**SİVAS CUMHURİYET ÜNİVERSİTESİ**
**Sosyal Bilimler Enstitüsü**
**Yönetim Bilişim Sistemleri Ana Bilim Dalı**

# DUYGU ANALİZİNDE FARKLI VEKTÖR TEMSİL YÖNTEMLERİ VE SINIFLAYICILARIN KARŞILAŞTIRILMASI

**Yüksek Lisans Tezi**

**Ayşegül ALBAYRAK**

**Sivas**
**Temmuz 2018**

SİVAS CUMHURİYET ÜNİVERSİTESİ
Sosyal Bilimler Enstitüsü
Yönetim Bilişim Sistemleri Ana Bilim Dalı

# DUYGU ANALİZİNDE FARKLI VEKTÖR TEMSİL YÖNTEMLERİ VE SINIFLAYICILARIN KARŞILAŞTIRILMASI

Yüksek Lisans Tezi

Ayşegül ALBAYRAK

**Tez Danışmanı**
Doç. Dr. Oğuz KAYNAR

Sivas
Temmuz 2018

## KABUL VE ONAY

| | |
|---|---|
| **Üniversite:** | **: Cumhuriyet Üniversitesi** |
| **Enstitü** | **: Sosyal Bilimler Enstitüsü** |
| **Ana Bilim Dalı** | : Yönetim Bilişim Sistemleri Ana Bilim Dalı |
| **Bilim Dalı** | : |
| **Tezin Başlığı** | **: Duygu Analizinde Farklı Vektör Temsil Yöntemleri ve Sınıflayıcıların Karşılaştırılması** |
| **Savunma Tarihi** | **: 23/07/2018** |
| **Danışmanı** | **: Doç. Dr. Oğuz KAYNAR** |

| | **Unvanı - Adı Soyadı** | **İmza** |
|---|---|---|
| **Jüri Başkanı** | **: Doç. Dr. Oğuz KAYNAR** | |
| **Üye** | **: Doç. Dr. Metin ZONTUL** | |
| **Üye** | **: Dr. Öğr. Üyesi Ahmet Gürkan YÜKSEK** | |

**Oy Birliği** ☑

**Oy Çokluğu** ☐

Ayşegül ALBAYRAK tarafından hazırlanan Duygu Analizinde Farklı Vektör Temsil Yöntemleri ve Sınıflayıcıların Karşılaştırılması başlıklı tez, kabul edilmiştir.

…./…./……...

**Prof. Dr. Ahmet ŞENGÖNÜL**
**Enstitü Müdürü**

# ETİK İLKELERE UYGUNLUK BEYANI

Cumhuriyet Üniversitesi Sosyal Bilimler Enstitüsü bünyesinde hazırladığım bu Yüksek Lisans/Doktora/Sanatta Yeterlik tezinin bizzat tarafımdan ve kendi sözcüklerimle yazılmış orijinal bir çalışma olduğunu ve bu tezde;

1- Çeşitli yazarların çalışmalarından faydalandığımda bu çalışmaların ilgili bölümlerini doğru ve net biçimde göstererek yazarlara açık biçimde atıfta bulunduğumu;

2- Yazdığım metinlerin tamamı ya da sadece bir kısmı, daha önce herhangi bir yerde yayımlanmışsa bunu da açıkça ifade ederek gösterdiğimi;

3- Başkalarına ait alıntılanan tüm verileri (tablo, grafik, şekil vb. de dahil olmak üzere) atıflarla belirttiğimi;

4- Başka yazarların kendi kelimeleriyle alıntıladığım metinlerini, tırnak içerisinde veya farklı dizerek verdiğim yine başka yazarlara ait olup fakat kendi sözcüklerimle ifade ettiğim hususları da istisnasız olarak kaynak göstererek belirttiğimi, beyan ve bu etik ilkeleri ihlal etmiş olmam halinde bütün sonuçlarına katlanacağımı kabul ederim.

24/07/2018

AYŞEGÜL ALBAYRAK

A. Albayrak

## ÖNSÖZ

Yüksek lisans tezim boyunca bilgi ve deneyimlerinden faydalandığım, benden yardımlarını esirgemeyen değerli hocam ve tez danışmanım Sayın Doç. Dr. Oğuz KAYNAR'a ve çalışmamda destekleri olan Sayın Arş. Gör. Yasin GÖRMEZ'e ve ilerlediğim bu yolda her zaman destek olup beni hiç yalınız bırakmayan aileme sonsuz teşekkürlerimi sunarım.

# İÇİNDEKİLER

# KISALTMALAR

**DA** : Duygu Analizi

**DDİ** : Doğal Dil İşleme

**KÇ** : Kelime Çantası

**NB** : Naive Bayes

**ME** : Maksimum Entropi

**LR** : Lojistik Regresyon

**KA** : Karar Ağaçları

**k-NN** : k- En Yakın Komşu

**DVM** : Destek Vektör Makineleri

**YSA** : Yapay Sinir Ağları

**MLP** : Multilayer Perception

**TF** : Terim Frekansı

**TDF** : Ters Doküman Frekansı

**CBoW** : Countinuous Bag of Word

**SG** : Skip-Gram

**HS** : Hiyerarchical Softmax

**NS** : Negative Sampling

# TABLOLAR LİSTESİ

# ŞEKİLLER LİSTESİ

# ÖZET

Son yıllarda bilişim teknolojileri hayatımıza hızlı bir şekilde girmiştir ve hayatımızı dijital cihazlar veya internet olmadan hayal etmek imkânsız hale gelmiştir. Bugün bloglar, yorum siteleri, sosyal medya platformları sadece kullanıcılara bilgi yayan bir kaynak değil aynı zamanda kullanıcıların birbirleriyle iletişim kurmalarını ve görüşlerini paylaştıkları ortamlar haline gelmişlerdir. Bu tür verilerin bir kısmı öznel olmasına karşın analiz ve karar desteği gibi çeşitli amaçlar için analiz edilebilir bilgiler içermektedir. Bu tür verileri kullanabilmek ve bu verileri işlemek için duygu analizi olarak da adlandırılan bir araştırma alanı ortaya çıkmıştır.

Duygu Analizi, bir metnin otomatik olarak sınıflandırılmasıyla bir konuşmacının veya bir yazarın belirli bir konuyla ilgili tutumunu(pozitif, negatif) belirlemeyi amaçlamaktadır. Etkin sınıflandırma sağlamak için etkili metin temsillerinin oluşturulması çok önemlidir. Bu nedenle tezin ana amacı duygu tahmininde uygulanabilecek yöntem ve teknikleri araştırmaktır. Tez kapsamında iki farklı dilde yazılmış iki veri seti kullanılarak uygulama gerçekleştirilmiştir. Farklı metin temsilleri oluşturma tekniklerinin ve farklı dillerde yazılmış metinlerin duygu sınıflamasında ki etkileri üzerinde çeşitli analizler gerçekleştirilmiştir.

**Anahtar Kelimeler:** Duygu Analizi, Makine Öğrenimi, Sınıflandırma, Kelime Çantası, N-gram, Word2vec

# ABSTRACT

In recent years, information technology has quickly entered our lives and it has became impossible to imagine our life without digital devices or the internet. Today, blogs, review sites, social media platforms are not only a source of information for users but also a place where users communicate with each other and share their opinions. While some of these of data are subjective, they contain analytical information for various purposes such as analysis and decision support. A research area has emerged which is named sentiment analysis in order to be able to use and process such data.

The sentiment analysis purposes to determine the attitude (positive, negative) of a speaker or an author on a particular topic by automatically classifying a text. Creating effective text representations is very important to provide effective classification. Fort his reason, the main purpose of the thesis is to investigate the techniques that can be applied to sentiment prediction. In the scope of the thesis the implementation was carried out by using two data sets written on two different language. Various analyzes have been carried out on the effects of different text representation techniques and texts written in different languages on sentiment classification.



**Keywords:** Sentiment Analysis, Machine Learning, Classification, Bag of Word, N-gram, Word2vec

# GİRİŞ

İnsanlar ve makineler arasındaki fark, insanların kişisel görüşlerini dile getirebilme yeteneğine sahip olmaları ve yapay zekânın temelinde yatan makinenin insanlar gibi davranması hayalidir. Görüşleri analiz eden bilgisayar dilbilimi alanı görüş madenciliği ya da duygu analizi (DA) olarak adlandırılmaktadır. DA, ürünler hizmetler ve hatta insanlar hakkındaki görüşlerin analizi ile ilgilenen Doğal Dil İşleme(DDİ) alanının bir parçasıdır. DA öncelikle pozitif ya da negatif duyguları ifade eden ya da ima eden görüşlere odaklanmaktadır.

Günümüzde görüş içerikli metinlerin elde edilmesi oldukça kolaylaşmıştır çünkü milyonlarca insan Twitter, Facebook gibi sosyal ağlar aracılığıyla farklı konular hakkındaki görüşlerini paylaşmakta ya da belirli bir web sitesinde kullandıkları ürünlerle ilgili yorum ve değerlendirmeler bırakmaktadır. Mikrobloglar düşünceleri paylaşmanın son derece popüler bir yoludur ve her gün çok miktarda mesaj yayınlanmaktadır. Bu nedenle mikrobloglar toplanabilen ve duyguların çıkarılması için oldukça yararlanılabilen zengin görüş içerikli mesaj kaynağı olarak düşünülebilir. Görüşlerin analizi tüm bilim alanlarında (politika, ekonomi, sosyal yaşam) önemli bir rol oynamaktadır. Örneğin; pazarlamada eğer satıcı belirli bir ürünün müşteri memnuniyetini bilirse ürün üzerindeki talebi tahmin edebilir. Politikacılar için de benzer şekilde, insanların onları destekleyip desteklemediklerini bilebilir.

Duygu sınıflandırma problemi yeni bir araştırma alanı değildir. Fakat araştırmanın odak noktası bugün çok rağbette olan sosyal medya araçlarının üzerinde değil, daha büyük metin belgelerinin analizi ve sınıflandırılması üzerindeydi.

Duygu analizi problemi için kullanılabilecek çeşitli teknikler mevcuttur. Temel yaklaşımlar makine öğrenmesine dayalı yaklaşımlar ve sözlüğe dayalı yaklaşımlardır. Makine öğrenimi yaklaşımı, belirli bir metnin duygusunu tanımlamak için uygulanacak olan eğitim sınıflandırıcı için veri kümesi kullanırken, sözlük tabanlı yaklaşım ise bir metnin pozitif veya negatif olup olmadığını tanımlamak için kelime ya da ifadelerin anlamsal yönelimlerini kullanır.

Duygu analizi (DA) için kullanılan tekniklerden birisi sözcüğe dayalı yöntemdir. Her terim için ilgili duygu puanlarına sahip terimlerden oluşan bir sözlük

kullanır. Terim tek bir kelime, ifade veya deyim ile ilişkilendirilebilir (Bosco, Bosco, Pilato, & Chiavetta, 2016: 159). Duygu, sözlükteki terimlerin varlığına veya yokluğuna dayalı olarak tanımlanır. Sözlüğe dayalı yöntemler, derlem tabanlı yaklaşım ve sözlük tabanlı yaklaşım olmak üzere iki temel yaklaşım içerir.

Sözlük tabanlı yaklaşımın ana fikri, bir belgedeki duyguyu çıkarmak için görüş kelimelerini içeren kelime veri tabanlarını kullanmaktır. Bu tür veri tabanlarına örnek olarak WordNet (Miller, 1995: 39-42), HowNet (Dong, Dong, & Hao, 2010:53-56), SentiWordNet (Musto, Semeraro, & Polignano, 2014: 60), SenticNet (Musto vd., 2014: 60), MPQA (Musto vd., 2014: 60) vb. verilebilir.

Bugüne kadar başta İngilizce olmak üzere farklı dillerde birçok duygu analizi çalışması yapılmıştır. Hu ve Liu (2004) araştırmalarında müşteri görüşleri ve değerlendirmelerinin sınıflandırılması üzerinde odaklanmışlardır, yani duygu içeren ürün özelliklerini çıkarmışlar, daha sonra bu özelliklere dayalı sınıflandırılmış cümleler çıkarılmış ve sonuç olarak ürün incelemelerinin özeti oluşturmuşlardır. Örneğin, değerlendirme bir kamera hakkında yapılmışsa, araştırmacılar kamera görüntüsü, kalitesi ve boyutu gibi özellikleri almış ve bu özellikleri kullanarak pozitif ve negatif kamera değerlendirmelerinde sınıflandırma yapmışlardır. Bir cümle için pozitif ya da negatif bir etiket atamak için önce araştırmacılar her ürün değerlendirmesinden kutupsal kelimeleri çıkarmışlardır ve bunun için sıfatları kullanmışlardır. Tahmin, eş anlamlı olarak kutupları aynı olan ve zıt anlamlı kelimenin tam tersi olan bir sıfatın kutbuna dayanmaktadır. Kelimelerin eş anlamlarını ve zıtlıklarını elde etmek için WordNet sözlüğünü kullanmışlardır. Uyguladıkları yöntem iyi sonuç vermiş, ortalama %84 doğruluk oranı elde etmişlerdir. Araştırma sonucu olarak sıfatların cümle kutupluluğunu tahmin edilmesinde etkili olabileceğini belirtmişlerdir.

Kim ve Hovy (2004) tarafından gerçekleştirilen çalışmada belirli bir konuyla ilgili metnin ve metin sahibinin duyguları araştırılmıştır. Çalışmada yazarlar birçok sınıflandırıcı uygulamıştır. İlk sınıflandırıcı kelimelerin kutuplarını elde etmek için cümledeki her kelimeye uygulanmıştır. İkinci sınıflandırıcı, görüş sahibinin ifade ettiği tüm cümlelerin kutupluluklarını tanımlamak için kullanılmıştır. Hu ve Liu (2004)'nin çalışmalarında olduğu gibi çekirdek kelime listelerini genişletmek için

WordNet kullanmışlardır. Ayrıca çalışma, kelimelerin belirsizliğini ortadan kaldırmak için kelimelerin pozitiflik ve negatiflik kuvvetinin tanımlanması gerektiğini vurgulamıştır.

Park ve Kim (2016) çalışmalarında çekirdek kelime listelerine dayalı eşanlamlı ve zıt anlamlı kelimeleri elde etmek için üç farklı sözlük (genel olarak yalnızca biri kullanılmaktadır) kullanan bir yöntem geliştirmişlerdir. Daha sonra, tweet sınıflandırması için genişletilmiş sözlük kullanmışlardır. Yazarlar, geleneksel sözlük tabanlı yöntemin yetersiz olduğunu önerdikleri yöntemin tweet'leri sınıflandırmaya olanak tanıdığını ifade etmişlerdir. Fakat önerilen yaklaşımın çeşitli dezavantajları vardır. Kelimelerin eş ve zıt anlamlarının toplanması çok fazla zaman gerektirmektedir. Ayrıca sözlükler genellikle daha biçimsel kelimeler içerir fakat tweetler biçimsel olmayan kelimeler ile doludur.

Genel olarak, sözlük tabanlı yaklaşımın en büyük dezavantajı alan ve bağlamsal özel kutup yönelimlerine sahip duygu kelimelerini tespit edememesidir(Medhat, Hassan, & Korashy, 2014). Bu nedenle derleme dayalı yaklaşım ortaya atılmıştır.

Bing Liu, derleme dayalı yaklaşımın iki durumda uygulanabileceğini belirtmiştir (2012). İlki, belirli görüş kelime kümesi kullanarak alan derlemindeki görüş kelimelerinin ve kutuplarının tanımlanmasıdır. İkinci durum, bir alan derlemi kullanan başka bir sözlükten belirli bir alan adına yeni bir sözlük oluşturmaktır. Bazı araştırmacılar görüş kelimeleri alan bağımlı olsa bile, aynı kelimenin içeriğe bağlı olarak zıt yönde olabileceğini düşünmektedir.

Hazivassiloglou ve McKeown (1997) tarafından yapılan araştırmalar, literatür de derleme dayalı teknik konusunda öne çıkmaktadır. Çalışmalarında derlemden sıfatların anlamsal yönelimini çıkaran bir yöntem önermişlerdir. Bu teknik, metinsel derlem ve çekirdek görüş kelimelerinin (sıfatların) kullanımı üzerine kuruludur. Görüş kelimelerini ortaya çıkarmak için özel dil kuralları derleme uygulanmıştır. Araştırmacılar, sıfatlar "and" bağlacıyla birleştirilirse kutuplarının aynı olduğunu varsaymaktadır. Fakat "but" bağlacının ters kutuplu sıfatları bağlamak için kullanıldığını belirtmişlerdir. Ayrıca "or", "either-or", "neither-nor" gibi bağlaçlar da kullanılmaktadır. Fakat bazen bu kurallar geçerli değildir. Bu nedenle, çalışmada

aynı zamanda, sıfatların kutuplarının aynı olup olmadığını kontrol edilmesi öngörülmüştür, bu amaçla log-doğrusal regresyon modeli kullanılmıştır. Tahmin aşamasından sonra sıfatlar arasındaki bağlantıyı sağlayan bir grafik elde edilmiş ardından sıfatlar pozitif ve negatif alt sınıflara bölmek için grafik üzerinde kümeleme yapmışlardır. Sonuç olarak, Hatzivassiloglou ve McKeown çalışmalarında % 90 oranında doğruluk elde etmişlerdir.

Yukarıda belirtildiği gibi, benzer duygu kelimeleri bağlama bağlı olarak farklı anlamsal yönelimlere sahip olabilir. Bu amaçla Ding ve diğ. (2008: 231-240), kullanıcılar tarafından iletilen duygunun yönelimini bulmak için bir yöntem önermişlerdir. Araştırmacılar, bazı sıfatların (çoğunlukla nicelik sıfatları, örneğin uzun, kısa vb.) bağlam-bağımlı olduklarını ve bunların kelimelerin kutuplarını değiştirebileceklerini vurgulamışlardır. Ürün özelliklerinin kutupluluklarını belirlemek için duygu kelimelerini cümledeki yönelimi ile birlikte ele almışlardır. Ding ve diğ. görüş sözlüğü olarak kelimeleri, cümleleri ve deyimleri kullanmışlardır. Sıfatlar ve zarfların listesini Hu ve Liu'nin çalışmasından (2004: 168-177) alarak fiil ve isimleri ekleyerek genişletmişlerdir. Bunlara ek olarak duygu içeren yaklaşık 1000 yeni deyim daha eklemişlerdir. Sözlük hazır olduktan sonra, cümlelerdeki her özellik için kutupsallık puanı tanımlamışlardır. Ek olarak "but" ifadesini içeren olumsuzlukları ve cümleleri işlemek için birçok dil kuralı uygulamışlardır. Ayrıca çalışma bağlam bağımlı duygu kelimelerinin belirlenmesi ile ilgili problemi çözmek için bütünsel bir yaklaşım getirmiştir. Bu amaç için cümle içi bağlaç tekniği, sözde cümle içi bağlaç tekniği ve cümleler arası bağlaç tekniği olmak üzere üç teknik önermişlerdir. Özetlemek gerekirse araştırmacılar, önerilen yaklaşımın daha önceki yöntemlere göre daha etkili ve iyi sonuç verdiğini belirtmiştir.

Derlem tabanlı yöntem, derlem içerisindeki kelimelerin kısıtlı olması nedeniyle tek başına kullanıldığında sözlük tabanlı yöntemden daha az etkilidir. Fakat bu yaklaşımın kullanılması alan ve bağlamsal özel sözlük oluşturmakta yardımcı olabilir.

Genel olarak, sözlük tabanlı yöntemlerin zaman karmaşıklığı ve doğruluğu bakımından performansı, sözlükte bulunan kelime sayısına bağlıdır, yani, kelime boyutu büyüdükçe performans önemli ölçüde azalır.

DA için kullanılabilecek ikinci teknik, aşağıda açıklanan denetimsiz ve denetimli makine öğrenme yöntemlerini içeren makine öğrenmesine dayalı yöntemlerdir.

Denetimsiz öğrenme yaklaşımı, girdi verilerinden yapıyı keşfetmek ve benzer kalıpları bulmak için etiketlenmemiş veri kümelerini kullanır. Denetimsiz öğrenme yaklaşımları etiketli verilerin, etiketsiz verilere oranla elde edilmesinin daha zor olduğu durumlarda kullanılmaktadır.

Denetimsiz öğrenme alanındaki en temel çalışmalardan biri Turney (2002) tarafından gerçekleştirilen çalışmadır. Değerlendirmeleri, önerilen (beğenilen) ve önerilmeyen( beğenilmeyen) sınıflara ayırmıştır. Bu çalışmada, etiket modellerine dayalı iki kelimeden oluşan ifadeler dikkate alınmıştır. Modeller, cümlelerdeki duygu ifadelerini yakalayacak şekilde tasarlanmıştır. Her bir ifade sıfat/zarf, isim/fiil birleşimidir (toplam 5 model önerilmiştir). Hangi ifadelerin alınması gerektiğine karar vermek için Part-of-speech (POS) etiketleyicisi uygulanmıştır. Değerlendirmelerden alınan ifadelerin anlamsal yönelimin hesaplanması için Pointwise Mutual İnformation (PMI) ve Latent Semantic Analysis yöntemleri kullanılmıştır. PMI iki terim arasındaki anlamsal benzerliği ölçmektedir. Modellere uyan bir ifade ilk terim olarak alınır ve referans kelimesi ikinci terim olarak alınır. Bir değerlendirme bir yıldız aldığında "poor" ve beş yıldız aldığında "excellent" olarak derecelendirme yapmak olağan olduğu için "excellent" ve "poor" kelimeleri referans kelimeler olarak kabul edilmiştir. Bir ifadenin anlamsal yönü "excellent" ifade ve "poor" ifade arasındaki fark olarak tanımlanmıştır. Anlamsal yönelim, bir ifade "excellent" referans kelimesiyle daha güçlü bir ilişkiye sahipse pozitiftir ve eğer ilişki "poor" ile daha güçlü ise negatiftir. Ortalama anlamsal yönlendirme pozitifse, değerlendirme önerilen olarak aksi halde önerilmeyen olarak etiketlenmiştir. Turney bu yöntem ile ortalama %74 oranında doğruluk elde etmiştir.

Rothfels ve Tibshirani (2010) film değerlendirmelerinin duygu sınıflandırması için denetimsiz bir yöntem uygulamıştır. Bu çalışmada Çince metinlerinin sınıflandırılması için Zagibalov ve Carroll tarafından önerilen yöntem (Zagibalov & Carroll, 2008: 1073-1080) uyarlanmıştır. Yöntem, belgelerden çıkarılan olumlu çekirdek kelime listelerini kullanır. Bu tür duygu kelimeleri (zarflar)

olumsuzlamalardan önce gelir veya olumsuzlama olmadan meydana gelebilir (en yaygın durum). Zagibalov ve Carroll'un çalışmalarında (2008:1073-1080) kullanılan pozitif çekirdek kelime listesi yinelemeli sınıflandırma kullanılarak zenginleştirilmiştir. Çalışmada sınıflandırılacak belgenin metni bölgelere bölünmüş ve her bölge noktalama işaretleri arasında bulunan metin parçasına karşılık gelmektedir. Daha sonra her bölgenin sınıflandırılması gerçekleştirilmiştir. Tüm metnin duygusu, belgede pozitif veya negatif bölgelerin baskınlığı ile tanımlanmıştır. Yani, pozitif bölgeler negatif bölgelere göre daha sık ortaya çıkarsa, belge pozitif aksi halde negatif olarak kabul edilir. Rothfels ve Tibshirani çalışmalarında çekirdek kelimeler olarak bigramlar, trigaramlar ve 4-gram modellerini kullanmayı denemişler ve çekirdek kelime listesini genişletmişlerdir. Fakat ilk ikisi, 4-grama kıyasla cümle içeriğini tam olarak koruyamamıştır. Çalışma sonucunda yeterli başarım oranı elde edilemediği görülmüştür. Başarımın artırılabilmesi için araştırmacılar sıfat öbeklerini çekirdek kümesi olarak kullanmışlardır. Fakat beklenildiği gibi doğruluğun iyileştirilmesi sağlayamamışlarıdır. Araştırmacılar puanlama yöntemini k-ortalama kümeleme yöntemine çevirmeye çalışmış ancak elde edilen sonuçlar önemli bir iyileştirme göstermemiştir. Son denemelerinde Turney (2002:417-424) tarafından önerilen ve bir cümlenin anlamsal yönelimini tahmin eden yaklaşımı uyarlamışlardır. Rothfels ve Tibhirani manuel olarak iki çekirdek kelime listesi (pozitif ve negatif) oluşturmuşlardır. Daha sonra, metindeki her kelime ile referans çekirdek kelimeler arasındaki anlamsal yönelim tahmini gerçekleştirmişlerdir. Uyarlanmış algoritma, her kelimenin anlamsal yönelimini ilk duygu puanı olarak kullanan yinelemeli bir sınıflamadır. Bu sefer %50,3 olan doğru sınıflandırma oranının %65,5'e çıktığı gözlemlemişlerdir.

Etiketli verilerin artan kullanılabilirliği, denetimli makine öğrenme yöntemlerinin duygu analizine uygulanmasında önemli bir rol oynamıştır. Bu yöntemler, bir dizi özellik biçiminde etiketli verileri temsil eder. Daha sonra özellikler etiketi bilinmeyen verilerin sınıflandırılmasında fonksiyon öğretmek için kullanılır. Genellikle kelime çantası (KÇ) modeli bir belgeyi özellik vektörü olarak temsil etmek için kullanılır (Tang, Kay, & He, 2016: 2508). Eğitim veri setini bir özellik vektörüne dönüştürmek için, eğitim verilerinden $N$ benzersiz kelimeli kelime grubu oluşturulması gerekir. Ayrıca, ikili özellik modeli, terim sıklığı (TF), terim

sıklığı- ters doküman sıklığı (TF-TDF) , bilgi kazancı, ki-kare gibi özellik modellerinden herhangi biri, bir özellik vektörünün oluşturulması için kullanılabilir. Veri kümesi bir vektör olarak gösterildikten sonra, sınıflandırıcı tarafından etiketlerin öğrenilmesi ve tahmini için kullanılabilir. Sınıflandırıcıyı eğitmek için farklı yöntemler kullanılabilir.

Metin sınıflandırması için kullanılan en yaygın ve basit yöntem Naive Bayes'tir. Model, özelliklerin bağımsız olduğu varsayımı ile Bayes teoremine dayanmaktadır(Tang vd., 2016: 2509). Naive Bayes sınıflandırıcısı, belgenin belirli bir sınıfa ait olma ihtimalini tanımlar. Geleneksel sınıflandırma algoritmalarından bir diğeri ise Maksimum Entropi (ME)'dir (Go, Bhayani, & Huang, 2009: 3). Sınıflandırma için bir başka yaklaşım Destek Vektör Makineleridir (DVM) ve birçok çalışmada uygulanmıştır (Go vd., 2009: 3) (Gautam & Yadav, 2014: 439) (Pang, Lee, & Vaithyanathan, 2002:82). Yöntem, uzayın belli sınıflara karşılık gelen alt uzaylara bölünmesini varsayar. İkili sınıflandırma açısından, eğitim aşamasının amacı, bir veri kümesini en iyi marjla iki sınıfa ayıran en iyi hiper düzlemi keşfetmektir.

Etiketli verileri kullanan yaklaşımın en önce gelen örneği, Pang ve diğ. tarafından yapılan çalışmadır(2002). Araştırmacılar, bir film değerlendirmesinin pozitif veya negatif olup olmadığını belirlemek için "Movie Review" veri seti üzerinde Naive Bayes (NB), Maksimum Entropi (ME) ve Destek Vektör Makineleri (DVM) sınıflandırıcılarını uygulamıştır. N gram, Part-of- speech (POS) etiketleri ve bunların kombinasyonları özellikler olarak kullanmışlardır ve üçlü çapraz doğrulama kullanmışlarıdır. En yüksek doğru sınıflandırma oranının %82,9 oranla destek vektör makineleri ile verinin 1-gram yöntemi ile temsil edildiğinde alındığı gözlemlemişlerdir.

Pang ve Lee (2004) "Movie Review" veri seti üzerinde bir başka çalışma gerçekleştirmişlerdir. Bu çalışmada yazarlar cümleleri öznel ve nesnel olarak sınıflandırmış ve nesnel cümleleri inceleme dışı bırakmışlardır. Önerilen iyileştirme yöntemi ile öznel cümlelerin, nesnel cümlelere kıyasla daha bilgi verici olduğunu ifade etmişlerdir. Önerdikleri bu yöntem ile doğru sınıflandırma yüzdesini %87'e çıkarmışlardır.

“Movie Review” veri seti üzerinde uygulanan bir başka çalışma Whitlelaw ve diğ. tarafından gerçekleştirilmiştir (2005). Çalışmalarında metnin anlamsal bağlantılarını daha iyi yansıtabilmek için Kelime Çantası (KÇ) vektör gösterimi ile Destek Vektör Makineleri eğitilerek %90,2 oranında doğru sınıflandırma oranı elde etmişlerdir.

Denetimli makine öğrenme yöntemleri yeterli sayıda etiketli veri bulunduğunda etkin bir şekilde çalışmaktadır. Fakat doğal dil işleme sürecinde veri setlerine ek açıklamalar eklemek oldukça maliyetlidir. Bu nedenle Zaidan ve diğ. ( 2007) çalışmalarında gönüllü gruplardan belirli bir ifadenin pozitif veya negatif olduğunu belirten ek açıklamalar eklemesi istemişlerdir. Belgelerin duygu kutbu etiketlerine ek olarak insanlardan elde edilen ek açıklamalar veri sınıflandırılması için kullanılmıştır. Yöntem, “Movie Review” veri seti üzerinde destek vektör makineleri ile uygulanmıştır. Çalışma sonucunda %92.2 oranında doğru sınıflandırma oranı elde etmişlerdir.

Tan ve Zhang (2008) çalışmalarında, dört özellik seçimi yöntemi (belge sıklığı ölçütü ki-kare ölçütü, karşılıklı bilgi ve bilgi kazancı) ve beş sınıflandırma yöntemini (centroid sınıflandırıcı, k-en yakın komşu, destek vektör makineleri, naive bayes ve winnow sınıflandırıcı) Çince metinler üzerinde uygulanmıştır. Sınıflandırma yöntemlerinin performansı duyarlılık, kesinlik ve F ölçütlerine dayalı olarak karşılaştırmışlardır ve en iyi sonuç bilgi kazancı özellik seçimi ve destek vektör makineleri sınıflandırma yöntemi ile elde etmişlerdir.

Bai (2011) “IMDb v1.0”, “IMDb v2.0”, “M&A News”, “Financial News” ve “Mixed News” gibi farklı veri setleri kullanarak duygu analizi problemi için Markow model sınıflandırıcılarına dayalı bir algoritma tasarlamıştır. Veri setlerindeki kelimeler arasındaki bağlantılar Markow model sınıflandırıcı ile belirlemiş ve daha sonra Tabu arama algoritması kullanılarak algoritmanın parametreleri iyileştirilmiştir. Geliştirilen algoritma klasik makine öğrenme algoritmaları ile karşılaştırmıştır (Destek Vektör Makineleri, Naive Bayes ve Maksimum entropi gibi). Geliştirilen algoritma ile “IMDb v1.0” veri setinde %92,7 oranında başarı oranı elde etmiştir.

Veri setinin dengesiz dağılımı sınıflandırma performansını olumsuz etkilemektedir. Veri seti dengesizliği problemini ortadan kaldırmak için Kang ve diğ (2012) çalışmalarında restoranlara ilişkin değerlendirmelerin bulunduğu veri seti üzerinde Naive Bayes yöntemini bigram ve unigram kullanarak iyileştirmiştir. İyileştirilmiş Naive Bayes yöntemi ile %84,4 doğruluk oranı elde etmişlerdir.

Son yıllarda birçok araştırmacı tarafından "Twitter'da Duygu Analizi" alanında çok fazla çalışma yapılmıştır. Pak ve Paroubek (2010) Tweetleri tarafsız, pozitif, negatif sınıflandırmak için bir model önermiştir. Çalışmalarında Twitter API kullanarak tweetleri toplayarak ve his simgelerini kullanılmasıyla bu tweetleri otomatik olarak etiketleyerek bir derlem oluşturmuşlardır. Bu derlemi kullanarak, N-gram ve POS etiketleri gibi özellikleri kullanan çok terimli Naive Bayes yöntemine dayalı bir duygu sınıflandırıcısı geliştirmişlerdir. Kullandıkları eğitim seti sadece his simgelerini içeren tweetleri içerdiği için daha az etkili olduğunu gözlemlemişlerdir.

Parikh ve Movassate (2009) ise tweetleri sınıflandırmak için Naive Bayes bigram modeli ve Maksimum Entropi modeli olmak üzere iki model uygulamışlardır. Çalışma sonucunda Naive Bayes sınıflandırıcısının Maksimum Entropi modeline göre daha başarılı olduğunu ifade etmişlerdir.

Go vd. (2009) distant supervision yöntemini kullanarak twitter verileri duygu analizi için bir çözüm önermişlerdir, söz konusu çalışmada eğitim verileri gürültülü etiketler olarak kullanılan his simgelerin bulunduğu tweetlerden oluşmaktadır. Naive Bayes, Maksimum Entropi ve Destek Vektör Makineleri kullanarak modeller oluşturmuşlar ve unigram, bigram ve POS etiketlerini sınıflandırıcı modellerin özellikleri olarak kullanmışlardır. DVM'nin diğer modellerden daha iyi performans gösterdiğini ve unigramın özellikler olarak daha etkili olduğunu belirtmişlerdir.

Barbosa ve Feng (2010) ise tweet'leri sınıflandırmak için iki aşamalı otomatik duygu analizi yöntemi tasarlamışlardır. İlk olarak tweeteleri nesnel veya öznel olarak sınıflandırdılar ve ikinci aşamada öznel tweetleri pozitif veya negatif olarak sınıflandırmışlardır. Kullanılan özellik alanı, kelimelerin ve POS etiketlerin kutupsallığı gibi özelliklerle birlikte retweet'leri, hashtag'ları, bağlantıları, noktalama işaretlerini ve ünlem işaretlerini içermektedir.

Liang ve Dai (2013) twitter verilerini toplamak için Twitter API kullanmıştır. Kamera, film, telefon gibi üç farklı kategoriye ait eğitim verileri olumlu, olumsuz ve görüş içermeyen olarak etiketlemişlerdir. Daha sonra görüş içeren cümleler filtrelemişlerdir. Çalışmada Unigram Naive Bayes modeli ve bağımsızlık varsayımını basitleştiren Naive Bayes modeli kullanmışlardır. Ayrıca karşılıklı bilgi ve Ki-kare özellik çıkarma yöntemini kullanarak duygu içermeyen özellikleri de kaldırmışlardır. Son olarak, bir tweet'in anlamsal yönelimi pozitif veya negatif olarak tahmin etmişlerdir.

Duygu sınıflamada özellik vektörlerinin oluşturulması için sık kullanılan ve etkili bir yöntem olan kelime çantası modelinin çok sıfırlardan oluşması bu nedenle de metinleri yeterince iyi temsil edememesi ayrıca modelin sadece kelimelerin belgede olup olmadığıyla ilgilenmesi yani kelimelerin anlamsal yönelimini göz ardı etmesi büyük bir dezavantajdır. Bu nedenle kelimeler arasındaki anlamsal ilişkileri de dikkate alan Word2vec ve Doc2vec gibi yöntemler metin ve duygu sınıflama çalışmalarında sıklıkla kullanılmaya başlanmıştır (Şahin, 2017). Venekoski vd. ( 2016) çalışmalarında bu sorunların sınıflandırma başarısını etkileyip etkilemediğini incelemiştir. Sosyal medya metinlerini tf-tdf ağırlıklı KÇ modeli, Word2vec kelime vektörleri ve Doc2vec paragraf vektörleri olmak üzere üç farklı yöntem kullanarak DVM ile sınıflanmışlardır. Ayrıca kelime köküne inme işlemi uygulayarak sınıflandırma başarısına olan etkisini de incelemişlerdir. Deneyler sonucunda KÇ tf-tdf için %77,3 word2vec için %77,8, doc2vec için %77,8 başarı elde etmişlerdir. Kelime köküne inme işlemi ile başarının %73,4'ten %77,2'e çıktığını gözlemlemişlerdir.

Jiang vd. (2016) ise çalışmalarında 1-5 aralığındaki otel ve restoran müşteri değerlendirme puanlarını olumlu ve olumsuz olarak ele almışlarıdır ve iki farklı veri seti oluşturmuşlardır(TripAdvisor ve Yelp). Her iki veri setini tdf ve tf-tdf ağırlıklandırmalar ile temsil etmişlerdir. Birinci veri setinden (TripAdvisor) KÇ modeli için %93, word2vec tdf için %93, word2vec tf-tdf çin %95 AUC değeri elde edilirken, ikinci veri kümesi(Yelp) için sırasıyla %87, %86, %90 AUC değeri elde etmişlerdir.

Bansal ve Srivastava (2018) çalışmalarında cep telefonu kategorisinde yazılan 400.000'den fazla tüketici yorumları veri setini kullanarak duygu analizi gerçekleştirmişlerdir. Metinleri Word2vec kelime vektörleri ile temsil etmişlerdir ve DVM, NB, LR, RF makine öğrenmesi algoritmaları ile sınıflamışlardır. En iyi sonucu CBoW mimarisinin kullanıldığı vektör gösterimi ile RF sınıflandırma algoritmasından %90,66 AUC değeri ile elde etmişlerdir.

Literatürdeki duygu analizi çalışmaları incelendiğinde daha çok sınıflandırma yöntemlerinin üzerinde odaklanıldığı görülmüştür. Bu tez çalışmasında iki ayrı çalışma gerçekleştirilmiştir. Birinci çalışmada İngilizce film yorumları ele alınarak bu yorumların analizine odaklanılmıştır. İkinci çalışmada ise Türkçe film yorumları veri seti üzerinden duygu analizi yapılmıştır. Duygu analizinin başarılı bir şekilde gerçekleşmesini sağlamak için kaliteli metin temsillerinin oluşturulması üzerine odaklanılmış olup TF ve TF-TDF ağırlıklandırma yöntemleri ile elde edilen Kelime Çantası modellinin yanı sıra Word2vec kelime vektörlerinin duygu sınıflandırılmasında ki etkisi ölçülmüştür. Bu tezin amacı farklı metin temsilleri oluşturma tekniklerinin hem İngilizce hem Türkçe metinlerin duygu analizinde ki etkilerini ölçmektir.

Bu amaçla tezin ikinci bölümünde, duygu analizi ve kullanılan yöntemler ele alınmış makine öğrenmesine dayalı duygu analizinde kullanılan çeşitli sınıflandırma yöntemleri incelenmiştir. Metin ön işleme, kelimelerin vektör ile temsilleri, Kelime Çantası ve Word2vec modelleri açıklanmıştır. 3. Bölümünde öncelikle üzerinde çalışmada kullanılan veri setleri hakkında bilgiler verilmiştir. Ardından metin önişleme ve farklı öznitelik çıkarma yöntemleri ile vektör uzay modellerinden oluşan veri seti temsilleri oluşturulmuştur. Oluşturulan bu veri seti temsilleri üzerinde sınıflama algoritmaları yardımıyla çeşitli deneyler gerçekleştirilmiştir. 4 bölümde elde edilen sonuçlar tartışılmış ve ilerde yapılabilecek çalışmalara ilişkin önerilerde bulunulmuştur.

# 2. DUYGU ANALİZİ

Görüş madenciliği olarak da bilinen Duygu Analizi (DA) 90'lı yıllardan beri incelenmektedir. Fakat 2000'li yıllarda DA farklı bilimsel alanlardaki önemi nedeniyle araştırmacıların ilgisini çekmiştir, ayrıca duygu analizinin araştırılmamış birçok sorunsalı vardır. Dahası duygu içeren verilerin geniş ölçüde bulunması, bu alandaki araştırmayı yeni bir aşamaya itmiştir. O yıllardan beri DA, Doğal Dil İşleme (DDİ) çalışmalarındaki en aktif araştırma alanı haline gelmiştir.

DDİ, belli bir bilgi elde etmek ya da belirli bir amaç için kullanılabilecek bilgi yapılarını türetmek amacıyla doğal dil ile yazılan cümlelerin bilgisayar tarafından anlaşılır bir hale getirilmesini sağlayan çalışmalardır (Chowdhury, 2003: 51). DDİ çalışmaları, harflerin seslerini ve bunların dil içerisinde nasıl kullanıldığının incelenmesi, kelimenin kökünü ve eklerinin doğru olarak tespit edilmesi, kelimelerin cümle oluşturmak için ne şekilde sıralanması gerektiğini, kelimelerin cümle içindeki yapısal görevlerinin tanımlanması, kelimelerin anlamlarını ve kelimelerin normal anlamından farklı anlamda kullanıldığı durumların incelenmesi gibi birçok konuyu kapsamaktadır. Kısacası DDİ, insanların konuştuğu dili bilgisayarlara öğretmektedir ve çeşitli uygulamalarla insan yaşamına kolaylıklar getirmektedir.

Duygu analizi sistemi her bir cümleyi veya belgeyi tam olarak anlamaya ihtiyaç duymaz, yalnızca metinlerin bazı yönlerini, yani pozitif ya da negatif duyguları, bu duyguların hedef varlıklarını ya da konularını anlamalıdır. Bu bağlamda duygu analizinin bir DDİ problemi olduğunu bilmek faydalıdır. DDİ araçları duygu analizi sürecini kolaylaştırmak için kullanılabilir ve Duygu analizinin daha doğru sonuçlar vermesine yardımcı olabilir.

Bing Lui'e göre (2012:7) "duygu analizi, ürünlerin, hizmetlerin, kuruluşların, bireylerin, sorunların, olayların, konuların ve bunların özellikleri gibi öğelere yönelik insanların fikirlerini, duygularını, değerlendirmelerini, tutumlarını ve hislerini analiz eden çalışma alanıdır."

Başka bir ifadeyle, duygu analizi, belirli belgelerdeki duyguları çıkarmak ve kategorilere ayırmak için görüş içeren metnin işlenmesini ele alır. Duygunun kutbu genellikle pozitif ya da negatif olarak ifade edilir (ikili sınıflandırma). Bununla

birlikte çok sınıflı sınıflandırma olabilir, duygu tarafsız bir etikete veya hatta çok pozitif, pozitif, tarafsız, negatif, çok negatif gibi genişletilmiş çeşitliliğe sahip olabilir. Ayrıca duygu etiketleri üzgün, kızgın, korkulu, mutlu gibi duygularla ilişkilendirilebilir.

Bazı araştırmacılar duygu analizi ve görüş madenciliği arasında farklılıklar olduğunu belirtmişlerdir (Tsytsarau & Palpanas, 2012: 481). Ancak, hem duygu analizi hem de görüş madenciliği temelde aynı çalışma alanını temsil eder. Türk Dil Kurumu tarafından *duygu*, belirli nesne, olay veya bireylerin insanın iç dünyasında uyandırdığı izlenim, nesneleri veya olayları ahlaki ve estetik yönden değerlendirme olarak tanımlanırken, *görüş* ise bir olay, varlık veya düşünce üzerinde varılan yargı, fikir olarak tanımlanmaktadır (TDK). Her ikisi arasındaki farkı ayırt etmek oldukça zor ve her iki tanımında diğerinin bazı elementlerini içerdiği görülmektedir. Örneğin; "mevcut politik durum hakkında endişeliyim" ifadesi bir duygudur oysaki "politika iyi gitmiyor" cümlesi bir görüştür.

Duygu analizi terimleri için bir örnek aşağıda verilmiştir,

$< cümle >= filmin\ hikayesi\ zayıf\ ve\ sıkıcıydı$

$< görüş\ sahibi > = < yazar >$

$< hedef > = < film >$

$< özellik > =< hikaye >$

$< görüş > = < zayıf >< sıkıcı >$

$< kutup > = < negatif >$

Matematiksel olarak bir görüş $(o, f, so, h, t)$ şeklinde gösterilebilir. Burada, $o =$ hedef , $f =$ hedef $o$'nun özellikleri, $so =\ o$ hedefinin $f$ özelliğine ilişkin yönü veya kutbu, $h =$ görüş sahibi, $t =$görüşün ifade edildiği zaman.

**Hedef:** Kişi, olay, ürün, kuruluş veya konu olabilir,

**Özellik:** Hedefin değerlendirildiği bir özelliği,

**Görüş yönü veya kutbu:** Bir özellik hakkındaki görüş yönelimi, görüşün pozitif, negatif veya tarafsız olup olmadığını gösterir,

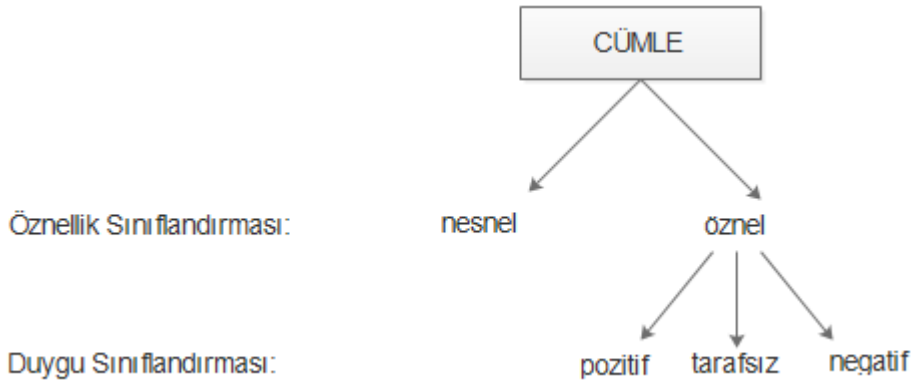
**Görüş sahibi:** Görüşünü ifade eden kişi veya kuruluş.

Duygu analizi, insanların özellikle kuruluşların ilgisini çeken gelişmekte olan bir alandır. Çünkü DA karar verme süreci için kullanılabilir. Bireyler, artık belirli ürün veya hizmetlerle ilgili olarak arkadaşlarından gelen görüşleri sormakla sınırlı değildir, bu bilgileri internette özgürce bulabilirler. Ayrıca kuruluşlar anket yapmaktan kaçınarak zamandan ve paradan tasarruf edebilmekte, bunun yerine Web 'den özgürce elde edilebilecek görüşleri işleme konusuna odaklanabilmektedirler. Fakat Web' den elde edilen bu veriler gürültülü verilerdir. Bu nedenle bu verilenden daha fazla yararlanılmak için metinlerden asıl anlamın çıkarılması önemlidir. DA bu tür gürültülü veriler ile başa çıkmak için farklı teknikler ve yaklaşımlar kullanmaktadır.

### 2.1. Metinlerde Duygu Tanımı

Öznel ve nesnel cümle kategorileri birçok metin işleme uygulaması için önemlidir. Duygu analizinde de ilk amaç öznel ve nesnel cümleleri ayırt etmektir. Eğer bir cümle nesnel olarak sınıflandırılmışsa başka hiçbir işleme gerek yoktur, cümle öznel olarak sınıflandırılmışsa onun kutbunun (pozitif, negatif, tarafsız) tahmin edilmesi gerekir. Başka bir deyişle duygu sınıflandırılması yapılmadan önce nesnel bilgi ifade eden cümleleri, öznel görüş ve düşünceleri ifade eden cümlelerden ayıran öznellik sınıflandırması yapılması gerekir (Wiebe, Bruce, & O'Hara, 1999).

Duygu analizi akışı Şekil 1'deki gibidir,




**Şekil 1.** Duygu Analizi İş Akışı

### 2.2. Duygu Analiz Seviyeleri

Daha önce de belirtildiği gibi, duygu analizinin amacı doğal dil metinlerinden öznel bilgileri otomatik çıkaran araçları tanımlamaktır. Duygu analizi uygularken ilk olarak, metnin ne anlama geldiği tanımlanmalıdır.

Duygu analizinde metin üç seviyede incelenir (Liu, 2012:10). Bu durum Şekil 2.'de grafik olarak gösterilmiştir.


**Şekil 2.** Duygu Analizin Farklı Seviyeleri

#### 2.2.1. Doküman Seviyesi

Doküman seviyesi duygu analizi, görüş bildiren tüm dokümanları pozitif ya da negatif bir görüş ya da düşünce olarak sınıflandırmayı amaçlamaktadır. Bu seviyede asıl amaç, bütün belgenin duygusunu tanımlamaktır (duygu bir konuda özetlenmelidir)(Pozzi, Fersini, Messina, & Liu, 2016: 7).

Liu'nun (2010) belirttiği gibi, doküman seviyesi duygu sınıflandırması, dokümanın tek bir varlık hakkında görüş ifade ettiğini ve görüşlerin tek bir görüş sahibinden olduğunu varsaymaktadır. Birden fazla varlığı değerlendiren veya karşılaştıran dokümanlarımız varsa doküman analiz seviyesi yeterli olmayabilir ve cümle seviyesinde duyu sınıflandırması uygulayarak daha fazla ayrıntı elde edilebilir.

### 2.2.2. Cümle Seviyesi

Cümle seviyesi duygu sınıflandırması, doküman seviyesi duygu analizine göre daha ayrıntılı bir görünüm vermektedir. Aynı zamanda doküman seviyesi analiz teknikleri cümlelere uygulanabilmektedir.

Bu analiz seviyesi cümlenin tek bir görüş sahibinden tek bir görüş ifade ettiğini varsaymaktadır(Liu, 2010). Ancak söz konusu bu durum her zaman böyle değildir. Liu'nun (2012) belirttiği gibi birçok karmaşık cümlenin farklı hedefler üzerinde farklı duyguları vardır (ör. "iPhone mükemmel bir kameraya sahip ancak pil ömrü ve güvenlik sorunları üzerinde çalışılması gerekiyor"). Ayrıca cümle seviyesi duygu sınıflandırmasının diğer dezavantajları, karşılaştırılmalı cümlelerin (ör."iOS Android'den daha iyi performans gösteriyor"), soru formunda formüle edilmiş cümlelerin (ör."iOS Android'den daha iyi midir?) ve görünen içeriğin tam tersi anlamına gelen alaylı cümlelerin (ör."iOS çok iyi!!!") içindeki görüşlerle baş edememesi gerçeğinden kaynaklanmaktadır.

Doküman seviyesi veya cümle seviyesi analizleri çoğu durumda iyi bir yaklaşım olmasına rağmen gerekli ayrıntı seviyesine ulaşamayabilirler. Bu gibi durumlarda varlık ve nitelik seviyesi analizi iyi bir alternatif sağlamaktadır.

### 2.2.3. Varlık ve nitelik seviyesi

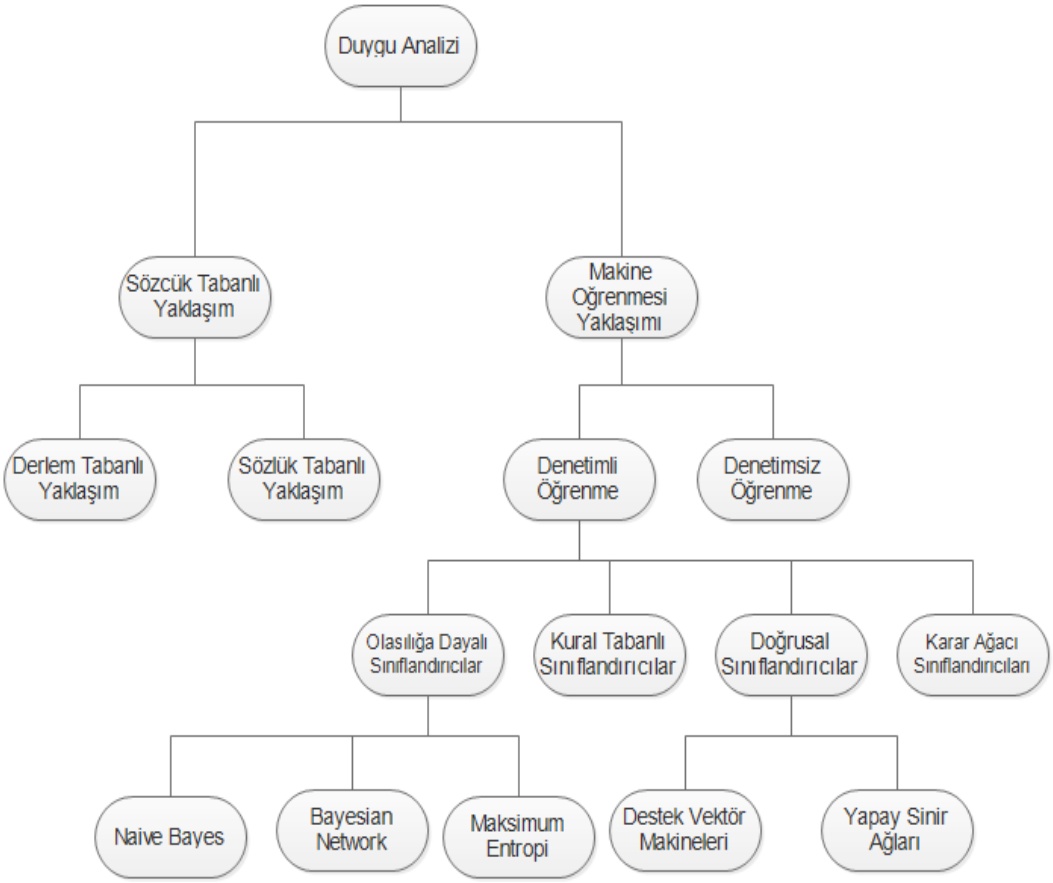
Önceki iki yaklaşım, tüm doküman veya her bir cümle tek bir varlığa tekabül ettiğinde çok iyi sonuç vermektedir. Fakat metinler çok yönleri olan farklı varlıklara tekabül edebilir ve her bir varlık veya her konu hakkında görüş farklı olabilir.

Varlık ve nitelik seviyesi duygu analizi daha önceleri özellik seviyesi olarak adlandırılmıştır (özellik tabanlı görüş madenciliği ve özetleme) (Hu & Liu, 2004) ve varlıklara ve/veya varlıkların yönlerine göre duyguları keşfetme amacına sahiptir. Feldman (2013), varlık ve nitelik seviyesi duygu analizini, belirli bir dokümanın içindeki tüm duygu ifadelerinin tanımlanmasına ve bunların yönünün vurgulanmasına odaklanan araştırma problemi olarak tanımlamıştır.

Varlık ve nitelik seviyesi duygu analizinin amacı, ürün değerlendirmelerinin tüm yönlerini tanımlamak ve tüm isim öbeklerini çıkarmak ve sık kullanılan isim öbeklerini elde etmektir (Hu & Liu, 2004).

### 2.3. Duygu Sınıflandırma Teknikleri

Duygu analizi sınıflandırma teknikleri esas olarak makine öğrenimi ve sözcük tabanlı yaklaşımlar olmak üzere ikiye ayrılmıştır (Medhat vd., 2014) (Şekil 3). Makine öğrenimi (MÖ) yaklaşımı, geleneksel makine öğrenimi algoritmalarını uygular ve dilsel özellikleri kullanır. Sözcük tabanlı yaklaşım, bilinen ve önceden derlenmiş duygu terimlerinin bir koleksiyonu olan duygu sözlüğüne dayanır. Sözcük tabanlı yaklaşım, duygu kutbunu bulmak için istatiksel ve anlamsal yaklaşımlarını kullanan derlem tabanlı yaklaşım ve sözlük tabanlı yaklaşım içerir. Bu tekniklerin daha ayrıntılı açıklaması aşağıdaki alt bölümde verilmiştir.




**Şekil 3.** Duygu Sınıflandırma Teknikleri

### 2.3.1. Sözcük Tabanlı Yaklaşım

Sözlük tabanlı yaklaşımlarda, duygu ifade eden kelimeler duygu sınıflandırma işleminin en önemli öğeleridir. Bunlar pozitif veya negatif duyguları ifade eden kelime veya birkaç kelimeden oluşan kelime grupları olabilir. Örneğin, iyi, harika vb. pozitif duygu kelimeleri beğenilen bir durumu ifade etmek için kullanılırken, kötü, korkunç vb. negatif duygu kelimeleri beğenilmeyen bir durumu ifade etmek için kullanılır. Bu duygu kelimeleri ve kelime gruplarının listesi duygu sözlüğü olarak adlandırılır.

Duygu kelimelerinin derlenip duygu sözlüğü oluşturulması için kullanılan üç temel yaklaşım vardır (Liu, 2012: 90):

- Manuel yaklaşım
- Sözlük tabanlı yaklaşım
- Derlem tabanlı yaklaşım

Manuel yaklaşım, bir kişi veya gönüllü gruplar tarafından kelimelerin toplanıp ardından kelimelerin ifade ettiği duyguların manuel olarak etiketlenmesidir. Bazı araştırmacılar bu yaklaşımı geçmişte seçmişlerdir. Taboada ve diğ. (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011) yaptığı çalışmasında otomatik olarak oluşturulmuş kelimeler için kesinlik eksiliği nedeniyle manuel olarak bir sözlük oluşturmaya karar verdiklerini belirtmişlerdir. Ancak manuel yaklaşım çok zaman alıcıdır ve nadiren kullanılır. Bu nedenle otomatik yöntemlerden kaynaklanan hataları önlemek için otomatik yaklaşımlarla birlikte kontrol için kullanılır.

#### 2.3.1.1. Sözlük Tabanlı Yaklaşım

Hu ve Lui (2004) çalışmalarında sözlük tabanlı yaklaşımın ana stratejisini sunmuşlardır. Sözlük tabanlı yaklaşımda ilk olarak duygu kelimelerinin küçük bir kümesini manuel olarak oluşturulur. Ardından, WordNet gibi çevrimiçi sözlüklerde kelimelerin eş ve zıt anlamları aranır ve bulunan yeni kelimeler bu kümeye eklenir ardından bir sonraki yinelemeli işlem başlar. Yinelemeli işlem yeni bir kelime bulamadığında durur. İşlem tamamlandıktan sonra hataları gidermek veya düzeltmek için manuel olarak denetleme yapılır.

Sözlük tabanlı yaklaşım çok sayıda duygu kelimesinin elde edilmesinde kolaylık sağlamaktadır fakat diğer bir taraftan farklı bağlamlardaki farklı anlamları olan görüş kelimelerini ayırt edememektedir. Örneğin, "artış" kelimesi "kar" içinse pozitif anlamdadır ancak "borç" söz konusuysa negatif anlam taşımaktadır. "artış" kelimesinin duygu yönelimi metinin içeriğine bağlı olarak değişmektedir.

Sözlük tabanlı yaklaşım, belirli bir alanın belirli özelliklerini temsil edemediğinden derlem tabanlı yaklaşım bu sorunun alternatif bir çözümüdür.

#### 2.3.1.2. Derlem Tabanlı Yaklaşım

Sözlük tabanlı yaklaşımın alan ve bağlamsal özel kutup yönelimlerine sahip görüş kelimelerini tespit edememesi büyük bir dezavantajdır. Derlem tabanlı yaklaşım, benzer kelimelerin bir bağlamda pozitif ve diğerinde negatif olabileceği problemi çözmeye çalışır. Derlem tabanlı yaklaşımdaki yöntemler, sözdizimsel veya birlikte-oluşum kalıplarına ve ayrıca büyük bir derlemdeki diğer görüşleri bulmak için görüş kelimelerinin bir çekirdek listesine dayanmaktadır(Liu & Zhang, 2012). Bu alandaki önemli gelişme, duygu kesinliği kavramını tanıtan Hatzivassiloglou ve McKeown (1997) çalışmasıdır. Çalışmalarına duygu sıfatlarından oluşan bir çekirdek kümesiyle başlamışlardır ve bunlara ek sıfat duygu kelimelerini ve yönelimlerini belirlemek için "AND, OR, BUT, EITHER-OR…" gibi bağlaçlar için bir dizi dil kuralları kullanmışlardır. Örneğin, "AND" ile bağlanmış sıfatların genellikle aynı duygu yönelime sahip olduğunu ifade etmişlerdir.

Liu ve Zhang'ın (2012) belirttiği gibi tüm kelimeleri kapsayacak kadar büyük bir derlem hazırlamak zor olduğu için tek başına derleme dayalı yaklaşım kullanıldığında sözlük tabanlı yaklaşım kadar etkili değildir. Ancak derlem tabanlı yaklaşım, belirli bir alana özgü derlem kullanarak alan ve bağlama özgü görüş kelimelerini ve yönelimlerini bulmaya yardımcı olabilecek önemli bir avantaja sahiptir.

### 2.3.2 Makine Öğrenimi Yaklaşımı

Duygu analizi çalışmalarında kullanılan ikinci yaklaşım makine öğrenimidir. Bir makine öğrenimi sistemi, matematiksel ve istatiksel yöntemler kullanarak mevcut eğitim verilerinden bilgi çıkarımı yapar daha sonra yeni verilerin çıktısını tahmin etmek için bu bilgiyi kullanır. Makine Öğrenme yöntemlerini genel olarak denetimli,

denetimsiz ve yarı-denetimli öğrenme yöntemleri şeklinde sınıflandırmak mümkündür. Hangi öğrenme yönteminin kullanılacağı, uygulanan problem tipine göre belirlenir. Bu çalışmada önceden etiketlenmiş duygu içeren metinlerin sınıflandırma problemi ele alınmıştır bu nedenle denetimli öğrenme yöntemleri kullanılmıştır. Denetimli öğrenmede girdi verileri etiketli bir eğitim setidir ve algoritma, bu eğitim verilerinin analiz edilmesiyle bir fonksiyon üretir. Yeni verilerin sınıflandırılması istenildiğinde algoritma, eğitim sürecinde elde edilen çıkarım fonksiyonu ile sınıflandırma işlemini gerçekleştirmektedir.

Bu bölümün devamında metin sınıflandırma probleminin tanımı ele alınmıştır. Ardından duygu analizi için kullanılan farklı makine öğrenimi yöntemleri açıklanmıştır.

#### 2.3.2.1. Sınıflandırma Problemi

Sınıflandırma problemi tanımlanırken, $C = \{c_1, c_2, \dots c_n\}$ sınıf etiketleri bilinen $X_1, X_2, \dots, X_N$ dokümanlarından oluşan $D = \{X_1, X_2, \dots, X_N\}$ etiketli bir eğitim seti belirlenir. Bu eğitim setini sınıflayan bir sınıflandırma fonksiyonu oluşturulur. DA'i çalışmalarında $C$ üç sınıftan oluşmaktadır $C = \{pozitif, negatif, tarafsız\}$. Daha sonra sınıfı bilinmeyen bir metnin sınıf etiketini tahmin etmek için tanımlanan sınıflandırma fonksiyonu kullanılır (Medhat vd. , 2014).

#### 2.3.2.2. Naive Bayes

Naive Bayes (NB) sınıflandırma yöntemi, basit ve yaygın olarak kullanılan sınıflandırma tekniğidir. Belirli bir özellik kümesinin belirli bir etikete ait olma ihtimalini tahmin etmek için Bayes teoremini kullanır.

Naive Bayes yöntemi, bir sınıfın her özelliğinin diğer özelliklerden bağımsız olduğunu varsaymaktadır. Yani tüm özellikler, sonuç olasılığını bağımsız olarak etkiler. Bu basitleyici varsayıma bağlı olarak Naive Bayes yöntemi özellikle girdi boyutunun yüksek olduğu durumlarda uygun bir yöntemdir (McCallum & Nigam, 1998).

Bir $n$ boyutlu uzayda ve $m$ farklı sınıflarda, bir örneğin üyelik olasılıkları $X = \{f_1, f_2, \dots, f_n\}$ aşağıdaki denklem ile hesaplanır (Demirci, 2014:11).

$$P(C_i|x) = \frac{P(C_i)P(x|C_i)}{P(x)} \qquad [1]$$

Denklem aşağıdaki şekilde yeniden yazılabilir.

$$P(C_i|x) = \frac{P(C_i)P(f_1, f_2, \dots f_n|C_i)}{P(f_1, f_2, \dots f_n)} \qquad [2]$$

$x$'in özelliklerinin bağımsız olarak dağıldığını belirten Naive varsayımı dikkate alındığında denklem aşağıdaki gibi yeniden yazılabilir.

$$P(C_i|x) = \frac{P(C_i)P(f_1|C_i, f_2|C_i \dots f_n|C_i)}{P(f_1, f_2, \dots f_n)} \qquad [3]$$

Olasılığı en yüksek sınıf, sınıf etiketi olarak belirlenir.

$$Sınıf = \arg\max(P(C_i|x) \qquad [4]$$

Payda sabit olduğu için göz ardı edilir ve denklem basitleştirilir.

$$P(Ci)P(ci|x) \propto P(Ci)P(Ci)P(f_1|C_i, f_2|C_i \dots f_n|C_i) \qquad [5]$$

#### 2.3.2.3. Lojistik Regresyon

Lojistik Regresyon (LR) sınıflandırma algoritması makine öğrenimi uygulamalarında etkili olduğu kanıtlanmış alternatif bir tekniktir(Berger, Della Pietra, & Della Pietra, 1996). LR algoritması, iki sınıflı bir problemde, sınıflar üzerinde bir dağılım bulmak için lojistik fonksiyonlar kullanmanın aynısıdır bu nedenle Lojistik regresyon (LR) yöntemi olarak bilinir.

LR sınıflandırması, Naive Bayes'in aksine özellikler arasındaki ilişkiler hakkında P(c|d) tahminini elde etmek için herhangi bir varsayımda bulunmaz(Appel, Chiclana, Carter, & Fujita, 2016). Söz konusu P(c|d) tahmini aşağıdaki eşitlikte gibi ifade edilir.

$$P_{LR} \coloneqq \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d, c)\right) \qquad [6]$$

Burada, $Z(d)$ bir normalizasyon fonksiyonudur ve $F_{i,c}$, özellik $f_i$ ve sınıf $c$ için aşağıdaki gibi tanımlanan bir özellik/sınıf fonksiyonudur.

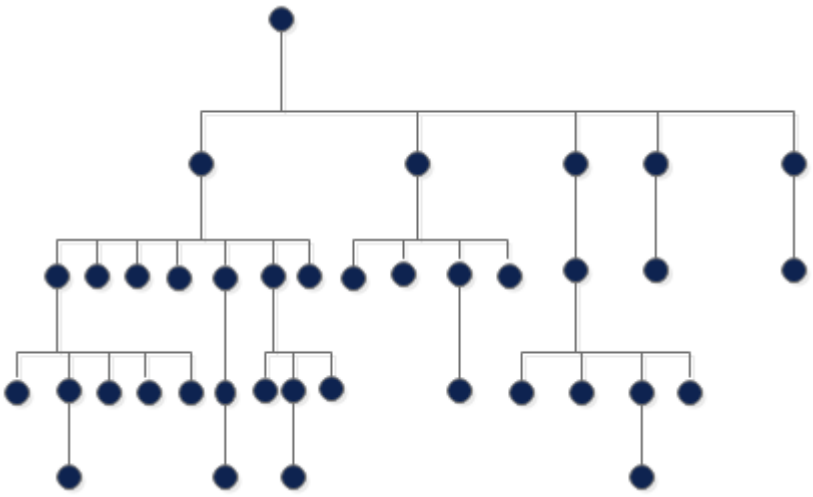
$$F_{i,c}(d, c') \coloneqq \begin{cases} 1, & n_i(d) > 0 \; ve \; c' = c \\ 0, & aksi \; halde \end{cases} \quad [7]$$

$\lambda_{i,c}$ , her bir özelliğin ağırlık değeridir ve $\lambda_{i,c}$ değerinin yüksek olması $f_i$ özelliğinin sınıf $c$ için güçlü bir gösterge olduğu anlamına gelir.

#### 2.3.2.4. Karar Ağaçları

Karar ağaçları hemen hemen her tür veriye adapte edilebilir bu nedenle makine öğrenme algoritmaları içerisinde en yaygın kullanılanlardan biridir. Karar ağacı sınıflandırıcısı, eğitim veri alanının hiyerarşik olarak sıralanmasını sağlar, söz konusu sıralamayı yaparken özellik değerleri üzerindeki bir koşul, verileri bölmek için kullanılır. Söz konusu koşul, bir veya daha fazla kelimenin veri setinde var olup olmamasıdır. Veriler herhangi bir istatiksel bilgi olmadan kolayca anlaşılabilen Şekil 4'deki gibi mantıksal yapı biçiminde sunulmaktadır. Karar ağaçları algoritması birçok hiyerarşik kategorik ayrım yapılabilecek problemlere oldukça uygundur.

Karar ağaçları, yinelemeli bölme denilen deneyimsel yöntem kullanılarak oluşturulur. Bir karar ağacının yapısı tüm veri kümesini temsil eden bir kök düğüm, hesaplamayı gerçekleştiren karar düğümleri ve sınıflandırmayı gerçekleştiren yaprak düğümlerinden oluşur. Eğitim aşamasında algoritma, etiketli eğitim verilerini sınıflarına ayırmak için yapılması gereken kararları öğrenir.



**Şekil 4.** Karar Ağacı Modeli

Bilinmeyen bir örneği sınıflandırmak için, veriler ağaçtan geçirilir. Her karar düğümünde giriş verilerinden seçilen bir özellik, eğitim aşamasında tanımlanan bir sabit özellikle karşılaştırılır. Genellikle her bir karar düğümünde gerçekleşen hesaplama, seçilen özelliği önceden belirlenmiş sabit özellik ile karşılaştırır, karar, özelliğin sabit özellikten büyük veya küçük olup olmadığına dayanılarak ağaçta iki yönlü bir bölünme oluşturur. Veriler, atanan sınıfı temsil eden bir yaprak düğümüne ulaşana kadar karar düğümlerinden geçer.

Random Forest, J48 yöntemi gibi karar ağacı algoritmasının birçok farklı uygulaması ve çeşidi vardır.

#### 2.3.2.5. K-En Yakın Komşu

*k*-en yakın komşu algoritması sınıflandırma ve regresyon için kullanılan temel bir makine öğrenme algoritmasıdır. Bu algoritma yeni bir örnek geldiğinde, eğitim setinde yer alan örnekler ile arasındaki benzerliğe göre sınıflandırmaktadır(Mitchell, 1997: 231).

k-NN algoritmasında, eğitim setinde yer alan örnekler n boyutlu vektörler halinde belirtilir. Her örnek n boyutlu uzayda bir noktayı temsil edecek biçimde tüm eğitim örnekleri n boyutlu bir eğitim uzayında tutulur. Yeni bir örnek ile karşılaşıldığında, eğitim setinden ilgili örneğe en yakın k tane örnek belirlenerek yeni örneğin sınıf etiketi, k en yakın komşusunun sınıf etiketlerinin çoğunluğuna göre atanır(Han, Kamber, & Pei, 2011:424).

Eğitim örnekleri arasında yer alan $x_1, x_2, \dots, x_n$, sınıflandırmak için verilen $x_q$ örneğine en yakın k tane örneği temsil etmek üzere, $x_q$ örneğinin sınıf etiketinin belirlenmesi aşağıdaki eşitlikte sunulmuştur.

$$f(x_q) = argmax_{v \in V} \sum_{i=1}^{k} \delta(v, f(x_i)) \qquad [8]$$

Burada $\delta(a, b)$ ifadesi a ile b birbirine eşitse 1, aksi durumda 0 değerini alır.

K-NN algoritmasının performansı için kritik öneme sahip olan noktalardan birisi örnekler arası yakınlığın nasıl ölçümleneceğidir. Yakınlık Öklid uzaklığı ya da bir başka uzaklık ölçütü kullanılarak hesaplanabilir.

#### 2.3.2.6. Destek Vektör Makineleri

Destek Vektör Makineleri (DVM) ilk olarak örüntü ve sınıflandırma problemlerinin çözümü için Vapnik ve Cortes tarafından geliştirilmiştir(1995). DVM iki sınıflı veya çok sınıflı problemin çözümüne odaklanır ve istatiksel öğrenme teorisine dayanmaktadır. DVM'de amaç sınıfları birbirinden ayıracak optimal hiper-düzlemin elde edilmesidir, başka bir değişle farklı sınıflara ait destek vektörleri arasındaki uzaklığı maksimize etmektir. DVM, doğrusal ve doğrusal olmayan olmak üzere iki durum için ele alınır.

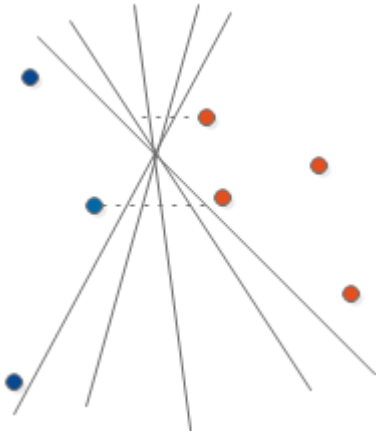
Doğrusal olarak ayrılabilen DVM, en büyük sınıra sahip Hiper-düzlemi bulmaya çalışır. Bu Hiper-düzlemin bulunabilmesi için veri setindeki tüm örneklerin aşağıdaki eşitsizlikleri sağlaması gereklidir (Soman, Loganathan, & Ajay, 2011: 126).

$$f(x_1) = (w, x_1) + b \geq +1 \qquad y_1 = +1 \qquad [9]$$

$$f(x_1) = (w, x_1) + b \leq +1 \qquad y_1 = -1 \qquad [10]$$

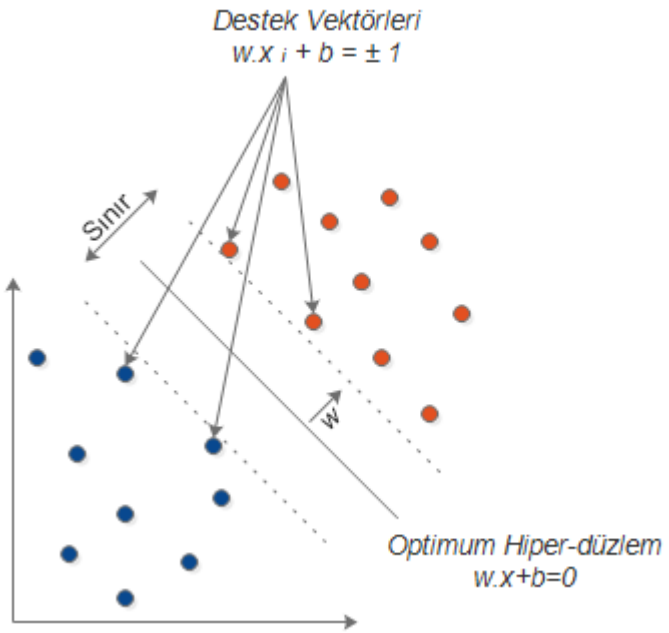
Burada (xi,yi) ikililerinden oluşan n boyutlu bir eğitim kümesi , $w$ hiper-düzlemin normal hali ve b sabit bir parametre değerini göstermektedir.

Veri setini ayıran düzlemlere ilişkin geometrik gösterim Şekil 5 'de verilmiştir. Şekilde görüldüğü gibi farklı sınıflara ait örnekleri birbirinden ayıran birçok doğrusal düzlem bulunabilir ancak DVM farklı sınıflara ait destek vektörleri arasındaki uzaklığı maksimize eden Hiper-düzlemi bulmayı amaçlar.



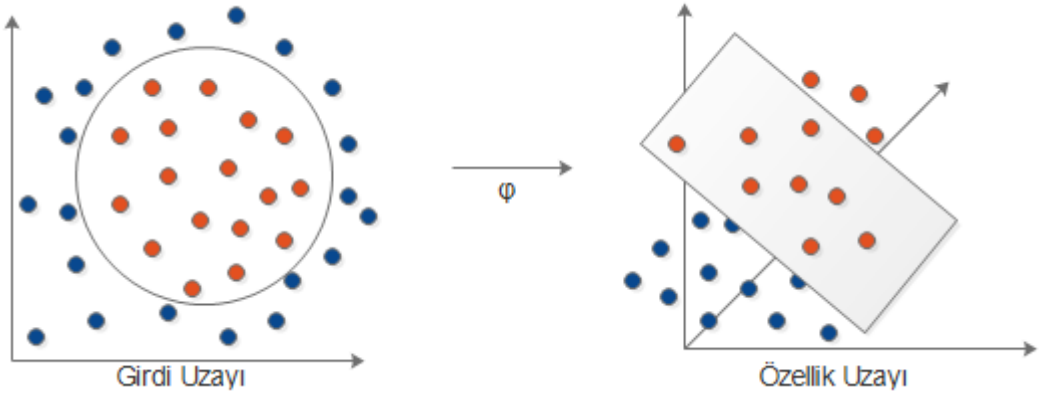
**Şekil 5.** Hiper Düzlemler

Hiper-düzleme en yakın olan her iki sınıfa ait örnekler destek vektörleri olarak adlandırılır ve bu düzlemler $w.xi + b = \pm 1$ şeklinde ifade edilir. Destek vektörleri ait olduğu sınıfın sınırını belirler ve hiper- düzleme paralel bir düzlem üzerinde yer alır(Burges, 1998). Şekil 6 'da optimum Hiper-düzlem ve destek vektörlerin geometrik gösterimi verilmiştir. Destek vektörlerinin üzerinde bulunduğu ve kesikli çizgilerle gösterilmiş düzleme sınır düzlemleri denir. Sınır düzlemlerinin tam ortasından geçen ve her iki düzlemede eşit uzaklıkta bulunan düzlem ise Hiper-düzlem olarak ifade edilir.




**Şekil 6.** Optimum Hiper-düzlem ve Destek Vektörleri

Doğrusal olmayan DVM, verinin doğrusal bir fonksiyonla sınıflandırılmadığı durumlarda kullanılır. Gerçek dünyada sınıflar doğrusal bir sınırla ayrılabilir olmayabilir böyle durumlarda Şekil 7' deki gibi iki boyutlu veri seti daha yüksek boyutlu özellik uzayına taşınarak veri setinin doğrusal ayrımı sağlanır.




**Şekil 7.** Kernel Fonksiyonu ile Verilerin Daha Yüksek Bir Boyuta Dönüştürülmesi

Doğrusal Olmayan DVM iki boyutlu girdi uzayının yüksek boyutlu uzaya taşınması için $\phi(x)^T\phi(x_i) = K(x, x_i)$ şeklinde bir kernel fonksiyonu kullanır(Küçüksille & Ateş, 2013). Literatürde kernel fonksiyonu olarak en sık kullanılan doğrusal, polinom, radyal tabanlı ve sigmoid kernelleridir. $x$ ve $y$ girdi vektörlerini ve $\gamma$,$d$ ve $c$ ilgili kernelin parametrelerini temsil etmek üzere, dört temel kernel fonksiyonu eşitlik [11],[12],[13],[14] ' de sunulmuştur.

Doğrusal Kernel: $K(x, y) = x^T y + c$ [11]

Polinom Kernel : $K(x, y) = (\gamma x^T y + c)^d, \gamma > 0$ [12]

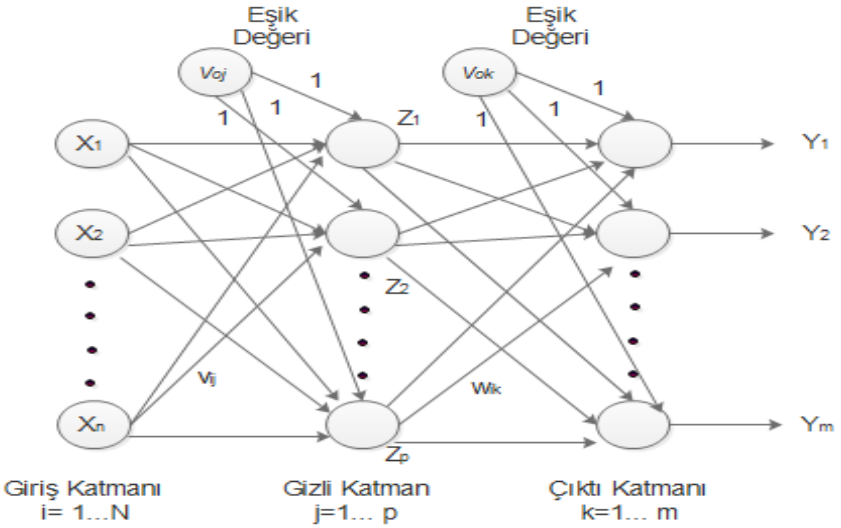
Radyal Tabanlı Kernel: $K(x, y) = exp(-\gamma \|x - y\|2), \gamma > 0$ [13]

Sigmoid Kernel: $K(x, y) = \tanh(\gamma x^T y + c), \gamma > 0$ [14]

DVM modelinden elde edilen sonuçlar seçilen kernel fonksiyonuna ve parametrelerin özelliklerine bağlıdır (Kavzoğlu & Çölkesen, 2010).

### 2.3.2.7. Yapay Sinir Ağları

Yapay Sinir Ağları(YSA) insan sinir sistemini taklit ederek öğrenmeyi hedefleyen makine öğrenmesi yöntemidir. Yapay sinir ağı modeli, Şekil 8'de gösterildiği gibi bir katmanda bulunan nöronların takip eden katmandaki nöronlara bağlanması ile oluşturulur. En sık kullanılan yapay sinir ağı modeli olan çok katmanlı algılayıcı sinir ağı (MLP) modeli; girdi katmanı, gizli katman ve çıktı katmanı olmak üzere üç farklı katmandan meydana gelmektedir. Girdi katmanı verilerin okunduğu katmandır. Her bir nöron bir özelliği temsil ettiği için özellik sayısı kadar nöron içermektedir. Çıktı katmanı ise sınıfların belirlendiği katmandır.




**Şekil 8.** Yapay Sinir Ağı Modeli

Bu katman oluşturulan modele göre tek bir nöron içerebileceği gibi sınıf çeşidi sayısı kadar nöron da içerebilmektedir. Gizli katman ise girdi katmanı ile çıktı katmanı arasında yer alan verilerin ara işleme maruz kaldığı katmandır. Gizli katman sayısı ve bir gizli katmandaki nöron sayısı tam olarak belli olmamakla birlikte eğitimin kalitesini etkileyen önemli iki faktördür(Arslan, Kaynar, & Yüksek, 2017). MLP modelinde öğrenme bir önceki katmandan takip eden katmana doğru yapıldığı için ileri beslemeli YSA olarak da bilinir. Kullanılan eğitim algoritması hatanın karesini en aza indirecek şekilde ağırlıkları güncellemeyi hedefler.

$$y_{i=}f\left(\sum_{j=1}^{n} x_j * w_{ji}\right) \quad [15]$$

Eşitlik 15, MLP modelinde ileri beslemeyi formüle etmektedir. Burada, $x_j$mevcut katmandaki j. nöronun değerini, $y_i$ takip eden katmandaki i. nörona aktarılan değeri, n mevcut katmandaki nöron sayısını, $w_{ij}$ mevcut katmandaki j. nörondan takip eden katmandaki i. nörona giden ağırlığı, $f$ ise aktivasyon fonksiyonunu (örneğin: gauss, softmax, sigmoid ) temsil etmektedir.

$$E(w) = \frac{1}{2}\sum_{k\in çıktı} (t_k - o_k)^2 \quad [16]$$

Eşitlik 16, geri beslemeyi formüle etmektedir. Burada, $k$, veri setindeki öernek sayısını, $t_k$ verilerin gerçek sınıfını, $o_k$ ise modelin üretmiş olduğu sınıf değerini temsil etmektedir.

### 2.4. Metin Önişleme

Verilerin önişleme tabi tutulması, metnin sınıflandırılması için temizlenmesi ve hazırlanması aşamasıdır. Çevrimiçi metinler doğal dil ile yazıldıkları için genellikle gürültülüdür ve HTML etiketleri, reklamlar gibi bilgilendirici olmayan bölümler içerir. Buna ek olarak metindeki pek çok kelimenin metnin genel yönelimi üzerinde bir etkisi yoktur. Bu kelimeleri kullanmak, problemin boyutsallığını artırır ve metindeki her kelimenin bir boyut olarak alınması nedeniyle kelimelerin sınıflandırılması daha zor hale gelir.

Verilerin doğru şekilde ön işleme tabi tutulması, metindeki gürültüyü azaltmaya ve sınıflandırıcının performansını artırmaya yardımcı olur ve sınıflandırma sürecini hızlandırır. Böylece çevrimiçi metinlerin duygu analizine yardımcı olur.

Metin önişleme süreci, metin temizleme, boşluk giderme, kısaltmaları genişletme, kelime köküne inme, durak kelimeleri kaldırma, olumsuzlama eklerinin değiştirilmesi gibi veri temizleme işlemleri ve son olarak özellik seçimi gibi birkaç adımı içerir. Metin önişleme sürecinin tümüne "dönüşümler" denilirken, gerekli

kalıpları seçmek için bazı işlemler uygulayan son adıma “filtreleme” denir(Feinerer, Hornik, & Meyer, 2008).

Duygu analizinde metin temizleme için uygulanan bazı önişleme yöntemleri şunlardır;

- Olumsuzlama eklerinin değiştirilmesi: Olumsuzlama ifadelerinin, çevrimiçi metinlerin duygusunun belirlenmesinde önemli bir rol oynamaktadır. İngilizce ’deki “won’t”, “can’t” ve “n’t” gibi ifadelerin sırasıyla “will not”, “cannot”, “not” ifadelerine dönüştürülmesidir.
- URL bağlantılarının kaldırılması: Bağlantı adresleri herhangi bir yararlı bilgi vermediğinden ve bağlantıların içeriği analiz edilmediğinden URL’lelerin kaldırılması, özellik boyutunu azaltabilir.
- Tekrarlanan harf içeren kelimeleri orijinal haline çevrilmesi: “çooooook” gibi tekrarlı harf içeren kelimeler, çevrimiçi metinlerde yaygın olarak görülür ve insanlar bu tarz kelimeleri duygularını ifade etmek için kullanırlar. Burada tekrar eden harf orijinal haline dönüştürülür.
- Numaraların kaldırılması: Genel olarak, duyguları ölçerken sayıların hiçbir anlamı yoktur ve metin içeriğini temizlemek için numaralar kaldırılır.
- Durak kelimelerinin kaldırılması: Durak kelimeler bir dildeki sıkça kullanılan kelimelerin listesini içerir ve genellikle tek başına kullanıldığında anlamlı bilgiler ifade etmezler. Çoğu araştırmacı, durak kelimelerinin duygu sınıflandırması probleminde olumsuz bir rol oynadığını düşünmektedir ve özellik seçimi öncesinde kaldırılması gerektiğini belirtmişlerdir (Jianqiang & Xiaolin, 2017). Durak kelimelerin kaldırılmasının klasik yöntemi, önceden derlenmiş kelime listelerine dayanan yöntemdir. Her dilde belirli durak kelimeleri olduğundan, tüm dil için standart bir liste mevcut değildir.
- Kelimenin köküne inme (Stemming) : Özellik uzayının boyutunu azaltmak için metinde bulunan tüm kelimelerin kök haline getirilme işlemidir.
- Kısaltmaların genişletilmesi: Kısaltmalar, çevrimiçi metinlerde yaygın olarak kullanılan bozuk yapıda kelimelerdir. Bu kelimeleri orijinal haline genişletmek gerekir.

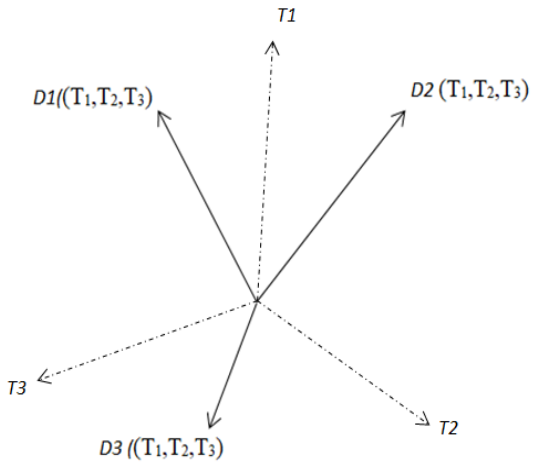
### 2.5. Metnin Vektörel Gösterim Modelleri

Makine öğrenme algoritmaları ham metinde doğrudan çalışmaz, iyi tanımlanmış sabit uzunlukta girdi ve çıktıları tercih eder. Bu nedenle metinler o metni temsil edecek sayısal vektörlere dönüştürülmelidir.

#### 2.5.1. Vektör Uzay Modeli (VUM)

Vektör uzay modeli (VUM), her belgenin bir vektör olarak temsil edildiği ve her bir boyutun ayrı bir terime(kelime) karşılık geldiği vektör temsilleridir. Belgede bir terim mevcutsa, vektördeki değeri sıfırdan farklıdır(Danışman & Alpkoçak, 2008). Bu modelde terimler önemlerine göre ağırlıklandırılmakta ve bu ağırlık değerleri kullanılarak metinler vektörel olarak ifade edilir.

VUM'de belgeler vektörler olarak temsil edildikten sonra iki vektör arasındaki açı değeri belgenin benzerliği olarak kullanılır. Vektörler arasındaki açı ne kadar büyükse belgeler arasındaki benzerlik o kadar az olacaktır. Şekil 9'da belgelerin terim uzayında ifade edilmesi gösterilmiştir.




**Şekil 9.** Vektör Uzay Modeli

#### 2.5.2. Kelime Çantası Modeli (Bag-of-Word)

Vektör gösterimi için kullanılan geleneksel yöntem Kelime Çantası (Bag-of-Word) modelidir. Kelime "çantası" olarak adlandırılır çünkü belgedeki tüm kelimeleri sırası veya yapısı hakkında herhangi bir bilgiyi dikkate almadan çantaya atar. Model yalnızca kelimelerin belgede olup olmadığıyla ilgilidir (ikili KÇ). Diğer

bir deyişle bu yaklaşımda her bir kelimenin sayısı bir özellik olarak dikkate alınır. Bu kelime sayımları, belgeleri karşılaştırmamıza, belge sınıflandırması ve konu modellemesi gibi uygulamalar için kelime benzerliklerini ölçmemize izin verir. İkili KÇ modeli aşağıdaki özellik fonksiyonu ile elde edilir.

Belge $d_i$ , $w \in d_i$ kelimelerinden oluşan bir kümeyi temsil etmek üzere;

$$f_{\mathrm{i}}(X) = \begin{cases} 1, & d_i, w_i \; kelimesini \; içeriyorsa \\ 0, & diğer \end{cases} \qquad [17]$$

KÇ modeli, kelimeleri tek tek özellik olarak kullanmak için yeterli olduğunu varsaymaktadır. Sonuç olarak bir cümleyi vektör olarak temsil etmektedir (eşitlik 18).

$$d_i = (w_{i1}, w_{i2}, \ldots w_{ij}, \ldots w_{in}) \qquad [18]$$

Burada, $w_{ij}$, $d$ cümlesindeki $wi.$ terimin ağırlığıdır. $n$,veri setindeki $|D|$ tüm kelimelerin sayısıdır. Terim ağırlığını tanımlama yöntemleri aşağıda açıklanmıştır.

Terim ağırlıklandırma (TA) , bir doküman vektöründe ki her terimin önemini ölçen ve bu terimin doküman sınıflandırılmasına ne kadar katkıda bulunduğunu belirten bir ağırlık değeri ($w_{ij}$) ile ilişkilendirilmesidir (Patra & Singh, 2013).

Terim ağırlığını tanımlamanın en yaygın yöntemi, kelime oluşumuna karşılık gelen ikili nitelikleri kullanmaktır. Bu yöntemde bir dokümanın bir terimi hangi sıklıkla içerdiği dikkate alınmaz önemli olan dokümanın terimi içerip içermediğidir. Eğer doküman terimi içeriyorsa ağırlığı 1 kabul edilir (eşitlik 17).

Terim Frekansı (TF) yöntemi; terim ağırlığı değerini ($w_{ij}$) , $d_i$ cümlesindeki $w_i$ teriminin kaç kere kullanıldığını hesaplar. TF yönteminde, metinde daha sık oluşan terimlerin daha önemli olduğu varsayılır (Çetin & Amasyalı, 2013).

$$w_{ij} = TF\,(w_i\, d_i) = \frac{n_i}{\sum n_k} \qquad [19]$$

Ters Doküman Frekansı (TDF) yöntemi; metinde bir terim ne kadar çok geçiyorsa o kadar az bilgi içerdiğini varsayar. $|D|$ veri setindeki toplam doküman sayısı, $|d_i \supset w_i|$ $w_i$ terimin göründüğü doküman sayısı olmak üzere TDF aşağıdaki eşitlik ile formüle edilir.

$$TDF(w, D) = \log \frac{|D|}{|d_i \supset w_i|} \quad [20]$$

TF-TDF yöntemi, bir kelimenin metinde görünme sayısına orantılı olarak artar, ancak bazı kelimeler diğer kelimelerden daha sık göründüğü gerçeğini kontrol etmek amacıyla veri seti gelenindeki terim frekansıyla normalleştirmeye çalışır.

$$w_{ij} = TF - TDF(w_i, d_i, D) = TF(w_i, d_i) \times TDF(w, D) \quad [21]$$

İkili KÇ modeli metinleri sırasız ve gramer kuralları dikkate almaksızın incelemektedir. Kelimeler tek tek incelendiğinde yeterli bilgiyi taşımazken ikili ve üçlü kombinasyonları daha anlamlı olabilmektedir. N-Gramlar kelimeleri bileşik kelime ve deyimler şeklinde incelemeye izin verdikleri için daha fazla bilgi kazancı sağlayabilmektedirler. N-gram modeli, bir sonraki kelimenin görülme olasılığının önceki n-1 kelimeye dayandığını varsayar. Buradaki n, tekrar derecesini ifade etmektedir. Gram ise bu tekrarın dizilim içerisindeki ağırlığını ifade etmek için kullanılır. $n = 1$ için n-gram modeline "unigram", $n = 2$ için "bigram" $n = 3$ için "trigram" denir. $n > 3$ N-gram modeller için sadece n sayısın sayısal değeri ile değiştirilir örneğin; 4-gram,5-gram gibi.

Cümlelerin kelimelerine ayrılması ile ilgili bir örnek "can kahve ve çay sever" ifadesi için n-gramlar aşağıdaki gibi elde edilebilir.

- Unigram : $[can, kahve, ve, çay, sever]$
- Bigram : $[[can\ kahve], [kahve\ ve], [ve\ çay][çay\ sever]]$
- Trigram : $[[can\ kahve\ ve], [kahve\ ve\ çay], [ve\ çay\ sever]]$

şeklinde gösterilebilir.

N-gram modeli bir cümlenin kelimelerini belli bir $n$ sayısına göre ayrıştırırken, karakter n-gram modeli harf dizisidir yani kelimelerin karakterlerini belli bir $n$ sayısına göre ayrıştırmaktadır. Karakter n-gramlar ise aşağıdaki şekilde elde edilmektedir.

- Bigram :

$[[du], [uy], [yg], [gu], [u\_], [-_a], [an], [na], [al], [li], [iz], [zi]]$

- Trigram

: $[[duy], [uyg], [ygu], [gu\_], [u\_a], [\_an], [ana], [nal], [ali][liz], [izi]]$

Duygu analizinde N-gramlar duygu içeren kelimeleri sıralı olarak elde etmemizi sağlayan yöntem olarak kullanılabilir. N-gramlar çıkarıldıktan sonra dokümandaki n-gram kelimelerinin ağırlıkları TF ve TDF ile hesaplanarak özellik olarak kullanılabilir.

### 2.5.3. Word2Vec Modeli

Word2vec, Tomas Mikolov ve diğ. tarafından geliştirilmiş yapay sinir ağı yapısını kullanan denetimsiz bir DDİ modelidir (2013). Word2vec, girdi olarak etiketsiz bir eğitim derlemi verildiğinde, derlem içindeki her bir kelime için anlamsal bilgilerini kodlayan vektör üretir. Bu vektörler iki ana nedenden dolayı oldukça kullanışlıdır. Birincisi iki kelime arasındaki anlamsal benzerliği, kelimelere karşılık gelen kelime vektörleri arasındaki kosinüs benzerliğini hesaplayarak ölçülebilir. İkincisi kelime vektörleri duygu analizi gibi çeşitli DDİ problemleri için özellik olarak kullanılabilir. Bu vektörlere ait anlamsal bilgi onları bu tür problemler için güçlü özellikler haline getirir.

Word2vec modeli insan sezgilerine oldukça uygundur. Örneğin, eş anlamlı olduğunu bildiğimiz kelimeler, kosinüs benzerliğine göre benzer vektörlere sahip olma eğilimdedir aynı şekilde zıt anlamlı kelimeler alakasız vektörlere sahip olma eğilimindedir. Dahası Word2vec kelime vektörleri arasında benzetme gibi bazı kelime ilişkilerini kullanır. Örneğin, vec(kral)- vec(erkek) + vec(kadın) = vec(kraliçe) gibi benzer anlamsal ilişkiye ait kelimeler üretebilir. Word2vec kelime vektörleri arasında aritmetik işlemler yapılmasına da olanak sağlamaktadır. Örneğin

vec(Paris)-vec(France)+vec(Turkey) vektör hesaplamasının sonucu diğer kelimelerin vektörlerine göre vec(Ankara) daha yakındır.

İki temel Wod2vec mimarisi vardır: Continuous Bag of Words (CBOW) ve Skip-Gram. Bu modellerin her ikisi de bir gizli katmana sahip basit sinir ağlarıdır. Bu modellerin her biri aşağıdaki bölümlerde ayrıntılı bir şekilde açıklanmıştır.

#### 2.5.3.1. Continuous Bag of Words (CBOW)

CBOW mimarisinde bir kelimenin belli bir pencere boyutu içindeki komşu kelimelerine (sağındaki ve solundaki kelimeler) bakılmaktadır ve ilgili kelime tahmin edilmeye çalışılmaktadır.

Şekil 10.'da gösterilen modelde $V$ kelime grubunun boyutu ve $N$ gizli katman boyutudur. Giriş vektörü $X = \{x_1, x_2, \dots, x_V\}$ one-hot kodlanmıştır yani $x_k = 1$ iken $k \neq k'$ için diğer tüm $x_{k'} = 0$'dır.

$W_{V\times N}$ Matrisi ile giriş katmanı ile gizli katman arasındaki ağırlıklar gösterilir. $W$'nin her satırı, giriş katmanındaki ilişkili kelimenin N boyutlu vektör temsilidir. $W$'nin $i$ satırı $V_w^T$'dir. Bir kelime göz önüne alındığında, $k \neq k'$ için $x_k = 1$ ve $x_{k'} = 0$ olduğu varsayılır ve Eşitlik 22 elde edilir,

$$h = W_x^T = W_{(k,\cdot)}^T \coloneqq v_{w_I}^T \qquad [22]$$

Burada, aslında $W$'nin $k$'ıncı satırı $h$' ye kopyalanmaktadır. $v_{w_I}$, $w_I$ girdi kelimesinin vektör gösterimidir. Bu gizli katman birimlerinin aktivasyon fonksiyonunun doğrusal olduğu anlamına gelir( ağırlıklandırılmış girdi toplamını doğrudan bir sonraki katmana iletir).

Gizli katman ile çıktı katmanı arasındaki ağırlık $N \times V$ matrisi olan $W' = \{w'_{ij}\}$ ağırlık matrisi ile gösterilebilir. Bu ağırlıklar kullanılarak, kelime grubundaki her kelimenin puanı $u_j$ hesaplayabiliriz.

$$u_j = {v'_{w_j}}^T h \qquad [23]$$

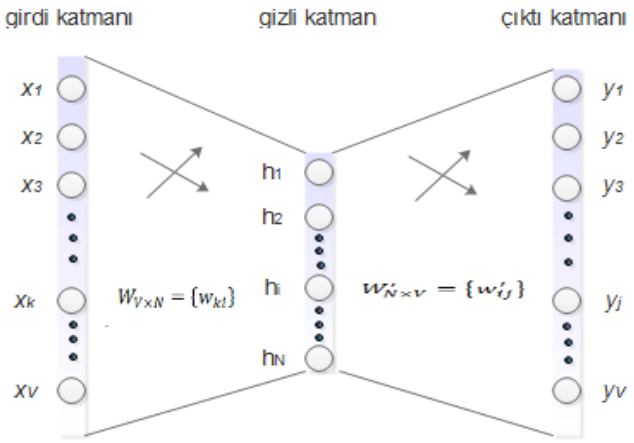
Burada, $v'_{w_j}$, $W'$ matrisinin $j$'inci sutundur. Çok terimli dağılım olan kelimelerin posterior dağılımını Softmax veya Negative Sampling sınıflandırma modellerinden birini kullanabilir.

$$p(w_j|w_I) = y_j = \frac{\exp(u_j)}{\sum_{j'=1}^{V} \exp(u_{j'})} \quad [24]$$

Burada, $y_j$, çıktı katmanındaki $j$'inci birimin çıktısıdır. Sonuç olarak Eşitlik 22, 23, 24'ten Eşitlik 25 elde edilir.

$$p(w_j|w_I) = \frac{\exp({v'_{w_j}}^T v_{w_I})}{\sum_{j'=1}^{v} ({v'_{w_{j'}}}^T v_{w_I})} \quad [25]$$

$v_w$ ve $v'_w$ kelime $w$'nin iki vektörel gösterimidir. $v_w$, girdi→gizli ağırlık matrisi olan $W$'nin satırlarından, $v'_w$, gizli→çıktı matrisi olan $W'$ sütunlarından elde edilir. $v_w$ ,$w$ kelimesinin "girdi vektörü" ve $v'_w$, "çıktı vektörü" olarak adlandırılır.




**Şekil 10.** CBoW Mimarisi

### 3.5.3.2. Skip-Gram

Skip-Gram mimarisi Mikolov ve diğ. (Mikolov vd., 2013) tarafından sunulmuştur. Skip-Gram, CBOW mimarisinin tam tersi şekilde hedef kelimeye

bakılarak komşu kelimeler tahmin etmektedir. Şekil 11. Skip-Gram mimarisini göstermektedir.

Giriş katmanındaki tek kelimenin giriş vektörünü ifade etmek için CBOW de olduğu gibi $v_{w_I}$ kullanılır ve böylece Eşitlik 31'deki $h$ gizli katman çıktılarının benzer tanımı elde edilir. Yani $h$, girdi kelimesi $w_I$ ile ilişkili, girdi→gizli ağırlık matrisinin $W$ bir satırını aktarmaktadır.

$$h = W_{(k,.)}^{T} \coloneqq v_{w_I}^{T} \quad [26]$$

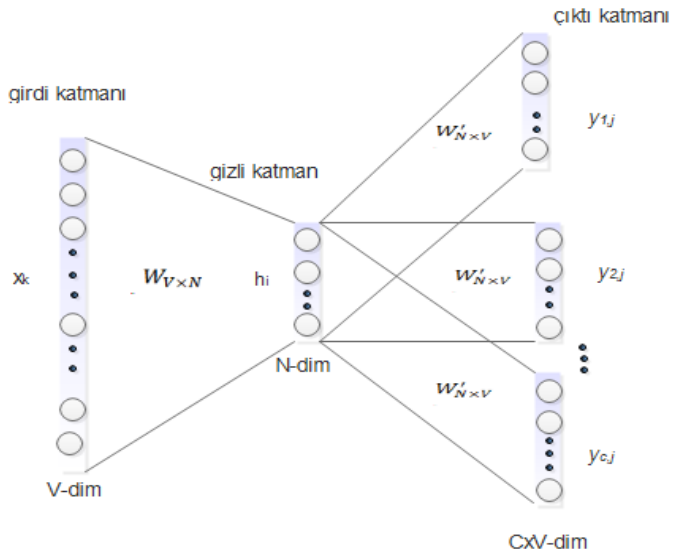
Çıktı katmanı üzerinde, tek çoklu terim dağılım çıkışı yerine $C$ çoklu terim dağılım çıkışı yapılır. Her çıktı gizli→çıktı matrisi kullanılarak hesaplanır:

$$p(w_{c,j} = w_{O,c}|w_I) = y_{c,j} = \frac{\exp(u_{c,j})}{\sum_{j'=1}^{v} \exp(u_{j'})} \quad [27]$$

Burada, $w_{c,j}$ ,çıktı katmanının $c$'inci panelindeki $j$'inci kelimedir. $w_{O,c}$ , çıktı bağlamındaki kelimelerin gerçek $c$'inci kelimesidir. Tek girdi kelimesi $w_I$'dir. $y_{c,j}$ , çıktı katmanının $c$'inci panelindeki $j$'inci birimin net girişidir. Çıktı katmanı panelleri aynı ağırlıkları paylaştığından sonuç olarak; $c = 1,2,3 \ldots C$ için

$$u_{c,j} = u_j = v'_{m_j}{}^{T} h \quad [28]$$

Burada, $v'_{m_j}$, $m_j$ kelime grubunun $j$'inci kelimesinin çıktı vektörüdür. Ayrıca $v'_{m_j}$, $W'$ gizli→çıktı matrisinin sütunundan elde edilir.

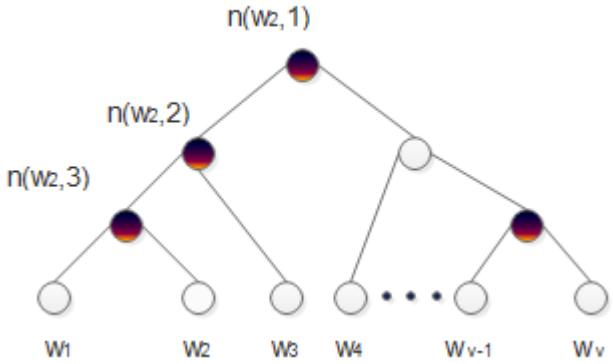



**Şekil 11.** Skip-Gram Modeli

CBOW ve Skip-Gram mimarilerinin her biri öğrenme için Hiyerarşik Softmax (Hierarchical softmax - HS) ve Negatif Örnekleme (Negative Sampling-NS) eğitim algoritmalarından birini kullanmaktadır.

#### 2.5.3.3. Hiyerarşik Softmax

Sinir ağı dil modelleri alanında ilk olarak Marin ve Bengio tarafından tanıltılmıştır( 2005). Hiyerarşik softmax, softmax hesaplamanın etkili yöntemidir. Model, kelime grubundaki tüm kelimeleri temsil etmek için ikili ağaç modeli kullanır(Şekil 12.). Her yaprak düğümü için kökten düğüme kadar uygun tek bir yol vardır ve bu yol yaprak düğümü tarafından temsil edilen kelimenin olasılığını tahmin etmek için kullanılır.




**Şekil 12.** İkili Ağaç Modeli

Şekil 11. hiyerarşik softmax modeli için ikili ağaç modelidir. Beyaz düğümler kelime grubundaki kelimelerdir. Siyah düğümler iç düğümlerdir.

Hiyerarşik softmax modeli kelimeler için bir çıktı vektörüne sahip değildir bunun yerine $V-1$ iç düğümlerinin her birinin çıktı vektörüne $v'_{n(w,j)}$ sahiptir.

$n(w,j)$ kökten $w$' ye uzanan yol zerindeki $j$'inci düğüm,

$L(w)$, bu yolun uzunluğu,

$ch(n)$, kök n'nin rastgele sabit çocuğu,

$v'_{n(w,j)}$ , iç düğüm $n(w,j)$'nin çıktı vektörü,

$h$, gizli katman çıktı değeri (Skip-Gram modelde $h = v_{m_I}$ ve CBoW modelde $h = \frac{1}{C}\sum_{C=1}^{C} C_{w_c}$ ) ve

$[\![x]\!]$, $x$ doğruysa 1, aksi halde -1 olmak üzere hiyerarşik softmax çıktı kelimesi olan bir kelimenin olasılığını aşağıdaki eşitlikteki gibi hesaplar;

$$p(w = w_0) = \prod_{j=1}^{L(w)-1} \sigma\left([\![n(w,j+1) = ch\left(n(w,j)\right)]\!]\right) . {v'_{n(w,j)}}^{T} h \qquad [29]$$

#### 2.5.3.4. Negatif Örnekleme

Negatif örnekleme (NS), hiyerarşik softmax'a alternatif olarak Word2vec modeli tarafından kullanılan Noise Contrastive Estimition (NCE)'nin bir çeşididir. NCE, iyi bir modelin lojistik regresyon aracılığıyla veriyi gürültüye göre ayırt etmesini gerektiğini varsayar. Mikolov ve diğ. NCE softmax'ın log olasılığını maksimuma çıkarabileceğini gösterse de, Word2vec modeli sadece yüksek kaliteli vektör gösterimleri ile ilgili olduğunu ve dolaysıyla vektör gösterimlerinin kalitesi korunduğu sürece NCE'nin basitleştirilebileceğini ifade etmişlerdir(2013). NS Eşitlik 30 ile tanımlanır.

$$E = -\log \sigma\left({v'_{w_o}}^T h\right) - \sum_{w_j \epsilon W_{neg}} \log \sigma\left({v'_{w_j}}^T h\right) \qquad [30]$$

Burada , $w_o$, çıktı kelimesidir( yani pozitif örnek) ve $v'_{w_0}$ çıktı vektörüdür. $h$, gizli katmanın çıktı değeridir; CBOW için $h = \frac{1}{C}\sum_{C=1}^{C} v_{w_c}$'dir ve Skip-Gram model için $h = v_{w_I}$'dir. $W_{neg} = \{w_j | j = 1, \ldots, K\}$, $P_n(w)$' ye dayanılarak örneklenen kelimelerin kümesidir yani negatif örnekler.

Word2vec modelinin çalışma süresini ve kalitesini etkileyen bir dizi parametre seçimi vardır. Bunlar Tablo 1' de gösterilmiştir.

**Tablo 1.** Word2vec Parametreleri

| **Parametre** | **Görevi** | **Değerler** |
|---|---|---|
| sg | Eğitim algoritmalarını tanımlar | sg=1 CBOW,<br>sg=0 Skip-Gram kullanır. |
| hs | Eğitimde kullanılacak öğrenme algoritması | hs=1 hierarchical softmax,<br>hs=0 negative sampling |
| alpha | Öğrenme oranı | Varsayılan değer 5 |
| size | Vektör uzunluğu | Varsayılan değer 100 |
| window | Pencere boyutu (dikkate alınacak komşu kelime sayısı) | Genellikle 5-10 arasındaki değerler |
| min_count | Frekansı düşük olan kelimeler dikkate alma | Varsayılan değer 10-100 arası |

### 2.6. Model Başarım Ölçütleri

Sınıflandırma algoritmalarının değerlendirilmesinde kullanılan kavramlar doğruluk, kesinlik, duyarlılık ve F-ölçütüdür. Ölçüt seçimi sınıflandırma algoritmaların performansının ölçülmesini ve karşılaştırılmasını büyük ölçüde etkilemektedir.

Bir sınıflandırma algoritmasının performansını özetlemek için karışıklık matrisi kullanılır. Matris, sınıflandırma algoritması tarafından yapılan tahmini ve gözlenen sınıflandırmalar hakkında bilgi içerir. Bu tür algoritmaların performansı genellikle matristeki veriler kullanılarak değerlendirilir.

Aşağıdaki Tablo 2'de iki sınıflı bir sınıflandırıcı için karışıklık matrisi gösterilmektedir.

**Tablo 2.** Karışıklık Matrisi

| | | Öngörülen | |
|---|---|---|---|
| | | Pozitif | Negatif |
| Doğru | Pozitif | TP | FP |
| | Negatif | FN | TN |

Karışıklık matrisindeki TP, FP, FN, TN girdileri sırasıyla;

TP (True Pozitif), pozitif olan aynı zamanda sınıflandırıcı tarafından pozitif olarak sınıflandırılmış örnek sayısını gösterir.

FP (False Pozitif), pozitif olan ancak sınıflandırıcı tarafından pozitif olarak sınıflandırılmamış örneklerin sayısını gösterir.

FN ( False Negatif), negatif olan ancak sınıflandırıcı tarafında negatif olarak sınıflandırılmamış örneklerin sayısını gösterir.

TN ( True Pozitif), negatif olan aynı zamanda sınıflandırıcı tarafından da negatif sınıflandırılmış yorumların sayısını gösterir.

Karışıklık matrisinin kendisi bir performans ölçütü değildir ancak hemen hemen tüm performans ölçütleri karışıklık matrisini ve içindeki sayıları temel almaktadır. İkili sınıf matrisi için birkaç standart terim aşağıdaki açıklanmıştır.

### 2.6.1. Doğruluk (Accuracy)

Sınıflandırma işleminde en çok kullanılan ölçümdür. Doğru olarak sınıflandırılmış örneklerin toplam örnek sayısına oranıdır.

$$doğruluk = \frac{TP + TN}{TP + FP + FN + TN} \quad [31]$$

Doğruluk oranı, verilerdeki hedef değişken sınıflarının neredeyse dengeli olduğu durumlarda başarılı bir ölçüttür. Ancak her sınıfta eşit olmayan sayıda gözlem varsa veya veri kümesinde ikiden fazla sınıf varsa yanıltıcı olabilir.

### 2.6.2. Kesinlik (Precision)

Sınıflandırıcı sonucunun kesinlik derecesidir. Pozitif olarak etiketlenen örneklerin sayısının pozitif olarak sınıflandırılmış toplam örneklere oranıdır. Yüksek kesinlik oranı, pozitif olarak etiketlenen örneğin gerçekten pozitif olduğunu gösterir (az sayıda FP).

$$kesinlik = \frac{TP}{TP + FP} \quad [32]$$

### 2.6.3. Duyarlılık (Recall)

Pozitif olarak etiketlenmiş örneklerin gerçekten pozitif olan örneklerin toplam sayısına oranıdır. Yüksek duyarlılık oranı, örneklerin doğru şekilde etiketlendiği anlamına gelir (az sayıda FN).

$$duyarlılık = \frac{TP}{TP + FN} \quad [33]$$

### 2.6.4. F-Ölçütü (F-Measure)

Kesinlik ve duyarlılık ölçütleri kullanılarak hesaplanmaktadır. Sistemin, kesinlik ya da duyarlılık yönüne doğru optimize edilmesinde kullanılmaktadır.

$$F - \text{ölçütü} = \frac{2 \times duyarlılık \times kesinlik}{duyarlılık + kesinlik} \quad [34]$$

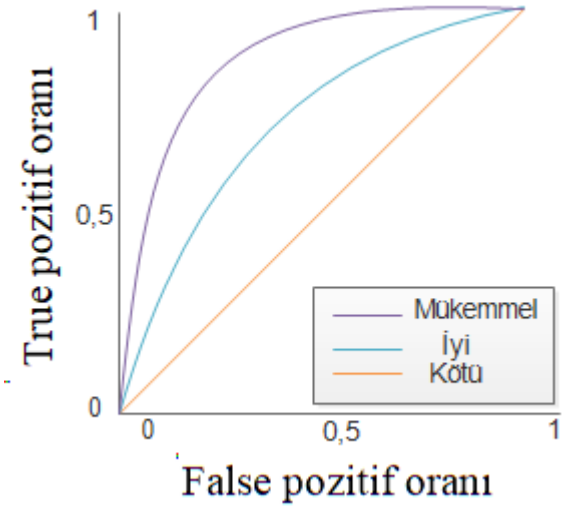
F-ölçütü her zaman kesinlik ve duyarlılık oranlarından küçük olana daha yakındır.

### 2.6.5. ROC Eğrisi

ROC grafikleri, TP oranının Y ekseni üzerinde çizildiği ve FP oranının X ekseni üzerine çizildiği iki boyutlu grafiklerdir. Bir ROC grafiği TP ile FP arasındaki dengeyi değerlendirmek için kullanılabilir. Yanlış değerlere sahip olmayan ideal bir testte ROC eğrisi (0,0), (0,1) ve (1,1) noktalarını birleştirmektedir. Performansı kötü

olan ROC eğrisi (0,0)'dan (1,1)'e kadar 45° açı yaparak uzanan köşegen şeklindedir (Şekil 13.). Genellikle ROC eğrisi bu iki durum arasında değişir.

Testlerin doğru karar vermede performansını değerlendirmede kullanılan ölçütlerden biri de ROC eğrisi altında kalan alandır. Bu alan ROC puanı (AUC) olarak tanımlanır. AUC sıfır FP ve sıfır FN ile en büyük 1 değerini alır. AUC 0 olduğunda ise herhangi pozitif bulunamadı anlamına gelir (Wikipedia).




**Şekil 13.** Performanslarına Göre ROC Eğrileri

# 3. DUYGU ANALİZİ UYGULAMA VE DEĞERLENDİRME

Daha önceki bölümlerde, duygu analizi üzerine yapılan bu çalışmayla ilgili gerekli arka plan tanımlanmıştır. Bu bölümde veri setlerimizi ve makine öğrenme yöntemlerini etkin kullanabilmek için her veri seti için gerekli tüm ön işleme süreci anlatılmıştır ve deneysel çalışmalar sunulmuştur. İlk uygulamada İngilizce film yorumları veri seti ele alınarak duygu analizi gerçekleştirilmiştir. İkinci uygulamada ise Türkçe film yorumlarını içeren veri seti kullanılmıştır.

## 3.1. Veri Analizi

### 3.1.1. Veri Setlerinin Açıklaması

Çalışmamızda İngilizce ve Türkçe film yorumlarından elde edilen iki farklı veri seti kullanılmıştır.

İngilizce film yorumları veri seti için gerekli olan pozitif ve negatif içerikli veri seti 50.000 IMDB film değerlendirmesinden oluşmaktadır. Değerlendirmelerin duygu puanları ikilidir, yani IMDB derecelendirmesi <5 ise duygu puanı 0 ve derecelendirme ≥7 ise duygu puanı 1 dir. Tek bir filmde 30'dan fazla yorum bulunmamaktır. 25.000 etiketli eğitim seti, 25.000 test seti ile aynı filmleri içermemektedir (Maas vd. , 2011).

Türkçe dilinde yazılmış pozitif ve negatif içerikli veri için Beyazperde.com' dan kullanıcıların filmler hakkında yaptıkları yorumlardan elde edilen veri seti kullanılmıştır. Bu veri seti pozitif sınıfta 5.331 veri negatif sınıfta 5.331 veri olmak üzere toplam 10.662 veriden oluşmaktadır. Bu web sitesinde yer alan film yorumları, yorumları yapan kullanıcılar tarafından 0-5 arasında derecelendirilmiştir. Derecelendirmesi $\geq 4$ ise pozitif veya derecelendirmesi $\leq 2$ ise negatif yorum olarak kabul edilmiştir (Demirtaş & Pechenizkiy, 2013).

Farklı metin gösterim tekniklerinin sınıflandırma başarısına olan etkilerini incelemek için iki farklı veri seti kullanılmıştır. Yukarıda açıklanan veri setlerinin farklı örnek boyutuna sahip olduğunu görebiliriz. Bu sayede kullanılan verinin büyüklüğünün kelime vektörlerinin kalitesine olan etkileri de incelenmiştir.

### 3.1.2. Veri Önişleme

Web'de yer alan yorum ve değerlendirmeler eksik veya hatalı, yapılandırılmamış metinsel veriler genellikle gürültülü verilerdir. Bu nedenle başarılı sınıflandırma sonuçları elde etmek için gürültülü verilerin temizlenmesi gerekir. Ön işleme sürecinin temelindeki amaç, verileri bir sonraki adımlar için hazırlamaktır. Bu tez çalışmasında, çalışmaya uygun şekilde bir veri seti oluşturulması için İngilizce ve Türkçe film yorumları veri setleri deney durumlarına göre aşağıdaki ön işleme yöntemlerine tabi tutulmuştur.

- Büyük ve küçük harflerin kullanımıyla ilgili uyumsuzluğu kaldırmak için tüm kelimeler küçük harfe çevrilmiştir. Bu adım kelimelerin anlamını etkilemediğinden, gerçekleştirilmemiş olsaydı, bazı kelimeler aynı kelime olarak kabul edilmez (örneğin, iyi ve İyi) ve sonuçları olumsuz etkileyebilir,
- Durak kelimelerinin kaldırılması,
- Boşluklar, noktalama işaretleri ve sayıların kaldırılması,
- Kelime köküne inme.

### 3.1.3. Metin Temsillerinin Oluşturulması

Metin tabanlı veriler üzerinde analizler yapabilmek için, ön işleme aşamasından sonra uygun forma getirilen veri, bu aşamada sayılaştırılır. Çalışmamızda kullanılan her iki veri seti için TF ve TF-TDF ağırlıklandırma yöntemleri ile elde edilen 1-Gram ve 2-Gram ve 1-2-Gram yöntemiyle oluşturulan klasik KÇ metin temsillerinin yanı sıra Word2vec kelime vektörleri çıkarılmıştır.

KÇ modellerde, sütunlar değerlendirmelerin mevcut kelimelerini ve değerler söz konusu bu kelimelerin frekanslarını temsil eden bir tablo ile gösterilir. Sonuç olarak, değerlendirmelerin bir koleksiyonu ( ön işleme adımından sonra) Tablo 3'de gösterildiği gibi ifade edilir, burada $n$ tane cümle, $m$ tane kelime bulunmaktadır. Her değerlendirme şu şekilde ifade edilmektedir.

**Tablo 3.** Değerlendirmelerin Gösterimi

| | $w_1$ | $w_2$ | | $w_m$ |
|---|---|---|---|---|
| değerlendirme$_1$ | $a_{11}$ | $a_{12}$ | ... | $a_{1m}$ |
| değerlendirme$_2$ | $a_{21}$ | $a_{22}$ | ... | $a_{2m}$ |
| ... | ... | ... | ... | ... |
| değerlendirme$_n$ | $a_{n1}$ | $a_{n1}$ | ... | $a_{nm}$ |

Word2vec kelime vektörleri oluşturmak için CBoW ve Skip-Gram (SG) olmak üzere iki farklı öğrenme mimarisi kullanmaktadır. CBoW, bağlamı verilen kelimeyi tahmin ederken, Skip-Gram bir kelimesi verilen bağlamı tahmin eder. Metin derlemi söz konusu bu iki model ile eğitilir. Eğitimden sonra her bir kelime bir vektör olarak gösterilir. Bu süreçte, Word2vec ilk olarak, metin derleminin eğitiminden bir kelime grubu oluşturur ve her kelimenin vektör gösterimlerini öğrenir. Son olarak yüksek boyutlu bir matris oluşturulur. Matristeki her satır, her bir eğitim örneğini ve sütunlar kelime vektörlerini temsil etmektedir. Sonuç olarak, bir kelime birden fazla benzerlik derecesine sahiptir.

Bu tez çalışmasında kelime benzerliğini test etmek için yalnızca en benzer iki kelimeyi ve bunların kosinüs mesafelerini listeleyen birkaç örnek Tablo 4 ve Tablo 5'de verilmiştir.

**Tablo 4.** İngilizce Veri Seti için Kelime Benzerliği

| Kelimeler | Kelime Benzerliği | | | |
|---|---|---|---|---|
| | 1.kelime | Kosinüs Mesafesi | 2.kelime | Kosinüs Mesafesi |
| Computer | Software | 0.6137 | Generated | 0.5866 |
| Queen | Princess | 0.6188 | Goddess | 0.6172 |
| Awful | Terrible | 0.7585 | Atrocious | 0.7302 |

**Tablo 5.** Türkçe Veri Seti için Kelime Benzerliği

| Kelimeler | Kelime Benzerliği | | | |
|---|---|---|---|---|
| | 1.kelime | Kosinüs Mesafesi | 2.kelime | Kosinüs Mesafesi |
| Bilgisayar | Yazılım | 0.6461 | Makine | 0.6053 |
| Kraliçe | Prenses | 0.6462 | İmparatoriçe | 0.6168 |
| Korkunç | Acımasız | 0.6885 | Korkutucu | 0.6880 |

Her bir kelimeye karşılık gelen Word2Vec değerleri çeşitli şekillerde bir araya getirilerek cümleyi temsil eden Word2Vec vektör temsillerinin elde edilmeleri gerekmektedir. Fakat Skip-Gram ve CBOW yöntemleri kullanılarak eğitilen Word2Vec vektörlerinin birleştirilmesi veri seti boyutunu oldukça artıracaktır. Derlem içerisinde yer alan her bir kelimenin 1x300 boyutunda olduğu dikkate alınırsa uzun bir cümle için bu vektörlerin birleşimi çok boyutlu vektörlere neden olacaktır. Ayrıca her bir cümledeki kelime sayısı farklı olacağından cümle temsilleri için elde edilecek vektörler birbirlerinden farklı boyutlarda olacaktır. Bahsedilen sorunların üstesinden gelmek amacıyla cümle temsillerinin elde edilmesi için çalışmamızda cümledeki her bir kelimeye karşı gelen Word2Vec değerlerini toplayarak ortalaması alınmıştır ve cümle temsili için bu ortalama vektörü kullanılmıştır.

### 3.2. Uygulama

Eğitim süreci için eğitim ve test setlerinin ayrılması gerekir. Bu amaç için basit bir yaklaşım veri setlerini rastgele eğitim ve test alt gruplarına (örneğin %60 ve %40) bölmektir. Ancak burada veri parçalanırken verinin dağılımına bağlı olarak modelin eğitim ve testinde bazı sapmalar ve hatalar oluşabilir. Bu hataları minimum seviyeye indirmek için K- katlamalı çapraz doğrulama tekniği kullanılır. Bu teknik eğitim veri setini rasgele k parçaya böler, k-1 parça eğitim için kullanılırken 1 parçada test seti için kullanılır ve k defa bu işlem tekrar edilir. Her tekrarlamada elde

edilen değerler toplanır, ortalaması alınır ve modelin performansı değerlendirilir. Bu tez çalışmasında kullanılan her iki veri seti için 10 kat çapraz doğrulama kullanılarak sınıflandırma işlemleri gerçekleştirilmiştir.

Çalışmamızda her bir özellik setinin sınıflandırma algoritmasının performansına etkisi bulmak amaçlanmıştır. 1-Gram, 2-Gram ve birli ve ikili kelime gruplarını içeren 1-2 Gram yöntemiyle oluşturulan KÇ metin temsillerinin ve Word2vec metin temsilleri yöntemi sonucu elde edilen vektörlerinin sınıflandırma başarısına olan etkileri test edilmiştir.

Sınıflandırma işlemi için çoğunlukla duygu analizi ile kullanılan Naive Bayes (NB), Lojistik Regresyon (LR), Karar Ağaçları(KA), K- En Yakın Komşu(k-EYK), Destek Vektör Makineleri(DVM), Yapay Sinir Ağları (YSA), kullanılmıştır. PYTHON programlama dili için ücretsiz yazılım makine öğrenme kütüphanesi olan Scikit-learn Kütüphanesi kullanılarak, NB, LR, KA, k-EYK, DVM ve Keras kütüphanesi kullanılarak YSA sınıflandırma algoritmalarının performansları karşılaştırılmıştır.

Sınıflandırma algoritmalarının performansını etkileyen unsurlardan biri de parametrelerinin doğru olarak belirlenmesidir. Naive Bayes sınıflandırma algoritması test edilecek veriye ait olan sınıfı tahmin etmek için Gauss dağılımını kullanmaktadır. Lojistik regresyon sınıfında yer alan C parametresi, düzenlileştirilmiş (regularization) için kullanılan $\lambda$ değerinin tersidir $C = \frac{1}{\lambda}$. DVM makinelerinde olduğu gibi daha küçük değerler daha güçlü düzenlileştirme sağlamaktadır. Çalışmamızda C değeri 1 seçilmiştir.

Karar ağacı sınıflandırma algoritmalarında bir kökten itibaren bölünmenin hangi kıstasa göre yapılacağı ağacın doğruluğunu etkileyen faktörlerdendir. Karar Ağaçları, bir düğümü iki veya daha fazla alt düğüme bölmek için birden fazla algoritma kullanır. Algoritma seçimi, hedef değişkenin tipine bağlıdır. En sık kullanılan algoritmalar, Entropi ve Gini'dir. Bu çalışmada dallanma kriteri olarak Gini kullanılmıştır. İç düğümleri bölmek için gereken minimum örnek sayısı 2, bir yaprak düğümünde bulunması gereken minimum örnek sayısı 1 olarak tanımlanmıştır.

k-NN algoritması kullanılarak yapılan sınıflandırma işleminde k değerinin uygun değerde seçilmesi çok önemlidir. En uygun k değerinin tespit edilmesi için 5'den 13'e kadar olan değerler kullanılmıştır, sonuçta en uygun k değeri 11 olarak belirlenmiştir. Ayrıca k-NN algoritmasının performansı için kritik öneme sahip noktalardan birisi örnekler arası yakınlığın nasıl ölçümleneceğidir. Çalışmamızda uzaklık ölçütü kapsamında Öklid uzaklığı kullanılmıştır.

DVM algoritmasında kernel fonksiyonu olarak lineer fonksiyonu tercih edilmiştir. Söz konusu kernel fonksiyonuna ilişkin C ceza parametresi 1 olarak belirlenmiştir.

Yapay sinir ağlarında 3 katmanlı ağ modeli tercih edilmiştir. Orta katmanda KÇ modeli için 500, Word2vec modeli için 200 gizli nöron, çıkış katmanda ise sınıf tahmin etmek için 1 nöron kullanılmıştır. Giriş katmanındaki nöron sayısı ise özellik sayısı ile aynıdır. Gizli katman için ReLu aktivasyon fonksiyonu kullanırken, çıktı katmanı için sigmoid aktivasyon fonksiyonu kullanılmıştır. Optimizasyon algoritması olarak Adam algoritması, Hata fonksiyonu için Mean Squared Eror( ortalama kareli hata) kullanılmıştır.

Çalışmamızda, sınıflandırma algoritmalarının performansını değerlendirmek için Kesinlik (P), Duyarlılık (R), F-ölçtü (F), Doğruluk (A) ve AUC değerleri dikkate alınmıştır.

### 3.2.1. İngilizce Film Yorumları Veri Seti İle İlgili Deneyler

Sınıflandırma işlemi öncesinde İngilizce film yorumları veri seti, deney durumlarına göre çeşitli ön işleme yöntemlerine tabi tutulmuştur. İlk olarak tüm yorumlardaki simgeler, noktalama işaretleri ve sayılar temizlenip tüm karakterler küçük harfe çevrilmiştir. Ardından PYTHON ortamında hazırlanan NLTK kütüphanesinden yararlanılarak Porter Stemmer ile her bir kelimenin kökleri bulunmuştur. Metin içerisindeki durak kelimeler (edat, bağlaç ve zamir) kelimeler kaldırılmıştır. 50.000 adet pozitif ve negatif yorumun bulunduğu veri setinden elde edilen farklı metin temsilleri sırayla programa verilmiştir.

İlk olarak TF ve TF-TDF terim ağırlıklandırma yöntemi ile elde edilen 1-Gram, 2-Gram ve 1-2-Gram terimlerinden oluşan metin KÇ temsillerinin performansı ölçülmüştür.

**Tablo 6.** İngilizce Veri Seti için 1-Gram Terimlerle Sınıflandırma Başarıları

| Yöntem | | P | R | F | A | AUC |
|---|---|---|---|---|---|---|
| TF | NB | 0.6874 | 0.66724 | 0.6580 | 0.6672 | 0.7562 |
| | LR | 0.8533 | 0.8531 | 0.8531 | **0.8531** | 0.9215 |
| | KA | 0.7214 | 0.7213 | 0.7213 | 0.7213 | 0.72136 |
| | k-NN | 0.6626 | 0.6620 | 0.6617 | 0.6620 | 0.7255 |
| | DVM | 0.8485 | 0.8482 | 0.8482 | 0.8482 | 0.9154 |
| | YSA | 0.8584 | 0.8325 | 0.8433 | 0.8460 | 0.8460 |
| TF-TDF | NB | 0.7288 | 0.72686 | 0.7262 | 0.7268 | 0.7673 |
| | LR | 0.8768 | 0.8767 | 0.8767 | 0.8767 | 0.9477 |
| | KA | 0.7159 | 0.7158 | 0.7158 | 0.7158 | 0.7158 |
| | k-NN | 0.6562 | 0.6557 | 0.6553 | 0.6557 | 0.7009 |
| | DVM | 0.8719 | 0.8718 | 0.8718 | 0.8718 | 0.9439 |
| | YSA | 0.8795 | 0.8882 | 0.8838 | **0.8833** | 0.8832 |

Tablo 6. incelendiğinde TF ağırlıklandırmalı 1-gram modeli için en yüksek değer LR sınıflandırma algoritması ile %85,31 oranında elde edilirken, TF-TDF tekniği ile YSA sınıflandırma algoritmasının %88,33 oranında doğru tahmin ettiği görülmektedir. Tablo 6.'daki değerler incelendiğinde TF tekniği yerine TF-TDF tekniğini kullanmak bazı sınıflandırma algoritmalarında doğru sınıflama performansını olumlu yönde etkilerken (YSA, DVM, LR, NB) bazı sınıflandırma algoritmalarında (KA, k-NN) %1 civarında bir performans kaybına neden olmuştur. Ayrıca YSA, DVM ve LR sınıflayıcılar her iki vektör temsilinde de %84 ile %87 aralığında birbirlerine yakın doğruluğa sahip en iyi sınıflayıcılar olmuşlardır. Her iki durumda da en düşük doğru sınıflama başarısı k-NN sınıflandırma algoritmasından elde edilmiştir.

**Tablo 7.** İngilizce Veri Seti için 2-Gram Terimlerle Sınıflandırma Başarıları

| Yöntem | | P | R | F | A | AUC |
|---|---|---|---|---|---|---|
| TF | NB | 0.7890 | 0.7876 | 0.7873 | 0.7876 | 0.8149 |
| | LR | 0.8098 | 0.8097 | 0.8096 | 0.8097 | 0.8930 |
| | KA | 0.7173 | 0.7168 | 0.7166 | 0.7168 | 0.7189 |
| | k-NN | 0.5996 | 0.5502 | 0.4837 | 0.5502 | 0.6207 |
| | DVM | 0.8013 | 0.8011 | 0.8011 | 0.8011 | 0.8852 |
| | YSA | 0.8064 | 0.8247 | 0.8154 | **0.8135** | 0.8134 |
| TF-TDF | NB | 0.7923 | 0.7923 | 0.7923 | 0.7923 | 0.8230 |
| | LR | 0.8225 | 0.8223 | 0.8223 | **0.8223** | 0.9061 |
| | KA | 0.7109 | 0.7108 | 0.71078 | 0.7108 | 0.7124 |
| | k-NN | 0.5860 | 0.5671 | 0.5405 | 0.5671 | 0.6047 |
| | DVM | 0.8156 | 0.8155 | 0.8154 | 0.8155 | 0.8999 |
| | YSA | 0.8117 | 0.8300 | 0.8207 | 0.8188 | 0.8187 |

TF tekniği ile 2-Gram terimlerle elde edilen metin temsilleri YSA sınıflandırma algoritması %81,55 oranında doğru tahmin ederken TF-TDF tekniği ile LR sınıflandırma algoritması %82,23 oranında doğru tahmin etmiştir. Genel olarak TF-TDF kullanımının sınıflama performansını artırdığı söylenebilir. 1-Gram yerine 2-Gram vektör temsili kullanmak performansı bazı sınıflandırma algoritmaların da artırmasına karşılık birçok sınıflandırma algoritmasın da performans azalmasına neden olmuştur. Dolayısıyla 2-Gram kullanmanın performansı olumlu yönde etkilemediği gözlemlenmiştir.

**Tablo 8.** İngilizce Veri Seti için 1-2 Gram Terimlerle Sınıflandırma Başarıları

| Yöntem | | P | R | F | A | AUC |
|---|---|---|---|---|---|---|
| TF | NB | 0.7771 | 0.7720 | 0.7709 | 0.7720 | 0.8372 |
| | LR | 0.8629 | 0.8628 | 0.8628 | 0.8628 | 0.9319 |
| | KA | 0.7236 | 0.7235 | 0.7235 | 0.7235 | 0.7235 |
| | k-NN | 0.6541 | 0.6523 | 0.6512 | 0.6523 | 0.7178 |
| | DVM | 0.8554 | 0.8553 | 0.8553 | 0.8553 | 0.9254 |
| | YSA | 0.8488 | 0.9123 | 0.8794 | **0.8750** | 0.8750 |
| TF-TDF | NB | 0.7949 | 0.7946 | 0.7946 | 0.7946 | 0.8356 |
| | LR | 0.8820 | 0.8819 | 0.8819 | 0.8819 | 0.9523 |
| | KA | 0.7216 | 0.7216 | 0.7215 | 0.7216 | 0.7216 |
| | k-NN | 0.6791 | 0.6782 | 0.6777 | 0.6782 | 0.7305 |
| | DVM | 0.8784 | 0.8783 | 0.8783 | 0.8783 | 0.9492 |
| | YSA | 0.8792 | 0.8934 | 0.8859 | **0.8850** | 0.8851 |

TF tekniği ile 1-2 Gram kombinasyonu ile elde edilen metin temsilleri için %87,50(YSA) ve TF-TDF tekniği ile %88,50(YSA) oranında doğru tahmin edildiği görülmüştür.

Benzer şekilde 1-2 Gram modelinde de 1-Gram ve 2-Gram modellerinde olduğu gibi TF-TDF ağırlıklandırmalı vektör temsilleri TF ağırlıklandırmalı vektör temsillerinden daha iyi sonuç vermiştir. Ayrıca TF-TDF ağırlıklandırmalı 1-2-Gram terimlerinin sınıflandırma başarısı 1-Gram ve 2-Gram TF-TDF ağırlıklandırmalı temsillerde sınıflandırma algoritmalarından bağımsız olarak artış göstermiştir. Aynı durum TF ağırlıkladırmalı metin temsilleri için söylemek mümkün değildir.

İngilizce film yorumları veri seti için uygulanan ikinci deneyde Word2Vec modeli kullanılmıştır ve öğrenme modeli kelime benzerliği kullanımında değerlendirilmiştir. Word2Vec modeli eğitilmeden önce ön işleme adımında yalnızca her eğitim örneği için simgeler, noktalama işaretleri ve sayılar temizlenip tüm karakterler küçük harfe çevrilmiştir.

Word2Vec modeli uygulamamıza yardımcı olması için PYTHON Kütüphanesi olan Gensim kullanılmıştır. Bu süreçte, öncelikle tüm 50.000 yorumdan oluşan etiketsiz eğitim verisi ile 25.000 yorumdan oluşan etiketli eğitim setine ilişkin yorumlar birleştirilerek 75.000 film yorumundan oluşan bir eğitim seti kullanılmıştır. Kelime vektörleri çıkarılırken CBoW ve Skip-Gram mimarisinin kullanıldığı çalışmada mimarilerin eğitimleri HS algoritması ve NS algoritması ile ayrı ayrı gerçekleştirilmiştir. Her iki mimaride de kelime pencere boyut (window) değeri 10, kelime vektör boyutu ise 300 olarak seçilmiştir. Daha sonra etiketli veri setinde yer alan yorumlar kullanılarak yorumlara ait kelime vektör temsilleri elde edilmiştir. Yorumlara ait Word2vec kelime vektör temsilerinde ilk olarak yorum içerisinde yer alan her bir kelimenin 1x300 boyutundaki Word2vec karşılığı olan yoğun(dense) vektörler elde edilmiş ardından bunların ortalaması alınıp ilgili yorumun Word2vec kelime vektör temsili oluşturulmuştur.

**Tablo 9.** İngilizce Veri Seti için Word2vec Kelime Vektör Ortalamaları ile Elde Edilen Sınıflandırma Başarıları (pencere boyut = 10, eğitim algoritması = HS)

| | | Vektör Boyutu (özellik sayısı=300) | | | | |
|---|---|---|---|---|---|---|
| Yöntem | | P | R | F | A | AUC |
| CBoW-HS | NB | 0.7596 | 0.7592 | 0.7591 | 0.7592 | 0.8393 |
| | LR | 0.8590 | 0.8589 | 0.8588 | 0.8589 | 0.9333 |
| | KA | 0.7079 | 0.7079 | 0.7078 | 0.7079 | 0.7079 |
| | k-NN | 0.7950 | 0.7910 | 0.7902 | 0.7910 | 0.8749 |
| | DVM | 0.8659 | 0.8657 | 0.8656 | **0.8657** | 0.9381 |
| | YSA | 0.8539 | 0.8534 | 0.8534 | 0.8534 | 0.9290 |
| SG-HS | NB | 0.7832 | 0.78294 | 0.7828 | 0.7829 | 0.8632 |
| | LR | 0.8687 | 0.8685 | 0.8685 | 0.8685 | 0.9409 |
| | KA | 0.7085 | 0.7084 | 0.7084 | 0.7084 | 0.7084 |
| | k-NN | 0.8078 | 0.8036 | 0.8030 | 0.8036 | 0.8858 |
| | DVM | 0.8766 | 0.8764 | 0.8764 | **0.8764** | 0.9458 |
| | YSA | 0.8760 | 0.8751 | 0.8750 | 0.8751 | 0.9458 |

**Tablo 10.** İngilizce Veri Seti için Word2vec Kelime Vektör Ortalamaları ile Elde Edilen Sınıflandırma Başarıları (pencere boyut = 10, eğitim algoritması = NS)

| | | Vektör Boyutu (özellik sayısı=300) | | | | |
|---|---|---|---|---|---|---|
| Yöntem | | P | R | F | A | AUC |
| CBoW-NS | NB | 0.7427 | 0.74194 | 0.7417 | 0.7419 | 0.8205 |
| | LR | 0.8563 | 0.85617 | 0.85615 | 0.8561 | 0.9309 |
| | KA | 0.7129 | 0.7128 | 0.7127 | 0.7128 | 0.7128 |
| | k-NN | 0.7954 | 0.7928 | 0.7924 | 0.7928 | 0.8749 |
| | DVM | 0.8633 | 0.8631 | 0.8631 | **0.8631** | 0.9357 |
| | YSA | 0.8590 | 0.8584 | 0.8584 | 0.8584 | 0.9336 |
| SG-NS | NB | 0.7787 | 0.7784 | 0.7783 | 0.7784 | 0.8585 |
| | LR | 0.8679 | 0.8677 | 0.8677 | 0.8677 | 0.9405 |
| | KA | 0.7154 | 0.7154 | 0.7153 | 0.7154 | 0.7154 |
| | k-NN | 0.8074 | 0.8034 | 0.8028 | 0.8034 | 0.8873 |
| | DVM | 0.8745 | 0.8743 | 0.8743 | 0.8743 | 0.9450 |
| | YSA | 0.8758 | 0.8753 | 0.8753 | **0.8753** | 0.9459 |

İngilizce film yorumları Word2vec vektör yöntemi için en yüksek başarı %87,64'lük bir doğruluk değeri ile HS yöntemi ile eğitilen Skip-Gram mimarisinin kullanıldığı vektör gösteriminden elde edilmiştir. Tablo 9 ve Tablo 10 'dan da görüleceği üzere Word2vec temsilleri ile yapılan deneylerde sınıflama başarısı en az TF ve TF-TDF temsillerinin kullanıldığı N-gram modelleri kadar başarılı sonuçlar alınmıştır. Hatta bazı sınıflandırma algoritmaları için Word2vec temsilleri N-gram temsillerine oranla çok daha yüksek başarı oranlarına sahip olmuşlardır. Örneğin, k-NN sınıflandırma algoritması 1-gram temsillerde ortalama %65, 2-gram temsillerde %55 civarında sınıflama başarısına sahipken Word2vec modellerinde ortalama %79 oranında bir başarım sergilemiştir.

### 3.2.2. Türkçe Film Yorumları Veri Seti İle İlgili Deneyler

Türkçe film yorumları veri seti için ön işleme aşamasında ilk olarak tüm yorumlardaki simgeler, noktalama işaretleri ve sayılar temizlenip tüm karakterler küçük harfe çevrilmiştir. Metin içerisindeki durak kelimeler (edat, bağlaç ve zamir) ve 3 harften kısa olan kelimeler kaldırılmıştır. 10.662 adet Türkçe film yorumları içeren veri seti TF ve TF-TDF değerlerinden oluşan 1-Gram, 2-Gram, 1-2Gram terimlerinden oluşan KÇ metin temsillerinin yanı sıra Word2vec kelime vektörleri kullanılmıştır ve Türkçe dili ile yazılan metinlerin duygu sınıflama başarısına olan etkileri değerlendirilmiştir.

Türkçe film yorumları veri seti için ilk olarak TF ve TF-TDF terim ağırlıklandırma yöntemi ile elde edilen 1-Gram, 2-Gram ve 1-2-Gram terimlerinden oluşan metin temsillerinin performansı ölçülmüştür.

**Tablo 11.** Türkçe Veri Seti için 1-Gram Terimlerle Sınıflandırma Başarıları

| Yöntem | | P | R | F | A | AUC |
|---|---|---|---|---|---|---|
| TF | NB | 0.8181 | 0.8164 | 0.8161 | 0.8165 | 0.8544 |
| | LR | 0.8404 | 0.8403 | 0.8401 | **0.8402** | 0.9250 |
| | KA | 0.7726 | 0.7724 | 0.7724 | 0.7726 | 0.7859 |
| | k-NN | 0.7203 | 0.7126 | 0.7098 | 0.7127 | 0.7887 |
| | DVM | 0.8355 | 0.8353 | 0.8352 | 0.8353 | 0.9187 |
| | YSA | 0.8190 | 0.8326 | 0.8253 | 0.8240 | 0.8241 |
| TF-TDF | NB | 0.8184 | 0.8165 | 0.8162 | 0.8166 | 0.8560 |
| | LR | 0.8300 | 0.8300 | 0.8298 | 0.8299 | 0.9152 |
| | KA | 0.7706 | 0.7702 | 0.7701 | 0.7703 | 0.7843 |
| | k-NN | 0.7573 | 0.7563 | 0.7559 | 0.7563 | 0.8311 |
| | DVM | 0.8267 | 0.8265 | 0.8264 | 0.8265 | 0.9106 |
| | YSA | 0.8244 | 0.8408 | 0.8323 | **0.8306** | 0.8305 |

Türkçe film yorumları verilerinin TF, TF-TDF ağırlıklandırmalı 1-Gram modeli için en yüksek başarı TF tekniği ile %84,02 (LR) oranında elde edilmiştir.

Tablo 11 incelendiğinde NB, k-NN ve YSA sınıflandırma algoritmaları için TF tekniği yerine TF-TDF tekniği kullanmak sınıflandırma performansını artırırken diğer 3 sınıflandırma algoritmalarında performans kaybı olduğu görülmektedir. İngilizce veri seti ile yapılan deneyler de olduğu gibi YSA, DVM ve LR sınıflandırma algoritmaları her iki teknikte de en iyi sınıflandırma algoritmaları olmuşlardır.

**Tablo 12.** Türkçe Veri Seti için 2-Gram Terimlerle Sınıflandırma Başarıları

| Yöntem | | P | R | F | A | AUC |
|---|---|---|---|---|---|---|
| TF | NB | 0.7551 | 0.7023 | 0.6858 | 0.7024 | 0.7109 |
| | LR | 0.7594 | 0.7401 | 0.7349 | **0.7400** | 0.8385 |
| | KA | 0.7349 | 0.7140 | 0.7073 | 0.7139 | 0.7880 |
| | k-NN | 0.6987 | 0.6799 | 0.6722 | 0.6801 | 0.7645 |
| | DVM | 0.7540 | 0.7294 | 0.7225 | 0.7294 | 0.8265 |
| | YSA | 0.8075 | 0.5991 | 0.6877 | 0.7282 | 0.7282 |
| TF-TDF | NB | 0.7538 | 0.7005 | 0.6836 | 0.7007 | 0.7091 |
| | LR | 0,7542 | 0.7352 | 0.7299 | **0.7351** | 0.8312 |
| | KA | 0.7356 | 0.7144 | 0.7076 | 0.7143 | 0.7884 |
| | k-NN | 0.7087 | 0.6913 | 0.6847 | 0.6915 | 0.7785 |
| | DVM | 0.7460 | 0.7209 | 0.7135 | 0.7209 | 0.8162 |
| | YSA | 0.7846 | 0.6319 | 0.6941 | 0.7234 | 0.7244 |

TF tekniği ile 2-Gram terimlerle elde edilen metin temsilleri için en yüksek başarı %74 oranında LR sınıflandırma algoritmasından elde edilmiştir. 1-Gram temsilleriyle karşılaştırıldığında 2-Gram temsillerinin daha düşük başarı sergilediği gözlemlenmiştir.

**Tablo 13.** Türkçe Veri Seti için 1-2 Gram Terimlerle Sınıflandırma Başarıları

| Yöntem | | P | R | F | A | AUC |
|---|---|---|---|---|---|---|
| TF | NB | 0.8242 | 0.8214 | 0.8210 | 0.8215 | 0.8652 |
| | LR | 0.8383 | 0.8381 | 0.8380 | **0.8381** | 0.9232 |
| | KA | 0.7672 | 0.7668 | 0.7667 | 0.7669 | 0.7825 |
| | k-NN | 0.7253 | 0.7220 | 0.7210 | 0.7223 | 0.7964 |
| | DVM | 0.8364 | 0.8361 | 0.8360 | 0.8361 | 0.9169 |
| | YSA | 0.8255 | 0.8246 | 0.8249 | 0.8248 | 0.8249 |
| TF-TDF | NB | 0.8240 | 0.8209 | 0.8204 | 0.8209 | 0.8648 |
| | LR | 0.8294 | 0.8292 | 0.8290 | **0.8292** | 0.9145 |
| | KA | 0.7667 | 0.7662 | 0.7662 | 0.7664 | 0.7816 |
| | k-NN | 0.7547 | 0.7543 | 0.7542 | 0.7544 | 0.8296 |
| | DVM | 0.8258 | 0.8255 | 0.8254 | 0.8255 | 0.9099 |
| | YSA | 0.8295 | 0.8212 | 0.8252 | 0.8261 | 0.8260 |

Tablo 13 incelendiğinde TF ve TF-TDF tekniği ile 1-2 Gram terimleriyle elde edilen metin temsillerinin 2-Gram terimlerine göre daha başarılı olduğu görülmüştür. Ancak aynı durum 1-Gram metin temsilleriyle karşılaştırma için söylenememektedir.

Genel olarak N-Gram temsillerinde en başarılı sınıflandırma algoritmalarının LR, DVM ve YSA olduğu söylenebilir. Benzer bir durum İngilizce veri seti ile yapılan deneylerde de gözlemlenmiştir.

Türkçe film yorumları veri seti için uygulanan ikinci deneyde ise Word2Vec modeli kullanılmıştır. Bu amaç için Türkçe dilinde yazılmış büyük bir derleme ihtiyaç vardır. Bu tez çalışmasında kelime vektörlerinin oluşturulmasında Türkçe dilinde yazılmış Wikipedia makalelerinden oluşan derlem kullanılmıştır. Kelime vektörleri oluşturulurken CBoW mimarisi HS ve NS algoritması ile ayrıca SG mimarisi HS ve NS algoritması ile eğitilmiştir. İngilizce veri seti için uygulanan Word2vec deneyleri ile kıyaslama yapabilmek için benzer şekilde kelime pencere boyut değeri 10, kelime vektör boyut değeri ise 300 olarak seçilmiştir. Etiketli Türkçe film yorumları veri setinde yer alan yorumlar kullanılarak yorumlara ait

temsiller elde edilmiştir. Türkçe film yorumları içerisinde yer alan her bir kelimenin 1x300 boyutundaki yoğun(dense) vektörlerin ortalaması alınarak ilgili yorumun Word2vec temsilleri oluşturulmuştur.

**Tablo 14.** Türkçe Veri Seti için Word2vec Kelime Vektör Ortalamaları ile Elde Edilen Sınıflandırma Başarıları(pencere boyutu =10, eğitim algoritması = HS )

| | | Vektör Boyutu (özellik sayısı=300) | | | | |
|---|---|---|---|---|---|---|
| Yöntem | | P | R | F | A | AUC |
| CBoW-HS | NB | 0.6797 | 0.6708 | 0.6663 | 0.6706 | 0.7481 |
| | LR | 0.7586 | 0.7585 | 0.7583 | **0.7584** | 0.8340 |
| | KA | 0.5924 | 0.5924 | 0.5922 | 0.5924 | 0.5926 |
| | k-NN | 0.6898 | 0.6885 | 0.6879 | 0.6886 | 0.7528 |
| | DVM | 0.7576 | 0.7575 | 0.7573 | 0.7574 | 0.8318 |
| | YSA | 0.7515 | 0.7437 | 0.7474 | 0.7490 | 0.7489 |
| SG-HS | NB | 0.6839 | 0.6782 | 0.6756 | 0.6782 | 0.7432 |
| | LR | 0.7585 | 0.7583 | 0.7581 | **0.7583** | 0.8349 |
| | KA | 0.5945 | 0.5945 | 0.5943 | 0.5945 | 0.5948 |
| | k-NN | 0.6968 | 0.6967 | 0.6964 | 0.6965 | 0.7582 |
| | DVM | 0.7582 | 0.7580 | 0.7578 | 0.7580 | 0.8323 |
| | YSA | 0.7499 | 0.7613 | 0.7548 | 0.7532 | 0.7532 |

**Tablo 15.** Türkçe Veri Seti için Word2vec Kelime Vektör Ortalamaları ile Elde Edilen Sınıflandırma Başarıları(pencere boyutu =10, eğitim algoritması = NS )

| | | Vektör Boyutu (özellik sayısı=300) | | | | |
|---|---|---|---|---|---|---|
| Yöntem | | P | R | F | A | AUC |
| CBoW-NS | NB | 0.6581 | 0.6572 | 0.6562 | 0.6567 | 0.7120 |
| | LR | 0.7474 | 0.7474 | 0.7473 | **0.7474** | 0.8209 |
| | KA | 0.5975 | 0.5974 | 0.5970 | 0.5972 | 0.5979 |
| | k-NN | 0.6733 | 0.6718 | 0.6710 | 0.6719 | 0.7277 |
| | DVM | 0.7468 | 0.7468 | 0.7467 | 0.7468 | 0.8190 |
| | YSA | 0.7256 | 0.7193 | 0.7219 | 0.7232 | 0.7231 |
| SG-NS | NB | 0.6863 | 0.6785 | 0.6750 | 0.6786 | 0.7510 |
| | LR | 0.7676 | 0.7673 | 0.7673 | **0.7674** | 0.8466 |
| | KA | 0.6079 | 0.6079 | 0.6079 | 0.6081 | 0.6094 |
| | k-NN | 0.7094 | 0.7092 | 0.7088 | 0.7090 | 0.7776 |
| | DVM | 0.7664 | 0.7662 | 0.7661 | 0.7662 | 0.8441 |
| | YSA | 0.7586 | 0.7615 | 0.7593 | 0.7590 | 0.7593 |

Tablo 14 ve Tablo 15 incelendiğinde Türkçe film yorumları Word2vec kelime vektörleri yöntemi için NS eğitim algoritması ile eğitilen CBoW mimarisinin kullanıldığı vektör gösterimi sınıflama performansını düşürürken, Skip-Gram mimarisinin kullanıldığı vektör gösterimi sınıflama performansını artırdığı görülmektedir. Dolayısıyla en yüksek başarı %76,74'lük bir doğruluk değeri ile NS yöntemi ile eğitilen Skip-Gram mimarisinin kullanıldığı vektör gösterimlerinden elde edilmiştir. En başarılı sınıflandırma algoritması, diğer deneylerde de olduğu gibi LR, DVM ve YSA olduğu görülmektedir. Fakat TF ve TDF yöntemlerin kullanıldığı N-gram modeller ile kıyaslandığında sınıflandırma algoritmalarının performanslarında düşüş olduğu görülmektedir.

Genel olarak Word2vec temsilleri ile yapılan deneylerde sınıflama başarısı beklenenin aksine TF ve TDF temsillerin kullanıldığı N-gram modeller kadar başarılı olamadığı gözlemlenmiştir.

# 4. SONUÇ VE ÖNERİLER

Bir metnin ifade ettiği öznel tutumun yani duyguların belirlenmesi için kullanılan duygu analizi, doğal dil işleme alanının önemli araştırma alanlarından biridir. Duygu analizinin amacı belirli bir metnin sınıfını (pozitif, negatif) belirlemektir.

Duygu analizi yönetim bilişim sistemleri bakış açısından değerlendirildiğinde işletmeler için oldukça faydalıdır. Örneğin, işletmeler her zaman ürün veya hizmetleri hakkında kamu veya tüketici görüşlerini bilmek istemektedir. Aynı zamanda potansiyel müşteriler bir ürünü satın almadan veya bir hizmeti kullanmadan önce mevcut müşterilerin görüşlerini öğrenmek istemektedir. Çoğu şirket sosyal medya platformlarında aktif haldedir ve markalarını ve hizmetlerini tanıtmak için bu platformları kullanmaktadır. Buna karşılık sosyal medya platformları sadece işletmelerin markalarını veya hizmetlerin tanıtabileceği bir platform değildir, aynı zamanda müşterilerin markalar hakkında konuştuğu ve markaların hedef müşterileri tarafından nasıl algılandığına dair bilgilerle dolu olduğu bir yerdir. Dolayısıyla duygu analizinden elde edilen bilgiler pazarlama stratejilerini optimize etmek için işletmelere fırsat sunmaktadır. Ayrıca işletmeler duygu analizi ile müşterilerin isteklerini karşılamak için kısa vadeli pazarlama kampanyası oluşturabilirler, duygu analizi uygulayarak kampanyalarını hedef kitlelerine daha da uygun hale getirebilirler. Bunlara ek olarak duygu analizi ürün kalitesinin geliştirilmesine ve pazar araştırmalarının tamamlanmasına yardımcı olur. Özellikle duygu analizine dayalı gerçekleştirilecek karar destek sitemleri yöneticilerin daha hızlı ve verimli ve akılcı kararlar almalarına yardımcı olacaktır. Bundan dolayı yazılı metinlerin duygu durumunun tespit edilmesi işletmeler için büyük öneme sahiptir.

Duygu analizinin başarılı bir şekilde gerçekleşmesi için etkili metin temsillerinin oluşturulması çok önemlidir. Duygu sınıflandırmada metin temsillerinin oluşturulması için en sık kullanılan yöntem kelimelerin anlamsal bağlamı dikkate almaksızın sadece kelimelerin belgede olup olmadığıyla ilgilen KÇ yöntemleridir. Fakat bu yöntemlerden elde edilen vektörlerin çok fazla sıfır değer içermesi metinleri yeterince iyi temsil edememesi sorununu da beraberinde getirmektedir. Bu tez çalışmasında N-gram yöntemleri ile elde edilen klasik KÇ metin temsillerinin yanı sıra son yıllarda metin madenciliği uygulamalarında gittikçe popülerlik kazanan

Word2vec kelime vektörleri de kullanılmış ve duygu sınıflandırma başarısına olan etkileri incelenmiştir. Bu amaçla çalışmada İngilizce ve Türkçe olmak üzere iki farklı dildeki film yorumları veri seti kullanılmıştır ve önerilen yöntemlerin farklı dillerdeki metinler üzerindeki etkisi de değerlendirilmiştir. Çalışma kapsamında sınıflandırma için hazır hale getirilen veriler üzerinde 6 farklı sınıflandırma algoritması (NB, LR, KA, k-NN, DVM, YSA) uygulanmıştır.

İngilizce film yorumları veri seti için yapılan deneylerde Word2vec kelime vektör temsil yöntemleriyle oluşturulan veri setlerinde, birçok sınıflandırma algoritması için KÇ modeliyle elde edilen başarı oranlarıyla birbirine yakın, hatta bazı durumlarda daha başarılı sonuçlar edilmiştir. Türkçe film yorumları veri seti içinse klasik KÇ modeli ile elde edilen metin temsillerin sınıflandırması daha başarılı olmuştur. Bunun ilk nedeni İngilizce film yorumları için kelime vektörlerinin oluşturulmasında kullanılan derlemin alana özgü olması fakat Türkçe film yorumları için kullanılan Wikipedia makalelerinin birçok farklı alanda yazılmış metinlerden oluşmasıdır. İkinci nedeni ise Türkçe dilinde yazılmış Wikipedia makaleleri derleminin dilbilgisi ve yazım kuralları açısından daha biçimsel ve doğru yazılmış kelimeler içerirken, doğal dille yazılmış Türkçe film yorumları veri setinin, dil bilgisi ve yazım kuralları açısından hatalı ya da eksik yazılmış kelimelerle dolu olmasıdır. Bu nedenle eksik ve hatalı yazılan kelimelerin eğitim için kullanılan Türkçe Wikipedia derleminde bulunamaması, dolayısıyla kelime vektörleri arasında ki benzerliğin beklenildiği gibi doğru hesaplanamamasına neden olmuştur. İki farklı dilde yazılan film yorumlarının değerlendirildiği bu çalışmada Word2vec kelime vektör temsillerinin, İngilizce günlük konuşma dili ile yazılan metinlerin duygu analizinde başarılı sonuçlar almak için kullanılabileceği gösterilmiştir. Türkçe metinler için yapılan çalışma sonucunda elde edilen sonuçların Türkçe'nin kurallı dil yapısı dikkate alınarak bir doğal dil işleme kütüphanesinin oluşturulması ve kelime vektör temsillerinin oluşturulmasında alana özgü derlemlerin kullanılması durumunda, Word2vec yönteminin sınıflandırma başarısının artacağı sonucuna varılmıştır.

Vektör temsil yöntemlerinden bağımsız olarak çalışmada en başarılı sınıflandırma algoritmalarının YSA, DVM ve LR yöntemleri olduğu gözlemlenmiştir. N-Gram'a dayanan vektör temsillerinde bu sınıflandırma

algoritmaları %85 ve üzerinde doğru sınıflama başarısı göstermişlerdir. Benzer şekilde Word2vec kelime vektör temsillerinde de en yüksek başarı söz konusu bu üç sınıflandırma algoritmalarından elde edilmiştir. DVM algoritmasının sınıflandırma sırasında diğer algoritmalara göre daha fazla zaman alması göz önüne alındığında yaklaşık olarak aynı başarımı gösteren YSA ve LR sınıflandırma algoritmaları, DVM sınıflandırma algoritmasına göre tercih edilebilir sınıflayıcılar olarak değerlendirilebilir. k-NN, KA ve NB sınıflandırma algoritmaları diğer algoritmalara oranla daha düşük sınıflandırma başarısı göstermişlerdir. Her iki veri seti içinde 2-Gram vektör temsilleri 1-Gram ve 1-2 Gram vektör temsillerine göre daha düşük başarım sergilemişlerdir. Bu durum veri setlerinde çok fazla sayıda birleşik kelimenin olmadığını göstermektedir. Ayrıca deney sonuçları incelendiğinde 1-Gram vektör temsilleri ve 1-2 Gram vektör temsilleri ile oluşturulan veri setlerinin yaklaşık aynı başarımı gösterdiği gözlemlenmiştir. Bu durum 1-2 Gram vektör temsillerinde tek kelimelik özniteliklerin iki kelimelik özniteliklere oranla daha etkili olduğunu göstermektedir. Öznitelik çıkarma maliyetleri göz önüne alındığında yaklaşık aynı başarıma sahip olduklarından 1-Gram vektör temsilleri, 1-2 gram vektör temsillerine göre daha tercih edilebilir bir vektör temsil yöntemi olarak değerlendirilebilir. Son olarak KÇ vektör temsil yönteminde TF yerine TF-TDF kullanmak birçok durumda sınıflandırma başarısını artırdığı görülmüştür.

Bu tez çalışmasında İngilizce metinler için her bir kelimenin duygu seviyesini gösteren SentiWordNet gibi bir sözlük bulunmadığından sözlük tabanlı yöntemlere ilişkin deneyler gerçekleştirilememiştir. Çalışmanın devamı olarak ileride Türkçe'nin kurallı dil yapısı dikkate alınarak kelimelerin eş anlamlarını, zıtlıklarını, anlamsal ilişkileri ve kutuplarını içeren bir sözlük yapısı ve doğal dil işleme kütüphanesi oluşturulduğunda sözlük tabanlı yöntemlerin Türkçe dilinde yazılmış metinlerin duygu analizinde kullanımı gerçekleştirilecektir. Ayrıca çeşitli vektör temsillerinin bir arada kullanıldığı hibrit vektör temsil yöntemlerinin kullanıldığı çalışmalar gerçekleştirilecektir. Bu çalışmalarda artan veri boyutunun getireceği maliyet ile veri karmaşıklığının sınıflandırma başarısı üzerindeki olumsuz etkilerini gidermek amacıyla çeşitli öznitelik seçim yöntemlerinin incelenmesi hedeflenmektedir. Bunlara ek olarak tez kapsamı dışında olduğu için yer verilmeyen ancak metin madenciliği üzerindeki uygulamaları giderek artan derin öğrenme yöntemlerinin

duygu analizinde kullanımı ele alınacaktır. İngilizce metinlerin sınıflandırılmasında başarılı sonuçlar elde CNN (Convolutional Neural Network)  LSTM (Long Short Term Memory) gibi derin sinir ağ modellerinin Türkçe metinlerin sınıflandırması ve duygu analizinde kullanımı ilgili çalışmalar yapılacaktır.

# KAYNAKLAR


Appel, Orestes, Chiclana Francisco, Carter Jenny ve Fujita Hamido. (2016). "A hybrid approach to the sentiment analysis problem at the sentence level". *Knowledge-Based Systems, 108*, 110-124.

Arslan, Halil, Kaynar Oğuz ve Yüksek Ahmet G. (2017). "Kurumsal Kolektif Süreçler için E-Posta İletilerinden Görev Keşfi ve Gerçek Zamanlı Görev Yönetim Sisteminin Geliştirilmesi". *Bilişim Teknolojileri Dergisi*, *10*(4), 381-388.

Bahadır, Ömer. (2007, Haziran). "Aritmetik Problemlerin Çözümlenmesi ve Simgelenmesi ". Doktora Tezi, *İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü*.

Bai, Xue (2011). "Predicting consumer sentiments from online text". *Decision Support Systems*, *50*(4), 732-742.

Bansal, Barka ve Srivastava Sangeet (2018). "Sentiment classification of online con sumer reviews using word vector representations". *Procedia Computer Scien ce*,132, 1147-1153.

Barbosa, Luciano ve Feng Junlan. (2010). "Robust sentiment detection on Twitter from biased and noisy data". *In Proceedings of the 23rd international conference on computational linguistics: posters* , 36-44.

Berger, Adam L., Della Pietra Stephen A., ve Della Pietra Vincent J. (1996). "A Maximum Entropy Approach to Natural Language Processing". *Computational linguistics*, *22*(1), 39-71.

Bosco, Giosuè L., Pilato Giovanni ve Chiavetta FFranco (2016). "A Lexicon-based Approach for Sentiment Classification of Amazon Books Reviews in Italian Language". *In: 2th International Conference on Web Information Systems and Technologies (WEBIST)*. 159-170.

Burges, Christopher J. C. (1998). "A Tutorial on Support Vector Machines for Pattern Recognition". *Data Mining and Knowledge Discovery*, *2*(2), 121-167.

Chowdhury, Gobinda G. (2003). "Natural language processing". *Annual Review of Information Science and Technology, 1*(37), 51-89. Competitions | Kaggle.

Cortes, Corinna ve Vapnik Vladimir (1995). "Support-Vector Networks". *Machine Learning, 20* (3), 273-297.

Çetin, Mahmut ve Amasyalı Fatih M. (2013). "Eğiticili ve Geleneksel Terim Ağırlıklandırma Yöntemleriyle Duygu Analizi". *In Proceedings of Signal Processing and Communications Applications Conference (SIU).*

Danışman Taner ve Alpkoçak Adil (2008). "Feeler: Emotion Classification of Text Using Vector Space Model". *In AISB 2008 Convention Communication, Interaction and Social Intelligence*, (*2*): 53-59.

Demirci, Sinem (2014). "*Emotion Analysis On Turkish Tweets"*. Yüksek Lisans Tezi, Orta Doğu Teknik Üniversitesi.

Demirtas, Erkin ve Mykola Pechenizkiy (2013). "Cross-lingual polarity detection with machine translation". *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining* ,9.ACM.

Ding, Xiaowen, Liu Bing ve Yu Philip S. (2008). "A holistic lexicon-based approach to opinion mining". *In Proceedings of the 2008 international conference on web search and data mining* , 231-240.

Dong, Zhendong, Dong Qiang ve Hao Changling (2010). "HowNet and Its Computation of Meaning". *In:Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations* , 53–56.

Feinerer, Ingo, Hornik Kurt ve Meyer David (2008)."Text mining infrastructure in r". *Journal of Statistical Software, 25*(5), 1-54.

Feldman Ronen (2013). "Techniques and Applications for Sentiment Analysis". *Commun. ACM, 56*(4), 82–89.

Gautam, Geetika ve Yadav Divakar (2014). "Sentiment analysis of twitter data using machine learning approaches and semantic analysis". *In:2014 Seventh International Conference on Contemporary Computing (IC3)* , 437-442.

Go, Alec, Bhayan Richa ve Huang, Lei (2009). "*Twitter sentiment classification using distant supervision"*. CS224N Project Report, Stanford, 1(12).

Han, Jiawei, Kamber Micheline ve Pei Jian (2011). "*Data Mining: Concepts and Techniques" (Third Edition)*. Elsevier.

Hatzivassiloglou, Vasileios ve McKeown Kathleen R. (1997). "Predicting the semantic orientation of adjectives". *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics* , 174-181.

Hu, Minqing ve Liu Bing (2004). "Mining and Summarizing Customer Reviews". *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* ,168–177.

Jiang, Suqi, Lewris Jason, Voltmer Michael ve Wang Hongning (2016). "Integrating rich document representations for text classification". *2016 IEEE Systems and Information Engineering Design Symposium (SIEDS)* , 303-308.

Jianqiang, Zhao ve Gui Xiaolin (2017). "Comparison research on text pre-processing methods on twitter sentiment analysis." *IEEE Access* 5, 2870-2879.

Kang, Hanhoon, Seong Joon Yoo, and Dongil Han (2012). "Senti-lexicon and improved Naïve

Bayes algorithms for sentiment analysis of restaurant reviews". *Expert Systems with Applications*, 39(5), 6000-6010.

Kavzoğlu, Taşkın ve Çölkesen İsmail. (2010). "Destek Vektör Makineleri Ile Uydu GörüntülerininSınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin Incelenmesi". *Harita Dergisi, 7*(144), 73-82.

Kim, Soo-Min ve Hovy Eduard. (2004). "Determining the Sentiment of Opinions". *Proceedings of the 20th International Conference on Computational Linguistics*, 1367. Association for Computational Linguistics.

Küçüksille, Ecir U ve Ateş Nurullah. (2013). Destek "Vektör Makineleri ile Yaramaz Elektronik Postaların Filtrelenmesi". *Türkiye Bilişim Vakfi Bilgisayar Bilimleri ve Mühendisliği Dergisi*, *6*(1), 18-25.

Liang, Po-Wei ve Dai Bi-Ru (2013). "Opinion Mining on Social Media Data". *2013 IEEE 14th International Conference on Mobile Data Management* ,( 2): 91-96.

Liu, Bing (2010). "Sentiment analysis and subjectivity". *Handbook of natural language processing*, (2): 627-666.

Liu, Bing (2012). "*Sentiment Analysis and Opinion Mining"*. *Synthesis lectures on human language technologies*, 5(1), 1-167.

Liu, Bing, & Zhang Lei. (2012)." A Survey of Opinion Mining and Sentiment Analysis". *Mining Text Data* , 415-463. Springer, Boston, MA.

Maas, Andrew L. , vd (2011). "Learning word vectors for sentiment analysis". *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, (1),142-150. Association for Computational Linguistics

McCallum, Andrew ve Nigam Kamal (1998)." A comparison of event models for naive bayes text classification". *In AAAI-98 workshop on learning for text categorization*, *752*(1), 41-48.

Medhat, Walaa, Hassan Ahmed ve Korashy Hoda (2014). "Sentiment analysis algorithms and applications: A survey". *Ain Shams Engineering Journal*, *5*(4), 1093-1113.

Mikolov, Tomas, Sutskever Ilya, Chen Kai, Corrado Greg ve Dean Jeffrey (2013). "Distributed representations of words and phrases and their compositionality". *In: Advances in neural information processing systems*, (2): 3111-3119.

Miller, George A. (1995). "WordNet: A Lexical Database for English". *Commun. ACM*, *38*(11), 39–41.

Mitchell, Tom M. (1997). "*Machine* Learning ",(1): 1–432. *McGraw-Hill Science/Engineering/Math*.

Morin, Frederic ve Bengio Yoshua. (2005). "Hierarchical Probabilistic Neural Network Language Model". *In Aistats*, (5): 246-252.

Musto, Catoldo, Semeraro Giovanni ve Polignano Marco (2014)." A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts". *Information Filtering and Retrieval*, 1-10.

Pak, Alexander ve Paroubek Patrick (2010)." Twitter as a Corpus for Sentiment Analysis and Opinion Mining". *In Proceedings of the Seventh Conference on International Language Resources and Evaluation* , (5), 1320-1326.

Pang, Bo ve Lee Lillian (2004). "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts". *In Proceedings of the*

*42nd annual meeting on Association for Computational Linguistics*, 271-278. Association for Computational Linguistics.

Pang, Bo, Lee Lillian ve Vaithyanathan Shivakumar (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing*,(10), 79-86. Association for Computational Linguistics.

Parikh, Ravi ve Movassate Matin (2009). "Sentiment analysis of user-generated twitter updates using various classification techniques ". CS224N Final Report, 118.

Park, Seongik ve Kim Yanggon (2016)." Building thesaurus lexicon using dictionary-based approach for sentiment classification". *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)* ,39-44.

Patra, Anuradha ve Singh Divakar (2013). "A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms". *International Journal of Computer Applications*, *75*(7), 14-18.

Pozzi, Federico A., Fersini Elisabetta, Messina, Enza ve Liu Bing (2016). "Sentiment Analysis in Social Networks*". Morgan Kaufmann.*

*Recevier Operating Characteristic - Wikipedia* (2016) 08 Haziran 2018 tarihinde *hhtps://en.wikipedia.org/wiki/Receiver_operating_characteristic* adresinden erişildi.

Soman, K. P., Loganathan, R., & Ajay, V. (2011). "*Machine Learning with SVM and Other Kernel Methods* ". PHI Learning Pvt. Ltd.

Şahin, Gürkan (2017). "Turkish document classification based on Word2Vec and SVM classifier". *Signal Processing and Communications Applications Conference (SIU)*, 2017 25th, 1-4. IEEE.

Taboada, Maite, Brooke Julian, Tofiloski Milan, Voll Kimberly ve Stede Manfred. (2011). "Lexicon-based Methods for Sentiment Analysis". *Comput. Linguist.*, *37*(2), 267–307.

Tan, Songbo ve Zhang Jin (2008). "An empirical study of sentiment analysis for chinese document". *Expert Systems with Applications*, *34*(4), 2622-2629.

Tang, Bo, Kay Steven ve He Haibo (2016). "Toward Optimal Feature Selection in Naive Bayes for Text Categorization". *IEEE Transactions on Knowledge and Data Engineering*, *28*(9), 2508-2521.

Tsytsarau, Mikalai ve Palpanas Themis (2012). "Survey on mining subjective data on the web". Data Mining and Knowledge Discovery, *24*(3), 478-514.

Turney, Peter D. (2002). "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews". *Proceedings of the 40th annual meeting on association for computational linguistics*, 417-424.

Venekoski, Viljami, Puuska Samir ve Jouko Vankka (2016). "Vector Space Representations of Documents in Classifying Finnish Social Media Texts". *Communications in Computer and Information Science*, 525-535. Springer, Cham.

Whitelaw, Casey, Garg Navendu ve Argamon Shlomo (2005). "Using appraisal groups for sentiment analysis". *Proceedings of the 14th ACM international conference on Information and knowledge management* , 625-631.

Wiebe, Janyce M. , Bruce Rebecca F. ve O'Hara Thomas P. (1999). "Development and use of a gold-standard data set for subjectivity classifications". *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 246-253.

Zagibalov, Taras ve Carroll John (2008)." Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Text". *Proceedings of the 22nd International Conference on Computational Linguistics* ,(1) , 1073-1080.

Zaidan, Omar F., Eisner Jason ve Piatko Christine D. (2007). "Using 'annotator rationales' to improve machine learning for text categorization". *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 260-267.

# ÖZGEÇMİŞ

## KİŞİSEL BİLGİLER


Adı Soyadı : **Ayşegül ALBAYRAK**

Uyruğu : **T.C.**

Doğum Tarihi ve Yeri : **1992/Sivas**

e-posta :aysegul.albyrk@gmail.com


## EĞİTİM

| **Derece** | **Kurum** | **Mezuniyet Yılı** |
|---|---|---|
| Lisans | **Cumhuriyet Üniversitesi** | **2015** |
| Yüksek Lisans | **Cumhuriyet Üniversitesi** | **2018** |

## İŞ TECRÜBESİ

| **Tarih** | **Kurum** | **Görev** |
|---|---|---|

## YABANCI DİL BİLGİSİ

**Yabancı Dilin Adı** **YÖKDİL (78,75)**