İNGİLİZCE VE TÜRKÇE TWİTTER MESAJLARININ WORD2VEC MODELİ İLE SINIFLANDIRILMASI

Abdullah Ammar KARCIOĞLU

Yüksek Lisans Tezi
Bilgisayar Mühendisliği Anabilim Dalı
Dr. Öğr. Üyesi Tolga AYDIN
2018

ATATÜRK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

İNGİLİZCE VE TÜRKÇE TWİTTER MESAJLARININ WORD2VEC MODELİ İLE SINIFLANDIRILMASI

Abdullah Ammar KARCIOĞLU

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

ERZURUM
2018

T.C.
ATATÜRK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TEZ ONAY FORMU

İNGİLİZCE VE TÜRKÇE TWİTTER MESAJLARININ WORD2VEC MODELİ İLE SINIFLANDIRILMASI

Dr. Öğr. Üyesi Tolga AYDIN danışmanlığında, **Abdullah Ammar KARCIOĞLU** tarafından hazırlanan bu çalışma **07/08/2018** tarihinde aşağıdaki jüri tarafından **Bilgisayar Mühendisliği Anabilim Dalı**'nda Yüksek Lisans Tezi olarak **oybirliği/~~oy çokluğu (.../...)~~** ile kabul edilmiştir.

Başkan	: Dr. Öğr. Üyesi Tolga AYDIN	*İmza*	:
Üye	: Dr. Öğr. Üyesi Bilal USANMAZ	*İmza*	:
Üye	: Dr. Öğr. Üyesi Hacı Mehmet GÜZEY	*İmza*	:

Yukarıdaki sonuç;

Enstitü Yönetim Kurulunun 16.../08.../2018... tarih ve 33.../...20..... nolu kararı ile onaylanmıştır.

Prof. Dr. Mehmet KARAKAN
Enstitü Müdürü

Not: Bu tezde kullanılan özgün ve başka kaynaklardan yapılan bildirişlerin, çizelge, şekil ve fotoğrafların kaynak olarak kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere tabidir.

ÖZET

Yüksek Lisans Tezi

İngilizce ve Türkçe Twitter Mesajlarının Word2Vec Modeli İle Sınıflandırılması

Abdullah Ammar KARCIOĞLU

Atatürk Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Tolga AYDIN

Teknolojinin gelişmesi ve internetin tüm dünyaya yayılmasıyla birlikte insanların dünyada meydana gelen değişimlerden her an haberdar olabilmesi için ve kendi düşüncelerini herkese paylaşabilmesi için sosyal medya platformları zamanla gelişmiş durumdadır. Tüm dünyada en çok kullanılan sosyal medya platformlarından biri olan twitter günlük hayatın en önemli parçalarından biri haline gelmiştir. Twitter ile kullanıcılar kendi duygu ve düşüncelerini paylaşarak veri madenciliği alanında sosyal medyada duygu analizi çalışmalarında kullanılabilecek önemli veri kaynaklarını oluşturmaktadırlar. Tez çalışmamızda duygu analizini metin sınıflandırma problemi olarak ele alınarak metin sınıflandırma teknikleri uygulanmıştır. Python programlama dilinde gerçekleştirilen bu çalışmanın ilk aşamasında, kullanıcıların paylaşmış oldukları Türkçe twitter mesajlarında metin temsili yöntemlerini kullanarak duygu analizi çalışması gerçekleştirilmiştir. Amaç, anlamsal ilişki tabanlı metin temsili yöntemlerinden Word2Vec modelinin duygu analizi çalışmalarında başarıya olan etkilerinin karşılaştırılmasıdır. 3 farklı modelin uygulandığı çalışmanın ilk aşamasında, 3. modelde Word2Vec modeline Rastgele Orman algoritması uygulanmasıyla %66,40 ile en yüksek başarı yüzdesi elde edilmiştir. Elde edilen sonuçlar scikit-learn kütüphanesine ait makine öğrenmesi algoritmalarını kullanıp performans metrikleri kıyaslanarak Türkçe doğal dil işleme çalışmalarına literatür katkı sağlanmıştır. Bu çalışmanın ikinci aşamasında ise, İngilizce ve Türkçe twitter mesajlarındaki etiketli verilerin sınıflandırılmasında Word2Vec modelinin uygulanması ve mesajlar üzerinde kök alma işleminin Word2Vec modeline olan etkisi araştırılmaktadır. Çalışmamızın ikinci aşamasında, İngilizce ve Türkçe olmak üzere iki farklı veri kümesi bulunmaktadır. Her bir veri setine twitter mesajlarının kökleri alınmamış hali ve kökleri alınmış hallerine öznitelik çıkarma yöntemlerinden kelime torbası (bag of words, BOW) ve Word2Vec modelleri uygulanmıştır.

Python programlama dilinde uygulanılan bu çalışmada, scikit-learn sınıflandırma algoritmalarından Doğrusal SVM ve Lojistik Regresyon uygulanarak başarı yüzdeleri kıyaslanmıştır ve duygu analizi sınıflandırmasında iyi sonuçlar ürettiği gösterilmiştir.

Ağustos 2018, 118 Sayfa

Anahtar Kelimeler: Twitter, Duygu Analizi, Metin Temsilleri, BOW, Word2Vec, Kelime Yerleştirmeleri, Makine Öğrenmesi

ABSTRACT

MS Thesis

CLASSIFICATION OF ENGLİSH AND TURKISH TWITTER MESSAGES USING WORD2VEC MODEL

Abdullah Ammar KARCIOĞLU

Atatürk University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering

Supervisor : Asst. Prof. Dr. Tolga AYDIN

With the development of technology and the spread of the internet all over the world, social media platforms have evolved over time so that people can be aware of the changes happening in the world at any moment, and that everyone can share their own thoughts. Twitter, one of the most used social media platforms around the world, has become one of the most important parts of everyday life. With twitter, users share their own feelings and thoughts to create important data sources that can be used in sentiment analysis work on the social media in the field of data mining. In the first phase of this study, which is implemented in python programming language, sentiment analysis was performed by using text representations in turkish twitter messages that users shared. The aim of the study, the performance effects of the term frequency-based Tf-Idf model(Bag-of-Words) and semantic relation based Word2Vec model are compared on sentiment analysis. In the first phase of this study, which applied 3 different models, in the third model, the highest accuracy percentage was obtained with 66.40% by applying Random Forest algorithm to Word2Vec model. The results obtained using the machine learning algorithms from the scikit-learn library compared the performance metrics and provided the literature contribution to turkish natural language processing studies. In the second phase of this study, the implementation of the Word2Vec model in the classification of labeled data in english and turkish twitter messages and the effect of root retrieval on messages on the Word2Vec model are investigated. In this phase, there are two different data sets: English and Turkish. For each data set, the roots of twitter messages have been unfounded and the roots have been extracted using attribute extraction methods, the bag-of-words(BOW) and the Word2Vec models have been applied.

In the second phase of this study, we compared the success percentages by applying Linear SVM and Logistic Regression from the scikit-learn classification algorithms in python programming language and demonstrated that it produces good results in sentiment analysis classification.

August 2018, 118 Pages

Keywords: Twitter, Sentiment Analysis, Text Representation,, BOW, Word2Vec, Word Embeddings, Machine Learning

TEŞEKKÜR

Bu çalışmada bilgilerini ve tecrübelerini esirgemeyen danışman hocam Sayın Dr. Öğr. Üyesi Tolga AYDIN'a, doğal dil işleme çalışmalarında ve Türkçe veri seti temininde yapmış olduğu katkılarından dolayı Sayın Dr. Öğr. Üyesi Gülşah Tümüklü ÖZYER'e, çalışma yaptığım sürece manevi desteklerini esirgemeyen mesai arkadaşım Arş. Gör. Emrah ŞİMŞEK'e, Arş.Gör. Mustafa Furkan KESKENLER'e ve çok kıymetli aileme teşekkür ederim.

Abdullah Ammar KARCIOĞLU
Ağustos Erzurum

İÇİNDEKİLER

SİMGELER ve KISALTMALAR DİZİNİ

Kısaltmalar

API	: Uygulama Programı Arayüzü
BOW	: Kelime Torbası
CRM	: Müşteri İlişkileri Yönetimi
DDİ	: Doğal Dil İşleme
DVM	: Destek Vektör Makineleri
KEYKSA	: K En Yakın Komşu Sınıflandırma Algoritması
MÖ	: Makine Öğrenmesi
NLTK	: Doğal Dil İşleme Kütüphanesi
POS	: Konuşma Parçası
RO	: Rastgele Orman
TF	: Terim Frekansı
IDF	: Ters Metin Frekansı
TF-IDF	: Terim Frekansı- Ters Metin Frekansı
LSA	: Gizli Anlamsal Analiz
PLSA	: Olasılıksal Gizli Anlamsal Analiz
LDA	: Gizli Dirichlet Tahsisi
TDA	: Tekil Değer Ayrışımı
Word2Vec	: Kelime Vektör Dönüşümü

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

Yapılan twitter istatistik verilerine göre (Aslam, 2018) günlük 100 milyon kullanıcıya ve ayda 500 milyondan fazla tweet mesajına sahip olan twitter ile kullanıcılar kendi duygu ve düşüncelerini tüm dünya ile hızlı ve güvenilir bir şekilde paylaşmaktadır (Xie vd. 2016).

Twitter; günlük kullanıcı sayısının ve yapılan paylaşımların çok fazla olması nedeniyle gün geçtikçe en önemli sosyal medya platformlarından biri haline gelmiştir. Sahip olduğu büyük veri seti sayesinde duygu analizi çalışmalarında araştırmacılar için önemli bir kaynak haline gelmiştir (Gemci ve Peker, 2013). Uzun içeriğe sahip diğer metinlerden farklı olarak, twitter mesajları 140 karakter ile sınırlandırılmıştır ve bu yüzden twitter bir mikro-blog servisi olarak adlandırılmaktadır (Kim vd. 2010).

Zaman içinde twitter'ın büyük bir sosyal medya platformu haline gelmesiyle, twitter sahip olduğu büyük veri seti sayesinde araştırmacıların çalışma alanlarının değişmesine neden olmuştur. Önceden araştırmacılar twitter'ın yapısına yoğunlaşarak bazı çalışmalar gerçekleştirirlerken daha sonra twitter verilerinden anlamsal bilgilerin çıkarılması üzerine çalışmalar gerçekleştirmişlerdir ve sosyal medyada duygu analizi çalışmalarına katkı sağlamışlardır (Kim vd. 2010).

Literatürde word2vec ile ilgili daha etkili sözcük temsilleri üretmek için kelime yerleştirmeleri önerilmiştir. Örneğin, Word2Vec modelinde, sabit bir bağlam penceresi içinde bir kelimenin görülme ihtimalini en üst düzeye çıkararak sözlükteki her bir kelime için sığ sinir ağından yoğun gerçek değerli bir vektör öğrenmenin mümkün olduğu gösterilmiştir (Yang vd. 2017). Sonuç olarak, anlamca benzer kelimelerin birbirine yakın olduğu gösterilmiştir.

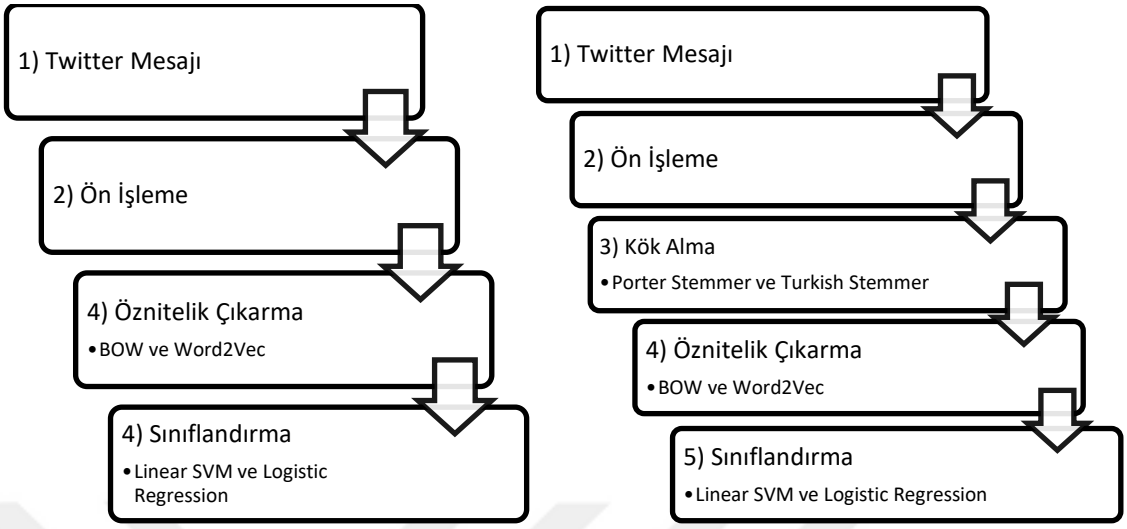
Türk dili üzerine literatürde doğal dil işleme çalışmalarının az olması nedeniyle, ayrıca anlamsal ilişki tabanlı çalışmaların fazla olmaması nedeniyle Şekil 1.1'de gösterildiği gibi çalışmamızın ilk aşamasında pozitif ve negatif etiketli Türkçe twitter mesajlarına metin temsili yöntemlerini uygulayarak duygu analizi çalışmasını gerçekleştirmek hedeflenmiştir. Bu çalışmanın ilk aşamasında metin temsili yöntemlerini kullanarak etiketli Türkçe twitter mesajlarındaki duygu analizi sınıflandırılması çalışmasında iyi sonuçların üretildiği gösterilmiştir.

Şekil 1.1. Çalışmamızın İlk Aşamasındaki Sistem Modeli

Ayrıca duygu analizi çalışmalarında bu metin temsillerinin ve sınıflandırma algoritmalarının başarılı sonuçlar üreteceği gösterilmiştir. BOW ile Word2Vec (geleneksel olan LDA ve PLSA yöntemlerinin aksine) metin temsillerini kullanarak her bir öznitelik çıkarma yöntemi için Doğrusal SVM, Doğrusal Regresyon, KNN, Karar Ağacı ve Rastgele Orman algoritmaları uygulanıp başarı metrikleri kıyaslanmıştır.

Literatürde konu modelleme üzerine LSA, PLSA, LDA gibi birçok algoritma ve yöntem geliştirilmiştir (Alghamdi ve Alfalqi 2015). Ancak Mikolov ve arkadaşları tarafından geliştirilen word2vec modeli ve uygulamaları son 4 yıldır bu model üzerinde ilginin giderek artmasına neden olmuştur ve ayrıca Word2Vec modeli tarafından öğrenilen sözcükler arasında anlamsal ilişki içerdiğini ve doğal dil işleme çalışmalarında çok kullanışlı olduğu gösterilmiştir (Rong 2014). Bu nedenle, çalışmamızın ikinci aşamasında Word2Vec modelini kullanmamızın amacı; Word2Vec modelinde kelimeler arasında anlamsal(semantik) ilişkinin bulunması nedeniyle bize twitter mesajlarının sınıflandırılmasında daha doğru sonuçlar üretecek olmasıdır. Ayrıca, veri setlerinin her birinin köklerini alarak bu veri setlerini kendi içlerinde kıyasladığımızda kök alma işleminin kelimeler arasındaki anlamsal ilişkiye olan etkisini incelemek amaçlanmaktadır.

Şekil 1.2'de gösterildiği gibi tez çalışmamızın ikinci aşamasında metinlerin basit temsillerini kullanarak seçmiş olduğumuz yöntemlerin etiketli mesajların sınıflandırılmasında iyi sonuçlar verebileceği gösterilmiştir. Önerilen sistem her biri belirli bir metin temsiline (Bag-of-Words veya Word2Vec) sahip üç temel sınıflandırıcıdan oluşmaktadır. Bu sınıflandırıcılarımız, verilerimizin kökleri alınmadan önceki hallerine ve kökleri alındıktan sonraki hallerine uygulanmıştır. Temel sınıflandırıcı olarak makine öğrenmesi algoritmalarından Doğrusal SVM ve Doğrusal Regresyon kullanılmıştır.



Şekil 1.2 Çalışmamızın İkinci Aşamasındaki Sistem Modeli

Tezin geri kalan kısmı şu şekilde düzenlenmiştir. Bölüm 2 kuramsal temelleri anlatmaktadır. Veri setlerimiz Bölüm 3'te sistemimizi, öznitelik çıkarıma yöntemlerimizi ve kullandığımız sınıflandırıcıları, ön işleme adımları, kök alma işlemleri ve kelime yerleştirmeleri hakkında bazı ayrıntıları açıklamaktadır. Araştırma bulgularını ve tartışma kısmını 4. bölümde sunmaktayız. Son olarak 5. bölüm, sonuçlarımızı ve gelecekteki çalışmalarımızla ilgili önerilerimizi özetlemektedir.

2. KURAMSAL TEMELLER

Veri madenciliği tekniklerini kullanarak verilerden anlamlı bilgiler elde edilmek amaçlanmıştır. Yapmış olduğumuz tez çalışması genel kapsamı itibariyle veri madenciliği uygulama alanına girmektedir. Veri madenciliği aşağıdaki gibi birçok kişi tarafından yapılan çalışmalarda şu şekilde tanımlanmıştır:

- Veri madenciliği, işlenmemiş olan ham verinin tek başına veremediği bilgiyi çıkarımsal olarak bulan, bir veri madenciliği süreci olarak belirtmiştir (Jacobs 1999).
- Büyük veri setleri içinden gelecekle ilgili tahminleri ve değerlendirmeleri bilgisayar destekli uygulamalar sayesinde sağlayan bir süreç olarak veri madenciliğini tanımlamıştır (Doğan ve Türkoğlu 2007).
- Veritabanı yönetim sistemini, istatistik, makine öğrenme, örüntü tanıma ile etkileşimli yeni bir disiplin ve büyük veritabanlarında tahmin edilemeyen ilişkilerin ikincil analizidir (Hand 1998).

Veri madenciliği, veri setleri arasındaki örüntülerin ya da düzenin, verinin analizi ve yazılım tekniklerinin kullanılarak elde edilmesidir (Şimşek 2012). Veriler arasındaki ilişki, kurallar ve özellikler bilgisayar yardımı ile belirlenir. Amaç, daha önceden fark edilmemiş veri örüntülerini tespit edip günlük yaşamda kolaylık sağlamaktır. Veri madenciliği, akıllı yöntemler sayesinde büyük miktarda veriden anlamlı bilgilerin elde edilmesi sürecidir (Şimşek 2012).

Tez çalışmamızda veri madenciliği uygulama alanı olarak sosyal medya platformlarından biri olan Twitter seçilmiştir. Twitter ile kullanıcıların yapmış oldukları paylaşımlar ile veri madenciliğinde kullanılabilecek önemli veri kaynakları oluşturulmaktadır. Twitter kaynaklı veri seti oluşturmak için aşağıdaki birçok neden kabul edilebilir;

- 140 karakterle sınırlı olan 'tweet mesajı' ile Twitter, kullanıcıların birbiriyle iletişim kurabildikleri güvenilir bir sosyal medya platformu olması.
- Farklı dil, kültür ve inanıştan milyonlarca insanın kullandığı twitter ile birbirinden farklı zengin içeriklere sahip mesajlara elde edilebilmesi.
- Çalışılabilecek birden çok konu alanına sahip olması ve belirlenen anahtar kelimeye özgü mesajların elde edilebilmesi.
- Aylık 328 milyon aktif kullanıcısı (Wolfe 2018) ile küresel sorunlara çözüm getirilip dünyanın gidişatına ışık tutabilecek güçte olması.

Veri madenciliğinin çalışma alanlarından birisi metin madenciliğidir. Metin madenciliği, temel olarak yapısal olmayan metinlerden bilgi içeren yapısal metinleri üretme işlemi olarak tanımlanabilir. Metinlerin işlenmesi ile anlamlı bilgilerin elde edilmesi için, veri ön işleme ve özellik çıkartımı gibi isimlerle adlandırılan bazı adımların gerçekleştirilmesi gerekmektedir. Bu aşamalardan sonra yapısal olmayan veriler, metin madenciliğinin kullanılacağı ve bilgisayarlar tarafından işlenen yapısal bir biçime dönüştürülebilir. Bu sayede büyük miktardaki veriler içerisinde bulunan değerli bilgiler keşfedilmiş olur. Üretilen anlamlı bilgiler kullanılarak, kurum ya da kuruluşların faydalanacağı çeşitli sonuçlara ulaşılabilir (Kılıç vd. 2016). Tez çalışmamızda metin tabanlı Twitter mesajları veri madenciliği teknikleriyle işlenerek duygu analizi çalışması yapmak hedeflenmiştir.

Tez çalışmamızda sosyal medya platformlarından biri olan Twitter'da doğal dil işlemenin alt çalışma alanlarından duygu analizi yapmak hedeflenmiştir. Doğal dil işleme kısaca, dil ile ilgilenen hesaplama bilimleri alanıdır (Albayrak 2011). Doğal dil işleme (DDİ) alanı, dilleri işleyen, anlayan ve oluşturan bilgisayarların nasıl oluşturulacağını araştırır. DDİ teknikleri, bilgisayarların elde ettikleri verileri dikkate alarak davranış kazanmasını sağlayan makine öğrenmesi algoritmalarını kullanır. DDİ birçok alt bölge ve görevi içerir. Otomatik özetleme, söylem analizi, makine çevirisi, ilişki çıkarımı, soru cevaplama bu alanların sadece birkaç örneğidir. Bu alanların gelişimi, diğer birçok alanda daha birçok alanın gelişmesiyle sonuçlanmaktadır (Albayrak 2011).

Uygulamada farklılık göstermesine rağmen temel işlem adımları ortak olan doğal dil işleme 4 ana başlık altında incelenebilir:

- **Sesbilim:** Harflerin seslerini ve bu harflerin dil içinde nasıl kullanıldığını inceler. Alfabesi olan dillerin kendi içinde harflerinin sesleri birbirlerinden farklıdır. Sesleri sözcüklere dönüştürerek konuşma dilini yazı diline çevirmeyi hedefler.
- **Biçim birim**: Dil kurallarına uygun biçimde her sözcük tek tek ele alınarak sözcüğün yapısını inceler. Sözcüklerin bütün parçaları işlemden geçirilerek eklerini, köklerini, bunlara ilişkin kurallarını ve bu yapıların sınıflandırılmalarını ele alır.
- **Sözdizimi:** Analizi tamamlanmış sözcüklerin bir araya getirilerek nasıl bir cümle yapısını ve bu cümle yapılarını kullanarak paragraf bütünlüğünün nasıl oluşturulacağını ele alır.
- **Anlambilim:** Anlambilim dilin dış dünya ile bağlantı kurmasını sağlar. Cümle yapısının anlaşılması ve bunun sonucunda eyleme geçilmesi bu aşamada olur. Dilde sözcüklerin dizilişlerinin cümlelere kazandırdığı anlamların incelenmesi ve bu yolla anlam kazandırılması temel işlevdir.

Duygu analizi üzerine literatürde; farklı ilgi alanlarına sahip kullanıcıların yorumlarından bir korpus oluşturularak bu veri seti üzerinde duygu analizi çalışması gerçekleştirilmiştir. Amaç; düşünce madenciliği alanına katkı sağlamak ve oluşturulmuş bir korpus ile duygu sınıflandırma üzerine bir model geliştirmektir. Temelde n-gram metodunu kullanarak linguistik analiz çalışması gerçekleştirmişler (Pak vd. 2010).

Twitter duygu analizi ile ilgili yapılan başka bir çalışmada, POS(Part of Speech) özelliklerinin mikro-blog alanında duygu analizi için yararlı olmayabileceğini gösterilmiştir (Kouloumpis vd. 2011). Çalışmada eğitim verilerini toplamak için hashtag'leri kullanmanın pozitif ve negatif ifadelere dayalı verilerde başarılı sonuçların ürettiği gösterilmiştir. Bununla birlikte, hangi yöntemin daha iyi eğitim verisi kullanılarak iyi sonuçlar üretmesinin özelliklerin türüne bağlı olabileceği anlatılmıştır. Sonuç olarak, mikro-blog servislerinin özellikleri eklendiğinde başarının düşeceği gösterilmiştir (Kouloumpis vd. 2011).

Duygu analizi üzerine literatürde; twitter üzerinden belirli markalar için yazılanların, o marka için iyi mi, kötü mü veya duygu belirtmeyen bir cümle mi sorularından, makine öğrenmesi yöntemlerini kullanarak, geri bildirim elde etme üzerine çalışmalar yapılmıştır (Nalçakan vd. 2015). Ayrıca makine öğrenmesi yöntemlerinden denetimli öğrenme yaklaşımı kullanılarak sosyal medyada twitter üzerine duygu analizi çalışması yapılmıştır ve bazı gıda firmalarının çeşitli ürünlerine ait yapılan yorumlardan oluşturulan veri setlerine makine öğrenmesi algoritmaları uygulanarak dengeli ve dengesiz veri setlerinin başarı yüzdeleri kıyaslanmıştır (Nizam ve Akın 2014).

Yelmen (2016) duygu analizinin fikir madenciliği ile ilgili çalışmalarını:"

1. Var olan verilerden özellik veya öznitelik seçiminin yapılması (Feature Selection)

2. Metinde okuyucuya düşünce bildiren kısımların çıkarılması (Subjectivity Extraction)

3. Metinden çıkarılan görüşün duygusal ilşkisinin belirlenmesi(Emotion Identification)

4. Düşünce kutuplarının belirlenmesi (Identifying Opinion Polarity)

5. Görüşlerin hedefinin çıkartılması (Target extraction)

6. Metinde geçen düşüncenin özetlenmesi (Opinion Summarization)

7. Kaynakları geliştirme ve sözlük-derlem oluşturma (Developing Resources)

8. Doğal dil işleme ile edinilen bilgilerin transferinin sağlanması (Learning Transfer)

9. Bir düşünceye yönelik kaynağının tespit edilmesi (Identifying Opinion Source)" 9 alt başlık olacak şeklinde listelemiştir (Yelmen 2016).

Twitter'da anlamsal analiz üzerine yapılan çalışmalara bakacak olursak, PLSA ve LDA yöntemlerini kullanarak twitter içeriklerinden anlamsal ilişki ile mesajların hangi konu başlıkları altında yoğunlaştığı gösterilmiştir (Kim vd. 2010). Ayrıca, LSA ve LDA yöntemlerini kullanarak duygu analizinde anlamsal ilişki çıkarmak için, anlam ve duygu bilgisini yakalayan metin temsillerini öğrenen bir vektör uzay modelini sunmuşlardır (Maas vd. 2011). Modelin olasılıksal temeli, yaygın olarak kullanılan matriks faktörizasyonu tabanlı tekniklerin ezici sayısına bir alternatif olarak kelime vektörünün indüksiyonu için teorik olarak doğrulanmış bir teknik verir.

Twitter'da anlamsal analiz üzerine yapılan çalışmalara bakacak olursak, PLSA ve LDA yöntemlerini kullanarak twitter içeriklerinden anlamsal ilişki ile mesajların hangi konu başlıkları altında yoğunlaştığı gösterilmiştir (Kim vd. 2010). Ayrıca, LSA ve LDA yöntemlerini kullanarak duygu analizinde anlamsal ilişki çıkarmak için, anlam ve duygu bilgisini yakalayan metin temsillerini öğrenen bir vektör uzay modelini sunmuşlardır (Maas vd. 2011). Modelin olasılıksal temeli, yaygın olarak kullanılan matriks faktörizasyonu tabanlı tekniklerin ezici sayısına bir alternatif olarak kelime vektörünün indüksiyonu için teorik olarak doğrulanmış bir teknik verir.

Literatürde LSA, PLSA ve LDA gibi konu modelleme yöntemleri geliştirilmiştir (Alghamdi ve Alfalqi 2015). Ancak Mikolov ve arkadaşlarının geliştirmiş olduğu anlamsal ilişki tabanlı yöntemlerden biri olan Word2Vec (Mikolov vd. 2013a;2013b) ile metinler üzerinde anlamlı bilgilerinin çıkarılması hedeflenmiştir (Ronga 2014).

3. MATERYAL ve YÖNTEM

3.1. Materyal

3.1.1. Twitter

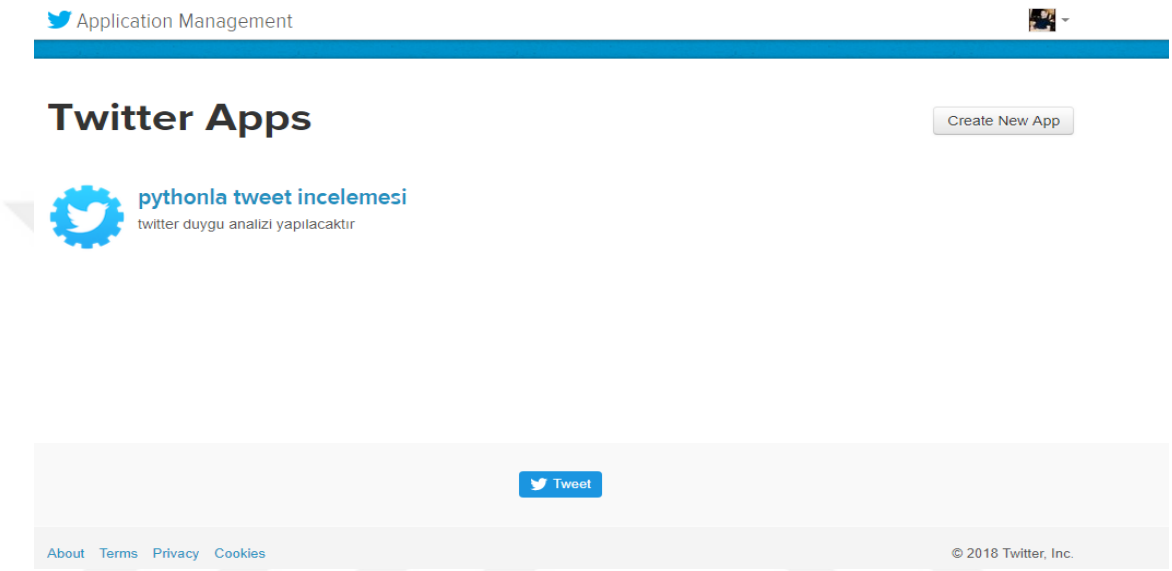
Teknolojinin gelişmesi ve internetin tüm dünyaya yayılmasıyla birlikte insanların dünyada meydana gelen değişimlerden her an haberdar olabilmesi için ve kendi düşüncelerini herkese paylaşabilmesi için sosyal medya platformları zamanla gelişmiş durumdadır. Sosyal medya platformlarından biri olan twitter günlük hayatın en önemli parçalarından biri haline gelmiştir. Twitter ile kullanıcılar kendi duygu ve düşüncelerini paylaşarak veri madenciliği alanında sosyal medyada duygu analizi çalışmalarında kullanılabilecek önemli veri kaynaklarını oluşturmaktadırlar. Twitter ile araştırmacıların çalışmalarında kullanabilecekleri içerikleri elde etmek için birçok programlama dili tarafından twitter api geliştirilmiştir.

3.1.2. Twitter API

Sosyal medya madenciliği ülkemizde çok popüler olmasa da dünya üzerinde birçok kişi tarafından ilgiyle takip ediliyor. Twitter, hızlı bir şekilde mesajlaşma platformu sağlamasının yanında program geliştirenler için bir arayüze de sahip. Bu arayüz sayesinde, profili halka açık olan kişilerin paylaştığı durumları takip etmenin yani sıra kimin kaç takipçisi var, bir post ne kadar paylaşılmış gibi detaylı analiz yapmak da mümkün olmuştur(Kurkcu 2014).

Şekil 3.1’de gösterildiği gibi, API (Application Programming Interface - Uygulama Programlama Arayüzü), Youtube, Facebook, Twitter, Google, WordPress gibi bir uygulamanın, servisin, platformun sahip olduğu yeteneklerin dışarıdan izin verilen sınırlandırmalar dahilinde kullanılabilmesini sağlayan bir arayüzdür (Aksan 2018).

Twitter, Facebook, Google, Wikipedia ve hatta GittiGidiyor gibi birçok büyük sistemin üzerine uygulama geliştirilmesi için API'ları vardır (Kotan 2014). Tweepy Modülü python'da twitter için geliştirmeler yapmaya yarayan bir kütüphanedir. Twitter, API'ı kullanılarak yazılmıştır. Tweet atmak, timeline'ı okumak, takipçiler, takip edilenler vs gibi birçok işlemi bu kütüphane üzerinden yapılabilir (Kotan 2014).



Şekil 3.1. Twitter API'nin Giriş Ekranı

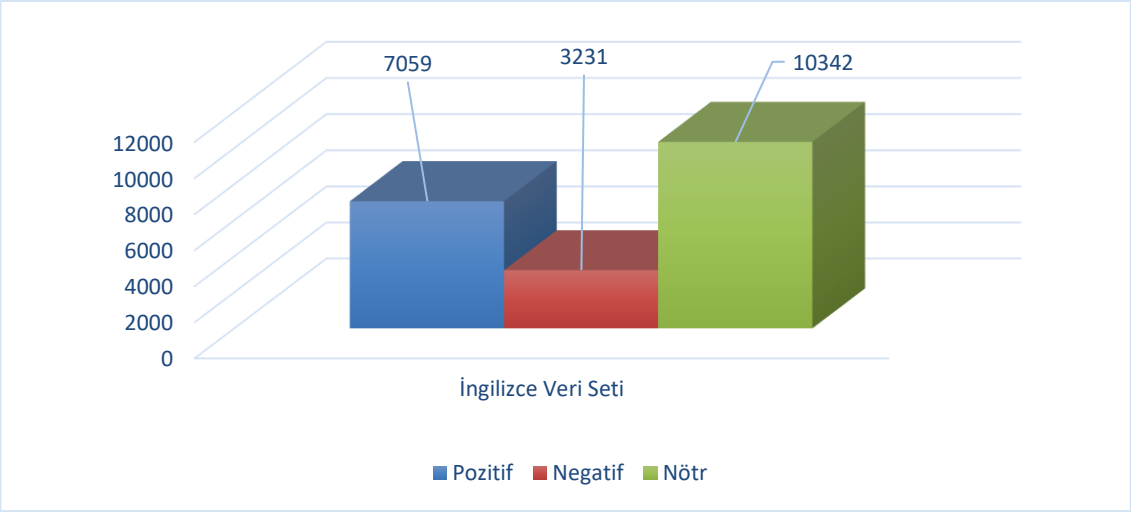
```
consumer_key = ""
consumer_secret = ""
access_token = ""
access_token_secret = ""


giris = tweepy.OAuthHandler(consumer_key, consumer_secret)
giris.set_access_token(access_token, access_token_secret)
```

Şekil 3.2. Twitter API İçin Gerekli Olan Anahtar Değerler

Twitter'dan verileri çekebilmemiz için oluşturduğumuz twitter api ile Şekil 3.2'de gösterildiği 4 adet anahtar şifre oluşturmuş oluruz. Bu anahtar değerleri python programlama dili ile çekeceğimiz mesajlar için kullandıktan sonra hangi konu üzerine çalışmak isteniyorsa anahtar kelime yazılıp o değer üzerine twitter mesajları çekilmiş olur.

3.1.3. İngilizce Twitter Veri Seti

Şekil 3.3.'de gösterildiği gibi, tez çalışmamızda duygu analizi sınıflandırılmasında pozitif (7059 adet), negatif (3231 adet) ve nötr (10342 adet) olmak üzere toplam 20632 adet İngilizce twitter mesajı (Kim 2017), veri seti olarak kullanıldı. Etiketler mesajların anlamına göre etiketlenmiş durumdadırlar.



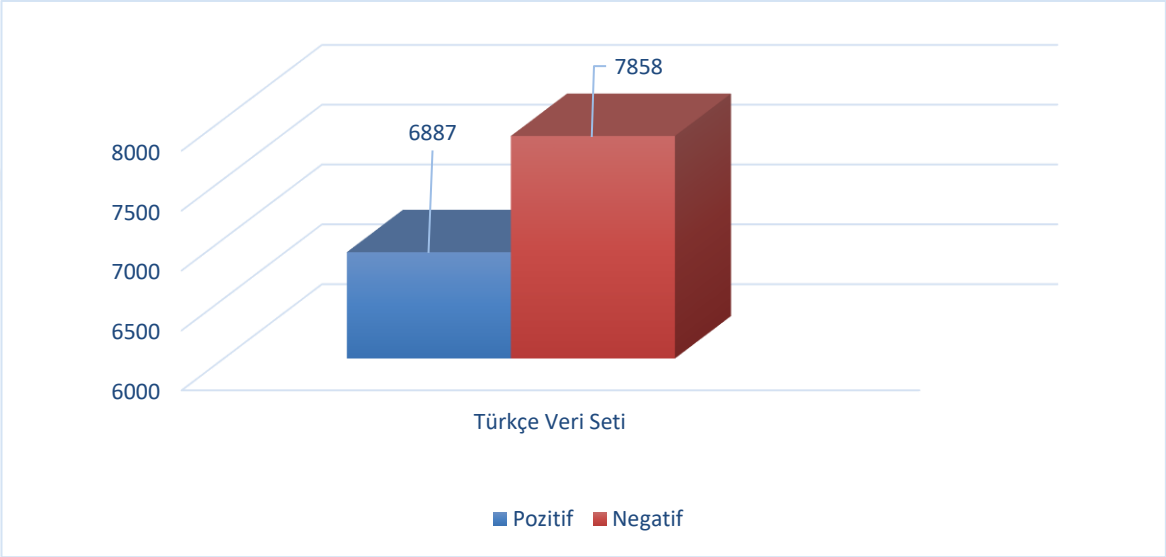
Şekil 3.3. Etiketli İngilizce Veri Seti Dağılımı

Çizelge 3.1. Etiketli İngilizce Veri Seti Örnekleri

619987808317407232	positive	A portion of book sales from our Harper Lee/Go Set a Watchman release party on Mon. 7/13 will support @CAP_Tulsa and the great work they do.
619971047195045888	negative	"If these runway renovations at the airport prevent me from seeing Taylor Swift on Monday, Bad Blood will have a new meaning."
619950566786113536	neutral	"Picturehouse's, Pink Floyd's, 'Roger Waters: The Walll - opening 29 Sept is now making waves. Watch the trailer on Rolling Stone - look..."

3.1.4. Türkçe Twitter Veri Seti

Twitter API ile his imgeleri (☺ ,☹ gibi mutlu veya üzgün ifadelerle oluşturulmuş) temel alınarak oluşturulmuş olan (Çoban vd. 2015) kaynağından aldığımız Şekil 3.4'de gösterilen pozitif (6887 adet) ve negatif (7858 adet) etiketli toplam 14657 adet Türkçe twitter mesajı), veri seti olarak kullanılmıştır.



Şekil 3.4. Etiketli Türkçe Veri Seti Dağılımı

Çizelge 3.2. Etiketli Türkçe Veri Seti Örnekleri

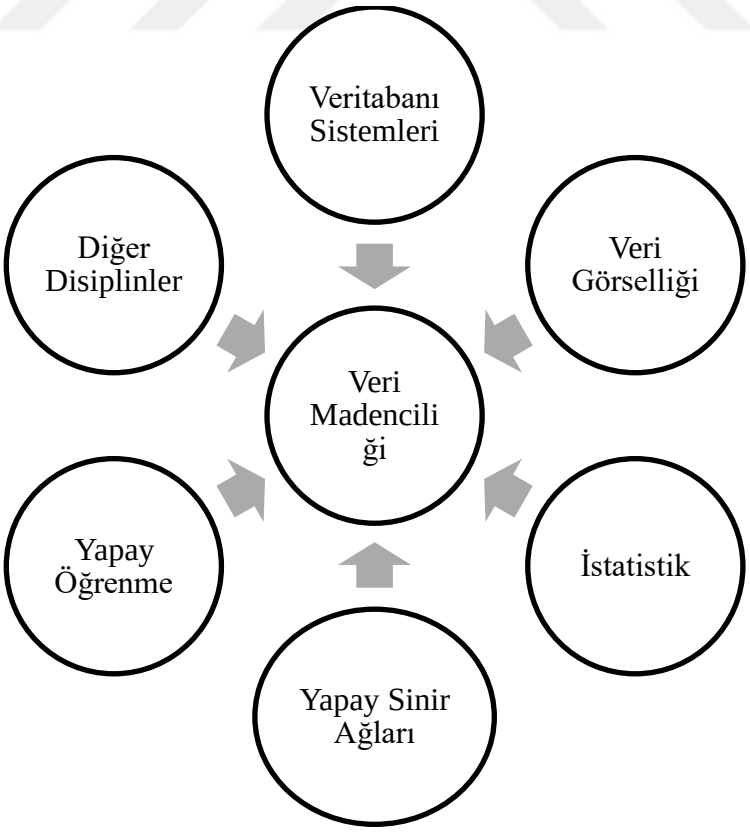
pozitif Böyle havalara böyle şeker gerek :) 0216 680 0365 #thelolishop #elyapimiseker #handmadecandy… http://t.co/EAmuOAVIK1
negatif camlarımıza veya balkonlarımıza lütfen ekmek ve su koyalım, dışarda gördüğümüz hayvanlara ekmek alalım :(http://t.co/1NgQG9xOXV

3.2. Yöntem

3.2.1. Veri Madenciliği

Veri madenciliği, büyük miktardaki verilerin içinden gelecek hakkında çıkarımlarda bulunarak elde edilen anlamlı kuralların bilgisayar programları aracılığıyla aranması ve analizinin yapılmasıdır (Savaş vd. 2012). Ayrıca veri madenciliği, tezimizde uyguladığımız Word2Vec modeli gibi veriler arasındaki ilişkileri inceleyerek gizli kalmış olan anlamsal ilişkilerin de çıkarılmasını sağlamaktadır. Bu işlemlerin uygulama alanı oldukça geniştir. Veri tabanı sistemleri, yapay sinir ağları, veri görselliği, yapay öğrenme, istatistik gibi çalışma alanları bulunmaktadır (Savaş vd. 2012).

Şekil 3.5'de diğer disiplinlerle ilişkisi verilen veri madenciliği, veri tabanlarında depolanan verilerden, bu veriler arasındaki ilişkileri analiz ederek hayata yönelik problemlerin çözülmesini hedeflemektedir (Savaş vd. 2012).

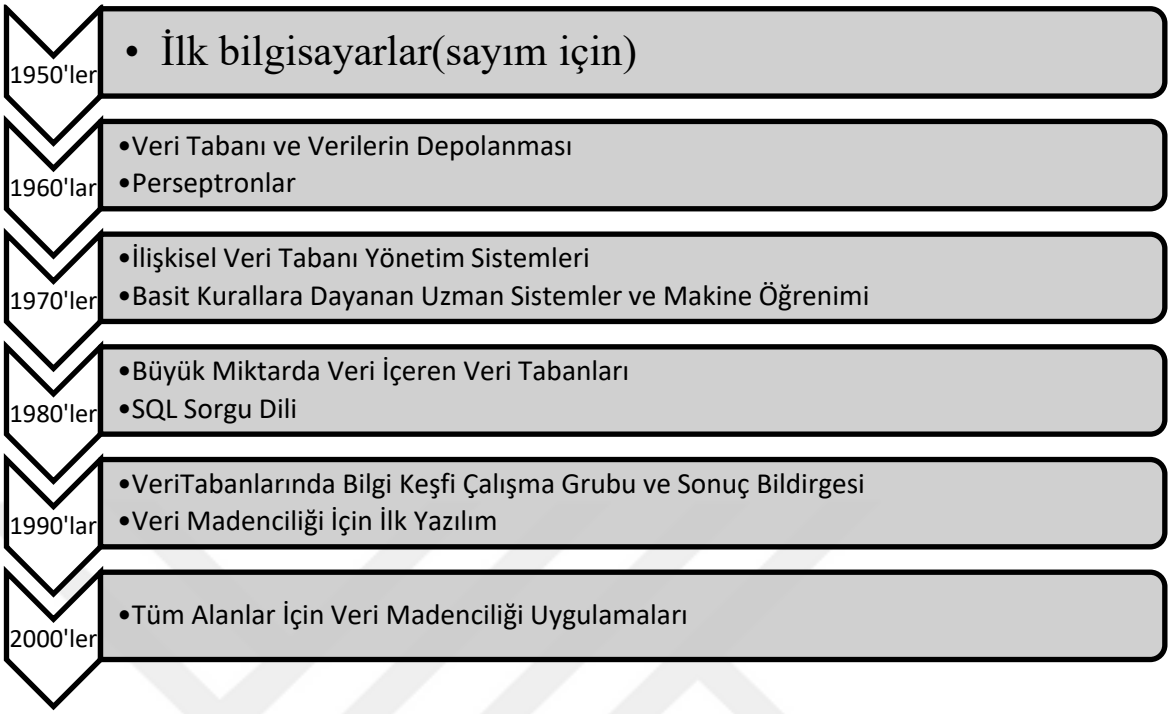


Şekil 3.5. Veri Madenciliği ve Disiplinler (Savaş vd. 2012)

3.2.1.1. Veri Madenciliği Gelişim Süreci

Günümüzde bilgisayarların hemen hemen tüm evlere girmesiyle ve internetin tüm dünyaya yayılmasıyla birlikte bilgi hızlı bir şekilde paylaşılır hale gelmiştir. Bu yüzden bilgilerin toplanması, saklanması ve yorumlanması için birçok uygulama ve donanım geliştirilmiştir. Tüm bu gelişim ve değişimler sonucunda geçmişten gelen bilgiler geleceğe aktarılabilir bir hale gelmiştir.

Gelişim süreci Şekil 3.6'da verilen veri madenciliği 1950'li yıllarda ilk bilgisayarlar ile birlikte hesaplamalar için kullanılmaya başlamıştır (Şimşek 2012). 1960'larda ise veri tabanı ve verilerin depolanması kavramı teknoloji dünyasında yerini almıştır. 1960'ların sonunda bilim insanları basit öğrenmeli bilgisayarlar geliştirebilmişlerdir. Minsky ve Papert, günümüzde sinir ağları olarak bilinen perseptron'ların sadece çok basit olan kuralları öğrenebileceğini göstermişlerdir (Adriaans ve Zantinge, 1997). 1970'lerde İlişkisel Veri Tabanı Yönetim Sistemleri uygulamaları kullanılmaya başlanmıştır. Ayrıca bilgisayar uzmanları basit kurallara dayalı uzman sistemler geliştirmişler ve basit anlamda makine öğreniminin gelişimine katkı sağlamışlardır. 1980'lerde ise veri tabanı yönetim sistemleri yaygınlaşmış ve mühendisliklerde, bilimsel alanlarda vb. alanlarda uygulanmaya başlanmıştır. Bu yıllarda şirketlerden, müşterilerden, rakiplerinden ve ürünlerinden elde edilen verilerden oluşan veri tabanları oluşturulmuştur. Bu veri tabanlarının içerisinde büyük miktarlarda veri bulunmaktadır ve bunlara SQL sorgulama dili ya da benzeri diller kullanarak ulaşılabilir. 1990'larda artık içindeki veri miktarı git gide artan veri tabanlarından, faydalı bilgilerin nasıl elde edilebileceği düşünülmeye başlanmıştır. Bu konu hakkında bir çok çalışma ve yayınlar yapılmıştır. 1989, KDD (IJCAI) 89 Veri Tabanlarında Bilgi Keşfi Çalışma Grubu toplantısı ve 1991, KDD (IJCAI)-89'un sonuç bildirgesi sayılabilecek Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop makalesinin KDD (Knowledge Discovery and Data Mining) ile ilgili temel tanım ve kavramları ileri atılması ile süreç daha da hızlanmış ve son olarak 1992 yılında veri madenciliği için ilk yazılım tasarlanmıştır. 2000'li yıllarda veri madenciliği sürekli gelişmiş ve pek çok alanda uygulanmaya başlanmıştır. Bu alana olan ilgi elde edilen faydalı sonuçlar görüldükçe artmıştır şeklinde açıklamıştır (Savaş vd. 2012).



Şekil 3.6. Veri Madenciliği Gelişim Süreci (Savaş vd. 2012)

3.2.1.2. Veri Madenciliği Kullanım Alanları

Veri madenciliğini büyük miktarda veriyi bünyesinde bulunduran her yerde kullanmak mümkündür. Karar verme sürecine ihtiyaç duyulan pek çok alanda veri madenciliği uygulamaları yaygın olarak kullanılmaktadır. Örneğin perakendecilik, borsa, pazarlama, biyoloji, telekomünikasyon, bankacılık, genetik, sigortacılık, sağlık, bilim, kriminoloji ve mühendislik, istihbarat, sağlık, endüstri vb. birçok alanda uygulamaları yapılmıştır(İnan 2003; Albayrak 2008; Akgöbek ve Çakır 2009).

3.2.1.3. Veri Madenciliğinde Karşılaşılan Problemler

Büyük hacimli verilerin depolanması ve işlenmesinde birçok problemle karşı karşıya kalınmaktadır. Veri madenciliği uygulamalarında karşılaşılan problemleri Savaş vd.(2012); “

- **Artık Veri:** Problemde elde edilecek olan sonuç için örnek kümede kullanılan gereksiz nitelikteki veridir.
- **Belirsizlik:** Kullanılan veri setindeki gürültülü verinin derecesidir.
- **Boş Veri:** Veri tabanındaki verilerde birincil anahtar değerinin kendisi de dahil herhangi bir değere eşit olmama durumudur.
- **Dinamik Veri:** İçeriği sürekli olarak değişen veridir.
- **Eksik Veri:** Veri kümesindeki bir değer için gerekli olan verinin verilmemiş olmasıdır.
- **Farklı Tipteki Veriler:** Sembolik, kategorik veri türleri ile kesirli sayılar, tamsayılar gibi farklı türdeki verilerdir.
- **Gürültülü ve Kayıp Değerler:** Sistem dışı hatalara gürültülü değerler denir. Veri girişinde eksik kalan değerler kayıp değerlerdir.
- **Veri Tabanı Boyutu:** Büyük boyutlu verilerin düşük seviyede çalışan sistemlerde kullanılma sorunudur.” şeklinde ifade etmişlerdir (Savaş vd. 2012).

3.2.1.4. Veri Madenciliği Süreçleri

Veri madenciliği sürecinde izlenen süreçler aşağıdaki gibi sıralanabilir (Shearer 2000):

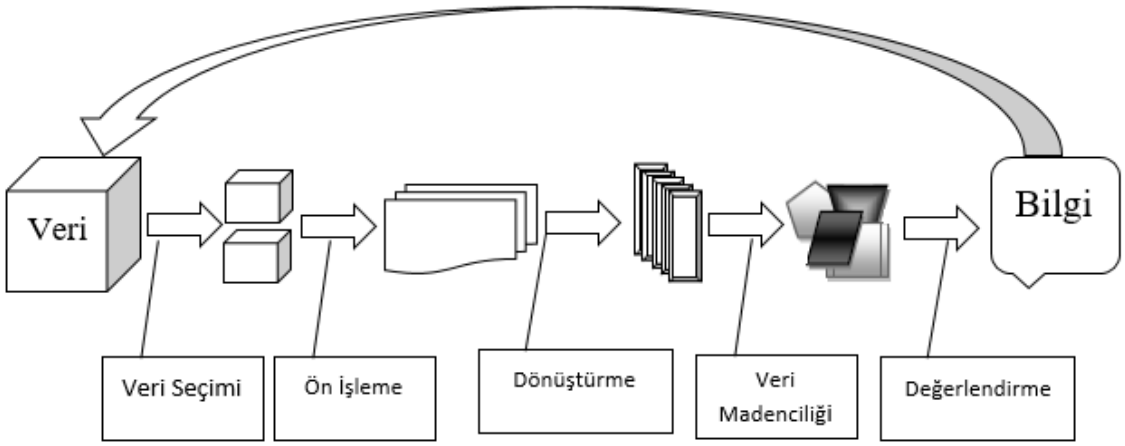
1. Problemin tanımlanması

2. Verilerin hazırlanması

3. Modelin kurulması ve değerlendirilmesi

4. Modelin kullanılması

5. Modelin izlenmesi



Şekil 3.7. Veri Madenciliğinde Bilgi Keşfi Süreci (Savaş vd. 2012)

Problemin tanımlanması: Veri madenciliği çalışmalarında yapılan çalışmanın hangi uygulama amacı için yapılacağının ve elde edilecek sonuçların başarı düzeylerinin nasıl değerlendirileceğinin belirlenmesidir.

Verilerin hazırlanması: Modelin oluşturulmasında oluşabilecek bir hata nedeniyle geri dönülmesi projenin uzamasına ve zaman kaybına neden olacaktır. Bu yüzden veriler doğru bir şekilde toplanır, değer biçilir, birleştirilip temizlenir ve dönüştürülür.

Modelin kurulması ve değerlendirilmesi: Tanımlanan probleme en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile sürekli yenilenen bir süreçtir.

Modelin kullanılması: Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulamada kullanılabileceği gibi bir başka uygulamanın alt parçası olarak da kullanılabilir.

Modelin izlenmesi: Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla elde ettikleri verilerde ortaya çıkan oluşan değişiklikler, kurulan modellerin devamlı olarak izlenmesini ve tekrardan düzenlenmesini gerektirecektir.

3.2.1.5. Veri Madenciliği Teknikleri

Veri madenciliği teknikleri genel olarak 3'e ayrılır (Akba 2014). Bunlar;

- Sınıflandırma
- Kümeleme
- Birliktelik Kuralları

Sınıflandırma: niteliklerin incelenmesi ile nesnenin önceden tanımlanmış bir sınıfa atamasıdır. Sınıf özelliklerinin iyi şekilde belirlenmesi gerekir. Sonuçlar önceden bilindiği için sınıflandırma denetimli öğrenme grubuna girer (Özçakır 2006). Sınıflandırma tekniğinde kullanılan başlıca algoritmalar:

- K-En Yakın Komşu,
- Karar Destek Vektörleri
- Yapay Sinir Ağları,
- Naïve-Bayes,
- Doğrusal Regresyon, Lojistik Regresyon,
- Karar Ağaçları

Kümeleme: Verilerin belli bir benzerlik kriterine göre gruplanması işlemine kümeleme denir (Şimşek 2012). Sınıflandırma algoritmalarına benzer olarak ortak özellikleri olan veriler aynı küme içerisinde yer alır. Çeşitli kümeleme algoritmaları ile alt kümeler bulunmaya çalışılır. Kümeleme algoritmaları olarak k- ortalamalar veya Kohonen şebekesi gibi istatiksel yöntemler kullanılmaktadır. Kümeleme modelinde, sınıfları bulunmayan veriler benzerlik-yakınlık kriterlerine göre gruplar halinde kümelere ayrılırlar. Küme içindeki elemanların benzerliği yüksek olmalı gerekirken kümeler arasında ise benzerliğin az olması gerekir (Şimşek 2012).

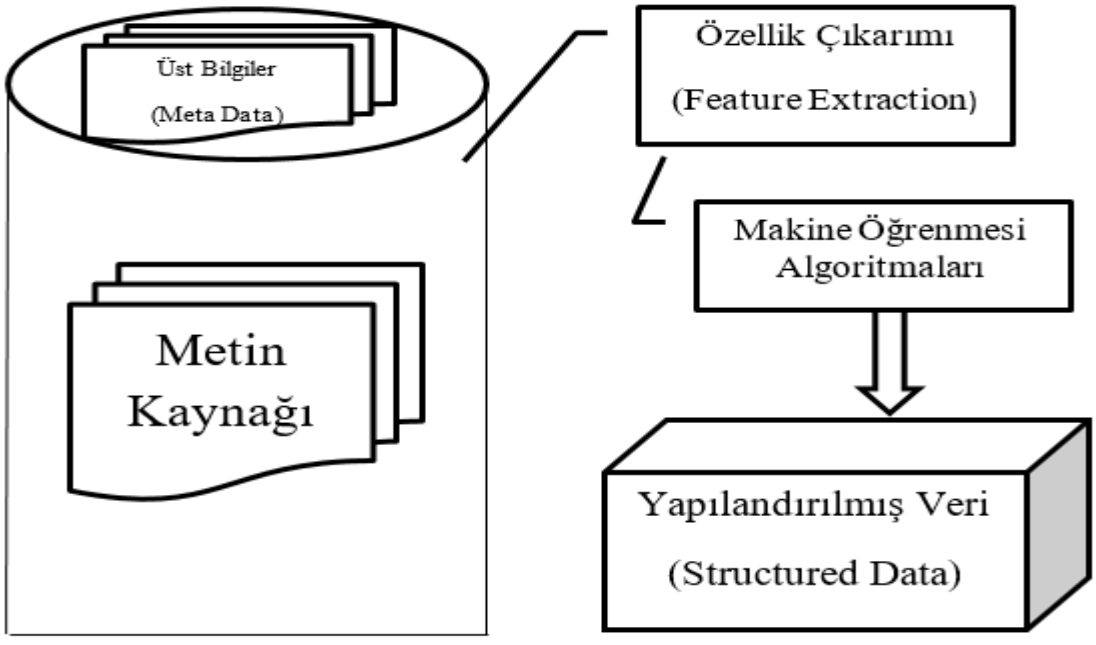
Birliktelik Kuralları: Veri sınıflama yöntemlerinden bir diğeri olan gözetimli olarak, yani sınıfların sayısı ve türleri bilinen veriler üzerinde yapılan sınıflamalar için kullanılan yöntemdir (Akba 2014). Market sektöründeki veriler üzerinde daha çok kullanılan bu yöntem aracılığıyla satın alınan yöntemler arasında ilişki kurarak daha kazançlı alışverişin nasıl yapılacağını öğretir.

3.2.2. Metin Madenciliği

Metin madenciliği çalışmaları metni veri kaynağı olarak kabul eden veri madenciliği (data mining) çalışması alanıdır ve diğer bir değişle metin üzerinden yapısallaştırılmış (structured) veri elde etmeyi amaçlar. Örneğin metinlerin sınıflandırılması, kümelenmesi, metinlerden konu çıkarılması, duygusal analizi, metin özetleme, varlık ilişki modellemesi gibi çalışmaları hedefler.

Metin madenciliği yöntemlerinin temelinde matematiksel ve istatiksel yöntemler bulunmaktadır. Metin madenciliği, yazar tanıma, metin sınıflama, fikir madenciliği, duygu analizi, anahtar kelime çıkartımı, başlık çıkartımı gibi farklı alanlarda da kullanılmaktadır (Kılınç vd. 2016).

Metin madenciliğinin metin tabanlı diğer bir çalışma alanı ise doğal dil işlemedir(DDİ). Doğal dil işleme çalışmaları daha çok yapay zekâ altındaki dil bilim bilgisine dayalı çalışmalarını kapsamaktadır. Metin madenciliği çalışmaları ise daha çok istatistiksel olarak metin üzerinden sonuçlara ulaşmayı hedefler. Metin madenciliği çalışmaları sırasında çoğu zaman doğal dil işleme kullanılarak özellik çıkarımı da yapılmaktadır (Şeker 2014).

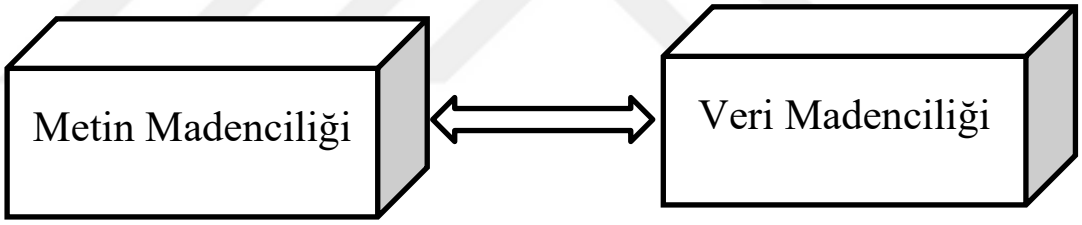


Şekil 3.8. Metin Madenciliğinde Veri İşleme Süreçleri (Şeker 2014)

Şekil 3.8. ‘de gösterildiği gibi, veri tabanından çekilen metinler üzerinden öznitelik çıkarımı yapıldıktan sonra sınıflandırma, kümeleme ve tahmin etme gibi makine öğrenmesi algoritmaları uygulanarak yapılandırılmış veri elde edilmiş olur.

Metin madenciliğinde doğal dil metinlerinden bilgi edinimi iki aşamada gerçekleşir; 1) Metinlerden anahtar ifadelerin elde edilmesi, 2)Elde edilen ifadelerin ilişkili olduğu kategorilere ayrılması. Metin madenciliği uygulamalarını ise şu şekilde listeleyebiliriz;

1. **Metnin Anlaşılması:** Örneğin, bir firma tarafından kullanıcıların şikayetleri baz alınarak şikayet metninin içerdiği anahtar içerik anlaşılarak sorun çözülebilecektir (Dolgun vd. 2009).
2. **Metin ile modelleme:** Metin madenciliğinde kullanıcı davranışları tahmin edilir. Metinden elde edilen içerik girdi değişkeni olarak kullanılır ve diğer bilgiler ile beraber tahmini model geliştirilir (Dolgun vd. 2009).



Şekil 3.9. Metin Madenciliğinde Süreçler Arasındaki İlişki (Dolgun vd. 2009)

Metin ve veri madenciliği arasında şekil 3.9.’de gösterildiği gibi interaktif bir ilişki vardır. Metin madenciliği süreçleri sonunda elde edilen yapısal veri, veri madenciliğinde kullanılarak o verinin yapısı hakkında incelemeler yapılmasını sağlar.

3.2.2.1. Metin Madenciliği Uygulama Alanları

Metin madenciliği uygulama alanlarını Dolgun (2009) aşağıdaki maddeler şekilde ifade etmektedir; “

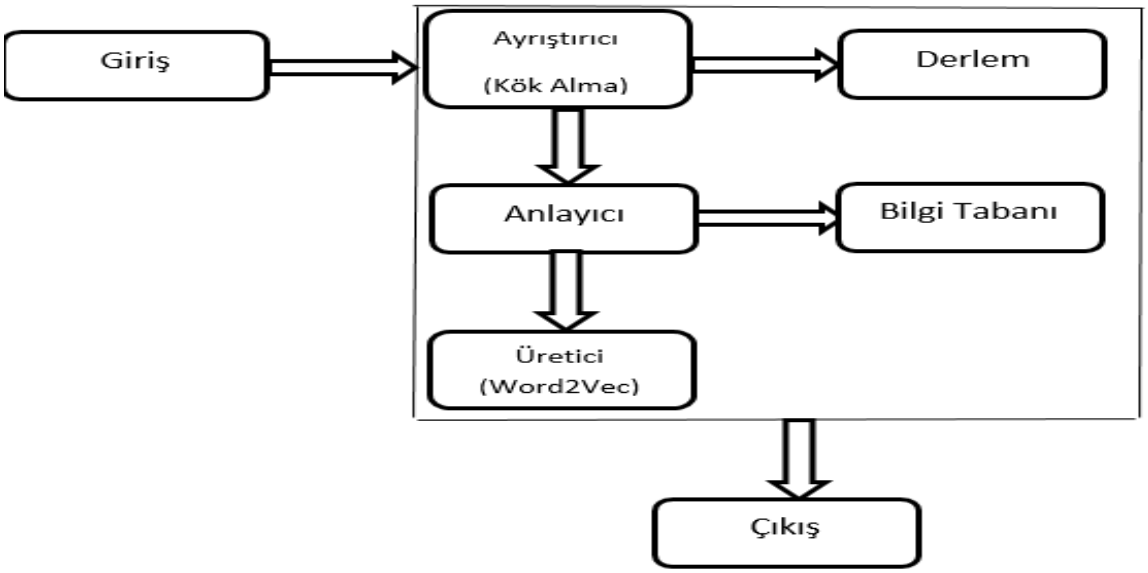
- **Müşteri İlişkileri Yönetimi** (Customer Relationship Management, CRM): Müşterilerin geçmiş zamana ait mesajlarından, çağrılarından ve anket verilerinden o müşterilerin bilgilerini elde ederek değerli bilgileri çıkartıp müşterilerin bu bilgilerine uygun satış tahmininde bulunulur.
- **Bilimsel ve Medikal Araştırmalar:** Sağlık sektöründeki metne dayalı hasta raporlarından, reçetelerden, araştırma sonuçlarından bilgiler elde ederek sağlık sektörünün geleceği hakkında çalışmalar yapılmaktadır.
- **Sahtekârlık (Fraud) Tespiti**: Kamu kurumlarında büyük çaptaki metin verilerinden anormallikler tespit edilerek sahtekârlıkların tespit edilmektedir.
- **Güvenlik/İstihbarat:** Askeri kurumlarda ve istihbarat kurumlarında kullanılmak üzere metinlerden kişiler arasındaki gizli bilgiler elde edilmektedir.
- **Pazar Araştırması:** Ekonomide pazarın etkisini ölçmek için basın bültenleri ve internet adreslerindeki metin belgeleriyle pazarın geleceği hakkında tahminler yaparak ticaret sektörüne öngörülerde bulunulmaktadır.” (Dolgun 2009)

3.2.3. Doğal Dil İşleme

Doğal dil işleme, asıl amacı, doğal bir dili çözümleme, anlama, yorumlama ve üretme olan bilgisayar sistemlerinin tasarımını ve gerçekleştirilmesini konu alan bir bilim ve mühendislik alanıdır. DDİ, yapay zekâ, biçimsel diller kuramı, kuramsal dilbilim ve bilgisayar destekli dilbilim gibi çok değişik alanlarda geliştirilmiş kuram, yöntem ve teknolojileri bir araya getirir. 1960’lı yıllarda yapay zekânın bir alt alanı olarak görülen bu konu, araştırmacıların ve gerçekleştirilen uygulamaların elde ettiği başarılar sonucunda artık bilgisayar bilimlerinin konusu olarak kabul edilmektedir (Delibas 2008).

DDİ, insan doğal dilini anlamak ve analiz etmek için kullanılan bilgisayar biliminin araştırma alanıdır. DDİ görevinin insan benzeri dil işlemesi olduğu ve yapay zekânın uygulama alanı içinde olduğu söylenebilir. Bir DDİ sistemi metni başka bir dile çevirebilir veya sorunun içeriğini işleyerek veya verilen metni açıklayarak soruları cevaplayabilir. Elbette, bir sistemin ulaşmaya çalıştığı birçok hedef vardır. DDİ'nin ana adımları morfolojik analiz, sözdizimsel analiz, semantik analiz ve söylem analizidir. Morfolojik analiz, her sözcüğün anlamını, soneklerini ve öneklerini, yani sözcüğün yapısını analiz ederek analiz eder. Sözdizimsel analiz, dilbilgisel yapıyı ve sözcüklerin birbirleriyle ilişkilerini analiz eder. Semantik analizde, sözdizimsel analiz ile oluşturulan yapılar anlamlandırılır. Söylem analizi, bir sonraki cümlenin anlamını etkileyen önceki cümleleri ele alır (Boynukalin 2012).

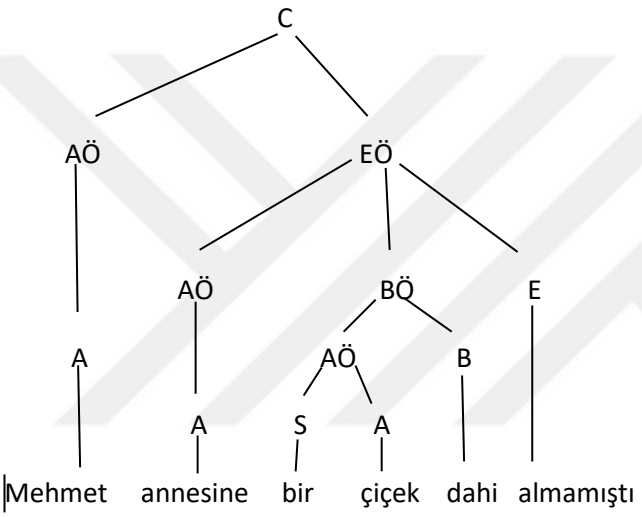
Şekil 3.10'daki gibi doğal dil işleme ayrıştırıcı(parser), sözlük(çalışmamızda derlem), anlayıcı, bilgi tabanı ve üretici olmak üzere 5 farklı temel eleman içerir.



Şekil 3.10. Doğal Dil İşleme Sistemlerinin Genel Blok Diyagramı (Delibas 2008)

Doğal dil işleme ayrıştırıcı(parser), sözlük(lexicon), anlayıcı, bilgi tabanı ve üretici olmak üzere 5 farklı temel eleman içerir. Bunların işleyişi şu şekildedir:

Ayrıştırıcı; Şekil 3.11'de örnek verildiği gibi, sözdizimsel olarak cümleyi analiz edip ayrıştırıcı ağacını oluşturur. Ayrıştırma alanında en yaygın tanınan yaklaşımlardan biri, öbek yapısal gramerlerdir. Bu yaklaşım Chomsky'nin üretimsel dönüşümlü dilbilgisi kuramına dayanır. Tümceleri öbeklere bölerek öbeklemeyi hedeflemektedir (Delibas 2008).



C: Tümce, AÖ: Ad Öbeği, EÖ: Eylem Öbeği, BÖ: Belirteç Öbeği,

S: Sıfat, A: Ad, E: Eylem, B:Belirteç

Şekil 3.11. Ayrıştırma Ağacı Örneği (Delibas 2008)

Sözlük, program tarafından tanınan tüm sözcükleri içeren yapıya denir. Ayrıştırıcı, sözlük ile sözdizimsel analiz yaparak çalışır. Sözlük, doğal dil işleme süreçlerinde program tarafından tanınması gereken kök ve anlamları bünyesinde bulundurur.

Anlayıcı bilgi tabanı ile birlikte cümlenin ne anlama geldiğini tespit etmeye çalışır.

Bilgi tabanı kavramsal olarak genel bilgi tabanı ve görev bağımlı bilgi tabanı olmak üzere iki alt ögeden oluşur. Anlayıcının temel görevi oluşturulan ayrıştırıcı ağacının bilgi tabanındaki karşılığını bulmaktır. Anlayıcı girilen cümleye uygun cevabı hazırlar. Doğal dil isleme alanında kullanılan en temel *üretici* sistem, belli kelime ve cümleler için depolanmış belli kalıpların kullanıcıya gösterilmesidir (Delibaş 2008).

3.2.3.1. Doğal Dil İşlemenin İlgi Alanları

Adalı(2012) Doğal dil işlemenin ilgi alanlarını:"

• Doğal bir dilin yazımında kullanılmak üzere yardımcı araçların geliştirilmesi

• Kullanıcı tarafından yapılan yazım hatalarının tespit edilip düzeltilmesi

• Metin içinde aranan ifadenin bulunması ve değiştirilmesi

• Optik okuyucu aracılığıyla basılı bir metnin okunması ve hatalarının düzeltilmesi

• Doğal dil işleme teknikleriyle verilen bir metnin otomatik olarak özetinin çıkarılması

• Metnin içinde gizli kalan ve kullanıcı için anlamlı olan bilginin çıkarılması

• Elde edilen bilgiye erişimin sağlanması

• Metnin okunup düzeltmeleri yapıldıktan sonra metnin anlaşılmasının sağlanması

• İnsan-bilgisayar etkileşimi için bilgisayarla konuşup bunun metne dönüştürülmesi

• Bilgisayara verilen herhangi bir metnin bilgisayar tarafından kullanıcıya okunması

• Sağır insanlarla iletişime geçmek için konuşulan kelimelerin metne dökülmesi

• Karşılıklı soru cevap yapılarının oluşturulması

• Bireyin gelişimine katkıda bulunmak için yabancı dil eğitim araçlarının geliştirilmesi

• Yabancı dil eğitiminde kolaylık sağlaması için yazma araçlarının geliştirilmesi

• Bir doğal dilden başka bir doğal dile çeviri sisteminin geliştirilmesi" şeklinde ifade etmiştir (Adalı 2012)

Doğal dil işlemenin diğer uygulama alanlarından bazıları diğer çalışmalarda şu şekilde ifade edilmiştir;

• Doğal dil işlemenin kullanım alanları çeviri, dilbilgisi analizi, veri tabanı yönetimi, belge yönetimi, konuşma tanımadır (Delibaş 2008).

• Doğal dil işlemenin kullanım alanlarından bazıları; makine çevirisi, otomatik metin çevri sistemi, bilgi kazanımı, internette metin tabanlı arama yapma, kelime işleme (gramere uygunluk açısından), Siri gibi yapay zekâ botları geliştirme, sınıflandırma ve kümele gibi işlemler ile metin işleme, verilen metnin özetinin çıkarılması (Şeker 2015).

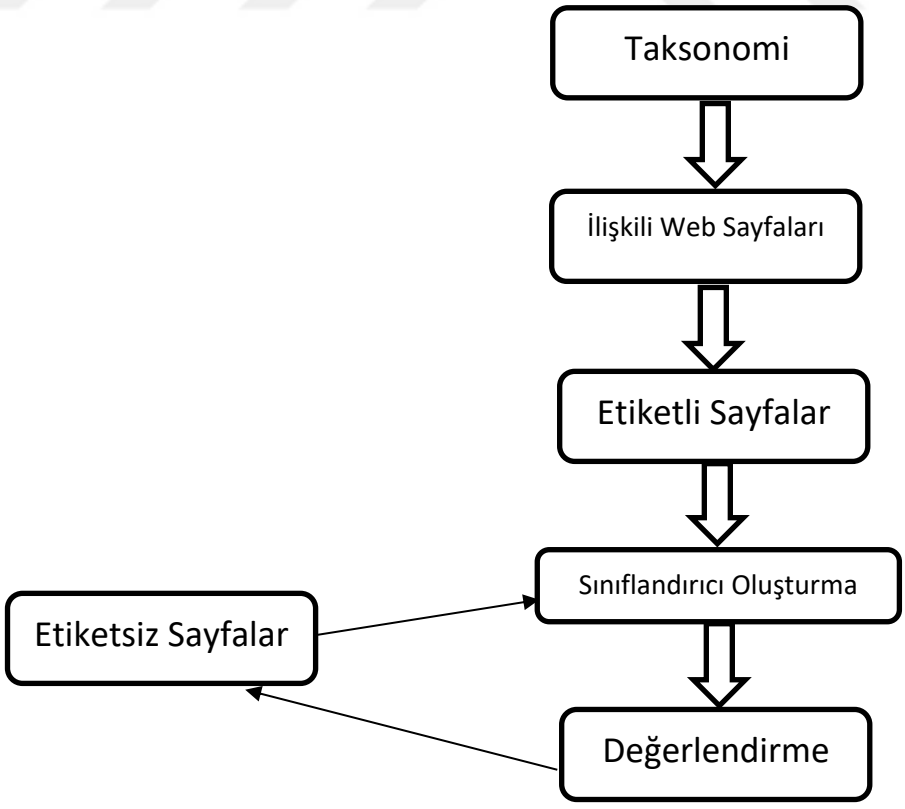
3.2.3.2. Doğal Dil İşlemenin Kısa Tarihçesi

DDİ çalışmaları 1940'lara dayanmaktadır. DDİ'nin ilk uygulaması, 2. Dünya Savaşı sırasında düşman kodlarını kırmak için geliştirilmiş bir Makine Çevirisi uygulamasıydı. 1950 yılında Alan Turing şimdi Turing Test (Turing 1950) olarak adlandırılan bir zekâ ölçütü önerdi. Ölçüt, bir bilgisayarın bir insanı bir yargıçla konuşmada taklit etme yeteneğidir. Dolayısıyla bir bilgisayarın zekâsının değerlendirilmesi DDİ teknikleri kullanılarak yapılabilir. 1957'de Noam Chomsky, evrensel bir dilbilgisi olduğunu ve bütün dillerin bu dilbilgisinin özelliklerine sahip olduğunu öne süren bir fikir (Chomsky 1957) ortaya koydu. 1966 yılında Otomatik Dil İşleme Danışmanı Komitesi (ALPAC) on yıl süren araştırmaların beklentileri karşılayamadığını gösteren bir rapor yayınladı. Rapordan sonra, DDİ ile ilgili araştırmalar uluslararası alanda çok fazla azaldı. 1970 yılında kurulan SHRDLU (Winogard 1971) yazılımı, bu durgunluktan sonraki başarılı çalışmalardan biridir. Kullanıcının taleplerini grafik bir blok dünyasında gerçekleştiren bir yazılımdır. Roger Schank ve meslektaşlarının bazıları insan kavramsal bilgisini inceledi. Albayrak (2011) tarafından bildirildiğine göre bunlardan biri, bilgi sistemleri teorisi olan Robert P. Abelson ile yapılan çalışmadır (Schank ve Albelson 1977) .

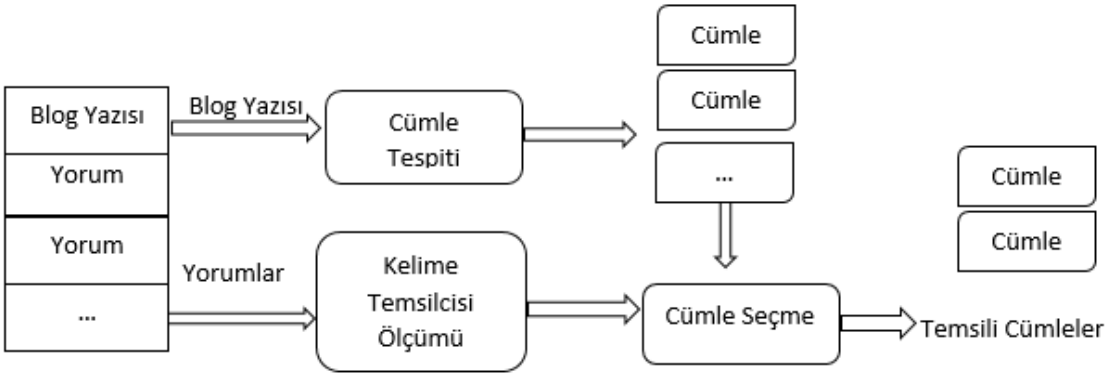
1980'lerin sonlarından itibaren doğal dil işlemeye makine öğrenmesi yaklaşımları uygulanmaya başlanılarak çalışma alanı genişletildi. 1990'lı yıllarda ise bilgisayar bilim ve teknolojisinin ilerlemesiyle DDİ yükselişe geçmiştir (Albayrak 2011).

3.2.3.3. En Gelişmiş Sistem

DDİ, birçok alanla ilgili şubelere sahip geniş bir alandır. Şekil 3.12-13'de örnek verildiği gibi, bugün DDİ tekniklerinin katkılarıyla pek çok çalışma sürdürülmektedir. Albayrak (2011) tarafından bildirildiğine göre bazı sanat örneklerini DDİ'nin yakın tarihli çalışmalarından biri olan görüşlerin özetlenmesinde düşük kaliteli ürün incelemelerinin tespit edilmesi sorununu ele almaktadır (Liu vd. 2007). Çalışma, bir sınıflandırma yöntemi kullanarak ürün gözden geçirmelerinin bilgi, okunabilirlik ve öznelliğini araştırmaktadır. Yaklaşım fikir madenciliği görevine uygulanır ve iki aşamalı bir çerçeve oluşturulur. Sınıflandırma yöntemini kullanan başka bir çalışma (Jin vd 2007), içerik reklamcılığı için web sayfalarının hassas sınıflandırmasını inceler. Araştırma, reklam sektörü için çok değerli olabilir. Bir yorumda bahsedilen bir şey, blog yazısı yorumlarında temsili cümleleri dikkate alarak blog özetlemesi yapılır (Hu vd 2007). Uygulamanın özellikleri, yorumlardan, yorum sıklıklarından ve terim sıklıklarından en az birinde bir sözcüğün görünümüdür (Albayrak 2011).



Şekil 3.12. İçerik Reklamcılığı İçin Web Sayfası Sınıflandırma Metodolojisi (Jin vd 2007)



Şekil 3.13. Cümle Çıkarımı İle Yorum Odaklı Blog Özetleme Metodolojisi (Hu vd 2007)

3.2.4. Duygu Analizi

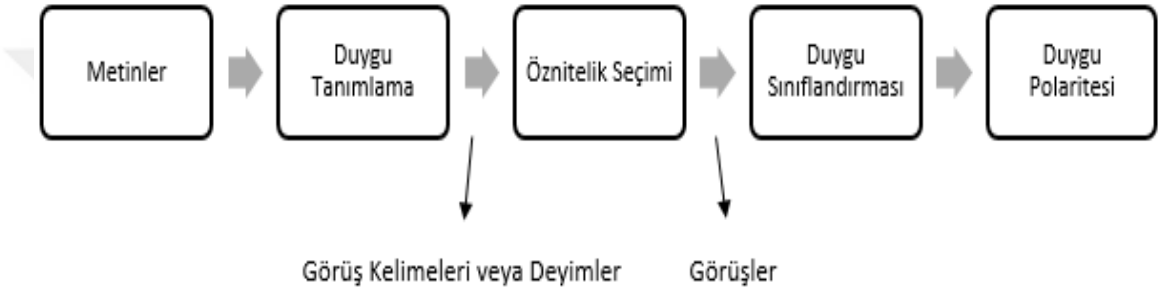
Doğal diller ile yazılmış olumlu, nötr, olumsuz vb. duygu ifadesi belirten durumların incelenmesi işlemine Duygu Analizi denir. Doğal dil işleme tekniklerinin ve makine öğrenmesi algoritmalarının uygulandığı duygu analizi çalışmalarında yazılan metinlerin dil bilgisi kurallarına uyması ve ayrıca anlam bütünlüğünü sağlayacak yapıda olması gerekmektedir.

İnsanlar gündelik hayatlarında konuşurlarken ve metinleri yazarken öznel ve nesnel olan iki farklı ifadeyi kullanırlar. Nesnel ifadeler genel geçer gerçekleri temsil ederken, öznel ifadeler görüşleri temsil ederler. Görüş anlamsal olarak çok kapsamlı olmasının yanı sıra olay, kişi ve bunların özellikleri hakkında olumlu, olumsuz ve nötr ifadeleri içermektedir (Thelwall vd. 2010).

Duygu analizi, kişilerin olaylar, hizmetler, ürünler, kurumlar, başka kişiler v.s. hakkındaki duygu ve düşüncelerini tespite yarar. Olumlu geri bildirimler firmayı teşvik ederken, olumsuz geri bildirimler caydırıcı görevler üstlenirler. Bazen firmalar kendileri ile diğer firmaları karşılaştırmak istediklerinde de duygu analizinden yardım alırlar. Bir twitter mesajında geçecek karşılaştırma iki firmanın birbirine göre durumunu bize verecektir (Takcı 2013).

İnternetin tüm dünyaya yayılmasıyla ve teknolojik ürünlerin gelişip mobil cihazları üretmesiyle birlikte, sosyal medya platformları duygu analizi çalışmalarında çok önemli kaynaklar haline gelmişlerdir. Bilimsel çalışmaların çoğu facebook gibi fonksiyonel detayı fazla olan sosyal medya araçları yerine Twitter gibi metin içeriğinin çok olduğu araçlara yoğunlaşmıştır (Ortigosa vd. 2013).

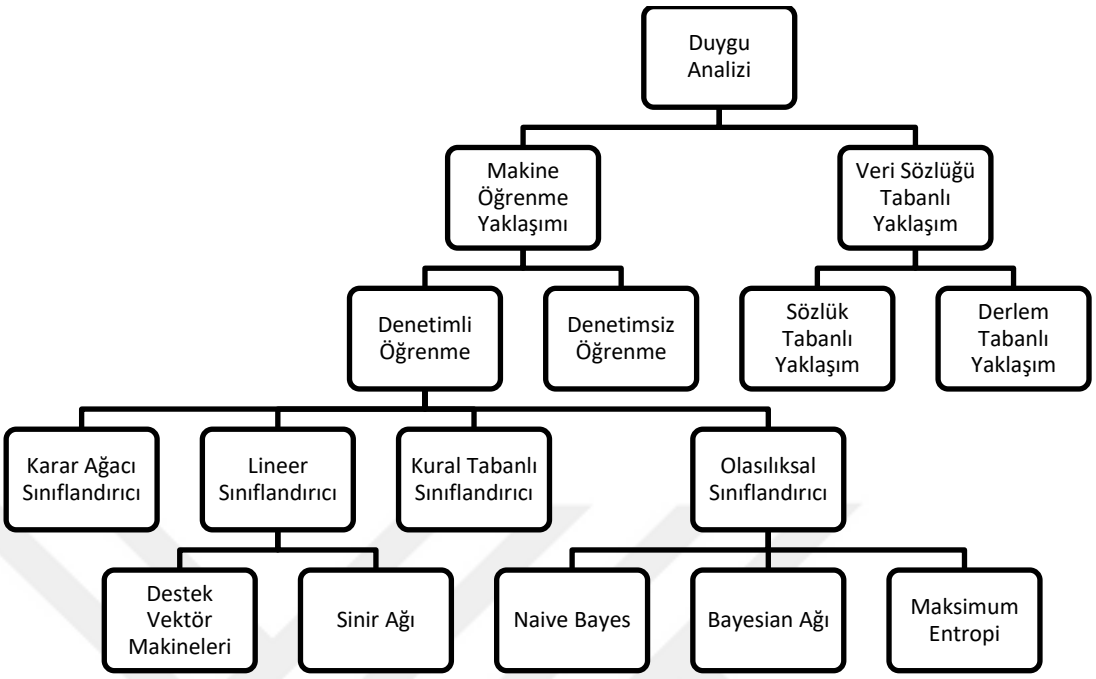
Şekil 3.14'de süreçleri verilen duygu analizi, veri madenciliği, makine öğrenmesii doğal dil işleme vb bilgisayar ve istatistik bilim alanlarında gelişmekte olan araştırma alanıdır.



Şekil 3.14. Duygu Analizi Süreci (Can ve Alatas 2017)

Duygu analizi çalışmalarında Şekil 3.15'de verildiği gibi iki farklı yaklaşım kullanılmaktadır. Bunlar makine öğrenmesi teknikleri ve sözlük tabanlı yaklaşım. Makine öğrenmesi denetimli ve denetimsiz öğrenmeye sahiptir. Sözlük temelli yaklaşımlar ise duygu sözlüğü temelli ve bütünce temelli olarak iki kısımda incelenir (İskender 2016).

Duygu analizi çalışmaları kullanılan çalışma alanına göre doküman seviyesinde, cümle seviyesinde, deyim seviyesinde, duygu seviyesinde sınıflandırılmaktadır (Yelmen 2016).



Şekil 3.15. Duygu Analizi Sınıflandırma Teknikleri (Can ve Alatas 2017)

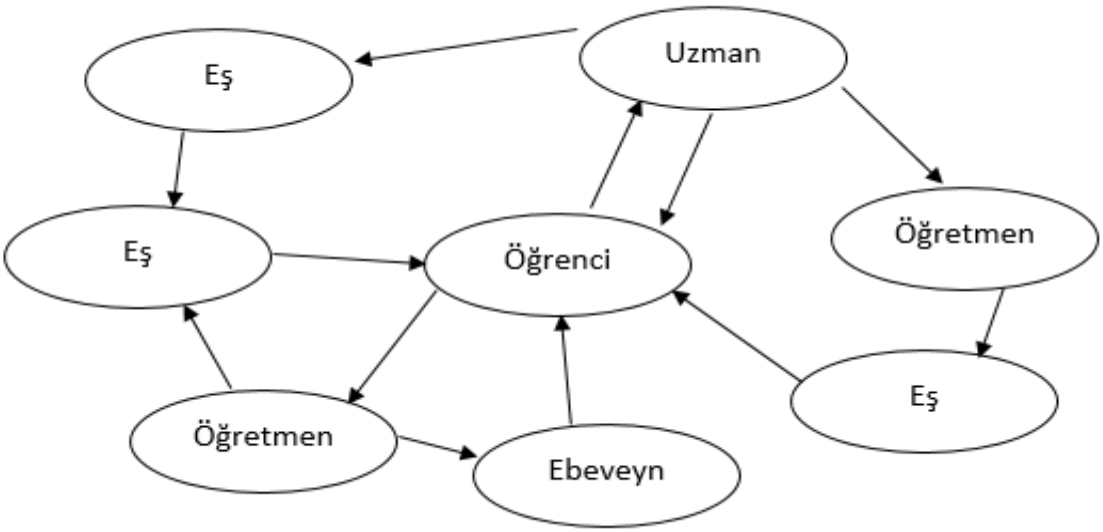
3.2.5. Sosyal Medya

Sosyal medya, internet araçlarını kullanarak insanların bilgi ve deneyimlerini diğer insanlarla daha verimli şekilde paylaşmayı ve tartışmayı sağlayan bir mecra olarak da tanımlanabilir (Demner-Fushman vd. 20058). Sosyal etkileşim için erişilebilir ve ölçeklenebilir bir ortam olan sosyal medya iletişim teknikleri kullanılarak sosyal etkileşimi sağlayan bir platformdur.

Sosyal medyaya erişim son yıllarda internetin yaygınlaşması ve geliştirilen uygulamalar sayesinde büyük ölçüde internet üzerinden gerçekleştirilmeye başlanmıştır (Şimşek 2012). Geliştirilen bu uygulamalar insanlara sadece iletişim ve bilgi paylaşımı olanağı değil eğlence ve iyi vakit geçirme imkânı da sunmaktadırlar. Genel olarak "sosyal ağlar" olarak tanımlanabilecek bu uygulamalar kişiler ve geniş kitleler hakkında büyük miktardaki verilere Internet üzerinden kolay bir şekilde erişim imkânı vermektedir şeklinde ifade etmiştir (Şimşek 2012).

Sosyal medyanın geleneksel medyadan farkı, sosyal medyada içerik çok önemlidir. İlgi çekici içerikler, kampanyalar, reklamlar dakikalar içerisinde binlerce kullanıcıya ulaşabilmektedir. Bu sayede kurumların hedef kitleye ulaşımı kolaylaşmaktadır. İçerik paylaşım konusunda; sosyal medyanın geleneksel medyaya göre en büyük avantajı ise hedef kitleye ulaşım oranını rahatlıkla tespit edebilmesidir (Anonim). Sosyal medya ile geleneksel medyayı ayıran bir başka unsur ise maliyetlerdir. Sosyal medyada ayrılan küçük bütçeler ile hedef kitlenize ulaşım çok daha kolaydır. Geleneksel medyada Gazete, Tv, Dergi Reklamları düşünüldüğünde hedef kitleye ulaşım için harcanan maliyet daha fazladır (Anonim).

Sosyal medya, kullanım açısından karmaşık bir yapıda görünse de basit bir iletişim simetrisine sahiptir. Örneğin; Şekil 3.16'da bir öğrenci sosyal medya üzerinden bir öğretmen ve bir uzmanla iletişim kurarken, aynı anda farklı kişiler bu öğrenciyle yine sosyal medya üzerinden iletişime geçebilmektedirler. Bu durum sosyal medyanın kullanım kolaylığını da açıklamaktadır (Vural ve Bat 2010).



Şekil 3.16. Sosyal Medya Kullanımı İletişim Grafiği (Dawley 2009)

Köksal ve Özdemir (2013) tarafından bildirildiğine göre, sosyal ağlar, video ve fotoğraf paylaşım siteleri, mikro-bloglar, film ve müzik siteleri gibi sosyal içerikli web sitelerinin önemli ortak özellikleri aşağıdaki gibidir:

1. Kişisel Profil: Sosyal medya platformlarına üye olmak için kullanıcı kendi kişisel bilgilerini girer. Böylelikle kişilerin bilgilerinden kimlerin kendisine üye olduğunu tanıyabilmektedir.

2. Online Bağlantı Kurma: Kişisel bilgilerini girerek sosyal medya uygulamasına üye olan kişiye bağlantı kurabileceği kişileri önererek bunu önceden email aracılığıyla bildirmektedir.

3. Online Gruplara Katılma: Sosyal medya platformlarından Facebook, LinkedIn, Flickr ve MySpace gibi sitelerde kişiler tarafından gruplar oluşturularak başka kullanıcıların da bu gruplara dahil edilmesi sağlanmaktadır.

4. Online Bağlantılarla İletişim Kurma: Teknolojinin gelişmesiyle birlikte sosyal medya uygulamaları aracılığıyla mesaj atarak, sesli ve görüntülü konuşma sağlayarak kullanıcıların birbirleriyle iletişim kurmasını sağlamaktadır.

5. Kullanıcıların Oluşturduğu İçeriği Paylaşma: Facebook, twitter, instagram gibi sosyal medya platformlarında kullanıcılar resim, video, haber paylaşarak diğer kullanıcıların yorum yapmasına ve beğenisine sunmaktadır.

6. Fikir ve Yorumda Bulunma: Sosyal içerikli sitelerde kullanıcıların oluşturduğu içeriği paylaşarak diğer kullanıcıların yorum yapmasına imkan sağlamaktadır.

7. Bilgi Edinme: Sosyal medya uygulamalarında kullanıcılar uygulama üzerinden kişi, yer, kurum gibi aramalarını yaparak bilgi edinmeyi amaçlamaktadırlar. Örneğin facebook'ta ve twitter'da kullanıcı hem çevrimiçi hem de çevrimdışı olmasına rağmen aradığı bilgiyi uygulama üzerinden bulanabilmekte ve LinkenIn ile aranan anahtar kelimeyle kişi, kurum, meslek veya iş hakkında bilgi verebilmektedir.

8. Kullanıcıları Sitede Tutma: Geliştirilen sosyal medya platformlarında kullanıcıları bu uygulamalarda uzun süreli tutabilmek için ve onların paylaşılan içeriklere karşı hızlı geri dönüş yapabilmeleri için birçok uygulama geliştirmişlerdir. Facebook' un pazarlama amaçlı kullanılabilecek "Market Place" uygulaması buna örnek gösterilebilir (Kim vd. 2010).

3.2.6. Makine Öğrenmesi

MÖ yaklaşımı, sözdizimsel ve dilbilimsel özelliklerin kullanılmasını yapan normal bir metin sınıflandırma problemi olan duygu analizini çözmek için kullanılan MÖ algoritmalarına dayanır (Can ve Alatas 2017).

Eğitim kayıtları seti D={ X1, X2, ..., Xn} her bir kayıt bir sınıfa etiketlenmiştir. Daha sonra bilinmeyen bir sınıf örneği için bu model bu sınıfa bir etiket tahmini yapmak için kullanılır. Zor bir sınıflandırma problemi bir örneğe sadece bir etiket atanmasıdır. Daha kolay bir sınıflandırma problemi ise bir örneğe etiketlerin olasılıksal bir değerinin atanmasıdır (Can ve Alatas 2017).

3.2.6.1. Denetimli Öğrenme

Denetimli öğrenmede oluşturulan model ile bir grup girdi değerine karşılık onlara ait hedef değerleri verilerek aralarındaki ilişkiyi öğrenmesi ve hedef değerlere en yakın çıktıların üretilmesi amaçlanır. Elde edilen en iyi model, yeni girdi değerleri için en yakın çıktıyı da verebilecektir (Atalay ve Çelik 2012). Geleceğe yönelik çıkarımlarda bulunmak için önceden verdiğimiz eğitim setlerine bakarak bunlardan eğitim kümesindekilerle benzeyenleri karşılaştırarak tahminde bulunur.

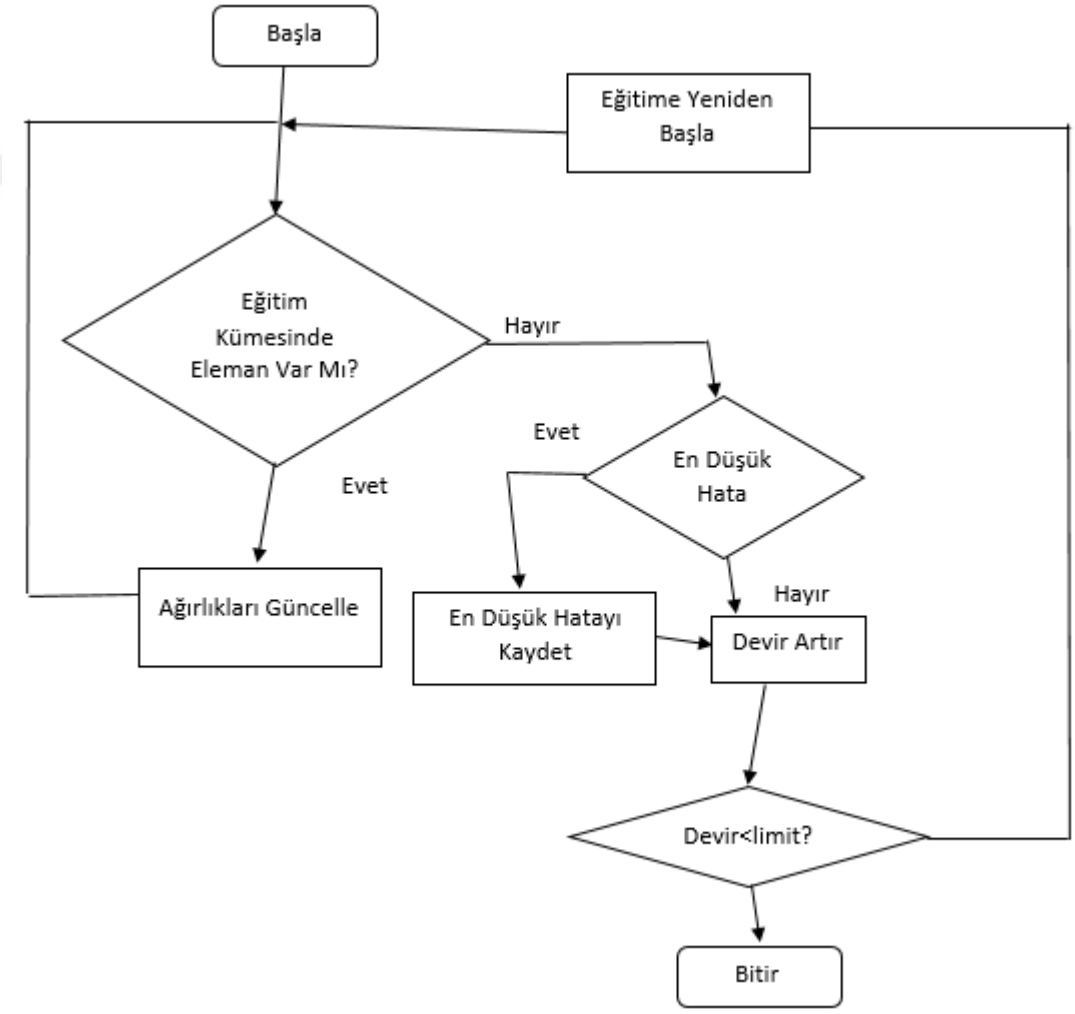
Denetimli öğrenme yönteminde şekil 3.17'de verildiği gibi makine eğitim kümesini ve yaptığı önceki yanlışları en düşük hatayı bulmak için kullanır. Burada amaç en düşük hata payı ile en isabetli tahmin yapabilmektir. Denetimli öğrenme sınıflandırma ve regresyon diye iki gruba ayrılır (Maini 2017).

Sınıflandırmada kullanılan sınıflandırma algoritmalarından şu şekildedir (Filiz 2017);

- K En Yakın Komşu Sınıflandırma Algoritması
- Mantıksal regresyon
- Ağaçlar
- Naive Bayes
- Destek Vektör Makineleri

Regresyonda kullanılan algoritmalar ise (Filiz 2017);

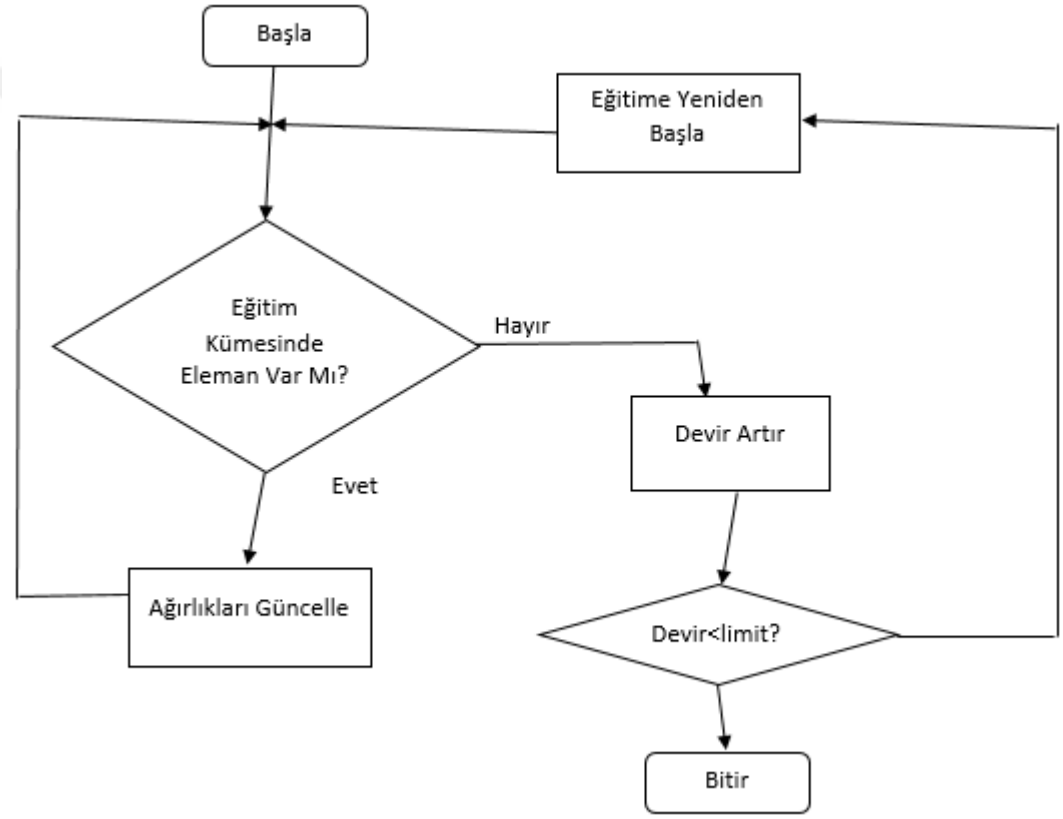
- Lineer Regresyon
- Polinomal Regresyon
- Karar Ağaçları
- Random Forest



Şekil 3.17. Denetimli Öğrenme Akış Şeması (Şeker 2008)

3.2.6.2. Denetimsiz Öğrenme

Elimizdeki verinin kaç sınıfa ayrıldığını, girdilerin hangi sonuçları ürettiğini bilmeden ham veriden bir anlam çıkarmaya denetimsiz öğrenme denir (Çınar 2017). Kümeleme işlemi örnek olarak verilebilir. Ayrıca denetimsiz öğrenme kendi içinde kümeleme ve ilişki analizi diye ikiye ayrılır. Şekil 3.18.'da görüldüğü gibi denetimli öğrenmedeki gibi hata payını minimuma düşürme işlemini gerçekleştirmez.



Şekil 3.18. Denetimsiz Öğrenme Akış Şeması (Şeker 2008)

3.2.6.3. Hata Matrisi ve Performans Metrikleri

Performans ölçümünde doğruluk (accuracy), anma (recall), duyarlılık (precision) ve F-ölçütü (F-measure) kullanılmaktadır. Model üzerindeki başarı ölçüldüğü sırada, başarı oranını doğru tespit edebilmek için öğrenme veri kümesini işleme almamak gerekir.

Öğrenme veri kümesi üzerinde model oluşturulduktan sonra var olan model eğitim kümesinde eğitildikten sonra test veri kümesinde sınanır (Kırmacı 2015). Bu sınama bizim çalışmamızda kullandığımız gibi aynı veri seti içinde de olabilir. Ayrıca çalışmamızda performans ölçüm değerleri olarak literatürde kullanılan değerler olan (Çoban vd. 2015) doğruluk(accuracy) ve anma(recall) değerlerini kullanılmıştır.

Hata Matrisi: Makine öğrenmesinde kullanılan sınıflandırma modellerinin performansını değerlendirmek için hedef niteliğe ait tahminlerin ve gerçek değerlerin karşılaştırıldığı hata matrisi sıklıkla kullanılmaktadır. Her ne olursa olsun sınıflandırma tahminleri Şekil 3.19'daki gibi dört değerlendirmeden birine sahip olacaktır (Şirin 2017):

1. Doğruya doğru demek (True Positive – TP) DOĞRU
2. Doğruya yanlış demek (True Negative – TN) YANLIŞ
3. Yanlışa doğru demek (False Positive – FP) YANLIŞ
4. Yanlışa yanlış demek (False Negative – FN) DOĞRU

	Tahmi Edilen Sınıf		
Asıl Sınıf		Sınıf = Evet	Sınıf = Hayır
	Sınıf=Evet	True Positive(TP) (10)	False Negative(FN) (10)
	Sınıf=Hayır	False Positive(FP) (20)	True Negative(TN) (100)

Şekil 3.19. Hata Matrisi (Joshi 2016)

TP: Doğru tahmin edilen pozitif değerlerdir; bu, gerçek sınıf değerinin evet olduğunu ve tahmin edilen sınıf değerinin de evet olduğu anlamına gelir. Örneğin gerçek sınıf değerinin bir yolcunun hayatta kaldığını ve tahmin edilen sınıfın da size aynı şeyi söylediğini göstermesidir (Joshi 2016).

TN: Doğru tahmin edilen negatif değerlerdir; bu, gerçek sınıf değerinin hayır olduğunu ve öngörülen sınıf değerinin de hayır olduğu anlamına gelir. Örneğin gerçek sınıfın bir yolcunun hayatta kalmadığını söylemesi ve tahmin edilen sınıf size aynı şeyi söylemesidir (Joshi 2016).

FP: Gerçek sınıfın hayır olduğu ve tahmin edilen sınıfın evet olduğu zaman geçerlidir. Örneğin gerçek sınıf bir yolcunun hayatta kalmadığını söylemesi ama tahmin edilen sınıf ınsize bu yolcunun hayatta kalacağını söylemesidir (Joshi 2016).

FN: Gerçek sınıf evet olduğu ama tahmin edilen sınıfın hayır olmasıdır. Örneğin gerçek sınıf değeri bir yolcunun hayatta kaldığını ve tahmin edilen sınıfın size yolcunun öleceğini söylemesidir (Joshi 2016).

Doğruluk(Accuracy): Doğruluk oranı model başarımı ölçümünde kullanıla en popüler yöntemlerden biridir. Formül 3.1'de gösterildiği gibi, hesaplama doğru sınıflandırılmış örnek sayısının (TP + TN), toplam örnek sayısına (TP + TN + FP + FN) bölünmesi ile yapılmaktadır (Yelmen 2016). Doğru sınıflandırmanın toplama bölümüdür. Yani; doğrular / toplam.

$$\text{Doğruluk} = (TP+TN)/(TP+FP+FN+TN) \qquad (3.1)$$

Doğruluk = (10+100) / (10+20+10+100) → 0,785

Anma(Recall): Formül 3.2'de gösterildiği gibi, doğru tahmin edilen pozitif gözlemlerin gerçek evet sınıfındaki tüm gözlemlere oranıdır. "Gerçekten hayatta kalan tüm yolculardan kaç tanesini etiketledik?" sorusuna cevap verir.

$$\text{Anma} = TP/(TP+FN) \qquad (3.2)$$

Anma = 10 / (10+10) → 0,5

Duyarlılık(Precision): Duyarlılık, formül 3.3'de gösterildiği gibi, doğru tahmin edilen pozitif gözlemlerin toplam tahmini pozitif gözlemlere oranıdır. Yüksek hassasiyet düşük FP oranı ile ilişkilidir (Joshi 2016).

Duyarlılık = TP/(TP+FP) Duyarlılık = 10 / (10+20) → 0,33 (3.3)

F1 Skoru: F1 skoru, formül 3.4'de gösterildiği gibi, Duyarlılık ve Anma değerlerinin ağırlıklı ortalamasıdır. Bu nedenle, bu puan hem FP'yi hem de FN'i dikkate alır. Sezgisel olarak doğruluk olarak anlaşılması kolay değildir, ancak özellikle eşit olmayan bir sınıf dağılımınız varsa F1 genellikle doğruluk metriğinden daha faydalıdır (Joshi 2016).

F1 Skor = 2*((Anma * Duyarlılık) / (Anma + Duyarlılık)) (3.4)

F1 Skor = 2*((0,5*0,33) / (0,5+0,33)) → 0,397

3.2.6.4. Kullanılan Makine Öğrenmesi Algoritmaları

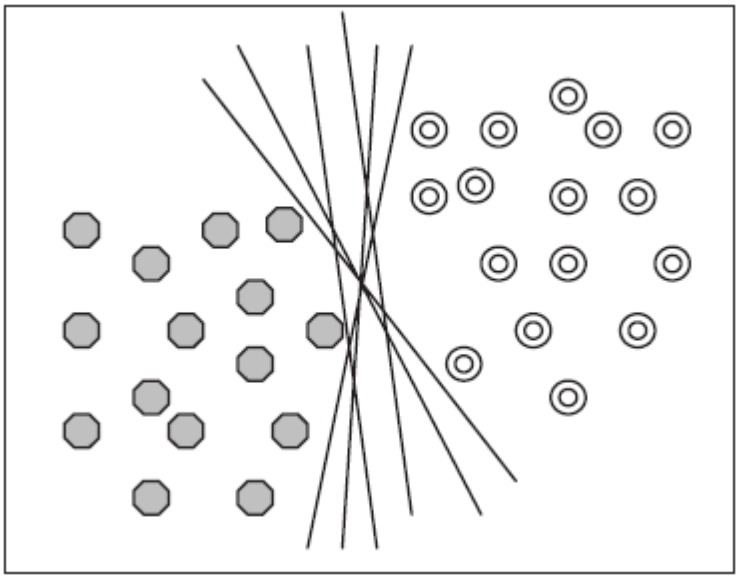
3.2.6.4.1. Destek Vektör Makineleri

Bir makine öğrenme algoritması olan destek vektör makineleri (DVM), sınıflandırma, kümeleme, yoğunluk tahmini ve regresyon kuralları üretmek için kullanılır. Lineer olmayan örnek uzayını, örneklerin lineer olarak ayrılabileceği yüksek boyuta aktarıp, farklı örnekler arasındaki maksimum sınırın bulunması esasına dayanır(Elmas 2012).

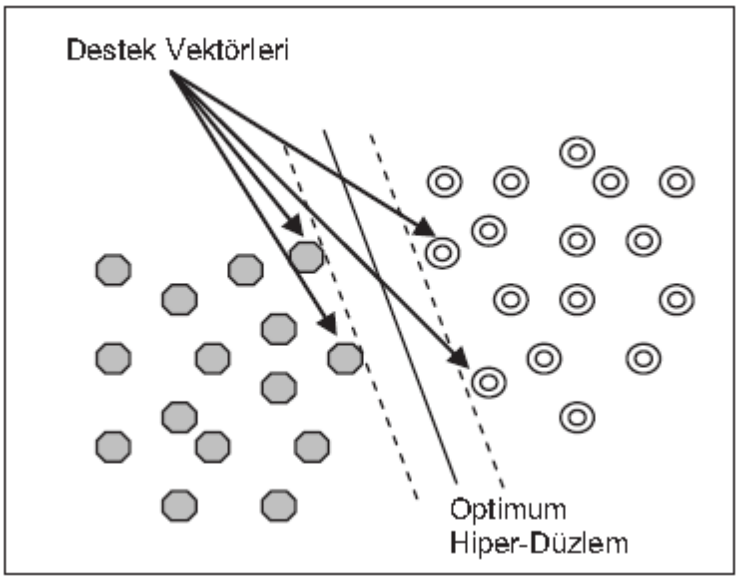
Doğrusal ayrılabilen DVM, Şekil 3.20'de gösterildiği gibi iki sınıflı verileri birbirinden ayırabilecek birçok hiper düzlem çizilebilir. Fakat DVM'nin amacı kendisine en yakın noktalar arasındaki uzaklığı maksimum seviyeye çıkaran hiper düzlemi bulabilmektir. Şekil 3.21'de ise sınırı maksimuma çıkararak en uygun ayrımı yapan hiper düzleme optimum hiper düzlem, sınır genişliğini sınırlandıran noktalara da destek vektörleri denilmektedir şeklinde görüşlerini ifade etmektedir (Yelmen 2016). Çizelge 3.3'de algoritmanın başlangıç değerleri gösterilmiştir.

Çizelge 3.3. DVM'de Seçilen Parametre Değerlerimiz

C=1.0	Kernel=rbf	Degree=3	Gamma=auto	coef()=0.0	Shrinking=True	Probability=False
Tol=0.001	Cache_Size=200	Class_Weight=None	Verbose=False	Max_Iter=-1	Decision_Func_Shape=ovr	Random_State=None



Şekil 3.20. İki Sınıflı Problem İçin Hiper Düzlemler (Kavzoğlu ve Çölkesen 2010)

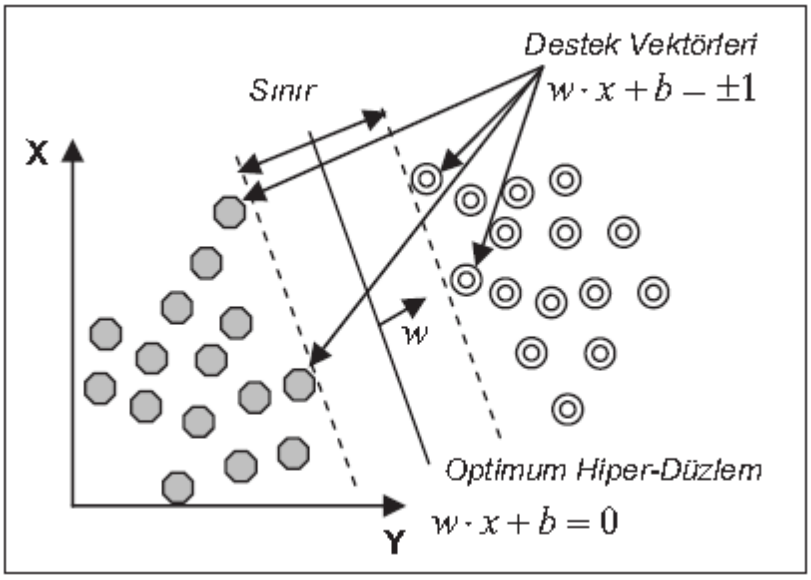


Şekil 3.21. Destek Vektörleri ve Optimum Hiper Düzlem (Kavzoğlu ve Çölkesen 2010)

Hiper düzleme ait eşitsizlikler formül 3.5 ve 3.6'da gösterildiği gibi,hakkında doğrusal olarak ayrılabilen ve iki sınıfı olan bir sınıflandırma işleminde DVM'nin eğitimi için k sayıda örnekten oluşan eğitim verisinin {xi, yi}, i = 1,.....,k = olduğu kabul edilirse, optimum hiper düzleme ait eşitsizlikler şekil 3.22'de gösterildiği gibi aşağıdaki şekilde olur: (Yelmen 2016)

$w.i + b \geq +1 \in y = +1$ (3.5)

$w.i + b \leq +1 \in y = -1$ (Yelmen 2016). (3.6)



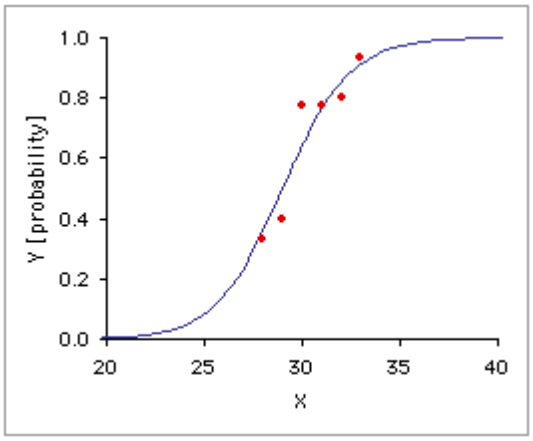
Şekil 3.22. Doğrusal Ayrılabilen Veri Setleri İçin Hiper Düzlemin Belirlenmesi (Kavzoğlu ve Çölkesen 2010)

3.2.6.4.2. Doğrusal Regresyon

Doğrusal regresyon, bağımlı değişken ve tahmin edici değişkenler arasındaki ilişkiyi analiz eden bir makine öğrenmesi yöntemidir (Çelik 2013). Şekil 3.23'de gösterdiği gibi, sınıf etiketleri arasında bir S tipi ayırıcı çizilir ve kenarları eşit olarak keser.

Doğrusal regresyon, ikili cevap verisini modellemek için kullanılan en yaygın yöntemdir. Yanıt ikili olduğunda, genellikle 1 ile bir başarı ve 1 bir başarısızlık gösteren 1 ile 1/0 şeklini alır.

Bununla birlikte, çalışmanın amacına bağlı olarak 1 ve 0'nın alabileceği gerçek değerler büyük ölçüde değişebilir (Hilbe 2011). Algoritmanın başlangıç parametre değerlerini (LinearRegression(fit_intercept=True, normalize=False, copy_X=True, n_jobs=1)) seçerek çalışmamızı gerçekleştirdik.



Şekil 3.23. Doğrusal Regresyon (Çelik 2013)

3.2.6.4.3. Karar Ağaçları

Bir karar ağacı, her düğümün bir özelliği (öznitelik) temsil ettiği, her bağlantının (dal) bir kararı (kural) temsil ettiği ve her yaprağın bir sonucu (kategorik veya sürekli değer) temsil ettiği bir ağaç yapısıdır (Sanjeevi 2017).

Karar ağacı, belirsizlik altındaki sıralı karar problemleri için model olarak kullanılabilir. Bir karar ağacı, alınacak kararları, meydana gelebilecek olayları ve karar ve olayların birleşimiyle ilişkili sonuçları grafik olarak açıklar. Karar ağacı modelleri, düğümler, dallar, terminal değerleri, strateji, belirli bir eşdeğer ve geri alma yöntemi gibi kavramları içerir. Gini algoritmasıyla karar ağacımızın parametre değerleri (criterion = "gini", random_state = 100, max_depth=100, min_samples_leaf=8) şeklinde seçilmiştir.

Şekil 3.24'de gösterildiği gibi, bir yüzme havuzu işletmesinin biletler üzerinde yapmış olduğu tarifeli satışın karar ağacıyla modellenmesi gösterilmiştir. Yüzücünün yaşına ve gün içindeki havuza girme saatine göre bilet kesilmektedir.

Şekil 3.24. Karar Ağacı Örneği

3.2.6.4.4. K En Yakın Komşu Sınıflandırma Algoritması(KEYKSA)

K en yakın komşu algoritması (KEYKSA), eğitim setinde test nesnesine en yakın olan bir grup k nesnesini bulur ve bu komşuluktaki belirli bir sınıfın üstünlüğüne dair bir etiketin atanmasına dayanır (Wu vd. 2008).

Bu yaklaşımın üç ana unsuru vardır. Bunlar; etiketli nesneler, örneğin bir dizi kaydedilmiş kayıt seti, nesneler arasındaki mesafeyi hesaplamak için bir mesafe veya benzerlik ölçüsü ve k, en yakın komşularının sayısıdır (Wu vd. 2008). Çalışmamızda algoritmanın komşuluk sayısı 3 olarak belirlenmiştir.

KEYKSA nesneler arasındaki mesafeyi hesaplamak için şekil 3.25.'de formül 3.7, 3.8 ve 3.9'da gösterildiği gibi Öklid, Manhattan, Minkowski olan 3 farklı fonksiyon kullanmaktadır.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Euclidean (3.7)

$$\sum_{i=1}^{k}|x_i - y_i|$$

Manhattan (3.8)

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

Minkowski (3.9)

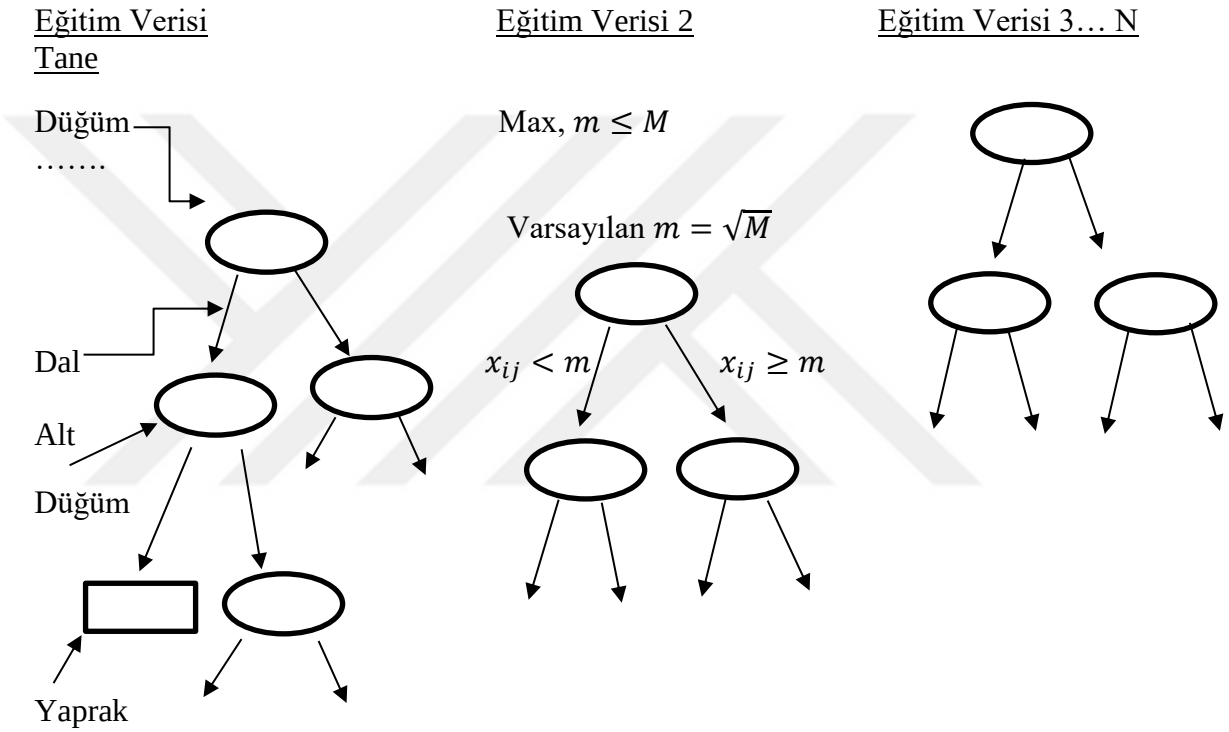
Şekil 3.25. KEYKSA Uzaklık Fonksiyonları(Atmaca 2017)

3.2.6.4.5. Rastgele Orman

Rastgele orman algoritması (RO), karar ağacı sınıflandırıcılarının bileşimini içeren bir yöntemdir. Karar ağacında olduğu gibi bu yöntemde de dallanma kriterlerinin belirlenmesinde en yaygın kullanılan teknik Gini indeksidir. Bu algoritmanın işleyişinde 2 farklı parametre esas alınmakta olup, bunlar sırasıyla her bir düğümde kullanılan değişken sayısı ve geliştirilecek olan ağaç sayısıdır (Küçükönder vd. 2015).

Ayrıca şekil 3.26'da gösterilen yapıya sahip olan RO algoritması bilinen makine öğrenme yöntemleri içerisinde daha gerçekçi bir tahmin geçerliliği ve model yorumlanabilirliği sağlar. Rastgele örneklemeyi ve topluluk yöntemlerdeki tekniklerin iyileştirilmiş özelliklerini içermesi nedeniyle RO yöntemi daha iyi genellemeler sunar ve geçerli tahminlerde bulunur (Fidancı 2017).

Bu açıklama yapmış olduğumuz çalışmanın ikinci aşamasında en yüksek doğruluk yüzdesini algoritmanın başlangıç değerlerini(RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0) kullanarak RO algoritmasıyla elde ettiğimiz sonucu doğrular niteliktedir.



Şekil 3.26. Rastgele Orman Yöntemine Ait Ağaç Yapısı (Akar ve Güngör 2010)

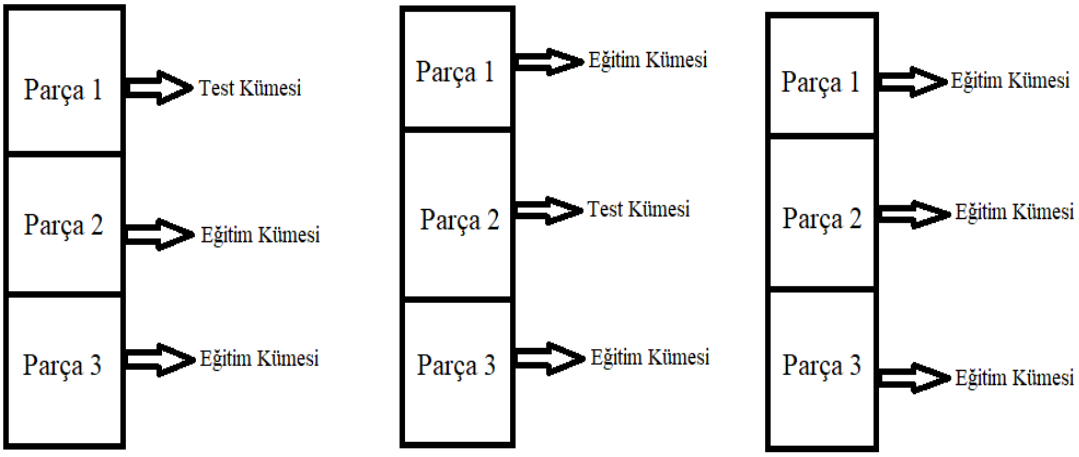
3.2.6.4.6. K-Kat Çapraz Doğrulama

Cross validation olarak bilinen çapraz doğrulama, tahmini model performansı tahmini için geliştirilmiş bir tekniktir. Veri seti, bu doğrulama tekniği içerisinde analiz için rasgele alt kümelere bölünmüştür. K-kat çapraz doğrulamada, veri seti rastgele olarak k alt kümelerine bölünür. Her alt set, test verileri olarak kabul edilirken, kalanlar eğitim verileri olarak kullanılır. Bu süreç, verilerin tüm kısımlarının bir kez doğrulama için kullanıldığı için k kere tekrarlanır.

Son olarak, tüm işlemlerin geçerlilik sonuçları tek ve son bir tahmin için ortalaması alınır (Çelik 2013). Çalışmamızda çaprazlama değerimizi 20 seçerek tüm veri setimizin %80'i eğitim seti, %20'si test seti olarak ayarlanmıştır.

Örneğin; k değerimiz 3 olsun ve elimizdeki veri sayısı toplam 300 olsun. Bu durumda 100'er parçalık veri kümesi 3 eşit parçaya ayrılır. Ayrılan parçaların 2'si eğitim verisi olurken 1 tanesi test verisi olur. Parçalardan hangisinin eğitim verisi hangisinin de test verisi olduğunun bir önemi yoktur. İlk parçayı test için ikinci ve üçüncü parçaları ise eğitim için seçtiğimizi varsayalım. Sistemimizi çalıştırdıktan sonra, örneğin amacımız sınıflandırma ise bunun sonucunu S1 olarak kabul ediyoruz. Şekil 3.27'de sırasıyla gösterildiği gibi, aynı işlemleri 2. ve 3. parçalar için de tekrarlıyoruz ve sonuçlarını S2, S3 olarak elde ediyoruz. Sonuçta aynı yöntemi, 3 farklı eğitim ve test verisi üzerinde çalıştırmış oluruz. Sistemin genel başarısını ise elde ettiğimiz bu 3 sonucun ortalamasını alarak hesaplayabiliriz. Burada ortalama alındığından ve toplamanın yer değiştirme özelliği olduğundan hangi parçadan çalıştırmaya başladığımızın bir önemi yoktur. Aşağıda 3.10'da ki formülde SF(test, eğitim) sınıflandırma fonksiyonunu, VK ise veri kümemizi, k ise çapraz doğrulama katsayımızı, t ise veri kümesi üzerinde seçilmiş olan eğitim kümesini, 1/k ise elde edilen sonuçların aritmetik ortalamasının alınmasını temsil etmektedir.

$$Sonuç = \frac{\sum_{i=0}^{k} SF(t_i, VK - t_i)}{k} \qquad (3.10)$$



Şekil 3.27. K-Kat Çapraz Doğrulama Veri Kümesi Bölünme İşlemi

3.2.7. Kelime Yerleştirmeleri

1990'lı yıllardan beri, vektör uzayı modelleri dağıtık semantikte kullanılmıştır. Bu süre zarfında, Latent Semantic Analysis (LSA) ve Latent Dirichlet Allocation (LDA)'da ve bunların geliştirilmiş türevleri de dahil olmak üzere, kelimelerin sürekli temsillerini tahmin etmeye yönelik birçok model geliştirildi. Kelime yerleştirmelerinin tarihçesi hakkında Bengio, 2003 yılında sözcük yerleştirmeleri terimini üretmiştir ve model parametreleriyle ortaklaşa bir sinirsel dil modelinde eğitmiştir (Ruder 2016). Önceden eğitilmiş sözcük yerleştirmelerinin kullanım kolaylığını ilk kez 2008'de Collobert ve Weston belirtti. İlkel metinleri Doğal dil işleme için birleştirilmiş bir mimari, kelime yerleştirmelerini aşağı akış görevleri için kullanışlı bir araç olarak değil, aynı zamanda aşağıdaki gibi birçok güncel yaklaşımın temelini oluşturan bir sinir ağı mimarisini de tanıttı. Nihai olarak kelime yerleştirmelerinin yaygınlaşmasıyla, Mikolov ve arkadaşları 2013 yılında, önceden eğitimli gömülmeleri sorunsuz bir şekilde eğitime ve kullanmaya olanak tanıyan bir araç seti olan Word2Vec'i tanıtmışlardır (Ruder 2016).

Kelime yerleştirmeleri temelde verilen metinlerin makineler tarafından algılanabilmesi için kelimelerin temsili için 0-1 formatında vektörlere dönüştürülmesine dayanır. Kelime yerleştirmelerine örnek bir cümle verecek olursak;

Cümle = “Word Embeddings are Word converted into numbers”

Bu cümlenin içindeki bir kelime “Embeddings” veya “numbers” vb. olabilir.

Bir *sözlük (dictionary)* cümle içinde tekrarsız geçen kelimelerin listesidir. Yani, bir sözlük şöyle görünebilir: [‘Word’,’Embeddings’,’are’,’Converted’,’into’,’numbers’]

Bir kelimenin bir vektör temsili, one-hot encode vektörü olabilir, sözcüğün bulunduğu konum 1 olur ve o kelimenin bulunmadığı diğer her yer de 0 olur. Bu formata göre “numbers” kelimesinin vektör temsili yukarıdaki sözlüke göre [0,0,0,0,0,1] ve dönüştürülmüş vektör temsili [0,0,0,1,0,0] 'dır.

3.2.7.1. Gizli Anlamsal Analiz(LSA)

LSA, Doğal Dil İşleme alanında kullanılan bir yöntem veya tekniktir. LSA’nın temel amacı, anlamsal içerik oluşturmak için metinler üzerinde vektör tabanlı temsil yaratmaktır. Vektör temsili ile LSA metinler arasındaki benzerliği hesaplar (Alghamdi ve Alfalqi 2015). Ayrıca Latent Semantic Analysis (LSA), verilerin yeniden düzenlenmesi için Tekil Değer Ayrışımı (TDA) kullanır.

TDA, vektör alanının tüm küçülmelerini yeniden yapılandırmak ve hesaplamak için matris kullanan bir yöntemdir. Buna ek olarak, vektör uzayındaki azalmalar hesaplanır ve en önemliden en az önemliye doğru düzenlenir (Alghamdi ve Alfalqi 2015).

LSA'daki en önemli adımlardan birincisi, çok sayıda ilgili metin toplanıp bunları belgelere ayırmaktır. İkincisi, terimler ve belgeler için birlikte-oluşum matrisini oluşturup, aynı zamanda belge x gibi hücre adını, terimler için boyutsal değer için y ve m terimlerini ve belgeler için n boyutlu vektörü belirtilir. Üçüncü olarak, her bir hücre hesaplanacaktır. Son olarak, TDA tüm azalmaları hesaplamak ve üç matris yapmak için büyük bir rol oynayacaktır şeklinde tanımlamaktadır (Alghamdi ve Alfalqi 2015).

Temel olarak, LSA belgelerin ve kelimelerin düşük boyutlu temsilini bulur. Elimizde d1….dm'e kadar döküman olduğunu ve belgelerden çıkarılan sözlükteki kelimelerin w1….wn'e kadar olduğunu varsayalım. Bunlardan bir tane mxn boyutunda doküman-terim matrisi oluştururuz. Kelimelerden hangi dökümanda kaç tane geçtiğini gösterir.

	kedi	köpek	bilgisayar	internet	tavşan
D1	1	1	0	0	2
D2	3	2	0	0	1
D3	0	0	3	4	0
D4	5	0	2	5	0
D5	2	1	0	0	1

Satır vektörlerinin nokta çarpımı, belge benzerliğidir, sütun vektörlerinin nokta çarpımı ise sözcük benzerliğidir. Sonuç olarak, LSA modelleri tipik olarak doküman-terim matrisindeki ham sayıları bir tf-idf puanı ile değiştirir. Ayrıca formül 3.11'da gösterildiği gibi matrisin toplam X boyutunu azaltmak için SVD kullanılır.

$$X = U \sum V^T \qquad (3.11)$$

Bu döküman vektörleri ve terim vektörleri ile kosinüs benzerliğiyle farklı kelimeler ve farklı dökümanlar arasındaki benzerliği bulabiliriz.

3.2.7.2. Olasılıksal Gizli Anlamsal Analiz (PLSA)

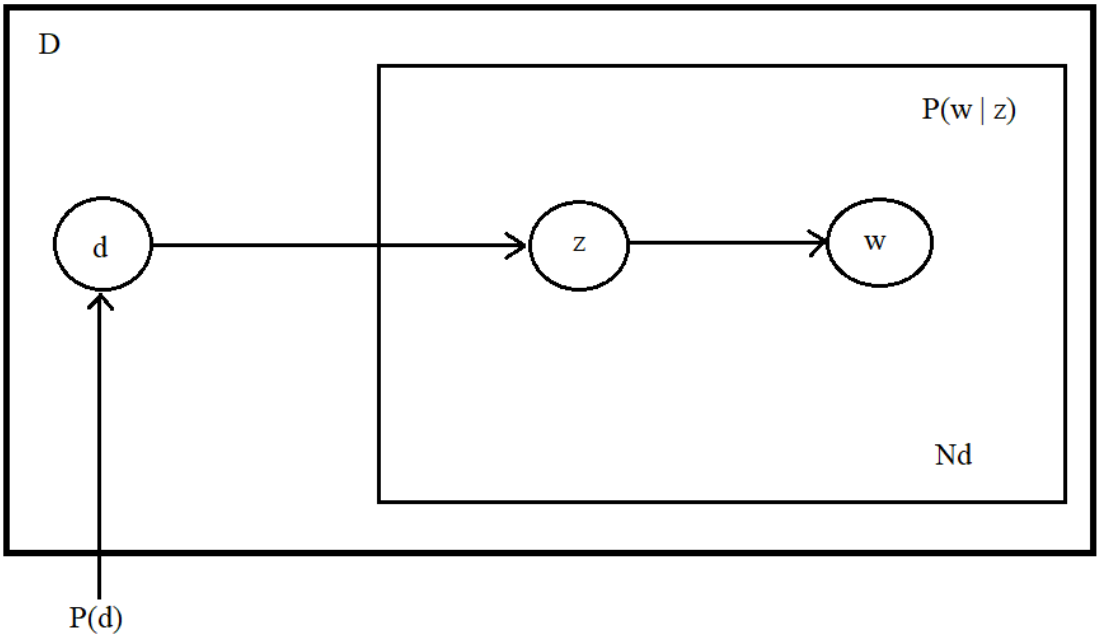
Olasılıksal Latent Semantic Analysis (PLSA), LSA yöntemini takiben LSA'ya dahil edilen bazı dezavantajları düzeltmek için yayımlanan bir yaklaşımdır. Thomas Hofmann 1999 yılında tanıttı (Hofmann 2001). PLSA, sayım verilerinin faktör analizi için istatistiksel gizli sınıf modeline dayanan doküman endeksleme işlemini otomatik hale getiren bir yöntemdir. Bu yöntem, üretken bir model kullanarak Latent Semantic Analysis (LSA) 'yi olasılıksal anlamda geliştirmeye çalışmaktadır. PLSA'nın asıl amacını Alghamdi ve Alfalqi (2015) bir sözlük veya eş anlamlılar listesine başvurmaksızın kelime kullanımının farklı bağlamlarını tanımlamak ve ayırmaktır.

İki önemli hususu içerir: Birincisi, çokseslilığı, yani birden çok anlamı olan kelimeleri belirginleştirmeye olanak tanır. İkinci olarak, ortak bir bağlamda paylaşılan kelimeleri gruplayarak tipik benzerlikleri ifşa eder şeklinde ifade etmişlerdir (Alghamdi ve Alfalqi 2015). PLSA modelinin Şekil 3.28'de verildiği gibi çalışma mantığı temel olarak sırasıyla şu şekildedir;

1) Olasılık P (d) ile bir d belgesi seçer,

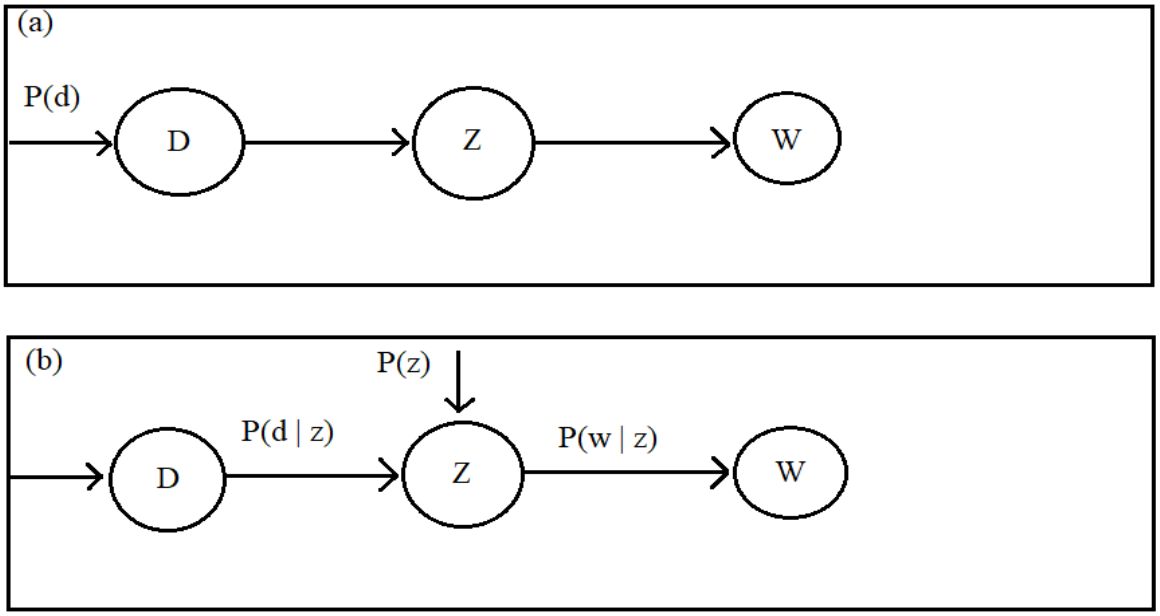
2) Olasılık P (z | d) ile gizli z sınıfını seçer,

3) Olasılık P (w | z) ile bir w kelimesi üretir.



Şekil 3.28. PLSA'nın Üst Seviye Görünümü (Alghamdi ve Alfalqi 2015)

PLSA bu yöntemi sunmak için iki farklı formülasyona sahiptir. İlk formülasyon, koşullu olasılıklar P (d | c) ve P (w | c) kullanılarak, gizli c sınıf'ından benzer şekillerde (w) ve (d) belgesinin elde edilmesine yardımcı olacak simetrik formülasyondur. İkinci formülasyon Şekil 3.29'da gösterildiği gibi asimetrik formülasyondur. Bu formülasyonda, gizli bir sınıf olan her d belgesi, koşullu olarak P (c | d) 'ye göre belgeye seçilir ve kelime P (w | c)' ye göre bu sınıftan üretilebilir.

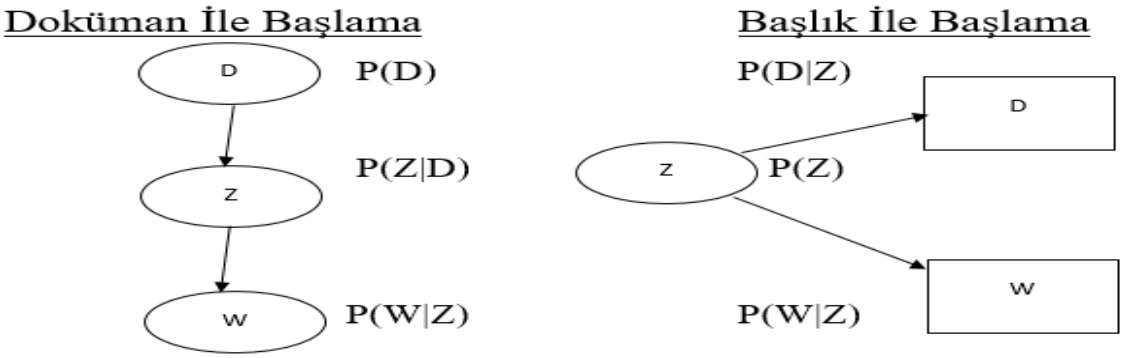


Şekil 3.29. Asimetrik (a) ve Simetrik (b) Parametreleme Görünüm Modeli (Alghamdi ve Alfalqi 2015)

PLSA, matrisleri kullanmak yerine olasılıksal bir metod kullanır. Belirli bir dökümanı ve sözcüğü birlikte görmenin ortak olasılığı formül 3.12'de verildiği gibidir:

$$P(d,w) = \sum_Z P(d)P(d|z)P(w|z) \qquad (3.12)$$

İlk parametrelendirmemizde P (d) ile başlayan ve P (z | d) ile konuyu oluşturup daha sonra P (w | z) ile kelimeyi üretmeye başladık. Bu parametrede P (z) ile konu(topic)'dan başlıyoruz ve daha sonra P (d | z) ve P (w | z) kelimelerini kullanarak belgeyi bağımsız olarak üretiyoruz.



PLSA’nın LSA’dan farkı; oldukça farklı görünmesine ve problemi çok farklı bir şekilde ele almasına rağmen, PLSA gerçekten LSA'nın zirvesinde olan konuları ve kelimeleri olasılıksal bir iyileştirmesini gerçekleştirmektedir. Çok daha esnek bir modeldir, ama birkaç problemi vardır. Bunlar:

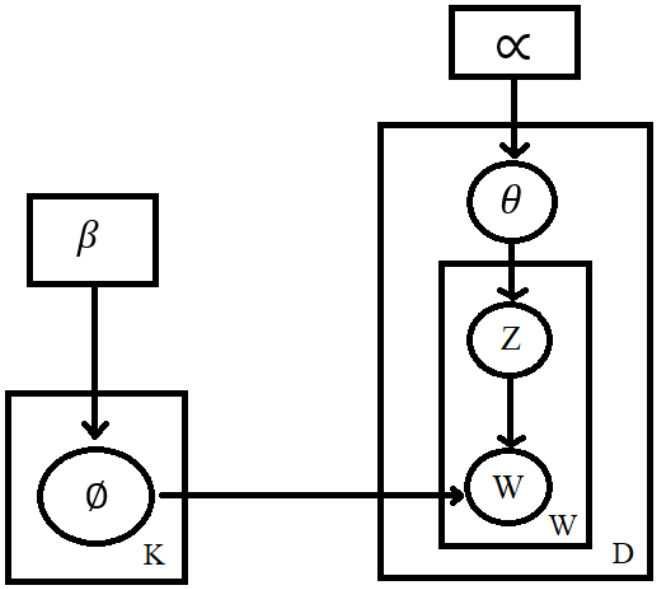
-P (D) modelinde parametrelerimiz yoktur, olasılıkları yeni belgelere nasıl atayacağımızı bilmiyoruz

-PLSA için parametrelerin sayısı, sahip olduğumuz belge sayısı ile doğrusal olarak büyür, bu yüzden overfitting’e eğilimlidir.

Genel olarak, insanlar LSA'nın verdiği temel performansın ötesinde bir konu modeli ararken, LDA'ya dönerler. Konu modelinin en yaygın türü olan LDA, bu konulara değinmek için PLSA'yı genişletmektedir (Xu 2018).

3.2.7.3. Gizli Dirichlet Tahsisi (LDA)

Gizli Dirichlet tahsisi (LDA), gözlem gruplarının, verilerin bazı bölümlerinin neden benzer olduğunu açıklayan gözlemlenmemiş gruplar tarafından açıklanmasına izin veren bir üretken modeldir. Şekil 3.30’da modeli verilen LDA, doğal dil işleme ve istatistiksel makine öğrenimi alanlarında büyük bir etki yapmıştır ve makine öğrenmesinde en popüler olasılıksal metin modelleme tekniklerinden biri haline gelmiştir (Tong ve Zhang 2016).



Şekil 3.30. LDA Modeli (Alghamdi ve Alfalqi 2015)

LDA, D belgelerinin her birini, her biri bir W kelimeli kelime dağarcığına göre çok terimli bir dağılımı açıklayan, gizli konular üzerinde bir karışım olarak modeller (Alghamdi ve Alfalqi 2015). Şekil 3.29. LDA modelinin grafik model sunumunu göstermektedir. LDA için üretken süreç aşağıdaki gibidir:

J belgesindeki her Nj kelimesi için

1) Bir konu seçin zij ~ Mult (θj)

2) Bir kelime seçin zij ~ Mult (φzij)

Bir dokümanda θj'deki konuların ve bir konudaki kelimelerin çoğullarının parametreleri Dirichlet önceliğine sahiptir.

Bu model metinlerde anlatılmak istenilen düşüncenin (duygu ve fikrin birlikte olduğu) başlıkta gizli de olsa var olacağı düşüncesini temel almaktadır. Ayrıca LDA'da metinlerin gizli başlıkları bir istatistiki modele dayandırılır ve metinlerdeki sözcüklerin dağılımlarından bu model oluşturulur (Blei vd. 2003).

Bu modele örnek verecek olursak;

Ben **balık** ve sebze yerim.

Balıklar evcil hayvanlardır.

Benim kedim **balık** yer.

Yukarıda örneği verilen cümlelerden iki başlık çıkartılır. Birincisi “yemek”, ikincisi ise “evcil hayvan”. Buna göre birinci cümle %100 yemek başlığı altında, ikinci cümle %100 evcil hayvan başlığı altında, üçüncü cümle ise %66 yemek başlığında, %33 evcil hayvan başlığı altında olacaktır.

Yukarıdaki örnekte görüldüğü gibi LDA iki aşamadan oluşmaktadır. Birincisi metinlerden başlıkların çıkarılması, ikincisi ise başlıklara göre metinlerin sınıflandırılmasıdır (Şeker 2016).

3.2.8. Metin Temsili Yöntemleri

Metin temsili, elimizde var olan verilen işlenmesi ve bilgisayar tarafından anlaşılır hale gelmesi için yapısal formata olan öznitelikleri elde edildikten sonra her kelime birer vektör yapısıyla ilişkilendirilmesidir. Kullanılan öznitelik çıkarma yöntemleri birer metin temsili yöntemi olarak kabul edilmektedir. Ayrıca, metin tabanlı dosya, belge vb. gibi verilerin otomatik olarak kategorize edilmesi, web sitelerinde arama işlemlerinin hızlandırılması, fikir madenciliği, otomatik özet çıkarma gibi birçok uygulama alanında kullanılmaktadır. Metin temsili twitter duygu analizi modellerinde kullanılan sözcüksel(lexical) öznitelikler için çok önemlidir. Son çalışmalar yoğun, düşük boyutlu ve gerçek değerli kelime gömülmesinin twitter duygu sınıflandırması için rekabetçi performans sağladığını göstermiştir (Ren vd. 2016).

3.2.8.1. N-Gram

N-Gram, n uzunluğundaki kelime dizisine denir. Kelime yaklaşımı paketinde, metin bir kelime topluluğu olarak ele alınır ve kelimelerin sırası kabul edilmez. Örneğin, "Ayşe, Betül'den daha güzel" ve "Betül Ayşe'den daha güzel" BOW yaklaşımında aynıdır. Bu, metinde büyük bir bilgi kaybına neden olur. N-gram modelinin kullanılması, özellikli kelime setlerinin özellik olarak kullanılmasıyla bilgi kaybını azaltır (Fürnkranz 1998).

Kelime seviye N-gram model, kelime torbası model ile yöntem olarak aynıdır (her iki modelde de öznitelikler kelimeler olur) ancak önişleme aşamasında N-gram modelde ayrıştırma dışında bir işlem uygulanmaz. Ayrıca karakter seviye N-gram modelde en düşük katar uzunluğu ikiden az olamaz (unigram çıkarılmaz) (Çoban 2016).

Çizelge 3.4. N-Gram Örneği

Cümle	Bugün okullar tatil olduğu için mutluyum.
Unigrams	{Bugün}, {okullar}, {tatil}, {olduğu}, {için}, {mutluyum}
Bigrams	{Bugün okullar}, {okullar tatil}, {tatil olduğu}, {olduğu için}, {için mutluyum}
Trigrams	{Bugün okullar tatil}, {okullar tatil olduğu}, {tatil olduğu için}, {olduğu için mutluyum}
Fourgrams	{Bugün okullar tatil olduğu}, {okullar tatil olduğu için}, {tatil olduğu için mutluyum}
Five-grams	{Bugün okullar tatil olduğu için}, {okullar tatil olduğu için mutluyum}

3.2.8.2. Kelime Torbası(BOW)

BOW modeli, Doğal Dil İşleme ve Bilgi Alışverişi metinlerini (belgeler, cümleler, sorgular ve diğerleri) temsil etmek için popüler bir yaklaşımdır. Bu modelde, önceden hazırlanmış bir sözlük göz önüne alındığında, bir metin bir dizi kelimeyle temsil edilir, bu gösterim ikili olabilir. Bir kelime, metinde geçiyor ise 1, aksi halde 0 alır (Jśnior vd. 2017).

İngilizcedeki Term Frequency – Inverse Document Frequency (Terim frekansı – ters metin frekansı) olarak geçen kelimelerin baş harflerinden oluşan terim basitçe bir metinde geçen terimlerin çıkarılması ve bu terimlerin geçtiği miktara göre çeşitli hesapların yapılması üzerine kuruludur. Klasik olarak TF yani terimlerin kaç kere geçtiğinden daha iyi sonuç verir. Kısaca TF-IDF hesabı sırasında iki kritik sayı bulunmaktadır. Bunlardan birincisi o anda ele alınan dokümandaki terimin sayısı diğeri ise bu terimi derlemde(korpusta) içeren toplam doküman sayısıdır (Yeşilyurt ve Şeker 2017).

Pang ve arkadaşları her bir kelimenin birer vektör olarak temsil edildiği kelime torbası (BOW) alanında öncülük etmişlerdir (Ren vd. 2016). Kelime torbası(BOW) modeli güçlü bir araçtır ancak iki problemi vardır. Birincisi, kelime gösteriminin yüksek boyutu ve yedek bilgi sınıflandırma performansını sınırlar. İkincisi, bu model kelimeler arasındaki anlamsal ilişkiyi yakalayamaz (Qiang vd. 2017). Ancak kelime yerleştirmeleri, vec(Berlin) - vec(Germany) + vec(France) = vec(Paris) gibi kelimeler arasındaki anlamsal ilişkiyi yakalayabilir (Pang vd. 2002). Çalışmamızda Word2Vec modelini kullanmamızın nedeni de budur.

TF ve TF-IDF hesaplamalarını Terim Frekansı (TF) burada terimin sıklığını ifade etmektedir. Formül 3.13, 3.14 ve 3.15'de gösterildiği gibi hesaplanırken dokümanda o terimin geçme sayısının (f), o terimin toplam dokümanlarda kaç kere geçmiş ise o sayıya (df) bölünmesiyle elde edilir. Ters Metin Frekansı (IDF) ise analiz edilen kelimenin ne kadar eşşiz (Unique) olduğunu hesaplanması işlemidir.

IDF hesaplanırken “N” değeri toplam doküman sayısını temsil eder. “df” ise hesabı yapılan kelimenin kaç dokümanda yer aldığını temsil eder. Cümledeki IDF değerinin yüksek olması demek o kelimenin kullanılmasının, kategorinin belirlenmesi için çok değerli olduğunu gösterir. Örneğin “çok” kelimesi bir cümlede tek başına bir duygu ifadesi içermemektedir. Bu nedenle hem olumlu cümle hem de olumsuz cümlenin içerisinde sık olarak geçebilir. TF-IDF yöntemi tam bu esnada bu kelimenin ağırlığını düşük bularak önemsiz olduğunun tespitinde büyük rol oynar (Akba 2014).

TF = f/df → TF Hesaplama (3.13)

IDF = Log(N/df) → IDF Hesaplama (3.14)

TF-IDF = TF * IDF → TF-IDF Hesaplama (3.15)

Tf-Idf terim ağırlıklandırmasına örnek verecek olursak;

Doküman A : The sky is blue. Doküman B : The sky is not blue.

Çizelge 3.5. Tf-Idf Örneği

	Tf		Idf	Tf-Idf	
	A	B		A	B
The	1	1	log(2/2)	0	0
Sky	1	1	log(2/2)	0	0
Is	1	1	log(2/2)	0	0
Blue	1	1	log(2/2)	0	0
Not	0	1	log(2/1)	0	1

Tf ile her bir dökümandaki kelimenin geçme sıklığını, Idf ile logaritmik hesaplamasını yaptıktan sonra Tf-Idf için bulunan bu iki değer çarpılır. Log(2/2) = Log(1) → 0’dır.

Tf-Idf için farklı bir örnek verecek olursak eğer elimizde üç adet cümle olduğunu kabul edelim, bunlar aşağıdaki gibi listelenmiştir;

- Muz sarı ya da yeşil renklidir genellikle sarı olur.
- Elma kırmızı ve yuvarlaktır, sarı renkte olanlar olabilir.
- Karpuzun içi kırmızı renklidir.

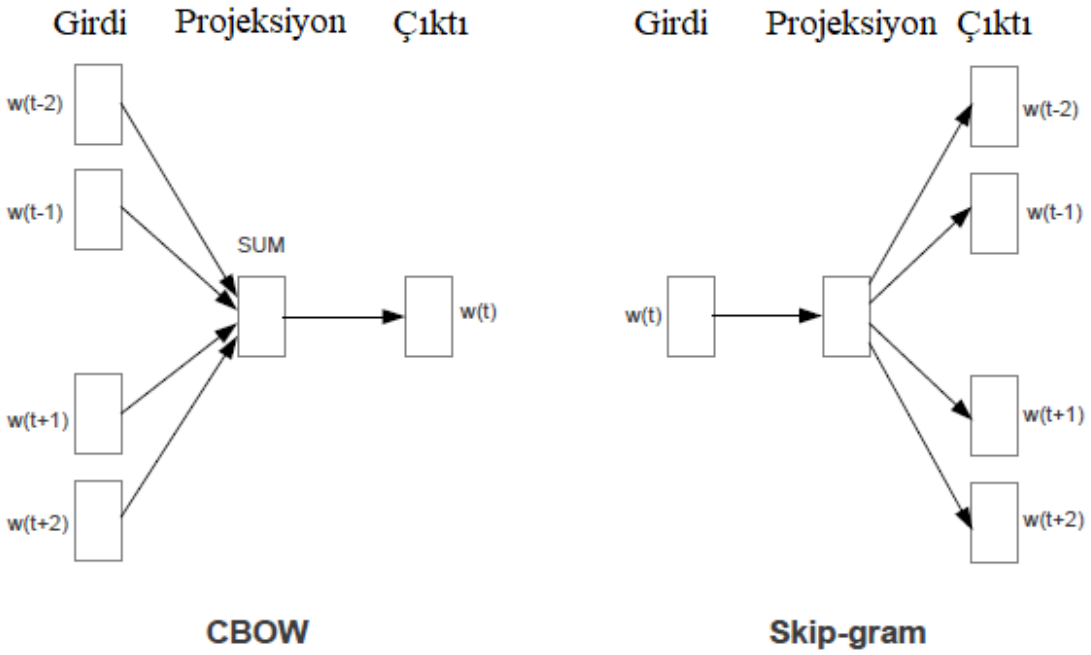
Sarı kelimesi için birinci cümle içinde hesaplama yapacak olursak, Tf=(2/8) → 0,25 ve Idf=(log(3/2)) → 0,18 olarak hesaplanır. Tf-Idf = 0,25*0,18 → 0,044 bulunmuştur. Burada cümle sayımızı artırıp 4 yapsaydık Idf değerimiz 0,3, Tf-Idf değerimiz 0,075 olacaktı. Yani sarı teriminin bütün metinler içinde seyrekliği arttı fakat birinci metin içindeki seyrekliği aynı kaldı dolayısıyla birinci metine olan etkisi artmıştır (Bavaş 2017).

3.2.8.3. Kelime Vektör Dönüşümü(Word2Vec)

Mikolov ve arkadaşları giriş katmanı, projeksiyon katmanı ve yakındaki kelimeleri tahmin etmek için bir çıktı katmanı içeren sığ bir sinir ağı mimarisi kullanarak her kelimenin vektör gösterimini öğrenmek için yeni geliştirilen Skip-Gram ve CBOW modellerine sahip kelime vektör dönüşümü ile kelime temsilini gerçekleştiren Word2Vec' modelini geliştirmişlerdir (Mikolov vd. 2013a; 2013b).

Word2Vec modeli doğrusal olmayan dönüşümlerden kaçınır ve bu nedenle eğitimi son derece verimli hale getirir(Ren vd. 2016).Bu sözlükteki milyonlarca kelimeden ve milyarlarca kelimeye sahip çok büyük veri setlerinden gömülü kelime vektörlerini öğrenmeyi sağlar (Ren vd. 2016).

Kelime vektörleri oluşturmak için en yaygın kullanılan modellerden biri olan Word2Vec, Şekil 3.31'de verildiği gibi "skip-gram(SG)" ve "continuous bag-of-words(CBOW)" olarak açıklanan iki model ile uygulanmaktadır (Nakov vd. 2016).



Şekil 3.31. CBOW ve Skip-Gram Modelleri(Mikolov vd. 2013a)

- ➔ CBOW, hedef kelimeyi çevreleyen bağlamsal kelimelerle tahmin etmek üzere eğitilir, örn. wi − 2, wi − 1, wi + 1, wi + 2 giriş kelimeleri alarak tarafından hedef kelime wi'yi tahmin eder(Güvenç 2016).
- ➔ Skip-gram modelinin eğitim amacı, bir cümle veya belgede çevresindeki kelimeleri tahmin etmek için faydalı olan kelime temsillerini bulmaktır((Mikolov vd. 2013b).
- ➔ Mikolov'a göre, Skip-gram sık olmayan kelimeler için daha yavaştır ancak küçük eğitim verileriyle iyi çalışır. Bu yüzden çalışmamızda Word2Vec modelinin default değeri olan skip-gram'ı küçük veri setimizde kullanmak üzere seçtik.

Python programlama dilindeki gensim kütüphanesi sayesinde Word2Vec modelini kullanabildiğimiz çalışmalarımızda, İngilizce veri setinin eğitilmesi için Word2Vec yöntemiyle eğitilerek oluşturulmuş 3 milyon kelimeye sahip 300 boyutlu "GoogleNews-vectors-negative300" (Mihăltz 2016) derlemi kullanıldı. Türkçe veri seti için de Boğaziçi Üniversitesi tarafından geliştirilen 416051 kelimeye sahip 300 boyutlu "wiki-tr" (Grave 2017) derlemi kullanılmıştır.

Çalışmamızda varsayılan değerleriyle uygulanılan Word2Vec modelinde, bir twitter mesajındaki tüm kelimelerin önceden eğitilmiş vektörleri olmadığından, rasgele başlatılan 300 boyutlu vektör belirlenilmiştir.

Word2Vec modeline örnek olarak 3 tane cümleyi ele alalım (Şeker 2016);

- James likes to play basketball.
- Julia likes basketball too.
- James also likes football.

Çizelge 3.6. Word2Vec İçin Terim-İndis Değerleri

Terim	İndis
James	1
likes	2
to	3
play	4
basketball	5
Julia	6
too	7
also	8
football	9

Yukarıda verilen her bir terim indeksini, terimlerin metinlerde geçme sayısına göre vektör haline dönüştürülmüş şekli aşağıdaki gibidir;

Çizelge 3.7. Word2Vec'de Kelimelerin Vektör Gösterimi

James	likes	to	play	basketball	Julia	too	also	football
1	1	1	1	1	0	0	0	0
0	1	0	0	1	1	1	0	0
1	1	0	0	0	0	0	1	1

Her bir terimin indeksine göre vektör yapısını çıkardıktan sonra her bir cümle ikilik tabanda bir vektör ile temsil edilmiştir. Örneğin ilk cümlenin ikilik tabandaki karşılığı 111110000 olmuştur.

Bu yöntemde çözüm olarak skip-gram modeli önerilmektedir. Tez çalışmamızda da skip-gram modeli kullanılmıştır ve formül 3.16'de gösterildiği gibi hesaplanmaktadır.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=-k}^{j=k}\log_p\left(w_{t+j}\middle|\, w_t\right) \tag{3.16}$$

K değerimiz pencere boyutunu temsil etmektedir ve skip-gram modeli bu boyut kadar kelime alarak işleme sokmaktadır. Ayrıca skip-gram'da kelime sayılarının tutulduğu u vektörü ile kelimelerin geçtiği içerik v vektörü birlikte güncellenir.

3.2.9. Sistem Mimarisi

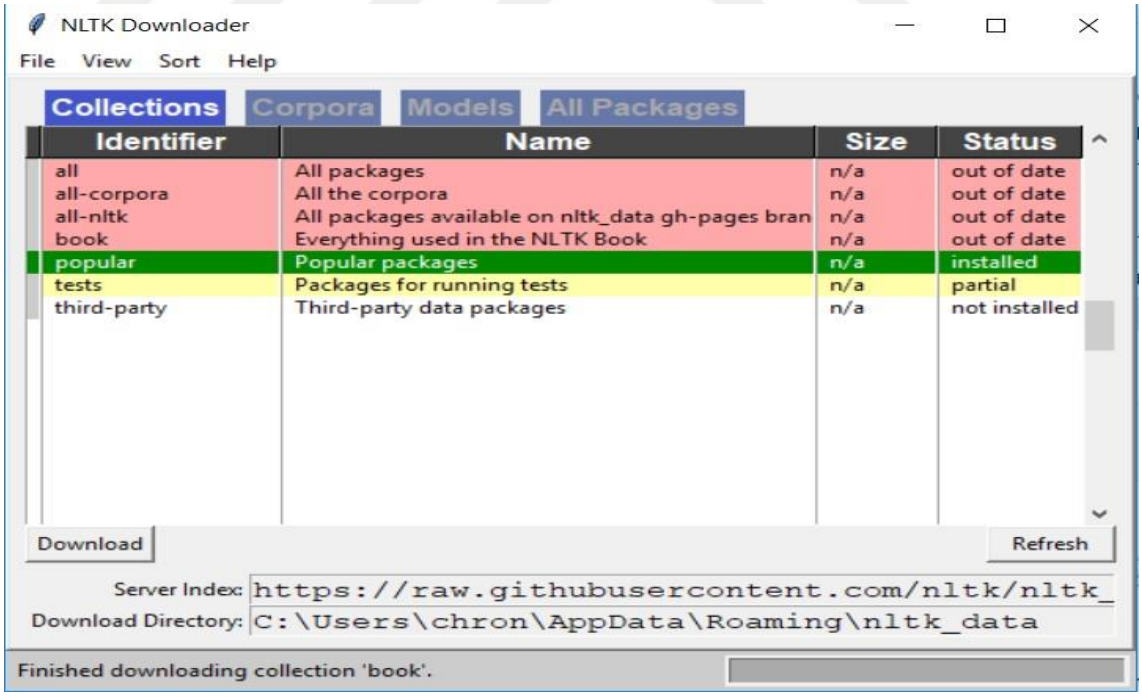
Çalışmamızda Windows 10 işletim sistemi kullanılmıştır ve CPU tabanlı bir mimariye sahip Intel Core İ7 2.20 GHz işlemcili, 64 bit işlemcili, 4 çekirdekli bir bilgisayarda Anaconda Navigator IDE'si içerisindeki Python 3.5.5 programlama dili kullanılarak çalışma gerçekleştirilmiştir. Doğal dil işleme kütüphanesi olan NLTK'dan, makine öğrenmesi algoritmalarını içeren scikit-learn kütüphanesinden, gensim kütüphanesinden, twitter mesajları üzerindeki ön işleme adımlarından, kök alma işleminden ve geliştirdiğimizden modellerden kısaca bahsedilecektir.

3.2.9.1. Python ve NLTK Kütüphanesi

Python programlama dili Guido van Rossum tarafından geliştirilmiş, yüksek seviye veri tipleri olan, nesneye yönelik, esnek, kolay öğrenilebilen bir programlama dilidir. İsmi genel kanı olan piton yılanından değil, BBC'de yayınlanan "Monty Python" adlı bir komedi dizisinden gelmektedir. Tez çalışmamızda Python kodları Windows 10 işletim sisteminde çalıştırılmıştır.

NLTK'nin yazılmasında Python dilini tercih edilme sebepleri, kolay öğrenilebilme, yazı dili ve mantık dilini şeffaf olması, gelişmiş katar kütüphanelerine sahip olması, nesneye yönelik programlamada veri ve yöntemlerin tekrar kullanılabilir olması, detaylı standart kütüphanesinin olması ve sayısal işlemeye uygun olmasıdır (Kobalas 2009).

Python programlama dili ile geliştirilmiş ve geliştirilmekte olan yaklaşık 120.000 satır kod, 70'e yakın derlem ve çok sayıda alt birimlerden oluşan açık kaynak bir araçtır. NLTK resmi sitesinden (www.nltk.org) nltk, matplotlib, numpy araçları indirilerek kurulum yapılabilir. Ayrıca python 3.3.5 sürümlerinden biri kurulmuştur. Şekil 3.32'de yüklenmesi gösterildiği gibi, tüm Nltk içinde yazılan modüller, matplotlip içerisinde grafik, şablon çıkarmak için sınıflar, numpy içinde ise sayısal hesaplama sınıfları bulunmaktadır. Bunlar kurulduktan sonra veri olarak, derlemleri, etiketleri, ağaç yapılarının içinde bulunduğu nltk_data indirilip kurulmuştur. Bu python/lib/site-packages/nltk/downloader.py'i çalıştırarak yapılmalıdır. Python komut satırından "import nltk" ile bu aracın genel modüllerini kütüphanemize dahil edilmiştir.



Şekil 3.32. NLTK Kütüphanesinin İndirilmesi

3.2.9.2. Scikit-Learn Kütüphanesi

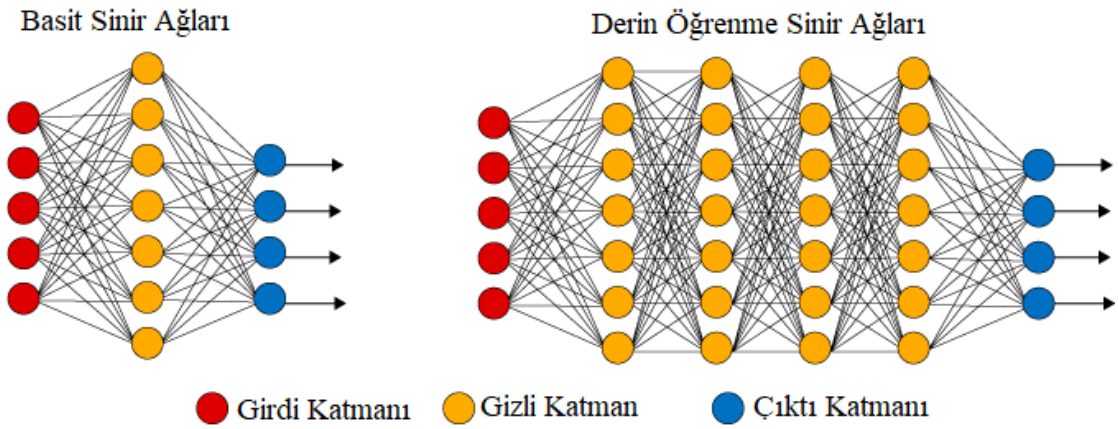
Scikit-Learn kütüphanesi, Python programlama dilinde denetimli ve denetimsiz öğrenme algoritmalarını sunan bir kütüphanedir (Brownlee 2014). Doğrusal regresyon, lojistik regresyon, karar ağaçları, rastgele orman gibi birçok temel yöntemi içerir.

Scikit-learn paketinin bu kadar popüler olmasının birkaç sebebini bunlardan ilki ihtiyaç duyacağınız temel yöntemlerin büyük bir kısmını içermesi. İkinci olarak, scikit-learn sayesinde veri analitiği uygulamalarını baştan sona yürütmenizin mümkün olmasını sayabiliriz. Verideki eksik değerleri doldurmak, öznitelik seçmek, çapraz doğrulama yapmak, sonuçları değerlendirmek için ayrı ayrı modüller sayesinde başka bir pakete ihtiyacınız kalmıyor.

Scikit-learn paketinin en güzel yanı basit bir API'a sahip olması sayesinde uygulayacağınız farklı yöntemler için farklı sözdizimleri öğrenmenizin önüne geçmesi. Tez çalışmamızda, fit/predict ya da fit/transform fonksiyonları sayesinde Doğrusal SVM, Doğrusal Regresyon, K En Yakın Komşu Algoritması, Rastgele Orman, Karar ağacı gibi makine öğrenmesi algortimaları uygulanmıştır. Eksik değerleri doldurmak, veriyi ölçeklendirmek gibi farklı adımlarda benzer fonksiyonlar kullanmak problemlerin çözümünü kolaylaştırmaktadır(Yüceloğlu, 2017).

3.2.9.3. Gensim Kütüphanesi ve Yapay Sinir Ağları

Gensim kütüphanesi python dilinde vektör uzaylarını modelleme ve başlık modelleme (topic modeling) özelliklerini barındıran ve uzun metinlerde verimli çalışabilecek şekilde geliştirilmiş olan bir kütüphanedir. Gensim kütüphanesi Python dilinin temel kütüphanelerinden NumPy ve SciPy kütüphanelerine ihtiyaç duymaktadır. Ayrıca Gensim kütüphanesinin kod yapısı incelenmiş olup, Hiyerarşik Dirichlet işlemi, Gizli Semantic Analiz(LSA) ve Gizli Dirichlet Tahsisi(LDA) gibi algoritmalarını içerdiği görülmüştür (Özaytaç 2018). Tez çalışmamızda içerisinde Word2Vec modelini de barındıran gensim kütüphanesi kullanılmıştır.



Şekil 3.33. Basit Sinir Ağı Modeli ve Derin Öğrenme Sinir Ağı Modeli (Tuzen 2017)

İnsan beyin sistemi örnek alınarak, nöronlar arasındaki ilişkilerin modellenmesiyle öğrenme, genelleme yapma ve tahmin etme gibi insana ait bazı temel görevleri gerçekleştirmek için geliştirilmiş olan mantıksal yazılımlar 'Yapay Sinir Ağları' olarak adlandırılmaktadır (Uslu 2016).

Yapay sinir ağlarının özellikleri; doğrusal olmamaları, paralel çalışmaları, eksik verilerle çalışması, çok sayıda parametre kullanmaları, hata toleransının ve esnekliğin olmasıdır (Uslu 2016).

Literatürde kullanılan en popüler kelime yerleştirme mimarisi Word2Vec olarak bilinmektedir (Yıldırım ve Yıldız 2018). Word2Vec mimarisinin genel yapısı Şekil 3.30'da gösterildiği gibi orta katmanı tek bir katmandan oluşmaktadır. Derin öğrenme modeli gibi görünsede, Word2Vec modeli aslında yapay sinir ağı modelidir, derin öğrenme modeli değildir (Yıldırım ve Yıldız 2018). Ayrıca Word2Vec modelinin derin öğrenme modeli olmayıp yapay sinir ağı modeli olmasının diğer bir nedeni ise; derin öğrenme modellerinin birden çok doğrusal olmayan(non-lineer) katmana sahip olması ve özellik hiyerarşilerini öğrenebilen makine öğrenimi algoritmaları olmasıdır. Ancak Word2vec'te orta katmanda aktivasyon fonksiyonu yoktur, Şekil 3.30'da gösterildiği gibi yansıtma(projection) katmanı bulunmaktadır. Yani, Word2Vec modelinde Şekil 3.33'de gösterildiği gibi modelin orta tabakasında birden fazla gizli katman bulunmamaktadır ve derin öğrenme modeline girmemektedir (Dwivedi 2017).

Python programlama dilinde derin öğrenme için geliştirilmiş Theno, Tensorflow ve Keras gibi kütüphaneler bulunmaktadır (Tuzen 2017). Tez çalışmamızda ise başlık modelleme kütüphanesi olan Gensim kütüphanesi ve içinde bulunan Word2Vec modeli kullanılmıştır.

Derin öğrenme sinir ağının basit sinir ağından farkından bahsedecek olursak, derin öğrenme büyük miktardaki veriler üzerinde çalışmaya uygun bir sistem mimarisine sahip olması, teknolojinin gelişmesiyle GPU ve işlem gücünün artması ve buna bağlı olarak çok sayıda gizli katman mimarisinde çalıştırıyor olması söylenebilir (Burhan 2017).

3.2.9.4. Ön İşleme

3.2.9.4.1. Dizge Parçalama

Veriler üzerinde öznitelik çıkarılmadan önce temizlenme işlemine tabi tutulur. Nasıl bir temizleme (cleaning) işleminin uygulanacağı; üzerinde çalışılan metnin türüne göre değişiklik gösterebilir. Eğer işlenen veri web sayfalarından oluşuyorsa html elementlerinin, köşe yazılarından oluşuyorsa özel karakterlerin, sosyal medya mesajlarından oluşuyorsa bazı özel terimlerin içerikten ayrıştırılması gerekir. Veri setlerimizde "@" karakteriyle başlayan kullanıcı adı, "#" diyez karakteriyle başlayan hashtag'leri, tüm rakamları, noktalama ve imge işaretlerini ve URL adreslerini kaldırdık. Bu işlemleri tamamladıktan sonra dizge parçalama(tokenize) işlemini gerçekleştirdik.

Dizge parçalamaya örnek olarak; '#gününnasıl Bugün hava güzel değil mi ? ☹ 'tweet mesajını ele alalım. Dizge parçalama yapılmadan önce noktalama işaretleri, diyez, emoji ifadeler gibi karakterler temizlendikten sonra dizgelerine ayrılmış hali şu şekildedir; 'Bugün hava güzel değil mi'.

3.2.9.4.2. Durak Kelimeleri Çıkarma

Durak kelimeler her dilde sıkça kullanılan, ancak tek başına kullanıldığında anlam ifade etmeyen kelimelerdir. Bu kelimeleri kaldırmadaki amaç daha temiz bir veri seti elde etmek ve öznitelikleri doğru çıkarabilmektir. Ekler 1 ve Ekler 2'de durak kelimeleri liste halinde verilmiştir.

İngilizce durak kelimelere örnek olarak 'a, about, above, after, again, against, all, am, an, and, any, are, aren't, as, at, be, because, been, before, being, below, between…' verilebilir.

Türkçe durak kelimelerine örnek olarak ise 'acaba, ama, ancak, artık, bazen, bazı, bazıları, bazısı, belki, gene, gibi, hangi, hangisi, elbette, için, ile…'verilebilir.

3.2.9.4.3. Tekrarlanan Harflerin Çıkarılması

Removing Repetitive Letters olarak bilinen tekrarlanan kelimelerin çıkarılması doğal dil işleme çalışmalarında kullanılan bir yöntemdir. Kullanıcılar tarafından girilen var olandan uzun kelimeler, öznitelik uzay boyutunun artmasına neden olacağından bunların elenmesi ve anlamlı kelimeler üzerinde doğal dil işleme çalışmalarının yürütülmesi hedeflenmiştir. Tekrarlanan kelimelerin çıkarılmasında dikkate alınacak tekrar sayıları İngilizce metinlerde 2, Türkçe metinlerde ise 1'dir (Türkmenoğlu ve Tantuğ 2014). Tekrarlanan harflerin çıkarılmasına örnek olarak;

Nasılsıııııınnnn → Nasılsın Hayıııır → Hayır

Hellllllooooooo → Hello Babyyyy → Baby

3.2.9.4.4. Kök Alma

Doğal dil işlemelerinde temel amaç öznitelik uzay boyutunun olabildiğince azaltılması ve doğru sonuçların elde edilmesidir. Kök alma işlemi de bu sebeptendir. Çalışmamızda kök alma işlemleri python programlama dilindeki farklı kütüphaneler tarafından gerçekleştirildi. İngilizce veri setinin kök alma işlemi python programlama dilindeki en çok kullanılan kök alma kütüphanelerinden Porter Stemmer() kullanıldı. Türkçe veri setinin kök alma işlemini gerçekleştirmek için de python programlama dilinde kök alma kütüphanesi olarak geliştirilmiş Turkish Stemmer (Tunçelli 2015) kullanıldı.
Türkçe ve İngilizce kelimelerin köklerini almaya örnek aşağıda gösterilmiştir;
"love loving lovingly loved lover lovely love" → "love love love love lover love love"
"demir demirci demircilik demirler" → "demir demir demir demir"
"bilgisizlikten" → "bilgisizlik+ten" → "bilgisiz+lik" → "bilgi+siz" → "bil+gi" → "bil"

3.2.9.4.5. Geliştirilen Modeller

Çalışmamızda farklı metin temsili yöntemlerini kullanarak 3 farklı model geliştirilmiştir. Tüm temel sınıflandırıcılar aynı veri kümesinde eğitim almıştır, ancak farklı metin yöntemleri uygulanmıştır. Birinci modelde; Tf-Idf tabanlı BOW metin temsili yöntemiyle öznitelik çıkararak her bir makine öğrenmesi algoritması uygulanıp sonuçlar elde edilmiştir. İkinci modelde; Word2Vec metin temsili yöntemi uygulanmıştır. Ancak mesajlar kelime yerleştirmelerinin ortalamasıyla temsil edilmiştir. Üçüncü modelde; Word2Vec yöntemi kullanılmıştır. Fakat bu modelde mesajların ağırlıklı kelime yerleştirmelerinin ortalamasıyla temsil edilmiştir.
Sistemimiz ilk aşamasında Türkçe twitter mesajları ön işlemeden geçirildikten sonra kökleri alınıp metin temsili yöntemleriyle öznitelikleri çıkarıldıktan sonra makine öğrenmesi algoritmalarıyla sınıflandırılmasını içermektedir. Python programlama dilinde gerçekleştirilen bu çalışmada, kullanıcıların paylaşmış oldukları Türkçe twitter mesajlarında metin temsili yöntemlerini kullanarak duygu analizi çalışması gerçekleştirilmiştir.

Amaç, metin temsili yöntemlerinden terim frekansı tabanlı kelime çantası (Bag-of-Words) modeliyle anlamsal ilişki tabanlı Word2Vec modelinin duygu analizi çalışmalarında başarıya olan etkilerinin karşılaştırılmasıdır. Aşağıda geliştirmiş olduğumuz her bir modelin sözde kodu gösterilmiştir.

Birinci Model İçin Sözde Kod:

BAŞLA

- Veri Setini Temizle
- Temizlenmiş Veri Setinin Köklerini Al/Alma ve Dizgelerine Ayır
- En Az Bir Kelimeye Sahip Her Tweet Mesajı İçin Bir Korpus Oluştur
- Her Bir Tweet Mesajını Etiketle
- Her Bir Etiket Sayısını Kullanıcıya Göster
- Veri Setinin %20'sini Test Kümesi, %80'i Eğitim Kümesi Olarak Belirle
- Tf-Idf İle Ağırlıklandırılmış BOW Modeliyle Test Kümesini ve Eğitim Kümesini Kendi İçinde Eğit
- Doğrusal SVM, Doğrusal Regresyon, K En Yakın Komşu, Rastgele Orman ve Karar Ağacı İle Ayrı Ayrı Doğruluklarını(Accuracy) Hesapla
- Her Bir Etiket İçin Ayrı Ayrı Anma(Recall) Değerlerini Hesapla
- Tüm Etikler İçin Ortalama Anma Değerini Hesapla
- Hesaplanan Değerleri Kullanıcıya Göster

BİTİR

İkinci Model İçin Sözde Kod:

BAŞLA

- Veri Setini Temizle
- Temizlenmiş Veri Setinin Köklerini Al/Alma ve Dizgelerine Ayır
- Word2Vec İçin GoogleNews-vectors-negative300.bin Korpusunu Tanımla
- Her Bir Tweet Mesajını Etiketle
- Her Bir Etiket Sayısını Kullanıcıya Göster
- Korpus İle Birlikte Çalışacak Olan Word2Vec Modelini Oluştur
- Veri Setinin %20'sini Test Kümesi, %80'i Eğitim Kümesi Olarak Belirle
- Word2Vec İle Her Bir Kelime Yerleştirmelerinin Ortalamasını Al
- Word2Vec Modeliyle Test Kümesini ve Eğitim Kümesini Kendi İçinde Eğit

- Doğrusal SVM, Doğrusal Regresyon, K En Yakın Komşu, Rastgele Orman ve Karar Ağacı İle Ayrı Ayrı Doğruluklarını(Accuracy) Hesapla
- Her Bir Etiket İçin Ayrı Ayrı Anma(Recall) Değerlerini Hesapla
- Tüm Etikler İçin Ortalama Anma Değerini Hesapla
- Hesaplanan Değerleri Kullanıcıya Göster

BİTİR

Üçüncü Model İçin Sözde Kod:

BAŞLA

- Veri Setini Temizle
- Temizlenmiş Veri Setinin Köklerini Al/Alma ve Dizgelerine Ayır
- Word2Vec İçin GoogleNews-vectors-negative300.bin Korpusunu Tanımla
- En Az Bir Kelimeye Sahip Her Tweet Mesajı İçin Bir Korpus Oluştur
- Her Bir Tweet Mesajını Etiketle
- Her Bir Etiket Sayısını Kullanıcıya Göster
- Korpus İle Birlikte Çalışacak Olan Word2Vec Modelini Oluştur
- Veri Setinin %20'sini Test Kümesi, %80'i Eğitim Kümesi Olarak Belirle
- Tf-Idf İle Eğitim Kümesindeki ve Test Kümesindeki Her Kelimeyi Ağırlıklandır
- Word2Vec Modeliyle Test Kümesini ve Eğitim Kümesini Kendi İçinde Eğit
- Word2Vec İle Ağırlıklandırılmış Her Kelime Yerleştirmelerinin Ortalamasını Al
- Doğrusal SVM, Doğrusal Regresyon, K En Yakın Komşu, Rastgele Orman ve Karar Ağacı İle Ayrı Ayrı Doğruluklarını(Accuracy) Hesapla
- Her Bir Etiket İçin Ayrı Ayrı Anma(Recall) Değerlerini Hesapla
- Tüm Etikler İçin Ortalama Anma Değerini Hesapla
- Hesaplanan Değerleri Kullanıcıya Göster

BİTİR

Tez çalışmamızın ikinci aşamasında ise İngilizce ve Türkçe twitter mesajları üzerinde kök alma işleminin Word2Vec modeline olan etkisinin incelenmesidir. Her bir öznitelik çıkarımı için en iyi sınıflandırıcıyı seçme süreci, Doğrusal SVM ve Doğrusal Regresyon seçimi sınıflandırıcılarla bir Grid Search yapılarak gerçekleştirildi.

4. ARAŞTIRMA BULGULARI ve TARTIŞMA

Çalışmamızın ilk aşamasında, Türkçe veri setimize doğal dil işleme aşamalarından olan ön işleme ve kök alma işlemlerini uygulandıktan sonra verilerimizin özniteliklerini çıkararak makine öğrenmesi algoritmalarıyla kıyaslanmıştır. Başarı metriklerimiz, literatürde kullanılan (Çoban 2016) duygu analizinde sınıflandırma ve kümelemelerin ne kadar doğru olduğunu göstermek için ve bu doğrulukların da kaç tanesinin getirmesi gereken doğrulukla uyuştuğunu göstermek için "doğruluk(accuracy)" ve "ortalama anma(average recall)" olarak belirlenmiştir. Ayrıca elde edilen sonuçların her birinin saniye cinsinden ne kadar işlem süresi tuttuğu hesaplanarak karşılaştırılmıştır. Tez çalışmamızın ilk aşamasındaki 2. ve 3. modellerinde eğitilen Word2Vec modelinin ve uygulanılan kelime yerleştirme ve ağırlıklandırma işleminin ne kadar işlem süresi tuttuğu da hesaplanmıştır.

Çalışma sonucu Word2Vec modeliyle oluşturmuş olduğumuz 3. modelde en yüksek başarı yüzdesini %66,40 ile elde edilmiştir. Ayrıca geliştirmiş olduğumuz iki farklı Word2Vec modeli ile başarının ve duyarlılığın uyguladığımız algoritmaların çoğunda artırılabileceği gösterilmiştir.

Çizelge 4.1. İlk Aşamadaki 1. Model Sonuçları

Sınıflandırıcılar	Accuracy	Average Recall
BOW + Doğrusal SVM	**%65,18**	%60,14
BOW + Doğrusal Reg.	%64,10	%59,56
BOW + Karar Ağacı	%59,66	%54,51
BOW+ KEYKSA	%57,04	%55,73
BOW + Rastgele Orman	%63,12	%58,82

Çizelge 4.1'de gösterildiği gibi, birinci modelimizde Tf-Idf ile ağırlıklandırılmış BOW modelinde duygu analizi çalışmasında en yüksek başarı oranı Doğrusal SVM ile elde edilirken en düşük başarı oranı ise Karar Ağacı algoritması ile elde edilmiştir.

Çizelge 4.2. İlk Aşamadaki 1. Model İşlem Süreleri

Sınıflandırıcılar	İşlem Süreleri (Saniye)
BOW + Doğrusal SVM	4,701
BOW + Doğrusal Reg.	12,451
BOW + Karar Ağacı	11,087
BOW+ KEYKSA	8,781
BOW + Rastgele Orman	33,286

Çizelge 4.2'de gösterildiği gibi, Tf-Idf ile ağırlıklandırılmış BOW modelinde duygu analizi çalışmasında en düşük işlem süresi Doğrusal SVM ile elde edilirken en yüksek işlem süresi ise Rastgele Orman algoritması ile elde edilmiştir. İlk aşamadaki 1. modelde en yüksek başarı oranı ve en düşük işlem süresi Doğrusal SVM ile elde edilmiştir.

Çizelge 4.3. İlk Aşamadaki 2. Model Sonuçları

Sınıflandırıcılar	Accuracy	Average Recall
Word2Vec+Doğrusal SVM	%51,20	%50,10
Word2Vec+ Doğrusal Reg.	%63,14	%62,72
Word2Vec+Karar Ağacı	**%66,12**	%65,47
Word2Vec+ KEYKSA	%65,77	%65,15
Word2Vec+Rastgele Orman	%65,92	%65,75

Çizelge 4.3'de gösterildiği gibi, kelime yerleştirmelerinin ortalamalarının alındığı Word2Vec ile ikinci modelimizde duygu analizi çalışmasında en yüksek başarı oranı Karar Ağacı ile elde edilirken en düşük başarı oranı ise Doğrusal SVM algoritmasında elde edilmiştir.

Çizelge 4.4. İlk Aşamadaki 2. Modelin İşlem Süreleri

Sınıflandırıcılar	İşlem Süreleri (Saniye)
Word2Vec+Doğrusal SVM	3,954
Word2Vec+ Doğrusal Reg.	0,257
Word2Vec+Karar Ağacı	2,273
Word2Vec+ KEYKSA	13,654
Word2Vec+Rastgele Orman	1,383

Çizelge 4.4'de gösterildiği gibi, kelime yerleştirmelerinin ortalamalarının alındığı Word2Vec ile ikinci modelimizde duygu analizi çalışmasında en düşük işlem süresi Doğrusal Regresyon ile en yüksek işlem süresi ise KEYKSA ile elde edilmiştir. Doğrusal Regresyon hızlı işlem süresi ve en yüksek başarı oranına yakın bir başarı oranı sunmuştur.

Çizelge 4.5. İlk Aşamadaki 3. Model Sonuçları

Sınıflandırıcılar	Accuracy	Average Recall
Word2Vec+Doğrusal SVM	%51,28	%50,10
Word2Vec+ Doğrusal Reg.	%63,51	%62,86
Word2Vec+Karar Ağacı	%66,18	%65,48
Word2Vec+ KEYKSA	%65,92	%65,30
Word2Vec+Rastgele Orman	**%66,40**	%65,31

Çizelge 4.5'de gösterildiği gibi, Tf-Idf ile ağırlıklandırılmış kelime yerleştirmelerinin ortalamasının alındığı Word2Vec ile üçüncü modelimizde duygu analizi çalışmasında en yüksek başarı oranını Rastgele Orman ile elde edilirken en düşük başarı oranı ise Doğrusal SVM algoritmasında elde edilmiştir.

Doğrusal SVM algoritmasının Word2Vec modelinde düşük başarı sonuçlarını verdiği gösterilmiştir. Rastgele Orman ve Karar Ağacı algoritmalarının ise Word2Vec ile modellenen duygu analizi çalışmalarında birbirlerine yakın yüksek sonuçları üretebileceği gösterilmiştir.

Çizelge 4.6. İlk Aşamadaki 3. Modelin İşlem Süreleri

Sınıflandırıcılar	İşlem Süreleri (Saniye)
Word2Vec+Doğrusal SVM	3,911
Word2Vec+ Doğrusal Reg.	0,232
Word2Vec+Karar Ağacı	2,916
Word2Vec+ KEYKSA	13,572
Word2Vec+Rastgele Orman	1,500

Çizelge 4.6'da gösterildiği gibi, Tf-Idf ile ağırlıklandırılmış kelime yerleştirmelerinin ortalamalarının alındığı Word2Vec modeli ile üçüncü modelimizde duygu analizi çalışmasında en düşük işlem süresi Doğrusal Regresyon ile en yüksek işlem süresi ise KEYKSA ile elde edilmiştir. Rastgele Orman algoritması ile hızlı işlem süresi ve en yüksek başarı oranı elde edilmiştir.

Çizelge 4.7. İkinci Aşamadaki İngilizce Veri Seti Sonuçları

	Kökleri Alınmamış İngilizce Veri Seti	**Kökleri Alınmış İngilizce Veri Seti**
Sistem Modelleri	**Average Recall**	**Average Recall**
BOW+ Doğrusal SVM	57,91%	56,66%
Word2Vec+ Doğrusal SVM	54,78%	52,35%
Word2Vec+ Doğrusal Regresyon	52,78%	50,00%

Çizelge 4.7'de gösterildiği gibi, çalışmamızın ikinci aşamasında ise, veri setlerimizi değerlendirmek, karşılaştırmak ve kıyaslamak için ortalama geri çağırım(average recall) kullanılmıştır. Yaptığımız çalışma sonucu, İngilizce twitter mesajlarında kök alma işlemi gerçekleştirildiğinde başarı yüzdelerinin düştüğü gözlemlenmiştir ve en yüksek başarı yüzdesi BOW+Doğrusal SVM modelinde %57,9 ile elde edilmiştir.

Çizelge 4.8. İkinci Aşamadaki İngilizce Veri Seti İşlem Süreleri

	Kökleri Alınmamış İngilizce Veri Seti	**Kökleri Alınmış İngilizce Veri Seti**
Sistem Modelleri	**İşlem Süresi (sn)**	**İşlem Süresi(sn)**
BOW+ Doğrusal SVM	9,685	9,020
Word2Vec+ Doğrusal SVM	6,742	8,401
Word2Vec+ Doğrusal Regresyon	5,196	4,624

Çizelge 4.8'de gösterildiği gibi, çalışmamızın ikinci aşamasında İngilizce veri setinde kök alma işlemi sonucu elde edilen başarı yüzdelerinin işlem süreleri gösterilmiştir. Kök alma işleminde Word2Vec+Doğrusal SVM ile başarı yüzdesi düşerken işlem süresinin arttığı, diğer sonuçlarda hem başarı yüzdelerinin hem de işlem sürelerinin düştüğü gözlemlenmiştir.

Çizelge 4.9. İkici Aşamadaki Türkçe Veri Seti Sonuçları

	Kökleri Alınmamış Türkçe Veri Seti	**Kökleri Alınmış Türkçe Veri Seti**
Sistem Modelleri	**Average Recall**	**Average Recall**
BOW+ Doğrusal SVM	60,47%	60,14%
Word2Vec+ Doğrusal SVM	50,10%	50,00%
Word2Vec+ Doğrusal Regresyon	50,28%	62,86%

Çizelge 4.9'da gösterildiği gibi, Türkçe twitter mesajlarında kök alma işlemi gerçekleştirildikten sonra, başarı yüzdelerinin sadece ilk iki modelde düştüğünü fakat son model olan Word2Vec+Doğrusal Regresyon'da %62,8 ile başarı yüzdesinin arttığı gözlemlenmiştir. Başarı metrik değerimizin %50 ve üzeri gelmesi, sunulan sistemin başarılı sonuçlar ürettiğini göstermektedir.

Çizelge 4.10. İkinci Aşamadaki Türkçe Veri Seti İşlem Süreleri

	Kökleri Alınmamış Türkçe Veri Seti	**Kökleri Alınmış Türkçe Veri Seti**
Sistem Modelleri	**İşlem Süresi (sn)**	**İşlem Süresi(sn)**
BOW+ Doğrusal SVM	4,895	4,701
Word2Vec+ Doğrusal SVM	3,970	3,954
Word2Vec+ Doğrusal Regresyon	0,295	0,232

Çizelge 4.10'da gösterildiği gibi, çalışmamızın ikinci aşamasında Türkçe veri setinde kök alma işlemi sonucu elde edilen başarı yüzdelerinin işlem süreleri gösterilmiştir. Türkçe metinlerde kök alma işleminin elde edilen sonuçlarda işlem süresini azaltacağı gösterilmiştir.

Çizelge 4.11. İlk Aşamadaki 2. Modelinin Word2Vec İşlem Süreleri

Yapılan İşlem	İşlem Süreleri (Saniye)
Türkçe Word2Vec Modelinin Eğitim Süresi	21,365
2.Modelde Kelime Yerleştirmelerinin Ortalamasının Toplam Süresi	0,281
2. Modelde Word2Vec Modelinin ve Kelime Yerleştirmelerinin Ortalamasının Toplam Süresi	21,689

Çizelge 4.11'de gösterildiği gibi, kelime yerleştirmelerinin ortalamalarının alındığı Word2Vec ile ikinci modelimizde duygu analizi çalışmasında Türkçe derlem ile kelimelerin eğitiminin işlem süresi, kelime yerleştirmelerinin ortalamasının toplam işlem süresi ve ikisinin birlikte aldığı toplam işlem süresi verilmiştir.

Çizelge 4.12. İlk Aşamadaki 3. Modelinin Word2Vec İşlem Süreleri

Yapılan İşlem	İşlem Süreleri (Saniye)
Türkçe Word2Vec Modelinin Eğitim Süresi	21,365
3.Modelde Ağırlıklandırılmış Kelime Yerleştirmelerinin Ortalamasının Toplam Süresi	8,316
3. Modelde Word2Vec Modelinin ve Ağırlıklandırılmış Kelime Yerleştirmelerinin Ortalamasının Toplam Süresi	29,589

Çizelge 4.12'de gösterildiği gibi, Tf-Idf ile ağırlıklandırılmış kelime yerleştirmelerinin ortalamalarının alındığı Word2Vec ile üçüncü modelimizde duygu analizi çalışmasında Türkçe derlem ile kelimelerin eğitiminin işlem süresi, ağırlıklandırılmış kelime yerleştirmelerinin ortalamasının toplam işlem süresi ve ikisinin birlikte aldığı toplam işlem süresi verilmiştir.

Ağırlıklandırma işleminin ikinci modelde word2vec modelinin eğitim süresinin sabit kaldığı, son iki işlem süresini arttırdığı gösterilmiştir. Ayrıca ağırlıklandırma işlemi yaparak ikinci modelde elde edilen en yüksek başarı oranı artırılıken tüm sonuçların işlem süresinin arttığı gözlemlenmiştir.

Çizelge 4.13. İkinci Aşamadaki 2. Modelinin Word2Vec İşlem Süreleri

Yapılan İşlem	İşlem Süreleri (Saniye)
İngilizce Word2Vec Modelinin Eğitim Süresi	28,885
2.Modelde Kelime Yerleştirmelerinin Ortalamasının Toplam Süresi	3,641
2. Modelde Word2Vec Modelinin ve Kelime Yerleştirmelerinin Ortalamasının Toplam Süresi	32,185

Çizelge 4.13'de gösterildiği gibi, kelime yerleştirmelerinin ortalamalarının alındığı Word2Vec ile ikinci modelimizde duygu analizi çalışmasında İngilizce derlem ile kelimelerin eğitiminin işlem süresi, kelime yerleştirmelerinin ortalamasının toplam işlem süresi ve ikisinin birlikte aldığı toplam işlem süresi verilmiştir.

Çizelge 4.14. İkinci Aşamadaki 3. Modelinin Word2Vec İşlem Süreleri

Yapılan İşlem	İşlem Süreleri (Saniye)
Türkçe Word2Vec Modelinin Eğitim Süresi	28,885
3.Modelde Ağırlıklandırılmış Kelime Yerleştirmelerinin Ortalamasının Toplam Süresi	22,007
3. Modelde Word2Vec Modelinin ve Ağırlıklandırılmış Kelime Yerleştirmelerinin Ortalamasının Toplam Süresi	50,040

Çizelge 4.14’de gösterildiği gibi, Tf-Idf ile ağırlıklandırılmış kelime yerleştirmelerinin ortalamalarının alındığı Word2Vec ile üçüncü modelimizde duygu analizi çalışmasında İngilizce derlem ile kelimelerin eğitiminin işlem süresi, ağırlıklandırılmış kelime yerleştirmelerinin ortalamasının toplam işlem süresi ve ikisinin birlikte aldığı toplam işlem süresi verilmiştir. Ağırlıklandırma işleminin üçüncü modelde word2vec modelinin eğitim süresinin sabit kaldığı, son iki işlem süresini arttırdığı gösterilmiştir. Ayrıca ağırlıklandırma işlemi yaparak üçüncü modelde elde edilen en yüksek başarı oranı artırılıken tüm sonuçların işlem süresinin arttığı gözlemlenmiştir.

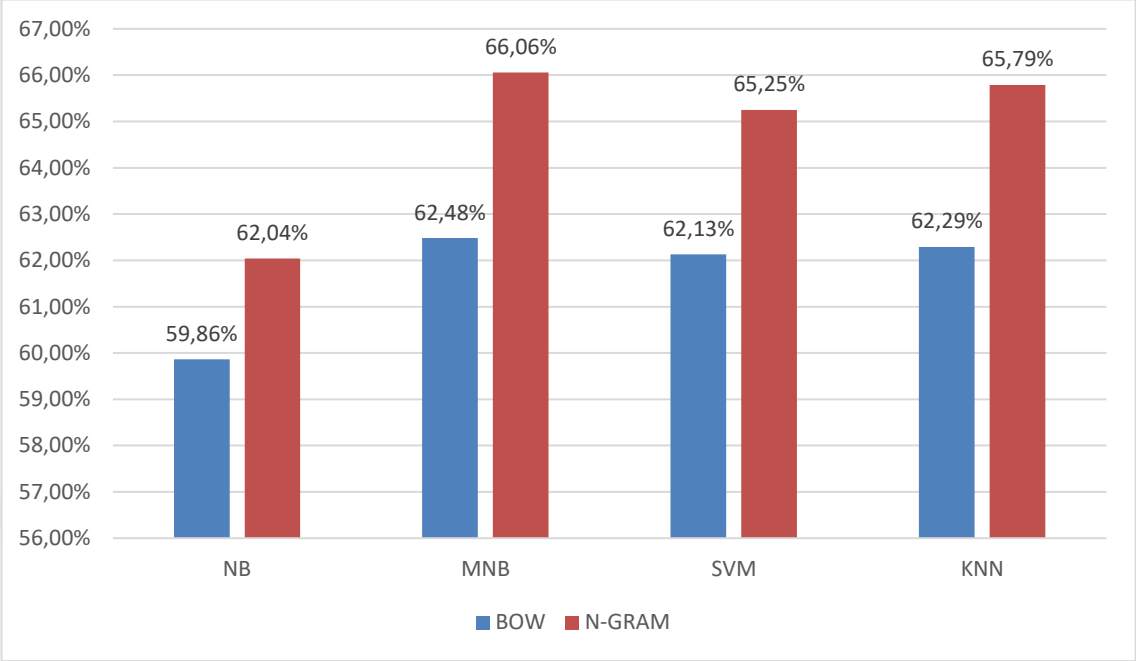
5. SONUÇ ve ÖNERİLER

Türkçe doğal dil işleme çalışmalarında python programlama dilinin sahip olduğu güçlü nltk kütüphanesi sayesinde duygu analizi çalışmasını gerçekleştirdiğimiz çalışmamızın ilk aşamasında, sosyal medya platformlarından biri olan twitter üzerine duygu analizi çalışması yapmak için farklı metin temsillerini kullanarak başarı yüzdeleri karşılaştırılmıştır. Anlamsal ilişkiye dayalı olan Word2Vec modelini kullanarak daha iyi sonuçların elde edilebileceği ve duygu analizi çalışmalarında sınıflandırıcıların etkili bir şekilde kullanılabileceği gösterilmiş. Ayrıca kullanılan Word2Vec modelini kendi içinde geliştirerek daha iyi sonuçların üretilebileceği de gösterilmiştir.

Literatürde yapılan çalışmalardan Şekil 5.1'de ön işlemden geçirilen Türkçe twitter mesajlarından çıkarılan özniteliklere ait istatistiksel sonuçlar gösterilmiştir. Öznitelik çıkarma sırasında BOW modeli için uygulanan kök alma, durak kelimelerin çıkarılması ve tekrarlanan harflerin çıkarılması işlemleri N-Gram model için uygulanmamış olup BOW modeline uygulanmıştır (Çoban vd. 2015). Yaptıkları çalışmada öznitelik çıkarma yöntemlerinin başarı değerlerini ölçmek için kullandıkları makine öğrenmesi yöntemlerinden her iki model için de en iyi sonucu Multinom Naive Bayes vermiştir. SVM sınıflandırmasında doğrusal çekirdek kullanılmış ve KEYKSA sınıflandırmasında en yakın komşu sayısı k=1 alınmıştır (Çoban vd.2015).

Çoban ve arkadaşlarının yapmış oldukları çalışmada (Çoban vd., 2015) BOW modelinde en yüksek başarı %62,48 ile elde edilmiştir. Ancak tez çalışmamızda uygulanılan BOW modeliyle başarı yüzdesini %4,32 arttırarak BOW modelinde en yüksek başarı yüzdesini %65,18 ile elde edilmiştir.

Geliştirmiş olduğumuz 3. modelde ayrıca elde ettikleri en yüksek başarı yüzdesini daha da arttırarak %66,40 ile elde edilmiştir. Word2Vec modelinin uygulanmasıyla başarı yüzdesinin artırılabileceğini göstermiş olduk.



Şekil 5.1. Öznitelik Çıkarma Yöntemleri Sonucu Sınıflandırma Başarıları (Çoban vd.2015)

Şekil 5.1'de gösterildiği gibi, Çoban ve arkadaşlarının yapmış oldukları çalışmada kullanmış oldukları öznitelik çıkarma yöntemlerinden BOW modelinde en yüksek başarı oranı %62,48 ile elde edilirken, Çizelge 5.1'de gösterildiği gibi, kendi tez çalışmamızda bu başarı oranını artırarak %65,18 elde edilmiştir. Ayrıca bu veri setine uygalanmayan tezimizde kullandığımız Word2Vec modeli uygulandığında en yüksek başarı oranı %66,40 ile elde edilmiştir.

Çizelge 5.1. İlk Aşamadaki En Yüksek Başarı Oranları

Sınıflandırıcılar	Accuracy	Average Recall
Birinci Model (BOW + Doğrusal SVM)	%65,18	%60,14
İkinci Model (Word2Vec+ KEYKSA)	%65,92	%65,30
Üçüncü Model (Word2Vec+Rastgele Orman)	**%66,40**	%65,31

Çizelge 5.2. İlk Aşamadaki En Yüksek Başarı Oranlarının İşlem Süreleri

Sınıflandırıcılar	İşlem Süreleri (Saniye)
Birinci Model (BOW + Doğrusal SVM)	4,701
İkinci Model (Word2Vec+ KEYKSA)	2,273
Üçüncü Model (Word2Vec+Rastgele Orman)	**1,500**

Çizelge 5.1 ve Çizelge 5.2'de gösterildiği gibi Türkçe twitter veri seti için yapmış olduğumuz duygu analizi çalışmasında sırasıyla geliştirdiğimiz her bir modelde başarı oranları artmıştır ve işlem süreleri azalmıştır. Her bir model için en yüksek başarı oranlarına karşılık gelen işlem sürelerini karşılaştıracak olursak eğer üçüncü modelde Word2Vec+Rastgele Orman ile 1,5 sn. işlem süresiyle elde edilmiştir.

Çalışmamızın ikinci aşamasında, twitter mesajlarında duygu analizini gerçekleştirmek için kelime temsili yöntemlerini kullanarak, kök alma işleminin Word2Vec modeline olan etkisi incelenmiştir. Yaptığımız çalışma sonucu kök alma işleminin İngilizce metinler üzerinde kelime yerleştirmeleri uygulandığında başarıyı düşürdüğü gözlemlenmiştir. Diğer bir yandan, Türkçe metinler için sunmuş olduğumuz ilk iki modelde kök alma işleminin başarıyı düşürdüğü, ancak Türkçe'nin sondan eklemeli(agglutinative) bir dil olmasına rağmen kök alma işleminin son modelde başarıyı arttırdığı tespit edilmiştir. Böylece kök alma işlemi uygulanarak Word2Vec modeliyle İngilizce metinlerde başarının azalabileceğini, Türkçe metinlerde ise uygulanılan modele göre başarının arttırılabileceğini gösterilmiştir.

Tez çalışmamızın ikinci aşamasında İngilizce ve Türkçe veri setlerinde kök alma işlemi sonucu elde edilen başarı yüzdelerinin işlem süreleri karşılaştırılmıştır. İngilizce veri setinde kök alma işleminde Word2Vec+Doğrusal SVM ile başarı yüzdesi düşerken işlem süresinin arttığı, diğer sonuçlarda hem başarı yüzdelerinin hem de işlem sürelerinin düştüğü gözlemlenmiştir. Türkçe metinlerde ise kök alma işleminin elde edilen sonuçlarda işlem süresini azaltacağı gösterilmiştir.

Her iki aşamadaki elde edilen başarı yüzdelerine karşılık elde edilen işlem sürelerinin karşılaştırılması yapılmıştır. İlk aşamadaki 1. modelde en yüksek başarı oranı ve en düşük işlem süresi Doğrusal SVM ile elde edilmiştir. İlk aşamadaki 2. modelde en düşük işlem süresi Doğrusal Regresyon ile en yüksek işlem süresi ise KEYKSA ile elde edilmiştir. Doğrusal Regresyon hızlı işlem süresi ve en yüksek başarı oranına yakın bir başarı oranı sunmuştur. İlk aşamadaki 3. modelde en düşük işlem süresi Doğrusal Regresyon ile en yüksek işlem süresi ise KEYKSA ile elde edilmiştir. Rastgele Orman algoritması ile hızlı işlem süresi ve en yüksek başarı oranı elde edilmiştir.

Çizelge 5.3. Veri Setlerinde Kullanılan Word2Vec Modellerinin İşlem Süreleri

Türkçe Veri Setinde Yapılan İşlem	İşlem Süreleri (Saniye)	İngilizce Veri Setinde Yapılan İşlem	İşlem Süreleri (Saniye)
Türkçe Word2Vec Modelinin Eğitim Süresi	21,365	İngilizce Word2Vec Modelinin Eğitim Süresi	28,885
2. ve 3. Modelde Ağırlıklandırılmış Kelime Yerleştirmelerinin Ortalamasının Toplam Süresi	0,281→8,316	2. ve 3. Modelde Ağırlıklandırılmış Kelime Yerleştirmelerinin Ortalamasının Toplam Süresi	3,641→22,007
2. ve 3. Modelde Word2Vec Modelinin ve Ağırlıklandırılmış Kelime Yerleştirmelerinin Ortalamasının Toplam Süresi	21,689→29,589	2. ve 3. Modelde Word2Vec Modelinin ve Ağırlıklandırılmış Kelime Yerleştirmelerinin Ortalamasının Toplam Süresi	32,185→50,040

Çizelge 5.3'de gösterildiği gibi, 416051 kelimeye sahip "wiki-tr" derlemi ile 3 milyon kelimeye sahip "GoogleNews-vectors-negative300" derleminin çalıştırılması sonucu işlem süreleri karşılaştırılmıştır. Derlemlerin boyutları 300 boyutlu olup birbirine eşit oldukları halde İngilizce derlemin sahip olduğu geniş kelime kapasitesinden dolayı daha çok işlem süresi almıştır. Ayrıca derlemdeki kelime sayısı arttıkça uygulanılan her yöntemde de işlem süresinin arttığı gösterilmiştir.

Duygu analizi çalışmaları birçok dilde gerçekleştirilmiştir. Bu çalışma alanında geliştirilmiş İngilizce derlemler ve proramlama dillerinin İngilizceye olan uygunluğu nedeniyle en yüksek doğruluk yüzdeleri İngilizcede elde edilmiştir. Diğer Avrupa dilleri de İngilizceyle benzer dil yapısına sahip olduğundan birbirine yakın doğruluk yüzdeleri elde edilmiştir. Türkçe, morfolojik yapı olarak sondan eklemeli bir dil olduğundan ve özne, yüklem, tümleç dizilişinin diğer Avrupa dillerinden farklı olduğundan dolayı Avrupa dilleri için uygulanmış olan bazı yöntemler, Türkçe için uygulandığında sınırlı düzeyde başarı sağlanmıştır (Akba 2014).

Gelecekteki çalışmalarımızda daha iyi sonuçlar elde etmek için aşağıdaki çalışmaları yapmayı hedeflemekteyiz;

- ✓ Daha büyük bir veri seti üzerinde çalışmak ve verilerin etiket sayısını arttırmak.
- ✓ Belirli bir anahtar kelimeye özgü çekilmiş veri setlerinde çalışmak.
- ✓ Türkçe doğal dil işleme çalışmalarında kullanılabilecek daha büyük boyutlu bir derlem geliştirmek.
- ✓ Pennington ve arkadaşlarının geliştirmiş olduğu Glob2Vec (Pennington vd. 2014) modeliyle modelimizi uygulayarak sistemimizi geliştirmek.

KAYNAKLAR

Adalı, E. (2012). Doğal Dil İşleme (Natural Language Processing). TÜRKİYE BİLİŞİM VAKFI BİLGİSAYAR BİLİMLERİ ve MÜHENDİSLİĞİ DERGİSİ, 5(2 (Basılı 6).Akbulut, S., " Veri madenciliği teknikleri ile bir kozmetik markanın ayrılan müşteri analizi ve müşteri segmentasyonu", Yüksek lisans tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, (2006).

Akba, F. (2014). Duygu Analizinde Öznitelik Seçme Metriklerinin Değerlendirilmesi: Türkçe Film Eleştirileri. Yüksek Lisans Tezi, Ankara

Aksan, C. E., 17 Mart 2018. API Nedir? Ne İşe Yarar?. https://ceaksan.com/tr/api-nedir/

Akgöbek, Ö. ve Çakır, F., (2009), "Veri Madenciliğinde Bir Uzman Sistem Tasarımı", Akademik Bilişim 09, 11-13 Şubat Harran Üniversitesi, Şanlıurfa, 801-806.

Albayrak, M., (2008), EEG Sinyallerindeki Epileptiform Aktivitenin Veri Madenciliği Süreci ile Tespiti, Doktora Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü.

Albayrak, N. B. (2011). Opinion and sentiment analysis using natural language processing techniques (Doctoral dissertation, Master's thesis, Fatih University).

Alghamdi, R., & Alfalqi, K. (2015). A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl.(IJACSA), 6(1).

Anonim, Sosyal medya nedir?. http://www.yazilimnet.com/tr/blog/12/sosyal-medya-nedir-

Aslam, S., 1 Ocak 2018. Twitter by the Numbers: Stats, Demographics & Fun Facts. https://www.omnicoreagency.com/twitter-statistics/ .

ATALAY, M., & ÇELİK, E. BÜYÜK VERİ ANALİZİNDE YAPAY ZEKÂ VE MAKİNE ÖĞRENMESİ UYGULAMALARI-ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING APPLICATIONS IN BIG DATA ANALYSIS. Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 9(22), 155-172.

Atmaca, K., 6 Mart 2017. KNN (K nearest neigborhood) algoritması, http://kenanatmaca.com/knn-k-nearest-neigborhood-algoritmasi/

Bavaş, E., 6 Ağustos 2017. Metin Madenciliğinde TF-IDF Kullanımı. http://erdoganb.com/2017/06/metin-madenciliginde-tf-idf-kullanimi/

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Boynukalin, Z. (2012). Emotion analysis of Turkish texts by using machine learning methods. M. Sc. Thesis, Middle East Technical University Ankara.

Brownlee, J., 16 Nisan 2014. A Gentle Introduction to Scikit-Learn: A Python Machine Learning Library. https://machinelearningmastery.com/a-gentle-introduction-to-scikit-learn-a-python-machine-learning-library/

Burhan, E., 20 Temmuz 2017. Derin Öğrenme (Deep Learning) Nedir?. https://www.slideshare.net/eburhan/derin-renme-deep-learning-nedir .

Can, Ü., & Alatas, B. (2017). Duygu Analizi ve Fikir Madenciliği Algoritmalarının İncelenmesi. Int. J. Pure Appl. Sci, 3(1), 75-111.

Chomsky, N. A., Syntactic Structures, Mouton, The Hauge, 1957.

Çelik, H. (2013). Sentiment analysis for Turkish Language (Doctoral dissertation, İstanbul Kültür Üniversitesi/Fen Bilimleri Enstitüsü/Bilgisayar Mühendisliği Anabilim Dalı).

Çınar, A. C., 21 Mayıs 2017. Gözetimli (Supervised) ve Gözetimsiz (unsupervised) öğrenme (learning) nedir? http://ahmetcevahircinar.com.tr/2017/05/21/gozetimli-supervised-ve-gozetimsiz-unsupervised-ogrenme-learning-nedir/ .

Çoban, Ö., Özyer, B., & Özyer, G. T. (2015, May). Sentiment analysis for Turkish Twitter feeds. In Signal Processing and Communications Applications Conference (SIU), 2015 23th (pp. 2388-2391). IEEE.

Çoban, Ö. (2016). METİN SINIFLANDIRMA TEKNİKLERİ İLE TÜRKÇE TWITTER DUYGU ANALİZİ. Yüksek Lisans Tezi, Erzurum

Dawley, L. (2009). Social network knowledge construction: Emerging virtual world pedagogy. On the Horizon, 17(2), 109-121

Delibas, A. (2008). Doğal Dil İşleme İle Türkçe Yazım Hatalarının Denetlenmesi (Doktora Tezi, Fen Bilimleri Enstitüsü).

Demner-Fushman, D. and Lin, J., Knowledge Extraction For Clinical Question Answering: Preliminary Results, Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains, July 2005, Pennsylvania, 9-13.

Dolgun, M. Ö., Özdemir, T. G., & Oğuz, D. (2009). Veri madenciliğiâ nde yapısal olmayan verinin analizi: Metin ve web madenciliği. İstatistikçiler Dergisi: İstatistik ve Aktüerya, 2(2).

Doğan, Ş., ve Türkoğlu,İ., (2007), " Hypothyroidi and Hyperthyroidi Detection from Thyroid Hormone Parameters by Using Decision Trees", Doğu Anadolu Bölgesi Araştırmaları Dergisi, Cilt 5, No 2, 163-169.

Dwivedi, V. P., 2 Haziran 2017. Can word2vec be considered deep learning?. https://www.quora.com/Can-word2vec-be-considered-deep-learning .

Elmas, M. (2012). Destek Vektör Makineleri ile Fiyat Tahminleri ve Kuyumculuk Sektöründe Bir Uygulama, Yüksek Lisans Tezi İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Filiz, F., 24 Mart 2017. Yapay Zeka Algoritmaları. https://medium.com/@fahrettinf/4-1-yapay-zeka-algoritmalar%C4%B1-fd64b8080748 .

Fidancı, A. S., 23 Ekim 2017. Random Forest Algoritması, https://www.slideshare.net/SezerFidanc/random-forest-algoritmas

Fürnkranz, J. (1998). A study using n-gram features for text categorization. Austrian Research Institute for Artifical Intelligence, 3(1998), 1-10.

Go, A., Bhayani, R., Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1, 12.

Grave, E, 3 May 2017. Pre-trained word vectors. https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

Güvenç, B. 2013. Machine learning methods in natural language processing, Yüksek Lisans Tezi, 2013, İstanbul

Hand, D.J., (1998), "Data Mining: Statistics and More?", The American Statistician, Cilt 52, 112-118.

Hilbe, J. M. (2011). Logistic regression. In International Encyclopedia of Statistical Science (pp. 755-758). Springer Berlin Heidelberg.

Hofmann, T., ―Unsupervised learning by probabilistic latent semantic analysis‖, Machine Learning, 42 (1), 2001, 177- 196.

Hu, M., Sun, A., and Lim, E. P., ―Comments Oriented Blog Summarization by Sentence Extraction‖, Proceedings of the Conference on Information and Knowledge Manegement, Portugal, November 2007, pp 901-904.

İnan, O., (2003), Veri Madenciliği, Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü.

İskender, E. (2016). Sosyal Medya Mesajlarında Müşteri Memnuniyetinin Fuzzy Sentiment Analizi İle Ölçülmesi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, sayfa. 6-7

Jacobs, P., (1999), "Data Mining: What General Managers Need to Know", Harvard Management Update, Cilt 4, No 10, 8.

Jin, X., Li, Y., Mah, T., and Tong, J., ―Sensitive Webpage Classification for Content Advertising‖, Proceedings of the First International Workshop on Data Mining and Audience Intelligence for Advertising, California, August 2007, pp 28-33.

Joshi, R., 9 Eylül 2016. Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures, http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

Jśnior, E. A. C., Marinho, V. Q., & dos Santos, L. B. (2017). NILC-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 611-615).

Kavzoğlu, T., & Çölkesen, İ. (2010). Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi. Harita Dergisi, 144(7), sayfa. 73-82.

Kılınç, D., Borandağ, E., Yücalar, F., Tunalı, V., Şimşek, M., & Özçift, A. (2016). KNN algoritması ve r dili ile metin madenciliği kullanılarak bilimsel makale tasnifi.

KIM, W., Jeong, O-R., Lee, S-W. (2010). "On Social Web Sites", Information Systems 35, 215-236, journal homepage: www.elsevier.com/locate/infosys.

Kırmacı B. (2015). Müzik Üst-Veri Tahmini için Türkçe Şarkı Sözü Madenciliği, Yüksek Lisans Tezi, Başkent Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul

Kim, Y., 21 Dec 2017. Twitter-Sentiment-Analysis. https://raw.githubusercontent.com/whyjay17/Twitter-Sentiment-Analysis/master/data/twitter-2016test-A.txt 10.01.2018.

Kobalas, T. 26 Mayıs 2009. NLTK Türkçe, http://nltk-turkce.blogspot.com.tr/

Kotan, A., Wed 01 October 2014. Python ile Twitter Kullanmak - Tweepy Modülü (Twitter API). http://ahmetkotan.com.tr/python-ile-twitter-kullanmak-tweepy-modulu-twitter-api.html 15.03.2018.

Köksal, Y., & Özdemir, Ş. (2013). Bir iletişim aracı olarak sosyal medya'nın tutundurma karması içerisindeki yeri üzerine bir inceleme. Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 18(1).

Kurkcu, A., Jul 30, 2014. Sosyal Medya Madenciligi—Python & Twitter. https://medium.com/@abdullahkurkcu/sosyal-medya-madenciligi-python-twitter-475166a9f2e1 18.03.2018.

Küçükönder, H., Vursavuş, K. K., & Üçkardeş, F. (2015). Determining The Effect of Some Mechanical Properties on Color Maturity of Tomato With K-Star, Random Forest and Decision Tree (C4. 5) Classification Algorithms. Turkish Journal of Agriculture-Food Science and Technology, 3(5), 300-306.

Liu, J., Cao, Y., Lin, C. Y., Huang, Y., and Zhou, M., ―Low Quality Product Review Detection in Opinion Summarization‖, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, June 2007, pp 334-342.

Maini, V., 19 Ağustos 2017. Machine Learning for Humans, Part 2.1: Supervised Learning. https://medium.com/machine-learning-for-humans/supervised-learning-740383a2feab .

Miháltz, M. 12 Mayıs 2016. word2vec-GoogleNews-vectors, https://github.com/mmihaltz/word2vec-GoogleNews-vectors

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016). SemEval-2016 task 4: Sentiment analysis in Twitter. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 1-18).

Ok, A. Ö., Akar, Ö., & Güngör, O. RASTGELE ORMAN SINIFLANDIRMA YÖNTEMĠ YARDIMIYLA TARIM ALANLARINDAKĠ ÜRÜN ÇEĢĠTLĠLĞĞĞNĠN SINIFLANDIRILMASI.

Ortigosa, A., Martin, J.M. ve Carro, M. R. (2013) Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior, sayfa.1-15.Takcı, H., 6 Ağustos 2013. DUYGU ANALİZİ (SENTIMENT ANALYSIS), http://verimadencisi.blogspot.com.tr/2013/08/duygu-analizi-sentiment analysis.html

Özaytaç, M. 24 Şubat 2018. Python'da Popüler Makine Öğrenmesi Kütüphaneleri, http://www.kodblog.net/2018/02/24/python-kutuphaneleri/

Özçakır, F. C., " Müşteri işlemlerindeki birlikteliklerin belirlenmesinde veri madenciliği uygulaması", Yüksek lisans tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü, İstanbul (2006).

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Qiang, J., Chen, P., Wang, T., & Wu, X. (2017, May). Topic Modeling over Short Texts by Incorporating Word Embeddings. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 363-374). Springer, Cham.

Ren, Y., Wang, R., & Ji, D. (2016). A topic-enhanced word embedding for twitter sentiment classification. Information Sciences, 369, 188-198.

Rong, X. (2014). word2vec parameter learning explained. arXiv preprint arXiv:1411.2738.

Ruder, S., 11 Nisan 2016. On word embeddings - Part 1, http://ruder.io/word-embeddings-1/

Sanjeevi, M., 6 Ocak 2017. Decision Trees Algorithms. https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1

Savaş, S., Topaloğlu, N., & Yılmaz, M. (2012). Veri madenciliği ve Türkiye'deki uygulama örnekleri. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 11(21), 1-23.

Şeker, Ş. E., 27 Ekim 2008. Gözetimsiz Öğrenme (Unsupervised Learning), http://bilgisayarkavramlari.sadievrenseker.com/2008/10/27/gozetimsiz-ogrenme-unsupervised-learning/

Şeker, Ş. E., 27 Ekim 2008. Gözetimli Öğrenme (Supervised Learning), http://bilgisayarkavramlari.sadievrenseker.com/2008/10/27/gozetimli-ogrenme-supervised-learning/ .

Şeker, Ş.E., 15 Haziran 2014. Metin Madenciliği (Text Mining). http://bilgisayarkavramlari.sadievrenseker.com/2014/06/15/metin-madenciligi-text-mining/

Şeker, Ş. E., 9 Ağustos 2015. Doğal Dil İşleme (Natural Language Processing) (Veri Bilimi Serisi 40). https://www.youtube.com/watch?v=a9I6MXgn8B4 .

Şeker, Ş. E., Duygu Analizi(Sentiment Analysis), YBS Ansiklopedi, Cilt 3 Sayı 3, Eylül 2016, http://ybsansiklopedi.com/wp-content/uploads/2016/09/duygu_analizi.pdf

Shearer, C., (2000), "The Crisp-DM Model: The New Blueprint for Data Mining" Journal of Data Warehousing, Cilt 5 No 4, 13-23.

Şimşek, M. U., 2012. Sosyal Ağlarda Veri Madenciliği Üzerine Bir Uygulama. Y. Lisans Tezi, Fen Bilimleri Enstitüsü, Ankara.

Şirin, E., 2 Temmuz 2017. Hata Matrisini (Confusion Matrix) Yorumlama, http://www.datascience.istanbul/2017/07/02/hata-matrisini-confusion-matrix-yorumlama/

Takcı, H., 6 Ağustos 2013. DUYGU ANALİZİ (SENTIMENT ANALYSIS), http://verimadencisi.blogspot.com/2013/08/duygu-analizi-sentiment-analysis.html .

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. ve Kappas, A. (2010). Sentiment strength detection in short informal text, Journal of the American Society for Information Science and Technology, 61(12), sayfa. 2544–2558.

Tong, Z., & Zhang, H. (2016). A text mining research based on LDA topic modelling. Jodrey School of Computer Science, Acadia University, Wolfville, NS, Canada, 10, 201-210.

Tunçelli, O., 11 Jan 2015. Turkish Stemmer for Python. https://github.com/otuncelli/turkish-stemmer-python on 12/01/2018.

Turing, A. M., ―Computing machinery and intelligence‖, Mind, 59, pp 433-460, 1950.

Tuzen, E. 18 Temmuz 2017. Python için 5 Muhteşem Derin Öğrenme Kütüphanesi, https://yapayzeka.ai/python-icin-5-muhtesem-derin-ogrenme-kutuphanesi/

Türkmenoğlu, C., Tantuğ, A. C., 2014. Sentiment Analysis in Turkish Media. ICML (International Conference on Machine Learning), Beijing

Uslu, M., Mayıs 2016. Yapay Sinir Ağları (YSA) Nedir ?. http://kod5.org/yapay-sinir-aglari-ysa-nedir/ .

Vural, Z., & Bat, M. (2010). YENİ BİR İLETİŞİM ORTAMI OLARAK SOSYAL MEDYA: EGE ÜNİVERSİTESİ İLETİŞİM FAKÜLTESİNE YÖNELİK BİR ARAŞTIRMA. Journal of Yasar University, 5(20)..

Winograd, T., Procedures as a Representation for Data in a Computer Program for Understanding Natural Language, MIT AI Technical Report 235, February 1971.

Wolfe, L., 31 Mayıs 2018. Twitter User Statistics 2008 Through 2017. https://www.thebalancecareers.com/twitter-statistics-2008-2009-2010-2011-3515899 .

Wu, X., Kumar, V., Ross Quinlan, J. et al. Knowl Inf Syst (2008) 14: 1. https://doi.org/10.1007/s10115-007-0114-2

Xu, J., 25 Mayıs 2018. How to easily do Topic Modeling with LSA, PLSA, LDA & lda2Vec. https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05 .

Yüceloğlu, B. 23 Kasım 2017. SCİKİT-LEARN İLE VERİ ANALİTİĞİNE GİRİŞ, http://www.veridefteri.com/2017/11/23/scikit-learn-ile-veri-analitigine-giris/

Yelmen, İ., 2016. Doğal Dil İşleme Yöntemleriyle Türkçe Sosyal Medya Verileri Üzerinde Duygu Analizi. Y.Lisans Tezi, İstanbul

Yeşilyurt, A. & Şeker, Ş. E. Metin Madenciliği Yöntemleri ile Twitter Duygu Analizi (Twitter Sentiment Analysis using Text Mining Methods), YBS Ansiklopedi, Cilt 4, Sayı 2,Haziran 2017

Yıldırım, S., Yıldız, T., Türkçe için karşılaştırmalı metin sınıflandırma analizi, Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, doi: 10.5505/pajes.2018.15931

ÖZGEÇMİŞ

Abdullah Ammar KARCIOĞLU, 1991 yılında Erzurum'da doğdu. 2005 yılında Erzurum Nevzat Karabağ Anadolu Öğretmen Lisesi'nde lise eğitimine başlayıp 2009 yılında aynı liseden mezun oldu. 2010 yılında Ege Üniversitesi Bilgisayar Mühendisliği bölümünü kazandı. 2015 yılında aynı üniversiteden mezun oldu. Eğitimine, 2015 yılında Erzurum Atatürk Üniversitesi Bilgisayar Mühendisliği bölümünde yüksek lisans yaparak devam etmektedir. 2017 yılında Recep Tayyip Erdoğan Üniversitesi, Mühendislik Fakültesi, Bilişim Sistemleri Mühendisliği'nde araştırma görevlisi olarak göreve başladı ve halen görevine devam etmektedir.