl S., Scharl A., Extracting and Grounding Context-Aware Sentiment Lexicons, *IEEE Intelligent Systems*, 2013, **28**(2), 39-46.
- [42] Liu B., Sentiment Analysis and Opinion Mining, *Synthesis Lectures on Human Language Technologies*, 2012, **5**(1), 1-167.
- [43] Qiu G., Liu B., Bu J., Chen C., Opinion Word Expansion and Target Extraction Through Double Propagation, *Computational Linguistics*, 2011, **37**(1), 9-27.
- [44] Ganu G., Elhadad N., Marian A., Beyond the Stars: Improving Rating Predictions Using Review Text Content, *12th International Workshop on the Web and Databases*, Rhode Island, USA, 28 June 2009.
- [45] Liu B., *Web Data Mining: Exploring Hyperlinks, Content and Usage Data*, 2nd Ed., Springer Science & Business Media, New York, 2008.
- [46] Thelwall M., A Web Crawler Design for Data Mining, *Journal of Information Science*, 2001, **27**(5), 319-325.
- [47] Dolgun M. Ö., Özdemir T. G., Oğuz D., Veri Madenciliğinde Yapısal Olmayan Verinin Analizi: Metin ve Web Madenciliği, *İstatistikçiler Dergisi: İstatistik ve Aktüerya*, 2009, **2**(2), 48-58.
- [48] Yüksel M. E., Turna Ö. C., Ertürk M. A., Bilgiye Erişim Sistemlerinde Veri Arama ve Eşleştirme, *Akademik Bilişim*, Muğla, Türkiye, 10-12 Şubat 2010.
- [49] Etzioni O., Cafarella M., Downey D., Popescu A. M., Shaked T., Soderland S., Yates A., Unsupervised Named-Entity Extraction from the Web: An Experimental Study, *Artificial Intelligence*, 2005, **165**(1), 91-134.

- [50] Blair-Goldensohn S., Hannan K., McDonald R., Neylon T., Reis G. A., Reynar J., Building a sentiment summarizer for local service reviews, *In WWW-3: Workshop on NLP in the Information Explosion Era*, 2008, **14**(2), 339-348.
- [51] Long C., Zhang J., Zhut X., A Review Selection Approach for Accurate Feature Rating Estimation, *The 23rd International Conference on Computational Linguistics*, Beijing, China, 23-27 August 2010.
- [52] Antonia A., Craig H., Elliott J., Language Chunking, Data Sparseness and the Value of A Long Marker List: Explorations with Word N-Grams and Authorial Attribution, *Literary and Linguistic Computing*, 2014, **29**(2), 147-163.
- [53] Baayen R. H., Hendrix P., Ramscar M., Sidestepping the Combinatorial Explosion: An Explanation of N-Gram Frequency Effects Based on Naive Discriminative Learning, *Language and Speech*, 2013, **56**(3) 329–347.
- [54] Gökdayı H., Türkiye Türkçe’sinde Öbekler, *Turkish Studies International Periodical For the Languages Literature and History of Turkish or Turkic*, 2010, **5**(2).
- [55] Türkmen H., Omurca İlhan S., Ekinci E., An Aspect Based Sentiment Analysis on Turkish Hotel Reviews, *ISEAIA*, 2015, **6**(2), 9-15.
- [56] Fokkink W., Grune D., Hond B., Rutgers P., Detecting Useless Transitions in Pushdown Automata, *VU University Departman of Computer Science*, DOI: arXiv:1306.1947.
- [57] <http://www.tdk.gov.tr/> (Ziyaret Tarihi: 3 Mayıs 2016).
- [58] Barracosa J. I. M. S., Mining Behaviors from Educational Data, Master Thesis, Institute of Science, Portugal, 2011.
- [59] Kanayama H., Nasukawa T., Fully Automatic Lexicon Expansion for Domain-Oriented Sentiment Analysis, *Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 22-23 July 2006, 355-363.
- [60] Kaji N., Kitsuregawa M., Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Russian, 17-22 June 2007.
- [61] Karamibekr M., Ghorbani A. A., Verb Oriented Sentiment Classification, *The 2012 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2012)*, Macau, China, 4-7 December 2012.
- [62] <https://github.com/ahmetaa/zemberek-nlp/> (Ziyaret Tarihi: 3 Mayıs 2016).
- [63] <https://github.com/yasserg/crawler4j/> (Ziyaret Tarihi: 3 Mayıs 2016).

- [64] <http://www.seleniumhq.org/> (Ziyaret Tarihi: 3 Mayıs 2016).
- [65] <https://developers.google.com/custom-search/json-api/v1/overview> (Ziyaret Tarihi: 3 Mayıs 2016).
- [66] <http://tools.ietf.org/html/rfc4627/> (Ziyaret Tarihi: 3 Mayıs 2016)
- [67] Nurseitov N., Paulson M., Reynolds R., Izurieta C., Comparison of JSON and XML Data Interchange Formats: A Case Study, Montana State University: Department of Computer Science, <http://www.cs.montana.edu/izurieta/pubs/IzurietaCAINE2009.pdf> (Ziyaret Tarihi: 4 Mayıs 2016).
- [68] <https://www.google.com.tr/trends/> (Ziyaret Tarihi: 3 Mayıs 2016).



KİŞİSEL YAYIN VE ESERLER

- [1] **Türkmen H.**, Omurca İlhan S., An Empirical Study for Turkish Sentiment Analysis by Machine Learning Methods, *International Conference on Advanced Technology & Sciences*, Antalya, Turkey, 12-15 August 2014.
- [2] **Türkmen H.**, Omurca İlhan S., Ekinci E., An Aspect Based Sentiment Analysis on Turkish Hotel Reviews, *ISEAIA*, 2015, **6**(2), 9-15.
- [3] Ekinci. E, **Türkmen H.**, Omurca İlhan S., Multi-word Aspect Extraction from User Reviews, *6th world conference on innovation and computer sciences*, Antalya, Turkey, 12-14 May 2016.
- [4] **Türkmen H.**, Ekinci E., Omurca ilhan S., A Novel Method For Extracting Feature Opinion Pairs For Turkish, *The 17th International Conference on Artificial Intelligence: Methodolgy, Systems, Applications*. Springer/LNCS (Kabul edildi.)

ÖZGEÇMİŞ

1991 yılında Almanya’da doğdu. İlk, orta ve lise öğrenimini Marmaris’de tamamladı. 2009 yılında girdiği Kocaeli Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü’nden 2013 yılında Bilgisayar Mühendisi olarak mezun oldu. 2013 yılında girdiği, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı’nda Yüksek Lisans öğrenimine devam etmektedir.

