T.C.
PAMUKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

KUTUPSALLIK SÖZLÜĞÜ VE YAPAY ZEKA YARDIMI İLE TÜRKÇE TWİTTER VERİLERİ ÜZERİNDE DUYGU ANALİZİ

YÜKSEK LİSANS TEZİ

HARISU ABDULLAHI SHEHU

DENİZLİ, OCAK - 2019

T.C.
PAMUKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

KUTUPSALLIK SÖZLÜĞÜ VE YAPAY ZEKA YARDIMI İLE TÜRKÇE TWİTTER VERİLERİ ÜZERİNDE DUYGU ANALİZİ

YÜKSEK LİSANS TEZİ

HARISU ABDULLAHI SHEHU

DENİZLİ, OCAK - 2019

KABUL VE ONAY SAYFASI

Harisu Abdullahi SHEHU tarafından hazırlanan "Kutupsallık Sözlüğü ve Yapay Zeka Yardımı ile Türkçe Twitter Verileri Üzerinde Duygu Analizi" adlı tez çalışmasının savunma sınavı 21/01/2019 tarihinde yapılmış olup aşağıda verilen jüri tarafından oy birliği / oy çokluğu ile Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans Tezi olarak kabul edilmiştir.

Jüri Üyeleri	İmza
Danışman Prof.Dr. Sezai TOKAT Pamukkale Üniversitesi	.....
Üye Doç.Dr. Adil ALPKOÇAK Dokuz Eylül Üniversitesi	.....
Üye Dr.Öğr. Üyesi Elif HAYTAOĞLU Pamukkale Üniversitesi	.....

Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 30/01/2019 tarih ve ..05/07... sayılı kararıyla onaylanmıştır.

.....

Prof. Dr. Uğur YÜCEL

Fen Bilimleri Enstitüsü Müdürü

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırmalarının yapılması ve bulgularının analizlerinde bilimsel etiğe ve akademik kurallara özenle riayet edildiğini; bu çalışmanın doğrudan birincil ürünü olmayan bulguların, verilerin ve materyallerin bilimsel etiğe uygun olarak kaynak gösterildiğini ve alıntı yapılan çalışmalara atfedildiğini beyan ederim.

Harisu Abdullahi SHEHU

IMZA

ÖZET

KUTUPSALLIK SÖZLÜĞÜ VE YAPAY ZEKA YARDIMI İLE TÜRKÇE TWİTTER VERİLERİ ÜZERİNDE DUYGU ANALİZİ
YÜKSEK LİSANS TEZİ
HARISU SHEHU ABDULLAHI
PAMUKKALE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

(TEZ DANIŞMANI:PROF. DR. SEZAİ TOKAT)

DENİZLİ, OCAK - 2019

Sosyal medya artık insanların duygularını etkilemede önemli bir rol oynamakta, insanların özellikle de tüketicilerin belirli bir konu, ürün veya fikir hakkında ne hissettiklerini analiz etmemize yardımcı olmaktadır. İnsanların düşüncelerini ifade etmek için kullandıkları güncel sosyal medya platformlarından biri Twitter'dır. Bu tez çalışmasında Twitter API'si kullanılarak Twitter'dan 13 bin tivit toplanmış ve kutupsallık sözlüğü ve makine öğrenmesi sınıflandırmaları yardımı ile duygu analizi yapılmıştır. Bu amaçla bu tez çalışmasında rasgele orman (random forest) ve destek vektör makineleri (support vector machines) olmak üzere iki farklı makine öğrenmesi yöntemi sınıflandırıcı olarak kullanılmıştır. Toplanan tivitler içeriğine göre pozitif, negatif veya nötr olarak etiketlenmiştir. Twitler üzerindeki duygu analizleri ham biçimdeki tivitler üzerinde, dizgecikler ve etkisiz-kelimeler (stop-words) çıkarıldıktan sonra oluşan veri üzerinde ve tivitlerin kökü bulunduktan sonra oluşan veri üzerinde olmak üzere üç farklı aşamada yapılmıştır. Bu aşamaların hepsinde ayrı ayrı duygu analizi yapılmıştır. Son olarak, kullanılan farklı yöntemler toplanan veriler üzerinde test edilmiştir. Ele alınan problem için destek vektör makinelerinin en kısa yürütme süresine sahip olduğu, rasgele orman yönteminin ham veriler üzerinde daha iyi performans gösterdiği, kutupsallık sözlüğü kullanan yöntemin performansının ise diğer yöntemlerde olmayan bir şekilde verilerin ham halinden köklerinin bulunduğu duruma doğru sürekli olarak iyileştiği gözlenmiştir

ANAHTAR KELİMELER: Duygu analizi, Twitter, Tivit, Türkçe, Kutupsallık sözlüğü, Sınıflandırma

ABSTRACT

SENTIMENT ANALYSIS OF TURKISH TWITTER DATA USING POLARITY LEXICON AND ARTIFICIAL INTELLIGENCE
MSC THESIS
HARISU ABDULLAHI SHEHU
PAMUKKALE UNIVERSITY INSTITUTE OF SCIENCE
COMPUTER ENGİNEERİNG

(SUPERVISOR:PROF. DR. SEZAI TOKAT)

DENİZLİ, JANUARY 2019

Social media is now playing an important role in influencing people's sentiment and also helps us to analyze how people particularly consumers feel about a particular topic, a product or an idea. One of the recent social media platforms to express thoughts is Twitter. In this thesis, a sum of 13K Turkish tweets had been collected from Twitter using the Twitter API and their sentiments are being analyzed using polarity lexicon and the use of machine learning classifiers. Random forests and support vector machines are the two kinds of classifiers that are adopted. The collected tweets are classified to be either positive, negative or neutral based on their contents and then their sentiments have been analyzed in three different phases both when the tweets are in raw form, after the tweets are converted into tokens and stop-words are being removed from them and also when the tweets are being stemmed. Finally, the different methodologies used have been tested and find out that support vector machines is the method with the shortest execution time, while random forests perform better on raw data before any manipulation of the data, the performance of the method using polarity lexicon increases continuously as the data being manipulated from raw up to stemmed data.

KEYWORDS: Sentiment analysis, Twitter, Tweet, Turkish, Polarity lexicon, Classification

İÇİNDEKİLER

Sayfa

ŞEKİL LİSTESİ

Sayfa

TABLO LİSTESİ

Sayfa

KISALTMALAR LİSTESİ

API	**:**	Application Programming Interface
NLP	**:**	Doğal Dil İşleme
SA	**:**	Duygu Analizi
ML	**:**	Makine Öğrenmesi
DT	**:**	Karar Ağacı
SVM	**:**	Destek Vektör Makineleri
PL	**:**	Kutupsallık Sözlüğü
RF	**:**	Rasgele Orman
NB	**:**	Naive Bayes
LR	**:**	Lojistik Regresyon
MCC	**:**	Matthews Korelasyon Katsayısı
MaxEnt	**:**	Maksimum Entropi
SynSet	**:**	Eş-anlamlılar Kümesi
ANN	**:**	Yapay Sinir Ağı
PCA	**:**	Principal Component Analysis
MSA-COSR	**:**	Multi-aspect Sentiment Analysis for Chinese Online Social Reviews
PMI-IR	**:**	Pointwise Mutual Information ve Information Retrieval
LSTM	**:**	Long Short-Term Memory
CNN	**:**	Evrişim Sinir Ağı
GDA	**:**	Gizli Dirichlet Ayrımı
ELM	**:**	Extreme Learning Machine

ÖNSÖZ

Öğrenme aşkım çocuklukta başladı. Ailem benim rol modelimdi, eğitimin değerini öğrenmek ve anlamak için ömür boyu süren bir coşku aşılamışlar.

Araştırma, bilgisayar mühendisliği anlayışlarını şekillendirmede güçlü bir etkiye sahiptir. Lisans günlerimde ilk araştırma yaptığımda ve ortaya sonuç çıkardığımda hayatımın coşkulu günlerinden biriydi. Harika hissetmiştim ve günün geri kalanı için çok mutlu olmuştum. O zaman her gün uyanmak istediğimi ve tutkulu iş yapmak istediğimi öğrendim, fark yaratacak bir araştırma. Bu araştırmayı "Kutupsallık Sözlüğü ve Yapay Zeka Yardımı ile Türkçe Verileri Üzerinden Duygu analizi" suyla ilgili yüksek lisans tezimi olarak gerçekleştirmem gerçekten bir ayrıcalıktır.

Gerçekte, beni sevgi ve anlayışla destekleyen hem annem HAFSAT IDRISS hem de babam SHEHU ABDULLAHI'dan güçlü bir destek almadan mevcut başarı düzeyime ulaşamazdım. Bu araştırma boyunca bana tavsiyesi ve rehberliği sağlayan danışma hocam PROF. DR. SEZAİ TOKAT'da ayrıca belirtmek isterim. Tüm sarsılmaz desteğiniz için hepinize teşekkür ederim.

1. GİRİŞ

İletişim bilginin değiş-tokuş edilmesidir. İletişim sürecinde, bir mesaj; belirli bir alıcıya veya bir grup alıcıya kaynak tarafından gönderilen bir iletişim veya iletişim talebinin ayrık bir birimidir. Gönderici, bir fikir geliştiren onu bir mesaja dönuştüren ve bir kanaldan alıcıya ileten kişidir. Alıcı ise göndericinin gönderdiği mesajı bir anlam kazandırmak için yorumlayan kişidir (Liu, 2015).

İletişimin amacı bir kişiden (gönderici) başka bir kişiye (alıcı) aktarılan bilgiyi anlamlandırmaktır. İletişimin temel araçlarından biri toplu izleyici kitlesine ulaşmayı amaçlayan kitle iletişim araçlarıdır. En yaygın kitle iletişim araçları dergiler, gazeteler, radyo ve İnternet'tir (Liu, 2015).

Sosyal medya kullanıcıların sanal topluluklar ve sosyal ağlar aracılığıyla bilgi, fikir, düşünce vb. oluşturmasına ve paylaşmasına olanak veren İnternet uygulamaları için kullanılan bir terimdir. Yıllar geçtikçe Web'deki sosyal medya sistemleri yeni katılımcı kültürümüzle sonuçlanan kitle katılmını sağlamak ve kolaylaştırmak için harika platformlar sağlamıştır (Liu, 2015).

İnternet farklı medya türleri aracılığıyla veri ileten bir ağ alt yapısıdır. İnternet tabanlı farklı sosyal medya platformları vardır. İnternet tabanlı sosyal medya platformlarında arkadaşlarla, ailelerle ve müşterilerle bağlantı kurulması genellikle sosyal ağ olarak adlandırılır.

İnternet tabanlı sosyal medyanın gelişimi bir kişinin yüzlerce hatta binlerce insanla iletişim kurmasını sağlamaktadır. Bu sayede sosyal medyayı kullananların sayısı her geçen gün artmaya devam etmektedir. Günümüzde, sosyal medya hiç tereddüt ve kısıtlama olmaksızın, kullanıcıların görüş ve düşüncelerini sosyal medya üzerinde yayınlamalarına izin vererek modern yaşamda önemli bir rol oynamaktadır.

Sosyal medya platformlarının bir kısmı kullanıcıların düşüncelerini kolaylıkla ayarlanabilen gizlilik seviyesiyle paylaşmalarına ve sadece arkadaşlarıyla etkileşime geçmelerine izin verirken, artık kullanıcılar geleneksel kitle iletişim araçlarından, Facebook ve Twitter gibi mikroblog sitelerine göç etmektedir (Pak, 2010).

İlk zamanlarında sosyal ağ siteleri sadece arkadaşlık veya karşı cinsle tanışma amaçlı bir ortam olarak kullanılmakta ve kabul görmekte iken zamanla sosyal medya platformları özellikler açısından yeniliklere, değişime ve çeşitliliğe uğramıştır.

Farklı amaçlarla kullanılan farklı sosyal medya platformları vardır. LOVOO, Tinder ve Bumble gibi karşı cinsle tanışmaya yönelik uygulamalar, WhatsApp, WeChat, Facebook Messenger, Viber, Google Allo ve Hangouts gibi çok amaçlı mesaj uygulamaları, Twitter, Facebook ve Google+ gibi çevrimiçi haber ve sosayl ağ uygulamaları, Microsoft News, Google News ve Flipboard gibi güncel haber uygulamaları, Azar, Chatroulette ve CamSurf gibi rasgele görüntülü konuşma uygulamaları, Twitter, Tumblr ve FriendFeed gibi mikroblog uygulamaları, Instagram, Flickr ve Pinterest gibi fotoğraf ve video paylaşım uygulamaları, Skype, Imo ve Google Duo gibi görüntülü sohbet uygulamaları vardır.

Twitter son zamanlarda kullanılan en pöpüler sosyal medya platformlarından biridir (Karabulut ve Küçüksille, 2018). Kullanıcıların başlangıçta sınırlı sayıda karakterlerden oluşan mesajlar göndermesine ve okumasına izin veren bir sosyal ağ sitesidir. Bu sınırlı karakter sayısı başlangıçta 140 karakter iken bu sayı Çince, Japonca ve Korece dışındaki diller hariç 7 Kasım 2017'de iki katına çıkarılmıştır. Gönderilen mesajlara tivit (tweet) denir. Twitter'in 2016 yılındaki bilgilere göre aylık 300 milyon aktif kullanıcısı vardır (Anastasia ve Budi, 2016). Twitter'da her gün yaklaşık olarak 500 milyon tivit atılmaktadır. Bu sayıların İnternet kullanımı yaygınlaştıkça artış göstermesi kaçınılmazdır.

Sosyal medyadaki milyarca veri, araştırmacıların veri analizi üzerine araştırma yapmaları için çok etkileyici bir ortam oluşturmaktadır. Belirli konularda görüş belirtmek için yaygın olarak kullanılan Twitter, kullanıcıların hashtag konusunu kullanarak belirli bir konuyla ilgili görüşlerini yayınlamalarına izin vermektedir. Mesela #politika, #endüstri, #bariş gibi konular üzerine tartışmak için her biri ayrı olarak politika, endüstri ve bariş yazıp görüş bildirilmesi örnek olarak verilebilir (Jain and Katkar, 2015).

Son yıllarda sosyal medya, siyasilerin seçim dönemlerinde kampanya yürütmeleri için önemli bir araç olmuştur. Örneğin 2008, 2012 ve 2016 yıllarındaki ABD başkanlık seçimleri sırasında, sosyal medya, seçim kampanyaları ve gençlerin seçime katılımı için kullanılmıştır (Kristin, 2011). Ayrıca 2009'da sosyal medya

özellikle siyasetciler ve siyaset ile ilgilenen insanlar tarafından seçim olaylarını tartışmak ve Alman genel seçimleri sırasında seçim kampanyası yapmak için kullanılmıştır (Jürgens et al, 2011).

Bazı şirketler ve iş kurumları da yaptıkları ticaretten fayda sağlamak için sosyal medyayı kullanırlar. Firmalar tarafından bir çok araştırmacıya, çeşitli sınıflandırma yöntemlerini kullanarak bir olay, ürün, endüstri, borsa vb. hakkında tahmin yapmak için tivit kullanarak araştırma yapma fırsatı verilmiştir (Jain and Katkar, 2015). Bu Twitter'da bulunabilecek büyük miktarda veri nedeniyle mümkün hale gelmektedir.

Bir kullanıcının profilini gönüllü veya reklam amaçlı paylaşım verilerini kullanarak oluşturma süreci, sosyal profilleme olarak bilinirken sosyal dinleme, genel bir stratejiye uygulanabilecek sosyal konuşmadan temel bilgiler edinme ile ilgilidir. Bu kullanıcılara içeriği oluşturmak için belirli bir konu veya anahtar kelimeler etrafında sohbet izleme süreci (Jackson, 2017).

Bu tezde, çeşitli konularda Türkçe Twitter verilerine sosyal dinleme yapılacaktır. Genellikle fikir madenciliği olarak adlandırılan duygu analizi, bir kişinin belirli bir metin parçasındaki görüşlerinden yararlanarak belirli bir konu, ürün, veya nesneye yönelik görüşünü hesaplama ve tanımlama yöntemidir (Anjaria and Guddeti, 2014). Duygular metin temelli mesajlar ve görüntüler sayesinde sosyal medya vasıtası ile ifade edilmektedir. Günümüzde Twitter, Facebook, Flickr ve LinkedIn gibi bazı sosyal medya platformları kullanıcıların görüşlerini herkese açık olarak yayınlamalarına izin vermektedir.

Türkçe dilleri en az 35 belgelenmiş dilden oluşan bir dil ailesidir. Türkçe toplumları Chuvash, Khalaj, ve Sakha dişinda Türkçe dillere fonoloji, mofoloji, ve söz diziminde birbirine yakın benzerlik göstermektedir. Türkçe dillerinin konuşulduğu ülkeler arasında Türkiye, Rusya, Azerbaycan, Kuzey Kıbrıs, Kazakistan, Kırgızistan, Türkmenistan, Özbekistan, Çin, İran, Afganistan, Irak, Bulgaristan, Bosna Hersek, Yunanistan, Romanya, Litvanya ve ayrıca son sanayi göçü sonucunda bir kaç Avrupa ülkesi de yer almaktadır.

Türkçe dili Güney Doğu Avrupa'da 15 milyon yerli konuşmacı ve Batı Asya'da 60–65 milyon yerli konuşmacı ile en çok konuşulan diller arasındadır.

Duygu analizi üzerinden bir çok çalışma yapılasa da Türkçe gibi başka dillerde de yapılmış çok az çalışma bulunmaktadır (Pang ve Lee, 2008; Etter ve diğ., 2016; Cummins ve diğ., 2018). Günümüzde, Türkçe için geliştirilen mevcut duygu analizi yöntemlerinin, Türkçenin bitişken (aglutinatif) bir dil olması nedeniyle, Türkçe söz konusu olduğunda, nadiren üretken ve etkin bir sonuç vermektedir (Sağlam ve diğ., 2016).

Yapısal olarak Türkçe'de 4 farklı tümce çeşidi vardır. Basit tümce; Tamamlanmış bir yargı bildirir ve içerisinde bir adet eylem veya eylem kümesi bulunur. Birleşik tümce; 1 adet Temel Tümce (TT) içeren ve anlamca TT'yi tamamlayan Yan Tümce (YT)'lerden oluşur. Sıralı tümce; Birden fazla TT içeren tümce çeşitleridir. Girişik tümce; n adet TT ve n adet YT içeren tümce yapısıdır (Çoşkun, 2013).

Yine Çoşkun (2013) tarafından yapılan çalışmada Türkçenin tümcenin öğeleri, yüklem, özne, nesne ve tümleç gibi öğelerden oluştuğunu açıklamıştır. Yüklem; Tümcede bir iş, oluş, hareket bildiren sözcük veya sözcük grubuna denir. Özne; Tümcede yüklemin bildirdiği iş, oluş, hareketi yapan veya o işle ilişkili olan öğedir. Nesne; tümce içerisinde öznenin yaptığı veya yüklem tarafından bildirilen iş veya oluşlardan etkilenen kavramlardır. Dolaylı tümleç; yüklemi yönelme, bulunma ve ayrılma açısından tamamlayan öğedir.

Şu anda İnglizce metinler için geliştirilen mevcut duygu analizinin, Türkçe dili söz konusu olduğunda daha az üretken sonuç verdiği gerçeğinden dolayı, bu tezin ardındaki temel motivasyon; yapılacak çalışmanın Türkçe metinler üzerinden duyarlılık analizi için kullanılması ve önerilen sistem akışlarının İngilizce metinlerde duygu analizi için kullanılan mevcut akışlarla karşılaştırılması için yapılacak olmasıdır.

Türkçe ve İnglizce arasındaki farklılıklardan bazıları Vural ve diğ. (2013) tarafından şu şekilde özetlenebilir:

Türkçe'de sözcükler yeni anlamlar üretmek için bir çok ek ile genişletiletilebilir. Bu genişletme ile ilgili bazı açıklayıcı örnekler Tablo 1.1'de verilmiştir.

Tablo 1.1: Türkçe sözcüklerinin yeni bir anlam üretecek şekilde nasıl genişletileceğine dair örnek.

Kelime	**Son-ek**	**İnglizce Anlam**
Yap		Do
Yapma	Yap-ma	Don't do
Yaptım	Yap-tı-m	I did
Yapıyorum	Yap-ıyor-um	I'm doing
Yapabilirim	Yap-abilir-im	I can do
Yapabilirdim	Yap-abilir-dim	I could have done
Yapamayabilirdim	Yap-amayabilir-di-m	I might not have been able to do

Eklenen son ek, bir kök kelimenin polaritesini değiştirebilir. Örnek Tablo 1.2'de verilmiştir.

Tablo 1.2: Kök kelimelerinin polaritesine değiştirme örneğin

Kelime	**Son-ek**	**İngilizce Anlam**	**Anlamsal polarite**
Merhametli	Merhamet-li	Merciful	Positif polarite
Merhametsiz	Merhamet-siz	Unmerciful	Negatif polarite

Bir cümlede kullanılan olumsuz görünen bir kelimenin farklı bir anlamı olabilir. Bu durum Tablo 1.3'te örneklendirilmiştir.

Tablo 1.3: Cümle içinde kullanıldığında kelimelerin anlamını değiştiren negatif kelimeler örneği.

Cümle	**İnglizce Anlam**
Boya **yapma** makinasi kullanarak boya yapabilirsiniz	You can paint using the painting machine
Buradan slayt **yapma** ve video düzenleme programını indirebilirsiniz	You can download slide and video editing program from here

Türkçede kelimeler, kelimeler içinde saklanan son ek tarafından reddedilebilir bu yuzden tüm olumsuzlukların ele alınması gerekir. Örnek Tablo 1.4'de verilmiştir.

Tablo 1.4: Türkçede gizli olumsuz kelimelere örnek.

Kelime	**Son-ek**	**İnglizce Anlam**
Saldırdı	Saldır-dı	Attacked
Saldırmadı	Saldır-**ma**-dı	Did not attack
Kırıldı	Kırıl-dı	Broken
Kırılmadı	Kırıl-**ma**-dı	Did not break

Zemberek kütüphanesi, yüksek doğrulukta sonuç elde etmek için analizde kullanılacak Türkçe verileri dönüştürme sürecinde için kullanılmıştır. Veri dönüştürme süreci temizleme, dizgecikleme (tokenization), kelimelerin kökünün bulunmasını (stemming) ve ayrıca veriden etkisiz-kelimelerin (stop-words) çıkarılmasını içerecektir (Akın ve Akın, 2016).

Önerilen sistem akış diyagramına ait performans bu çalışmanın sonunda gösterilecektir.

1.1 Literatür Taraması

1990'lardan bu yana, farklı dillerdeki bir çok çalışma duygu analizi alanında gerçekleştirilmiştir. Bu çalışmaların hepsi farklı hedeflerle gerçekleştirilmiştir ve bunlar öznel sınıflandırma, duygusal sınıflandırma, istenmeyen jest tespiti, fikir özeti ve metni çıkarma vb. içermektedir. (Ghang ve diğ., 2013) ve bu çalışmalardan bazıları aşağıda gösterilmektedir.

Belirli bir grup, Türkçe siyaset haberlerindeki duygular üzerinde çalışmıştır (Kaya ve diğ., 2012). Siyasi haberlerden oluşan bir veri seti oluşturmak için farklı haber sitelerinden makaleler kullanılmıştır. Kullanılan veri seti, makine öğrenmesi temelli bir yaklaşımla yapılandırılmış ve aynı zamanda, yalnızca politik alandan gelen verilerden olduğu için alana bağımlıdır. Elde ettikleri bulgular, maksimum entropi ve N-Gram dil modelinin destek vektor makineleri (SVM) ve Naive-Bayes yönteminden üstün olduğunu göstermiştir. Araştırmada kullanılan tüm yaklaşımlar %65 - %77 bir doğruluk düzeyine ulaşmıştır.

Aynı grup, aynı alanda Türkçe duygu verilerinden duygu sınıflandırması yaptıkları başka bir araştırma yürütmüşlerdir (Kaya ve diğ. 2013). Kullanılan yöntemlerin performansını artırmak için etiketlenmemiş Twitter verilerden etiketli politik verilere dönüştüren öğrenen bir yapı uygulamışlardır. Amaçları, tüm dokümanın konusu ne olursa olsun pozitif ve negatif olup olmadığını belirlemektir. N-Gram dışında bir önceki yıl kullanılan aynı makine öğrenmesi tekniklerini kullanarak doğrulukta %26'ya varan bir artış gözlemlemişlerdir.

Türkçe duyarlılık sözlüğünü üretmek için duygu analizi çalışması bir tez çalışmasında gerçekleştirilmiştir (Uçan, 2014). Türkçe duyarlılık sözlüğü, İnglizceden Türkçeye çevrilerek üretilmiştir. Bazı film şirketlerinin performanslarını belirlemek için Destek Vektor Makineleri (SVM), atanmış polariteye sahip 27,000 Türkçe kelime içeren bir sözlük ile kullanılmıştır.

Bir başka tez çalışmasında, film incelemesinde duygu analizi yapılmıştır (Eroğlu, 2009). Bu çalışmada; film yorumları http://rec.arts.movies.reviews film eleştirileri, http://rottentomatoes.com ve http://beyazperde.com gibi çeşitli Web sitelerinden toplanmaktadır. Analiz yapmak için Destek Vektor Makineleri (SVM) kullanılmıştır. Her ne kadar bu çalışma kapsamlı bir Türkçe duyarlılık sözlüğü geliştirmemiş olsa da, konuşmanın bir kısmının etkileri, sözcüklerin ve olumsuzlama son ekinin yorumların duyguları üzerindeki ettikleri sırasıyla analiz edilmiştir.

Yukarıda bahsedilen iki tez (Uçan, 2014; Eroğlu, 2009) tek bir araştırmada birleştirilmiştir (Türkmenoğlu ve Tantuğ, 2014). Bu araştırma, iki tez çalışmasını, sözlük tabanlı ve makine öğrenmesine dayalı duygu analizleri arasında bir karşılaştırma önermesi açısından bir araya getirmektedir ve Türkçe resmi olmayan metinlerinin performansını değerlendirmek için hem kısa (Twitter veri kümesi) hem de uzun (film veri kümesi) kullanılmıştır. Sözlük, İngilizce kelimeleri Türkçe'ye çevirerek elde edilmiş ve bu yöntemle elde edilen en iyi sonuç Twitter veri seti kullanılarak% 75.2 iken, film veri seti kullanılarak% 79'luk bir sonuç elde edilmiştir. Öte yandan, NB, SVM ve J48 Decision Trees, verileri sınıflandırmak için ML teknikleri olarak kullanılmaktadır. SVM, Twitter veri kümesini kullanarak,% 85'lik bir doğrulukla diğer sınıflandırıcılardan daha iyi performans gösterirken, SVM ve NB'den daha iyi performans sergileyen J48 sınıflandırıcısı, film veri kümesini kullanarak % 89,5'lik bir doğruluk elde etmiş gibi görünüyor.

Türkçe bloglarında metne olumlu ve olumsuz kutuplar atan bir model, ürün ve hizmetlere genel bir bakış sunmak için tasarlanmıştır (Aytekin, 2013). Geliştirilen model, Naive-Bayes yaklaşımına dayalı yarı-denetimli (etiketli ve etiketsiz veri kümesi) öğrenmeyi kullanmaktadır. Bu araştırmadaki kutupsal kelimeler İngilizce'den çevrilmiş ve elde edilen doğruluk, farklı vakalarda% 64 - 84 arasında değişmektedir.

Bu çalışmada, Türkçe metin belgelerinde denetimsiz duygu analizi için bir çerçeve sunulmuştur (Vural ve ark, 2013). Çalışma, polarite sözlüğünü çevirerek İngilizce için SentiStrength adlı bir duyarlılık analiz kütüphanesinin kişiselleştirilmesini içermektedir. SentiStrength (Thelwall ve arkadaşları, 2012), İngilizce metne olumlu ve olumsuz bir puan veren bir duyarlılık analiz kütüphanesidir. Daha sonra polarite, metnin Türkçe'ye İngilizce'den polarite sözlüğünü çevirerek cümlelere bölünmesinden sonra her cümleye atandı. Zemberek, ön işlemede, çevirme, yazım denetimi, olumsuzlama çıkarımı ve ASCII'nin Türkçe'ye dönüştürülmesinde kullanılır. Elde edilen sonuç, değerlendirmelerin pozitif ve negatif (iki yönlü) olarak sınıflandırılmasında %76 doğrulukta olduğu bildirilmiştir.

Belirli bir nakliye şirketine dayalı bir çalışma gerçekleştirilmiştir (Çoban et al, 2015). Amaç, müşterilerinin twitter tivitine göre müşteri memnuniyetini analiz etmektir. Tarafsızlığı belirlemek yerine (tivit ne kadar tarafsızdır), araştırma tivitin pozitif mi yoksa negatif mi olduğunu belirlemek için iki şekilde gerçekleştirilmiştir. Ön işleme yöntemi kullanıldıktan sonra 20K cümleden oluşan ancak 14,777 ile biten bir veri ile başlamışlardır. Performansı belirlemek için SVM, NB, multinomial NB ve k-NN gibi farklı yöntemler kullanılmış ve Multinomial Naive-Bayes, %66.06 doğrulukla daha doğru sonuç vermiştir.

Benzer şekilde, başka bir alana özgü bir çalışma bir otele dayalı olarak gerçekleştirilmiştir (Oğul ve Ercan, 2016). Bu çalışmada, bir roc işletim karakteristiğinin (ROC) eğri altındaki alanı (AUC) çalışmanın sonucunu belirlemek için kullanılmış ve girdi olarak dönem matrisinin TFIDF matrisinden daha iyi bir sınıflandırma sonucu elde ettiği bulunmuştur. En iyi sonucun, hem olumlu hem de olumsuz yorumlarda AUC değeri %89 olan rasgele orman (RF) sınıflandırıcısı kullanılarak elde edildiği de gözlenmiştir.

Bir Türkçe metindeki duyguları analiz etmek için yapılan bazı çalışmalar ve duygu analizini gerçekleştirmek için kullanılabilecek veri setini oluşturmak için yapılan çalışmalar vardır ve bu çalışmalardan bazıları aşağıda sunulmuştur.

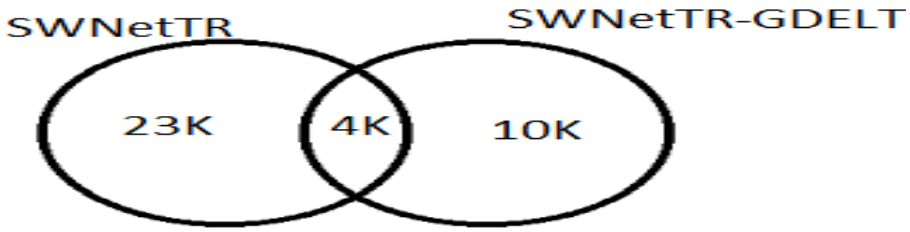
Bu çalışmada bireylerden elde edilen veriler yeni bir veri seti oluşturmak için toplanmıştır (Tocoğlu ve Alpkocak, 2018). Daha sonra oluşan veri kümesi ikiye ayrıldı; ham ve doğrulanmış veri kümesi. Ayrıca, 5 karakter ve Zemberek ya da sözlük-temelli Türkçe gövdeleyici (stemmer)'den sonra daha isabetli olduğu ispatlanmış sabit ön-eki olan iki farklı gövdeleme yöntemi, her bir veri setine uygulanmakta ve toplam dört farklı veri kümesi oluşturulmaktadır. Oluşturulan veri setinde Naive-Bayes, karar ağacı (DT), rasgele orman (RF) ve güncellenmiş SVM gibi çeşitli makine öğrenme algoritmaları çalışılmış ve SVM sınıflandırıcısının daha yüksek bir sonuç verdiği ve doğrulanmış veri seti ile eğitilen modelin, olmayan eğitimli modelden daha yüksek bir sonuç verdiği sonucuna varılmıştır.

Bu araştırma Türkçe için ilk polarite sözlüğünü oluşturmak ve diğer diller için de bunu yapmak için yarı otomatik bir yaklaşım önermiştir (Dehkharghani et al, 2015). Geliştirilen söz dizimi, yaklaşık 15.000 Synset'ten oluşan Türkçe WordNet'teki tüm Synsets (eş anlamlılar kümesi) için polarite puanını üçlü olarak (pozitif, negatif ve nötr / objektif) içermektedir.

Yaklaşık 27.000 kişiden oluşan SentiTurkNet adlı gelişmiş polarite sözlüğünün inşasında üç İngilizce ve bir Türkçe kaynağının kombinasyonu kullanılmaktadır. Kullanılan üç İngilizce kaynak English WordNet (Miller, 1995), SentiWordNet (Baccianella ve arkadaşları, 2010) ve senticNet (Cambria ve arkadaşları, 2014) ve kullanılan Türkçe kaynak WordNet'dir (Bilgin ve diğerleri, 2004). Weka'da üç farklı algoritma kullanılarak uygulanan bir sınıflandırıcı daha sonra geliştirilmiş sözlüğün performansını belirlemek için kullanıldı. Üç (3) sınıflandırıcının tüm özellikleri ve sınıflandırıcı kombinasyonu kullanılarak elde edilen en iyi doğruluğa ulaşmıştır; nearest neighbor (NN), sequential minimal optimization (SMO) ve logistic regression (LR) beraber kullandıktan sonra sonuç 91.11% ulaşmıştır.

Türkçe dilinde ilk kutupsallık sözlüğünü (Dehkharghani et al, 2015) oluşturmak için yapılan araştırmanın bir uzantısı olan bir başka çalışmada, başka bir grup sözlüğü temelli duygu analizi üzerinde bir araştırma yürütülmüştür (Sağlam et

al, 2016). Bu çalışmada, bir kelimenin veya cümlenin kutupluluğu, tek tek sözcüklerin veya deyimlerin kutupluluğunun toplamı olarak kullanılmıştır. Bu çalışma büyük Türkçe haber sayfasının veritabanı ile başlamıştı ve bu veritabanın URL'sı GDELT'den alınmıştır. Önce ham veriler alınmıştır sonradan metinde olan cümlelerin kökünü bulmak için Zemberek kullarak aldığı HTML sayfaları ayrıştırmıştır. Daha sonra her bir kelimeye, GDELT veri tabanından elde edilen kutupsallık değerleri kullanılarak bir puan verilmiştir. Sonuç SWNetTR-GDELT olarak adlandırılmıştır ve 14,000 civarında Türkçe sözcükten oluşuyor. Denemede kullanılan veriler SWNetTR-PLUS olarak adlandırılmıştır ve SWNetTR-GDELT'de bulunan ancak SWNetTR'de bulunmayan neredeyse 10 bin benzersiz kelime eklenerek oluşturulur. Aşağıdaki Şekil 1.1, SWNetTR-PLUS'taki kelime sayısını ve bunların nasıl oluşturulduğunu göstermektedir (Sağlam et al, 2016).



Şekil 1.1 SWNetTR-PLUS ve ilgili kaynak sözcükler (Sağlam ve diğ., 2016).

Yeni sözlük bu veriler kullanılarak test edilmiş ve sonuçlar rapor edilmiştir. Sonuç, Türkçe haberlerin polaritesinin belirlenme doğruluğu %60.6'dan %72.2'ye artırıldığını göstermiştir. Özünde, bu yöntem, bir metnin sıralanmamış kelimelerden oluşan bir sözcük olarak temsil edildiği, kelime-torbası yaklaşımıdır.

1.2 Tezin Amacı

Sosyal medyanın yardımıyla, insanların duyguları artık bir hükümet veya bir kurum lehine veya aleyhine etkili olabilmektedir. Twitter, insanların düşüncelerini ifade etmeleri için yaygın olarak kullanılan sosyal medya platformlarından biri olmuştur (Jain and Katkar, 2015). Son yıllarda, duygu analizi, konuşma tanıma alanındaki en önemli araştırma alanlarından biridir (Tyagi and Chandra, 2015).

Türkçe'nin bitişken (aglutinatif) bir dil olması nedeniyle ve bu özelliğe sahip dillerin karmaşıklıklarından dolayı insanların duygu analizi yapması zorlaşmaktadır. Bu tezin amacı, Türkçe tivitlerinin pozitif, negatif ve nötr duygularını iki farklı yöntemle (kutupsallık sözlüğü ve sınıflandırma) üç farklı aşamada hem tivitler ham olduğunda (hiç veri dönüştürme yapılmadan önce), dizgecikleme yapıldıktan ve gereksiz kelimeler çıkarıldıktan sonra ve son olarak tivitlerin kökü bulunduktan sonra analiz edilmesini sağlamaktır.

Tezde kullanılacak olan Türkçe sözlüğü, Hu ve Liu, (2004) tarafından yıllar boyunca derlenen yaklaşık 6800 olumlu ve olumsuz kelimeyi içeren karşılaştırmalı İngilizce veri kümesi elle Türkçeye çevrilerek ve uyarlanarak geliştirilmiştir. Bu tezin sonunda, kullanılan yöntemler arasında en iyi performansı gösteren yöntem, hangi tür verilerde ve hangi yöntemlerin en hızlı şekilde yürütüldüğünün analizi yapılacaktır. Ayrıca, pozitif, negatif ve nötr sınıfın her birinde en çok kullanılan kelimeler belirlenecek ve çubuk grafikler ve kelime bulutu kullanılarak gösterilecektir.

2. **DUYGU ANALİZİ**

Son zamanlarda yaygınlaşan metin sınıflandırma alanlarından biri de duygu analizidir (Cesarano et al, 2006; Sleator and Temperley, 1991; Subrahmanian ve Reforgiato, 2008).

Çoğumuz için karar verme aşamasında 'diğer insanlar ne düşünür' önemli bir bilgi parçasıdır. İnternet gelmeden önce çoğu insan arkadaşlarından tavsiye isterdi ya da onlara yerel seçimde kime oy vereceğini söylerdi. Düşünceler çözüm bulmuş fakat tartışmaya açık sonuçları işaret eder. Bir fikir her zaman doğru olmayabilir ve kanıtlanmamış olabilir. Duygu, bir kişinin duygularını yansıtan yerleşik bir görüş önermektedir örneğin onun feminist düşünceleri iyi bilinir (Pang ve Lee, 2008).

Duygu analizi ya da fikir madenciliği, insanların ifadelerine ve tutumlarına yönelik görüşlerine ilişkin hesaba dayalı bir çalışmadır. Görüşler bir etkinlik, organizasyon, birey ya da konu hakkında olabilir. (Kiprono ve Abade 2016).

Liu (2015) duygu analizi insanların görüşlerini, duygularını, değerlendirmelerini, özniteliklerini ve duyguları kurumlara ve bunların yazılı metinde ifade ettikleri özniteliklere göre analiz eden çalışma alanı olarak tanımlanmıştır.

Genel olarak, duygu analizi ve fikir madenciliği birbirlerinin yerine kullanılmasına rağmen bazı araştırmacılar duygu analizi ve fikir madenciliği kısmen farklı olduğunu söylemeye başlamışlardır. Fikir madenciliği insanların düşüncelerini analiz edip onu açığa çıkarır, duygu analizi ise bir metni analiz ederek oradaki duygusal ifadeler ortaya koyar (Can ve Alatas, 2017).

Genellikle konuşmada, düşünce analizi konuşmacının ya da yazarın bazı konular çerçevesinde tavırlarını tanımlamayı amaçlar. Çoğu durumda, konular tekrarlarla kaplıdır. Mesela çoğu Afrika, Avrupa, Asya ülkelerinde temel olarak ses ve veri düzenleme işi yapan MTN (haberleşme şirketi) gibi bir şirket yeni arama tarifesini arttırma ya da başlatma kararı alabilir ve meydana gelen bu değişiklik hakkında insanların yorum yapmasını bekleyebilir.

Duygu analizi, bir metinde ifade edilen duyguları tanımlar ve sonra onu analiz eder; fikir madenciliği ise insanların bir ürün veya bir şey hakkındaki fikirlerini ortaya çıkarır ve onu analiz eder. Duygu analizinin amacı, fikirleri bulmak, ifade ettikleri duyguları tanımlamak ve daha sonra karar almada kullanılmak üzere kutuplarını sınıflandırmaktır. Bu nedenle duygu analizi makine öğrenmesi ve sözlük temelli yaklaşımdan oluşmaktadır.

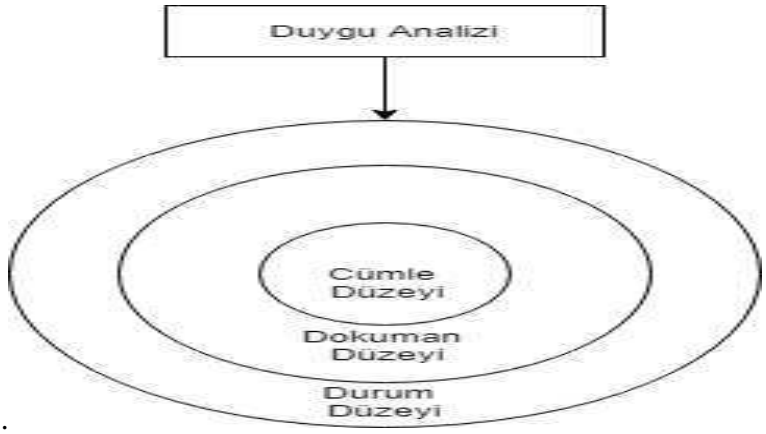
Duygu analiz sistemleri, genel olarak bilgiye dayalı (Cambria ve diğ., 2013[a]) ve istatistiksel olarak kategorize edilebilir (Cambria ve diğ., 2013[b]). Bilgiye dayalı duygu analizi sistemlerinin kullanımı, başlangıçta metindeki ifadelerin ve kutupluluğun tanımlanması için daha popüler olsa da yakın zamandaki duyarlılık analiz araştırmacıları istatistiksel tabanlı (makine öğrenmesi) duygu analizi yaklaşımlarını kullanmaya yönelmişlerdir.

Makine öğrenmesi modeli, istatistik modelleri az nitelikli az miktarlı verilerle uğraşırken var olan program uygulamalarına güvenmeden veriden öğrenilebilen bir algoritmadır ve böylece uyum göstermenin ortaya çıkabilme şansı vardır. Sonuçları tahmin etmek için değişkenler arası ilişkileri bulmakla ilgilidir. Makine öğrenmesi modeli ve istatistik modelinin aksine bilgi tabanlı modeller kullanıcının daha iyi sonuçlar üretmek için yorum modelini koruması için öncelikli bilgi alanından faydalanılması önerilen modellerdendir (Liu, 2015).

Duygu analizi (DA) bir sistem için farklı seviyelerde gerçekleştirilebilir. Temel bir DA'nın görevi, aşağıdaki şekil 2.1'de gösterildiği gibi bir metnin polaritesini farklı seviyelerde sınıflandırmaktır.

2.1 Duygu Analizi Seviyeleri

Şekil 2.1, duygu analizinin gerçekleştirilebileceği farklı düzeyini göstermektedir.



Şekil 2.1: Duygu analizi seviyeleri

2.1.1 Cümle Düzeyinde Duygu Analizi

Cümle düzeyinde duygu analizi, her cümlenin duygularını tek tek analiz eder. Bu aşamada ilk önce cümlenin öznel mi yoksa nesnel mi olduğunu tespit edilir. Ardından eğercümle öznel ise, olumlu ya da olumsuz sınıfa ait olup olmadığını analiz etmeye devam edilir (Medhat ve diğ., 2014). Amaç, bir cümlede ifade edilen görüşün olumlu, olumsuz veya tarafsız olup olmadığını belirlemektir (Liu ve diğ., 2014).

2.1.2 Belge Düzeyinde Duygu Analizi

Belge düzeyinde duygu analizi, verilen bir dokümanın metnini analiz eder ve bu analiz sonucunun olumlu ya da olumsuz duygu değeri gösterip göstermediğini belirler (Behdenna ve diğ., 2018). Bu aşamada belli bir konu üzerine yazılmış belgeyi işler ve belgedeki metni analiz ederek belgenin pozitif veya negatif bir polariteye sahip olup olmadığını belirler.

2.1.3 Durum Düzeyinde Duygu Analizi

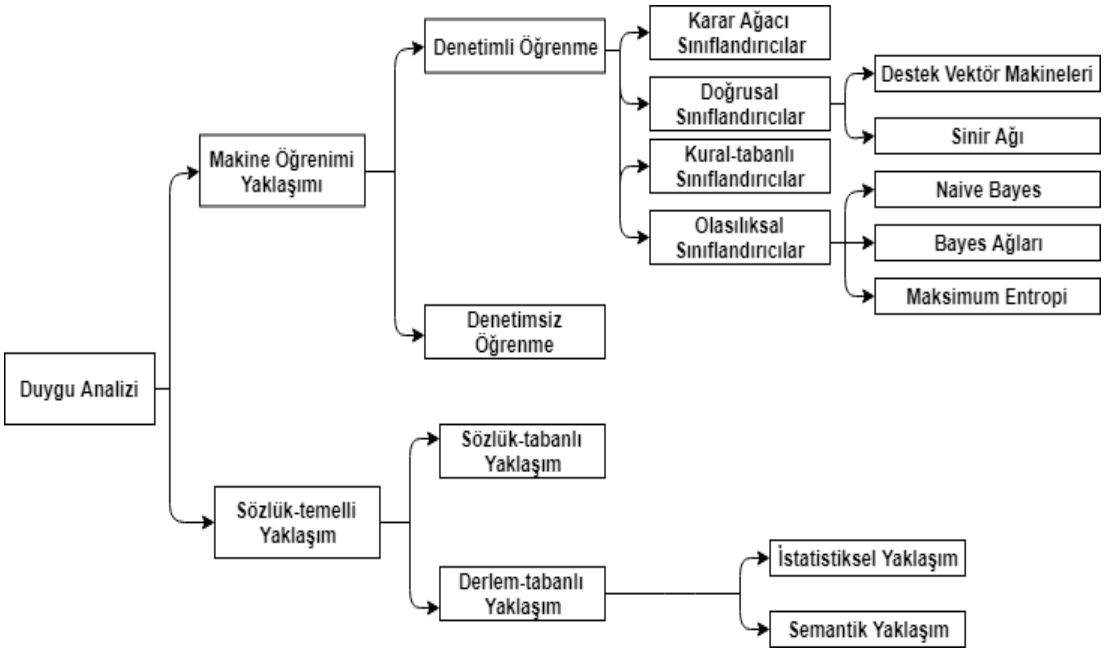
Durum düzeyinde duygu analizi (DDDA), duygu analizinin tüm çıkış yönlerini ele almayı amaçlamaktadır. DDDA'nın amacı, her yönüyle ifade edilen

varlıkların ve duyguların (olumlu veya olumsuz olsun) yönünü tanımlamaktır (Kumara, 2015).

Bu tezde belge düzeyinde duygu analizi kullanılacaktır. Çünkü her tivit duygu analizi yapılmadan önce indirilir ve bir belge olarak kaydedilir.

2.2 Duygu Sınıflandırma Teknikleri

Duygu sınıflandırma teknikleri, kabaca makine öğrenim (ML) yaklaşımı, sözlük tabanlı yaklaşım ve hibrit yaklaşım olarak sınıflandırılabilir (Maynard ve Funk, 2011). Makine öğrenim yaklaşımı, makine öğrenimi algoritmalarının kullanımını içerir. Sözlük tabanlı yaklaşım, bilinen ve önceden derlenmiş teknik terimlerin bir koleksiyonunu ifade eden duygu sözcüklerine dayanır. Hibrit yaklaşım ise her iki yaklaşımı (makine öğrenimi ve sözlük tabanlı) birleştirerek bir sonuç elde etmeye dayanır. Duygu sınıflandırma teknikleri Şekil 2.2'de verilmiştir.



Şekil 2.2: Duygu sınıflandırma teknikleri (Medhat ve diğ., 2014)

ML yaklaşımını kullanan metin sınıflandırması, kabaca denetimli ve denetimsiz öğrenim olararak 2 ye bölünebilir. Denetimli öğrenim, çok sayıda etiketli veri kümesinin kullanılmasına izin verirken denetimsiz öğrenim, toplanmamış veri kümelerinin kullanımını içerir. Denetimsiz öğrenim, etiketli veri setininin bulunması zor olan durumlarında kullanılır.

Sözlük tabanlı yaklaşım, metni analiz etmek için kullanılan duygu sözlüğünü bulmayı içerir ve ikiye ayrılır: sözlük tabanlı yaklaşım ve Derlem tabanlı yaklaşım. Sözlük tabanlı yaklaşım, duygu kelimelerinin köklerini bulmaya ve daha sonra eşanlamlı ve zıt anlamlı sözcükleri araştırmaya dayanır. Derlem tabanlı yaklaşım ise bir duygu kelimesenin köküyle başlar ve daha sonra bağlamsal yönelimlerle benzer duygu kelimelerini bulmaya yardımcı olmak için büyük bir dizinde başka görüşler bulur (Medhat ve diğ., 2014).

2.2.1 Makine Öğrenimi Yaklaşımı

Makine öğrenim yaklaşımı, dilbilimsel veya dinamik özelliklerden yararlanır ve duygu analizini normal bir metin sınıflandırması olarak çözmek amacıyla algoritmaların kullanılmasına başvurur. Sınıflandırma modeli, temel kayıttaki etiketlerden birinin özelliği ile ilgilidir ve model, bilinmeyen sınıfın her örneği için bir sınıf etiketini tahmin etmek için kullanılabilir (Kiprono ve Abade, 2016).

2.2.1.1 Denetimli Öğrenme

Denetimli öğrenme, etiketli belgenin varlığına bağlıdır. Tablo 2.1'de (Vaghela ve Jadav, 2016) denetimli bir öğrenme yöntemi kullanılarak gerçekleştirilen birkaç çalışmanın temsilini göstermektedir.

Tablo 2.1: Denetlenen öğrenme teknikleri kullanılarak gerçekleştirilen önceki çalışmaların özeti

Kaynakça	Teknik	Veri kümesi	Veri kümesi boyutu	Doğruluk
Ay Karakuş ve diğ. (2018)	CNN LSTM CNNLSTM	Film incelemesi	4,000	%97.62 %96.57 %98.07
Güven ve diğ. (2018)	GDA stage1 GDA stage2 GDA stage3	Twitter	4,000	%60.4 %70.5 %76.4

Tablo 2.1 (devam): Denetlenen öğrenme teknikleri kullanılarak gerçekleştirilen önceki çalışmaların özeti

Coban ve diğ. (2018)	SVM ELM	Twitter	10,000	%74'ye kadar %70'ye kadar
(Pang ve diğ., 2002)	SVM NB	Film incelemesi	1,400	%82.9 %81.5
Tripathy ve diğ., 2015	SVM NB	Film incelemesi	2,000	%94 %89.5
Da Silva ve diğ., 2014	RF SVM LR NB	Sanders Twitter Stanford Twitter OMD Twitter HC Twitter	7,660	%84.89 %87.2 %76.81 %78.35
Shahana ve Omman, 2015	NB	Müşteri yorumu	2,000	%92.37
Go ve diğ., 2009	MaxEnt NB SVM	Twitter	1.6 million (training) 359 (test)	%83 %82.7 %82.2
Anjaria ve Guddeti, 2014	SVM NB MaxEnt ANN SVM + PCA	Twitter	100,000	%88 %84 %83 %77 %93
Chaovalit ve Zhou, 2005	3-fold validation	Film incelemesi	221	%85.54
Anastasia ve Budi, 2016	SVM NB DT	Twitter	126,405	%72.97 %61.25 %72.97
Islam, 2016	NB	Facebook	200	%83
Jain ve Katkar, 2015	RF NB KNN BayesNet	Twitter	210, 252	%65.67 %60.32 %96.64 %48.96

Literatürde çok sayıda denetimli öğrenme sınıflandırıcısı vardır ve bunların bazıları aşağıda açıklanmaktadır.

2.2.1.1.1 Olasılıksal Sınıflandırıcılar

Olasılıksal sınıflandırıcılar, sınıflandırma için karışım modellerini kullanır. Bu modelde, her bir sınıfın aynı karışımın bir bileşeni olduğu varsayılır. Her bir karışım bileşeni, bu bileşen için belirli bir terimi örnekleme olasılığını sağlayan bir üretken modeldir. Bu tür sınıflandırıcılar üretken sınıflandırıcılar olarak da adlandırılabilir. En ünlü olasılık sınıflandırıcılarından bazılarına örnek olarak Naive Bayes, Bayes Ağı ve Maksimum Entropi verilebilir.

2.2.1.1.2 Kural-tabanlı Sınıflandırıcılar

Kural tabanlı sınıflandırıcılar, kuralları belirleyen, öğrenen, saklayan, yöneten veya işleyen herhangi bir makine öğrenme yöntemini kapsamayı amaçlamaktadır ((Bassel ve diğ., 2016; Weiss ve Indurkhya, 1995). Kural tabanlı bir makine sınıflandırıcısının tanımlayıcı özellikleri, bir dizi ilişkisel kuralın tanımlanması ve kullanılması olarak tanımlanabilir. Diğer bir deyişle, kural tabanlı sınıflandırıcılar "if-then" kalıplarının kullanılmasını ifade eder ve aşağıdaki formda ifade edilebilir: “IF condition THEN conclusion”

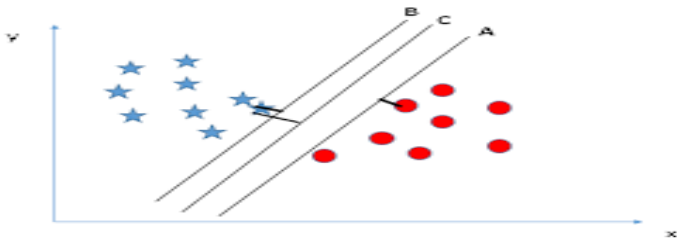
2.2.1.1.3 Doğrusal Sınıflandırıcılar

Doğrusal bir sınıflandırıcı, nesnenin hangi gruba (sınıf) ait olduğunu tanımlamak için nesne özelliklerini kullanan bir sınıflandırıcı türüdür. Özellik değerleri olarak bilinen nesne karakteristikleri özellik vektörü olarak adlandırılan bir vektörde makineye sunulur. Doğrusal sınıflandırıcılar, belge sınıflandırması ve daha genel olarak birçok özellik ve değişkenle ilgili problemler için iyi çalışır. Lineer olmayan sınıflandırıcılara göre doğruluk seviyelerine ulaşabilir yapıdadırlar ve triaja ile kullanıma daha az zaman harcarlar. (Yuan ve diğ., 2012).

Günümüzde doğrusal sınıflandırıcıların bir çok çeşidi vardır; Bunlardan bazıları sinir ağları (SA) ve destek vektör makinesi (SVM) (Cortes ve Vapnik, 1995; Vapnik, 1995) olarak örnek verilebilir. Destek vektör makinesi aşağıdaki bölümde ele alınmıştır.

2.2.1.1.3.1 Destek Vektör Makineleri

Bir destek vektör makinesi, ayırıcı bir hiper düzlem tarafından resmen tanımlanmış ayırt edici bir sınıflandırıcıdır. Başka bir deyişle, etiketli bir eğitim verisi verildiğinde, algoritma yeni örnekleri kategorize eden optimal bir hiper düzlem çıkarır. İki boyutlu uzayda, hiper düzlem bir düzlemi iki parçada bölen bir çizgidir ve her sınıfta her iki tarafta da uzanmaktadır (Patel, 2017). SVM'nin ana ilkesi, farklı sınıfları en iyi şekilde ayırabilen, arama alanındaki satır ayırıcılarını belirlemektir. Şekil 2.3'te iki sınıf "x ve o" ve üç hiper-düzlem "A, B ve C" vardır. Hiper-düzlem C, sınıflar arasında en iyi ayrımı sağlar, çünkü verilerin herhangi birinin C'ye olan normal uzaklığı en büyüktür ve bu nedenle maksimum ayırma marjını temsil eder.



Şekil 2.3: Bir sınıflandırma probleminde destek vektör makinesini kullanımı (Kubat, 2015)

SVM'ler birçok uygulamada kullanılmaktadır. Bu uygulamalar kendi aralarında kaliteye göre sınıflandırılmaktadırlar.

Li ve Li (2013), SVM'leri bir duyarlılık polarite sınıflandırıcısı olarak kullanmışlardır. Mikro-blog platformlar hakkında fikirlerin kompakt bir sayısal özetini sunan bir yapı önermişlerdir. Geliştirdikleri bir mekanizmanın, gerçek zamanlı olarak bir işletmenin farklı yönleri hakkındaki dış görüşlerini izlemeye yönelik bir izleme sistemi kurarak karar vericileri desteklemek için piyasa istihbaratını (MI) etkili bir şekilde keşfedebileceğini kanıtladılar.

Chen ve Tseng (2016) ayrıca çift yapıya sahip çok merkezli SVM tabanlı yaklaşım kullanmışlardır: Ürün incelemelerindeki bilgilerin kalitesini bir sınıflandırma problemi olarak değerlendirmek için bir yöntem önermişlerdir. Elde ettikleri sonuçlar, son teknoloji yöntemlerden çok daha iyi performans göstermiştir ve aynı zamanda kullandıkları yöntemlerin ilgili değerlendirmeleri doğru bir şekilde sınıflandırabildiğini de göstermişlerdir.

2.2.1.1.4 Karar Ağacı Sınıflandırıcıları

Karar Ağacı sınıflandırıcıları, örnek veri alanının bir hiyerarşik ayrışmasını sağlar ve kullanılan veriler, öznitelik değerleri üzerinde bir koşulu kullanarak bölünür (Quinlan, 1986). Yüklem veya koşul, bir veya daha fazla kelimenin varlığıdır. Bölümlendirme, sınıflandırma amacıyla kullanılan yaprak düğümlerinin minimum sayıda kayıt içermesine dek ardışık olarak yapılır.

2.2.1.2 Denetimsiz Öğrenme

Denetimsiz öğrenme, etiketlenmemiş veri kümeleriyle ilgilenen etiketli belgenin varlığına bağlı olmayan bir öğrenme biçimidir. Kullanıcı bir sınıf örneği sağlamadan verileri analiz eder. Tablo 2.2'de (Vaghela ve Jadav, 2016) denetimsiz bir öğrenme metodu kullanılarak yürütülen birkaç çalışmanın temsili gösterilmektedir.

2.2.2 Sözlük-temelli Yaklaşım

Bu yöntem, belirli bir içeriğin genel değerlendirme puanına karar vermek için kutupluluk değeri tarafından açıklanmış çeşitli kelimeler kullanır. Olumsuz görüş içeren kelimeleri bazı istenmeyen durumları ifade etmek için kullanılırken, olumlu görüş içeren kelimeleri bazı istenen durumları ifade etmek için kullanılır. Bu tekniğin en güçlü varlığı, herhangi bir eğitim verisi gerektirmemesidir. En zayıf noktası ise, duygu sözcüklerinde çok sayıda kelime ve ifadenin yer almamasıdır (Symeonidis, 2018).

Duygu içeren kelime listesini toplamak için üç temel yaklaşım vardır. Manuel yaklaşım çok zaman alıcıdır ve tek başına kullanılmaz. Otomatik kontrolden kaynaklanan hataları önlemek için genellikle son iki kontrol ile birlikte çalışır. Konu ile ilgili iki otomatik yaklaşım aşağıdaki alt bölümde açıklanmıştır.

Tablo 2.2: Denetimsiz bir öğrenme tekniği kullanılarak gerçekleştirilen önceki çalışmaların özeti

Kaynakça	Teknik	Veri kümesi	Veri kümesi boyutu	Doğruluk
(Chaovalit and Zhou, 2005)	Semantik yönelim	Film incelemesi	1,400	%77
(Khan et al, 2014)	İfadeler Kelime çantası SentiWordNet	Twitter	2116	%80 yukarı
(Fu et al, 2013)	MSA-COSR	Sosyal Yorumlar	2000	%91.23
(Turney, 2002)	Semantik yönelim + PMI-IR	Film incelemesi, Bankalar İncelemesi, Otomobil İncelemesi, Seyahat Değerlendirmesi	410	%74
(Lin and He, 2009)	Ortak duygu konusu	Film incelemesi	1,049	%84.6

2.2.2.1 Sözlük-tabanlı Yaklaşım

Bilinen yönelimlerle küçük bir duygu kelimesi koleksiyonu manüel olarak toplanır. Daha sonra, bu koleksiyon, iyi bilinen bir şirket olan WordNet (Miller ve diğ., 1990) ile eş anlamlıları ve zıt anlamlıları için araştırma yapılarak büyütülür. Yeni bulunan kelimeler kök listesine eklenir ve bir sonraki iterasyona geçilir. Yeni bir sözcük bulunamadığında rekürsif olan bu işlem durdurulur. İşlem

tamamlandıktan sonra hataları gidermek veya düzeltmek için manuel inceleme yapılır.

Özel alan ve içerik yönelimlerine sahip kelimeleri bulmadaki yetersizlik sözlük temelli yaklaşımın önemli bir dezavantajı olmuştur.

Qui ve diğ. (2010) yaptıkları bir projede sözlük temelli yaklaşımı kullanarak reklamlarındaki duygu cümlelerini tanımlamıştır. Bu süreç sonunda reklam ilgi düzeyini ve kullanıcı deneyimini iyileştirmek için bir reklam stratejisi önermişlerdir. Ek olarak, reklamların anahtar kelimelerini çıkarma ve reklam seçimiyle ilgili önerilen yaklaşımın etkinliği göstermişlerdir.

2.2.2.2 Derlem-tabanlı Yaklaşım

Derlem-tabanlı yaklaşım, içeriğe özel yönelimlerle duygu kelimeleri bulmaya yardımcı olur. Bu süreçte büyük bir dizindeki diğer duygu kelimelerini bulmak için duygu kelimelerinin bir kök listesi ile birlikte ortaya çıkan sözdizimsel kalıplara veya kalıplara bakılmasını amaçlar.

Hatzivassiloglou ve McKeown (1997) korpus tabanlı yaklaşımı temsil etmişlerdir. Bu sürece tohum duygu sıfatlarının bir listesi ile başladılar ve bunları, ek sıfatla ilgili duygu kelimelerini ve yönelimlerini belirlemek için bir dizi dilsel kısıtlama ile birlikte kullandılar. Daha sonra, sıfatlar arasındaki bağlantılar ile bir grafik oluşturdular ve grafik üzerinde pozitif ve negatif olmak üzere kümeleme yaptılar.

Cruz ve Troyano (2013), taksonomi düzeyindeki duyguların çıkarılması için bir taksonomi temelli yaklaşımı temsil etmekte ve bunları bir taksonomi sınıflandırması haline getirmektedir. Bu taksonomi, bir nesnenin bölümlerinin ve niteliklerinin anlamsal bir temsilidir. Bulguları, alandan bağımsız yaklaşımlarla ilgili olarak, doğru duygu çıkarma sistemleri oluşturmak için alanın önemini ortaya koymuştur.

Yalnızca Derlem-tabanlı yaklaşımı kullanmak sözlük temelli bir yaklaşım kadar etkili değildir, çünkü tüm İngilizce kelimeleri kapsayacak kadar büyük bir dizin hazırlamak her ne kadar zor olsa da bu yaklaşım, etki alanına ve içeriğine özel

duygu sözcüklerini bulmaya yardımcı olabilecek derece önemli bir avantaj içerir. Korpus tabanlı yaklaşım, aşağıdaki alt bölümlerde gösterildiği gibi istatistiksel yaklaşım veya semantik yaklaşım kullanılarak gerçekleştirilir.

2.2.2.2.1 İstatistiksel Yaklaşım

Eşdizimlilik kalıpları veya kök duygu kelimelerinin bulunması istatistiksel teknikler kullanılarak yapılabilir. Bu, Fahrni ve Klenner (2008) tarafından önerildiği gibi, bir derlemde sıfatların birarada oluşu kullanılarak zıt kutupların türetilmesiyle de yapılabilir. Bu süreçte dizine eklenen tüm belgeler dizisini sözlük yapısının derlemi olarak kullanmak da mümkündür.

İstatistiksel yöntemler, SA ile ilgili birçok uygulamada kullanılmaktadır. Bunlardan birisi de rastlantısallığın istatistiksel testini yaparak değişiklik tespit eden Runs testidir.

Hu ve diğ. (2012), incelemelerin müşteriler tarafından yazıldığı durumlarda, incelemelerin yazılma tarzının müşterilerin farklı geçmişleri nedeniyle rasgele olacağını varsaymışlardır. Bu nedenle bunu kanıtlamak için Amazon.com'dan Kitap incelemeleri üzerinde çalışmışlardır ve ürünlerin yaklaşık %10,3'ünün çevrimiçi yorum değişikliği tabi olduğunu keşfetmişlerdir.

Latent Semantic Analysis (LSA), bir dizi belge ile bu belgelerdeki terimler arasındaki ilişkilerini, belgelere ve terimlere yönelik anlamlı bir dizi kalıbını oluşturmak için kullanılan istatistiksel bir yaklaşımdır (Deerwester ve diğ., 1990).

Cao ve diğ. (2011), metinlerinin anlamsal özellikleri bulmak için LSA'yı kullanmışlardır. İşlerinin amacı, bazı incelemelerin neden pek çok yardım oyu aldığını, diğerlerinin ise çok az ya da hiç oy almadığını anlamaktır. Semantik özelliklerin, diğer özelliklerden daha etkili olduğunu göstermişlerdir.

2.2.2.2.2 Semantik Yaklaşım

Semantik yaklaşım duyguların değerlerini doğrudan verir ve kelimeler arasındaki benzerliği hesaplamak için farklı prensiplere dayanır. Bu yöntem,

anlamsal olarak birbirine yakın olan sözcüklere benzer duygu değerlerini verir. Örneğin WordNet, duygu polaritesini hesaplamak için kullanılacak kelimeler arasında farklı türlerde semantik ilişkiler sağlar. WordNet, ilk setin eş anlamlı ve zıt anlamlı bir şekilde genişletilmesi ve daha sonra bilinmeyen bir kelimenin duygu polaritesinin bu kelimenin pozitif ve negatif eşanlamlılarının bağıl sayısı ile belirleyerek, sahip olduğu duygu kelimelerinin bir listesini elde eder. (Kim ve Hovy, 2004).

Semantik yaklaşım, Maks ve Vossen (2012) tarafından sunulan bir çalışma olarak SA'da kullanılmak üzere birçok uygulamada fiiller, isimler ve sıfatların tanımlanması için bir sözlük modeli oluşturmak için kullanılmıştır. Modelleri, her aktör için ayrı durumları ifade eden bir cümledeki aktörler arasındaki ayrıntılı öznellik ilişkilerini tanımlamıştır. Sonuç olarak, konuşmacının öznelliğini güvenilir bir şekilde tanımlayabildiğini kanıtlanmıştır.

Semantik yöntemler, Wenhao ve arkadaşları (2012) tarafından sunulan çalışma olarak SA süreçleri için istatistiksel yöntemlerle de birleştirilebilir. Yapılan bir çevrimiçi incelemede, ürün zayıflığını bulmak için her iki yöntem de kullanılmıştır.

3. KUTUPSALLIK SÖZLÜĞÜ VE YAPAY ZEKA YARDIMI İLE TÜRKÇE TWITTER VERİLERİ ÜZERİNDE DUYGU ANALİZİ İÇİN ÖNERİLEN SİSTEM AKIŞ YAPILARI

Bu tez, Türkçe atılan tivitler üzerindeki duyguları öngörmek için bir mekanizma sunmaktadır. Bunu yapmak için iki farklı yöntem kullanılmıştır.

3.1 Kutupsallık Sözlüğü

İlk yöntem, kelimelerden oluşan bir sözlük yardımıyla kelimelerin yapılarına ayrıştırıldığı tivitler ile eşleştiren kutupsallık sözlüğünün (PL) kullanımını içerir. Tivitler bu sözlükteki kelimelerle eşleştirildikten sonra bulunan sonuçlara göre pozitif, negatif veya nötr olarak sınıflandırılır. Eğer tivitler eşleştikten sonra polarite sonucu 0'ın üzerinde oluşursa pozitif, polarite sonucu 0'ın altında oluşursa negatif ve sonuç tam olarak 0 ise nötr olarak kabul edilir.

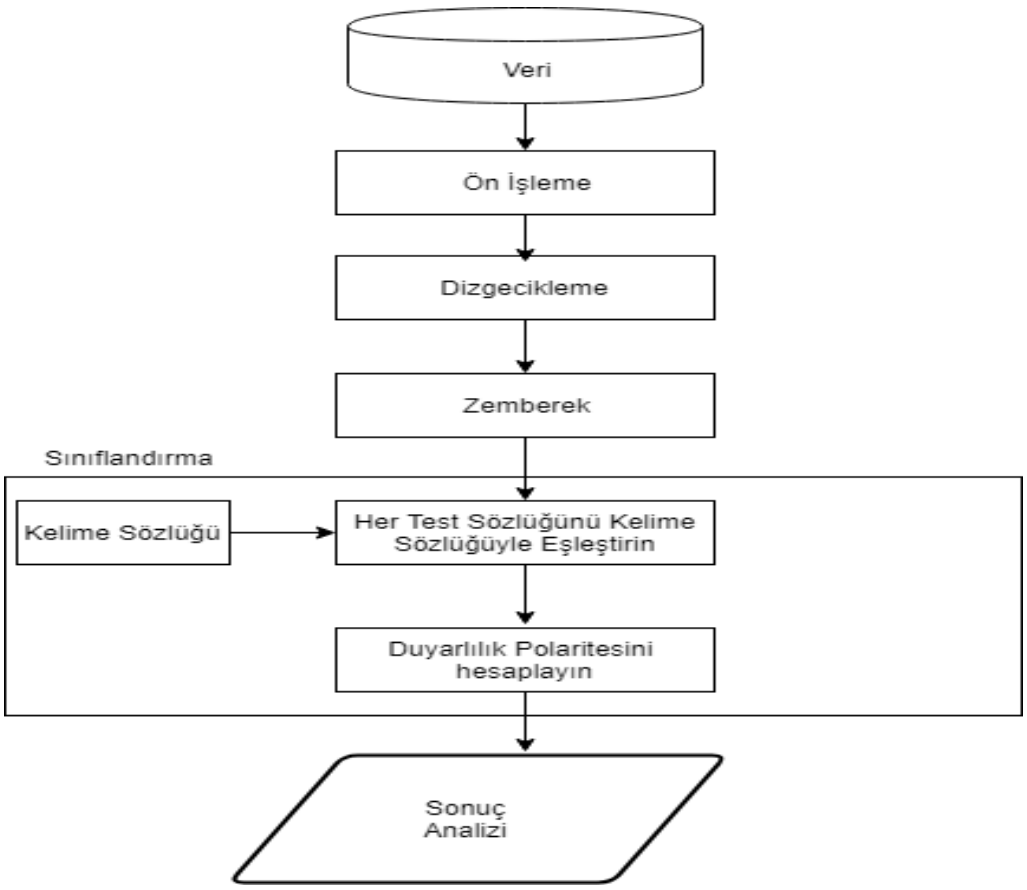
Şekil 3.1, PL ile kullanılan yöntemin sistem akışını göstermektedir ve ayrıntılı olarak şekil bloğunun her birini açıklamaktadır.

3.1.1 Veri Toplama

Veri setleri (uygulama ve test), Twitter arama API'sininin 3.4.3 R sürümü ile Twitter'dan toplanmıştır. Tezin bütün süreçlerinde Türkçe tivitler kullanılmıştır. Tivitler iki şekilde toplanmıştır.

Twitter arama API tamamen twitlerin indeks değildir, o sadece son zamanlarda yapılan twitlerin indektir. Şimdilik o indeks son 6 – 9 gün kamsıyor aşağı formulde gözüktüğü gibi. Ilk parametre kullanıcının hasat etmek istediği Tivitlerin konu oluyor, ikinci parametre de Tivitlerin sayısı gösteriyor, son parametre ise kullanicinin hasat etmek istediği Tivitlerin dili işaretliyor.

```
tweets <- searchTwitter(search.string, n=no.of.tweets, lang="tr")
```



Şekil 3.1 Kutupsallık Sözlüğü için duygu analizi süreç akışı

Aşağıdaki verilen bu kod betiği, belirli bir zaman dilimi içinde belirli bir tarihten belirli bir tarihe kadar atılan tivitleri almak için kullanılabilir.

```
tweets <- searchTwitter(search.string, n=no.of.tweets, lang = "tr",
                        since = '2018-1-12', until = "2018-1-13")
```

Yukarıdaki kod satırında da görüldüğü gibi, arama API'sine "since" ve "until" anahtar kelimesinde iki parametre ekleyerek tivitleri 6-9 gün içinde değil, istediğimiz belirli bir zaman aralığında almak için de kullanabiliriz

3.1.2 Ön İşleme

Bazen, doğrudan twitter'den elde edilen tivitler kullanılabilir bir formatta değildir ve kullanılabilir bir formata dönüştürmek için çeşitli ön işleme yöntemleri

uygulanır. Ön işleme yöntemi tivitleri temizler ve kullanıcıların kimliklerinin, bilgilerinin, tivitlerin küçük harf dönüştürmesi, twitter kimliğinin kaldırılmasını, tırnak vb. gibi özel karakterlerin de tivitlerden kaldırılmasını ve Mulki ve arkadaşları tarafından (2018) yaptığı gıbı tivitlerin gibi tivitlerden gereksiz kelimeler çıkarılmasını içerir. Ön işleme yöntemi ayrıca, tivitlenmiş metinlerin ve şifrelerin yeniden atılmış (retivit) tivitlerin tivitlerden kaldırılmasını da içerir. Tüm ön işleme yöntemleri uygulandıktan sonra, geriye kalan sadece tivitlenmiş metinlerdir ve analiz edilecek olan şeydir.

3.1.3 Dizgecikleme

Belirtgeleme veya dizgeciklere ayırma olarak da adlandırılan dizgecikleme (tokenization) karakter sırasını veya belge ünitesini dizgecik (token) olarak adlandırılan parçalara ayırma işlemidir ve aynı zamanda noktalama, vb. gibi bazı karakterleri de atabilir. Dizgecikleme uygulandıktan sonra, geriye kalan değerler aynı karakter dizisini içeren sınıfıdır.

3.1.4 Zemberek

Zemberek, Türkçe dilinde hazırlanmış açık kaynak kodlu bir doğal dil işleme (NLP) altyapısıdır. Güncel versiyonu ile yazım denetimi, morfolojik ayrıştırma, kaynak oluşturma, kelime seçimi, kelime önerme, sadece ASCII harfleriyle yazılmış sözcüklerin dönüştürülmesi ve hecelerin çıkarılmasında temel NLP işlemleri sağlamaktadır (Akın ve Akın, 2007). Bu tezde, Türkçe verilerin dönüştürülmesi işlemlerinde Zemberek kullanılacaktır.

3.1.4.1 Gövde

Gövde, genellikle kelimelerin sahip olduğu eklerinin istenilen hedefe ulaşmasını zorlaştıran ve çoğu zaman türetme eklerinin kaldırılmasını içeren kaba bir sezgisel sürece işaret eder. Tablo 3.1, Türkçe kelimelerin bir örneğini ve bunlara gövdeleme işleminin uygulanmasından sonra nasıl değiştiklerini göstermektedir.

Tablo 3.1: Türkçe bazı sözcüklerde ayıklama örneği

Kelime	**Kök bulunduktan sonra**
Alanında	Alan
Birleşmiş	Birleş
Ucuz	Ucuz
Eklemek	Ekle
Anlatılmak	Anlat

Bu projede kullanılan gövde metodu, kullanılan kelimenin sadece gövde bulmayı değil, aynı zamanda kelimenin birden fazla gövde sahip olduğu durumlarda kelimenin gerçek gövde ayırt edilmesini de amaçlar. Sözcüğün gövde, bağlamda nasıl kullanıldığına bağlı olarak kök sonuçlarında birden fazla yeniden yazılabilir. Örneğin, Türkçe dilinde çoğul olduğunu belirten "ler" son eki ile biten bazı kelimeler kök sonuna 3 kez yeniden yazılırken, "lik" son ekiyle biten bazı kelimeler, söz konusu duruma bağlı olarak kök sonuna yalnızca 2 kez yazılır. Tersi de mümkündür. Aşağıdaki Tablo 3.2 ve 3.3, birden fazla köke sahip olan kelimelerin bir örneğini ve gövde sonucunda bir defadan fazla yeniden yazılan kelimeleri göstermektedir.

Tablo 3.2 Birden fazla gövde içeren kelimelerin bir örneği

Kelime	**İnglizce Anlam**	**Gövde**	**İnglizce Anlam**
Gözlükçü	Optician	Gözlük Göz	Glasses Eye
Birlik	Union	Birlik Bir	Union One
Yemek	Food	Yemek Ye	Food Eat
İçine	Into	İçine İç	Into Inner
Kötülük	Wickedness	Kötülük Kötü	Wickedness Bad

Tablo 3.3 Birden fazla yeniden yazılan kelimelerin örneği

Kelime	**İnglizce Anlam**	**Yeniden yazılanların sayısı * Kök**	**İnglizce Anlam**
Güzellik	Beauty	2 * güzel	Beautiful
Güzeller	Beauties	3 * güzel	Beautiful
Çalışmalar	Studies	3 * çalış; 3 * çal	Work; Steal
Kötülük	Wickedness	1 * kötülük; 2 * kötü	Wickedness; Bad
Gözlükçüler	Opticians	2 * gözlük; 2 * göz	Glasses; Eye

Bütün bu değişiklikler, bir sonraki alt bölümde tarif edilecek olan dilin morfolojisine göre gerçekleşir.

3.1.4.1.1 Türkçe Dili Morfolojisi

Türkçe dilinin tamamen sondan eklemeli olması ve son ekinin yalnızca dil eki türü olması nedeniyle doğal diller içerisinde özel bir yeri vardır. Aslında Türkçe dilini bilen herkesin, belirli bir sözcüğün kök olduğunu bilmese bile bir kelimeyi kolayca analiz edebileceği söylenmiştir. Aşağıda Türkçe dili fonolojik kurallarına örneğin dili etkileyen önemli faktörler verilmiştir. (Chief ve diğ., 2014).

(her hangi bir kelime)lerim → (her hangi bir kelime)ler-im.

"ler" çoğul ekidir ve "im" ilk kişi tekildir. Aşağıda, Türkçe diliyle ilgili kurallardan bazıları verilmiştir.

1. Eklerin hepsi sondan eklemelidir.
2. Çoğul bir son ek iyelik ekini takip edemez.
3. Türkçe bir son ek, eklendiği kelimede ses uyumu sağlamak için çoklu biçimbirimciğe sahip olabilir.
4. Türkçe'de her sesli harf ayrı bir heceye işaret eder.
5. Türkçe'de, tek heceli kelimeler çoğunlukla kökün kendisidir.

6. Bir sözcüğün ad kökenli fiil ekleri varsa, her zaman sözcüğün sonunda görünürler. İsim eklerinin ya da isim son eklerinin yokluğunda kökün kendisini takip ederler.
7. Türkçede "-lar" son eki hem ad kökenli fiil eki olarak (üçüncü kişi çoğul şimdiki zaman) hem de isim son eki olarak (çoğul) kullanılabilir
8. Türkçe'de kelimeler 'b', 'c', 'd' ve 'ğ' harfleriyle bitemez. Bununla birlikte, sesli bir harf ile başlayan bir son ek, 'p', 'ç', 't' veya 'k' ile biten bir kelimeye eklendiğinde, son ünsüz, 'b', 'c', 'd' veya 'ğ' olarak dönüştürülür.

Eklerin farklı kategorileri aşağıdaki Tablo 3.4'te gösterilmiştir.

Tablo 3.4: Son ek sınıfları

Sınıf	**Tip**
Nominal fiil ekleri	Enfeksiyon
Türetilmiş Son ekler	Yapım
İsim son ekleri	Enfeksiyon
Zaman ve kişi fiil ekleri	Enfeksiyon
Fiil ekleri	Enfeksiyon

3.1.4.1.1.1 Son-ek Biçimbirimsel Değişikliği

Son-ek biçimbirimsel değişikliği bir kelimenin anlamını değiştirmez, sadece iyi ses oluşturmak için kullanılırlar. Bir ekin parantez içinde bir harfi varsa, o zaman ihmal edilebilir. Benzer şekilde, eğer bir son-ek büyük harf içeriyorsa, o son-ekin biçimbirimsel değişikliğini sahip olduğu söylenebilir. Tablo 3.5, son ek biçimbirimsel değişikliğinın, Tablo 3.6 ad kökenli fiil ekleri, Tablo 3.7 İsim son-ekleri, Tablo 3.8 türeten ek örnekleri örneğini göstermektedir.

Tablo 3.5: Son ek biçimbirimsel değişikliğinın örneği

Harf	**Biçimbirimsel Değişikliği**
U	ı, i, u, ü
C	c, ç
A	a, e
D	d, t
I	ı, I

Tablo 3.6: Ad kökenli fiil ekleri örneği

a/a	**Son-ek**	**a/a**	**Son-ek**
1	-(y)Um	2	-sUn
3	-(y)Uz	4	-sUnUz
5	-lAr	6	-md
7	.-n	8	-k
9	-nUz	10	-DUr
11	-cAsInA	12	-(y)DU
13	-(y)sA	14	-(y)mUş
15	-(y)ken		

Tablo 3.7: İsim son ekleri örneği

a/a	**Son-ek**	**a/a**	**Son-ek**
1	-lAr	2	-(U)m
3	-(U)mUz	4	-(U)n
5	-(U)nUz	6	-(s)U
7	-lArI	8	-(y)U
9	-nU	10	-(n)Un
11	-y(A)	12	-nA
13	-DA	14	-nDA
15	-Dan	16	-nDAn
17	-(y)la	18	-ki
19	-n(c)A		

Tablo 3.8: Türeten ek örnekleri

a/a	**Son-ek**	**a/a**	**Son-ek**
1	-lUk	2	-CU
3	-Cuk	4	-lAş
5	-lA	6	-lAn
7	-CA	8	-lU
9	-sUz		

3.1.4.1.1.2 Ünlü uyumu

Bu kontrol, son iki sesli harfinin ünlü uyumun kurallarına uyup uymadığını kontrol eder. Ünlülerin uyumu hakkında kısa bir açıklama aşağıda verilmiştir. Türkçe ünlü uyumu sistemi, ünlülerin ön ve yuvarlak olan iki özelliği ile karakterize ettiği iki boyutlu sesli uyum sistemleridir. Her özelliğin kendi kuralları vardır.

1. Büyük ünlü uyumu kuralları: Türkçedeki ünlüler, üretildikleri yere göre ikiye ayrılır. 'E', 'i', 'ö' ve 'ü' gibi ince üretilen sesler ağzın önünde ve 'a', 'ı', 'o' ve 'u gibi kalın üretilen ünlü boğaza yakın oluşturulur. Kurallara göre, bir kelime hem kalın hem de ince ünlüleri içeremez. Bu, sözcüklerin büyük ünlü uyumu kuralına uymaları için farklı biçimler alabilmesinin nedenlerinden biridir.

2. Küçük ünlü uyumu kuralları: Türkçedeki ünlüler, üretirken dudakların yuvarlanıp yuvarlanmadığına göre ikiye ayrılır. ‘o’, ‘ö’, ‘u’ ve ‘ü’ gibi ünlüler yuvarlanırken, ‘a’, ‘e’, ‘ı’ ve ‘i’ gibi ünlüler yuvarlanmaz. Uyum kurallarına göre, bir sözcüğün herhangi bir hecesinde yuvarlak ünlü ‘o’, ‘ö’, ‘u’ ve ‘ü’ varsa bunu izleyen ilk hecede ‘a’, ‘e’,’u’ veya ‘ü’ ünlüsü bulunmalıdır.

3.1.4.1.1.3 Son Ünsüz

Türkçe'de, ünlü ile biten bazı kelimelerin sonuna ünlü ile başlayan başka bir ek getirilmesi durumunda ek ile kök arasına bir ünsüz harf eklenir. Bu ünsüz harfler birleştirme ünsüzleri olan ‘y’, ‘n’ veya ‘s’ olabilir. Son ekten önce bir birleştirme ünsüzünün eklenmesi durumunda, son ekin gösterimi parantezle çevrelenen isteğe bağlı ünsüz ile başlar (örneğin (y) Um, - (n) cA). Bu tür son ekler için, eğer bir birleştirilmiş ünsüz varlığı varsa, birey kökün bir sesli harf ile bitip bitmediği kontrol etmelidir.

Son-ekten önce 'y' ünsüzü yoksa, son-ekin yalnızca kök kısmı (ör. -Um) kaynaştırma amaçlı seçilir. Eğer bir 'y' ünsüzü varsa ve bir sesli harften önce gelmişse, 'y' bir birleştirme ünsüzü olarak kabul edilir. Eğer 'y' den hemen önce bir ünsüz varsa, 'y' ekinin gerekliliğine karar verilir. Böyle bir durumda, fazla

yönlendirmeyi önlemek için imleç ilerlemez. Son durumda ki özellik, 'lityum' gibi başka bir dilden (Lityum elementi anlamına gelir) gelen kelimelerde ortaya çıkabilir. Sesli uyumunun kontrolü yapılmazsa, kelime 'lit' şeklinde kesilir, çünkü '- (y) Um' kendisine yapıştırılmış bir son-ek olarak kabul edilir. Ancak Türkçenin morfolojik kurallarına göre bu kelimenin son sözü 'lityum' değil 'litim' olurdu.

3.1.4.1.1.4 Seslerin Birleştirilmesi

Ünsüzlerin birleştirilmesi kuralı gibi, ünsüzler ile başlayan bazı ekler için de ünlü birleştirmesi vardır. '- (U) muz' son ekinde olduğu gibi ünsüz ile biten bir gövdeye ünsüz ile başlayan bir son ek eklenebilir. Böyle bir durumda, telaffuz ve kolaylık için kök ve gerçek son ek (örneğin, '-mUz') arasına bir U eki ('ı', 'i', 'u' veya 'ü') eklenir. Aşağıda verilen Tablo 3.9 birleştirilen sesler örneğini göstermektedir.

Tablo 3.9: Birleştirilen sesler örneği

Kelime/Analiz	**İnglizce Anlam / Kök**
Kalelerimizdekilerden	From the ones at one of our castles
Kale-lAr-UmUz-DA-ki-lAr-DAn	Kale
Çocuğuymuşumcasına	As if I were her child
Çoçuk-(s)U-(y)mUş-(y)-Um-cAsInA	Çoçuk
Kedileriyle	With their cats
Kedi-lAr-(s)U-(y)lA	Kedi
Çocuklarımmış	Someone told me that they were my children
Çoçuk-lAr-(U)m-(y)muş	Çoçuk
Kitabımızdı	It was our book
Kitap-UmUz-(y)DU	Kitap

3.1.5 Kelime Sözlüğü

İngilizce sözcükler için elde edilmiş olan ve bu tezde kullanılacak kelime sözlüğü, uzun yıllardır derlenen ve 6800'den fazla pozitif – negatif İngilizce kelimeyi içeren bir kelime haznesidir (Hu ve Liu, 2004). Bu tez çalışmasında Türkçe

cümleler üzerinde analizler yapabilmek amacı ile tüm kelimeler, Google Çeviri ve Turing yardımı ile Türkçe kelimelere tek tek elle çevrilmiştir.

3.1.6 Test Sözlüğü Kelime Sözlüğü Eşleştirme

Tüm tivit ön işleme, dizgecikleme ve gövde oluşturma aşamasından sonra, bir tivit içindeki her dizge, sözlükteki bir sözcükle eşleşir.

3.1.7 Duyarlılık Polaritesinin Hesaplanması

Bir tivitin duygu polaritesi, kelimelerin kelime sözlüğündeki sözcüklerle eşleştirilmesinden sonra hesaplanır. Uygulama aşamasında bir kelime pozitif kelimelerin listesinde bulunursa, kelime pozitif, aksi durumda tam tersi olarak işaretlenir.

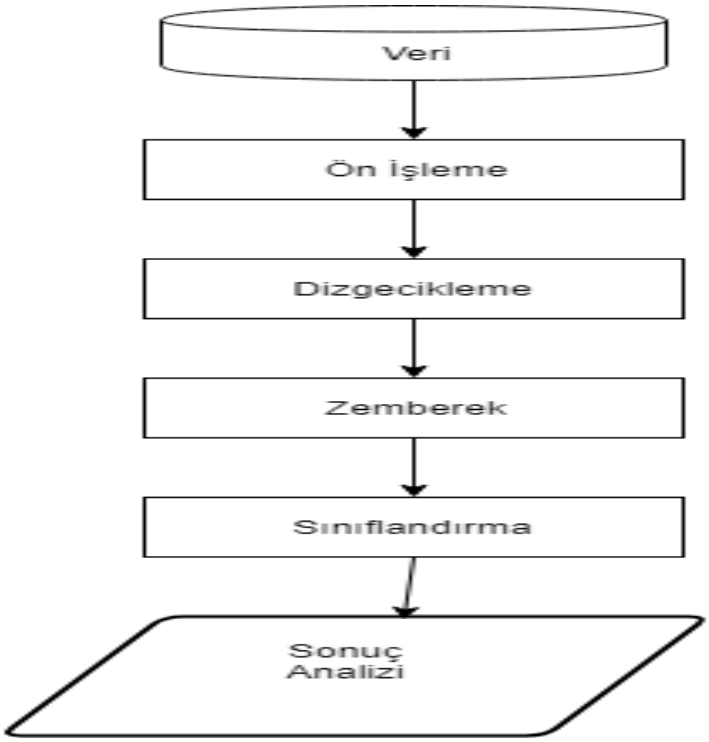
3.1.8 Sonuçların Analizi

Son olarak, bir tivitin pozitif, negatif veya nötr olarak sınıflandırılabilmesi için kutupluluk açısından analiz edilmesi gerekir. İçerisindeki pozitif kelimelerin sayısı, negatif kelimelerin sayısından fazla olduğu söylenirse, bu tivit pozitif olarak sınıflandırılır. Eğer, içerisindeki negatif kelimelerin sayısı, pozitif kelimelerin sayısından fazla olduğu söylenirse negatif olarak sınıflandırılır. Ve bir tivit, içindeki pozitif sözcüklerin sayısı, içindeki negatif kelimelerin sayısına eşitse veya hem olumlu hem de olumsuz kelimelerde eşleşme bulamazsa, tarafsız olarak sınıflandırılır.

3.2 Yapay Zeka

İkinci yöntem, tivitleri pozitif, negatif veya nötr olarak sınıflandırmak için farklı algoritmalara sahip makine öğrenmesi sınıflandırıcılarını kullanmaktır.

Şekil 3.2, yapay zeka kullanılan yöntemin önerilen sistem akışını göstermektedir ve yapıyı ayrıntılı olarak açıklamaktadır.



Şekil 3.2 Yapay zeka için duygu analizi süreç akışı

İlk dört adımda, önceki yönteme uygulanan aynı prosedürler aynı şekilde bu yönteme de uygulanır.

3.2.1 Sınıflandırma

Makine öğrenmesinde sınıflandırma, önemli veri sınıflarını tanımlayan ya da gelecekteki eğilimleri tahmin edebilen bir veri analizi biçimidir; sınıflandırıcı olarak modelleri kullanarak kategorik (ayrık, sırasız) etiketlerin işlevlerini önceden tahmin eder (Han ve diğ. 2012).

Makine öğrenimi duygu analizini daha kolay hale getirir ve bu durumda tivit sınıfını pozitif, negatif veya nötr olarak tahmin edecek bir model ortaya koyar.

Makine öğrenimi gerçekleştirebilmek için iki farklı algoritma kullandık. Bu tezde sınıflandırmada kullanılacak bu iki farklı algoritma hakkında daha detaylı kısa bilgi aşağıdaki alt bölümde verilmiştir.

3.2.1.1 Destek Vektör Makineleri Kullanarak Sınıflandırma

Destek vektör makinesi (SVM), hem sınıflandırma hem de regresyon problemlerini çözmek için kullanılabilecek ilişkili öğrenme algoritmasına sahip bir denetimli makine öğrenmesi (ML) modelidir. Regresyon için kullanılabilmesine rağmen, SVM daha çok sınıflandırma problemleri için kullanılmaktadır. SVM ile ilgili daha fazla bilgi, Bölüm 2.2.1.1.3.1'de görülebilir

3.2.1.2 Rasgele Orman Algoritması

Rasgele orman (RF) algoritması, bir grup-öğrenme algoritmasıdır. Bunun arkasındaki motivasyon, az sayıda özelliği olan az sayıda karar ağacının oluşturulması, hesaplama açısından ucuz bir süreçtir. Birçok küçük, zayıf karar ağacı paralel olarak inşa edildiğinde, tek bir güçlü ağaç, ortalama ya da çoğunluk oyu alarak inşa edilebilir. Diğer bir deyişle, RF algoritması denetimli bir öğrenme algoritmasıdır ve isminden de anlaşılacağı gibi, bir şekilde bir orman yaratma ve bunu rasgele yapma ile ilgilidir. Ağaç sayısı ve sonuç arasında doğrudan bir ilişki vardır; ormandaki ağaç sayısı ne kadar büyükse, sonuç o derece doğru olacaktır. Pratikte, RF başarı düzeyi yüksek bir öğrenme algoritması olarak kabul edilmektedir.

Bunlara ek olarak aşırı uygunluk, birçok algoritma için en kötü sonuçları ortaya çıkaran sorunlardan biridir. Ancak Ormandaki algoritmalar, ormanlarda yeteri kadar ağaç varsa sınıflandırıcı modeline aşırı uyum sağlamayacaktır. RF algoritmaları hem sınıflandırma hem de regresyon görevleri için kullanılabilen, kategorik değerler için modellenen ve eksik değerleri ele alan bir algoritma türüdür. Tüm sınıflandırma alanı için en iyi algoritma olmamasına rağmen, geçmiş yıllarda birçok kişi tarafından en çok tercih edilen sınıflandırıcı olmuştur.

4. UYGULAMA SONUÇLARI

Bu tez de elde edilen sonuç ve gelecekteki çalişma aşağıdaki alt bölümde verilmiştir.

4.1 Performans Ölçütleri

Bir duyarlılık analiz sisteminin doğruluğu, insan kararları ile iyi bir fikir birliği içerisinde olmalıdır. Bu nedenle bir duyarlılık analiz sisteminin performansını değerlendirmek için sadece analizler yeterli değildir. Bu aşamada kesinlik, hassasiyet ve F1-skoru gibi diğer değerlendirme kriterlerini de kullanacağız. Değerlendirme kriterlerini ölçmek için, aşağıda gösterilen gibi bir karmaşa matrisinin oluşturulması gerekmektedir. Tablo 4.1 karışıklık matrisinin örneği göstermektedir.

Tablo 4.1: Karmaşa matrisi

		Doğru Sınıf (Deney tarafından onaylandığı gibi)	
		pozitif	negatif
Öngörülen Sınıf (Test tarafından tahmin edildi)	pozitif	TP True Pozitif	FP False Pozitif
	negatif	FN False Negatif	TN True Negatif

Aşağıdaki denklem 6.1 ve 6.2, sırasıyla doğruluk ve yanlış sınıflandırma oranını hesaplamak için kullanılabilir.

$$Doğruluk = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \quad (6.1)$$

$$Yanlış\ sınıflandırma\ oranı = \frac{FP+FN}{\text{TP+FP+FN+ TN}} \quad (6.2)$$

4.1.1 Kesinlik

Kesinlik, tahmin edilen pozitif vakaların oranıdır. Diğer bir deyişle kesinlik, doğru kabul edilen pozitif tahmin sayısının pozitif tahminlerin toplam sayısına bölüm oranıdır. Bunu hesaplamak için kullanılabilecek formül:

$$Kesinlik = \frac{\text{TP}}{TP+FP} \quad (6.3)$$

olarak verilebilir.

4.1.2 Hassasiyet

Hassasiyet, doğru olarak tespit edilen pozitif vakaların oranıdır. Diğer bir deyişle, doğru pozitif tahmin sayısının toplam pozitif sayısına bölüm oranıdır ve

$$Hassasiyet = \frac{\text{TP}}{TP+FN} \quad (6.4)$$

denklemi ile hesaplanabilir.

4.1.3 F1-Skoru

F1 puanı, gerçek pozitif (Hassasiyet) ve kesinliğin ağırlıklı ortalamasıdır ve

$$F1\ Skoru = 2 * \frac{\text{kesinlik*hassasiyet}}{\text{kesinlik +hassasiyet}} \quad (6.5)$$

denklemi kullanılarak hesaplanabilir.

4.1.4 Diğer Performans Ölçütleri

Bir duygu analiz sisteminin performansını ölçmek için kullanılabilecek bir dizi başka değerlendirme yöntemi de mevcuttur. Her ne kadar bu tezde performansın

ölçülmesi için sadece doğruluk, kesinlik, Hassasiyet ve F1 puanını kullanacak olsak da, aşağıdaki değerlendirme kriterlerine de bir göz atalım.

- MCC (Mathews Korelasyon Katsayısı)

MCC, korelasyon matrisindeki tüm dört değer kullanılarak hesaplanabilen bir korelasyon katsayısıdır ve

$$MCC = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (6.6)$$

denklemi kullanılarak hesaplanabilir.

- ROC Eğrisi

ROC eğrisi, bir sınıflandırıcının tüm eşikler üzerindeki performansını özetleyen bir grafiktir. Y Ekseni'nde çizilen Gerçek Pozitif Hız ile X ekseninde çizilen Yanlış Pozitif Hız'a göre oluşturulur.

4.2 Örnek Veri

Twitter'den 13.000 veri örneği indirildi ve sonra veriler temizlendi, 435 verinin bir kısmı gereksiz olarak belirlendi ve 12.565 veriden arkaya kalan bütün veriler silindi. Kalan 12.565 veri içeriğindekilerden dolayı hem pozitif-negatif ya da nötr olarak el ile etiketlendi. Başkent Üniversitesi, Türkiye (Hayran ve Sert, 2017) den alınan karşılaştırmalı değerlendirme olan 2000 (1000 tanesinin her biri pozitif ya da negatif) veri alındı, indirilen veriden alınan 1000 nötr veri de eklendiğinde toplamda 3000 veri doğrulama için kullanılmış oluyor. Kalan 11,565 veri sonucu bulmak için analiz için kullanıldı. Analiz için kullanılan veri her biri pozitif, negatif, veya nötr olmak üzere 3.500 kelime oluşturuyor.

Şifreleme ve 'acaba, ben, benim, bize, var, ve, seni, son' gibi diğer 205 gereksiz Türkçe kelimenin çıkarılması gibi birkaç ön işleme faaliyetlerinden sonra analiz için kullanılan 10.500 veri bir diğer 21.500 veri daha oluşturmak için kullanıldı ve 10.500 veri elde ettik. Daha sonra ön işleme verilerini kelimelerin kökenlerini bulmak için kullandık. Sonuç olarak, şifreleme ve gereksiz kelimelerin çıkarılmasından sonraki verilerle orijinal veriler ve doğrulama için kullanılan 3000 veriyi bağladığımızda başka bir 10.500 veri oluşturmak toplamda 34.500 veri yapıyor. Veriler Tablo 4.2 gösterildiği gibi farklı konulardan indirildi.

Tablo 4.2 İndirilen veri sayısı ve ilgili konular

Tarih	Konu	İndirilen veri sayısı
26/02/2018	#merhamet	160
27/02/2018	#bokoharam #namaz #isis	272
28/02/2018	#tecavuz #darbe	139
01/03/2018	#sendeyaz #spor	178
19/08/2018	#bayram	1241
19/08/2018	#dolar	364
19/08/2018	#NelerOluyorCHP	1000
19/08/2018	#Pazar	403
19/08/2018	#dünya	8
19/08/2018	#politika	128
19/08/2018	#ekonomi	1000
19/08/2018	#eğlence	587
19/08/2018	#pist	4
19/08/2018	#yaz	610
19/08/2018	#hava	81
19/08/2018	#hayat	475
19/08/2018	#spor	430
19/08/2018	#seçim	21
19/08/2018	#parti	81
19/08/2018	#Yahudi	50
19/08/2018	#MutluOlacaksınDeseler	551
19/08/2018	#TevfikFikret	254

Tablo 4.2 (devam): İndirilen veri sayısı ve ilgili konular

20/08/2018	#AK Parti 6	8
20/08/2018	#islam	42
20/08/2018	#iphone	434
20/08/2018	#america	121
20/08/2018	#özil	504
20/08/2018	#kultur	28
20/08/2018	#insan	315
20/08/2018	#toprak	54
20/08/2018	#tatlı	181
20/08/2018	#cennet	75
20/08/2018	#cehennem	25
20/08/2018	#küfür	128
20/08/2018	#mutlu	367
20/08/2018	#beğen	237
24/08/2018	#millet	153
24/08/2018	#siyaset	252
24/08/2018	#magazin	39
24/08/2018	#FenerinMaçıvar	1000
24/08/2018	#BilmemFarkındaMısın	1000
	Toplam	13,000

4.3 Simulasyon Sonuçları

Analiz, 3.4.3 R sürümü ile yapılmıştır ve ilk durumda 3.000 ikinci durumda ise 10.500 veri kümesinin kullanıldığı iki farklı simulasyon ortamında

değerlendirilmiş ve sonuçlar buna göre belirlenmiştir. Sınıflandırma yönteminin ilk durumunda, 750 veri (her biri pozitif, negatif ve nötr) eğitim verisi olarak kullanılmıştır ve 250 (her biri pozitif, negatif ve nötr) veri ise test verisi olarak kullanılmıştır. Bunun yanında 3,000 veri (her birinde 1000 veri olmak üzere pozitif, negatif ve nötr) PL yönteminde sonucu bulmak için kullanılmıştır. Orijinal veri kümesine ham veri kümesi denir, etkisiz, gereksiz kelimeler çıkarıldıktan ve dizgecikleme gerçekleştirildikten sonra elde edilen veri kümesine etkisiz-kelime (stop-word) veri kümesi ve son olarak da verilerin kökü bulunduktan sonra elde edilen veri kümesine gövdelenmiş (stemmed) veri kümesi denir. Elde edilen sonuçlar aşağıdaki tablolarda sunulmuştur.

Tablo 4.3: İlk veri kümesindeki (a) PL, (b) SVM, (c) RF algoritması, kullanılarak ham verilerden elde edilen sonuç

(a)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	317	114	250
	Negatif	254	455	178
	Nötr	429	431	572

(b)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	116	1	3
	Negatif	43	198	81
	Nötr	91	51	166

(c)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	182	3	5
	Negatif	10	218	52
	Nötr	58	29	193

Tablo 4.3a, birinci veri kümesinin ham verileri üzerinde PL kullanılarak elde edilen sonucu göstermektedir. Bu tabloda, PL'nin, pozitif ve negatif verilere kıyasla nötr verilerde daha iyi olduğu görülmektedir. Aynı zamanda PL'nin, nötr verilerin sınıflandırılması söz konusu olduğunda %50'den fazla bir doğruluk sağladığı görülmektedir.

Tablo 4.3b, ilk veri kümesinin ham verileri üzerinde SVM kullanılarak elde edilen sonucu göstermektedir. Bu tabloda SVM'nin negatif veriler üzerinde hem olumlu hem de nötr verilerden daha iyi performans gösterdiği görülmektedir. Nötr verilerin %66'sından fazlası SVM tarafından doğru şekilde sınıflandırılırken, pozitif verilerin sadece %46.4'ünün doğru şekilde sınıflandırıldığı görülmektedir.

Tablo 4.3c, ilk veri kümesinin ham verileri üzerinde RF kullanılarak elde edilen sonucu göstermektedir. Bu tabloda, RF algoritmasının negatif verilere %87.2'lik bir doğrulukla sahip olduğu, hem pozitif hem de nötr veriler yerine, negatif veriler üzerinde daha iyi performans gösterdiği görülmektedir. Ayrıca bu tabloda, RF kullanılarak yapılan sınıflandırılmalarda %72'den fazla bir doğruluk performansının elde edildiği söylenebilir.

Tablo 4.4: İlk veri kümesindeki ham verileri kullanarak elde edilen sonucun performansı

	Yanlış sınıflandırma oranı	Kesinlik	Hassasiyet	F1 skoru
PL	0.552	0.459	0.448	0.453
SVM	0.36	0.707	0.64	0.672
RF	0.209	0.809	0.791	0.8

Tablo 4.4, ilk veri setinin ham verileri üzerinde elde edilen tüm sonuçların performansını göstermektedir. Tablodan RF'nin en iyi performansa sahip yöntem olduğu açıkça görülebilir. Genel durumda RF %79.1'lik bir hassasiyetla ile en yüksek performansı elde etti, ardından %64 ile SVM ve onun da ardından %44.8 ile PL geldi.

Tablo 4.5: İlk veri kümesindeki (a) PL (b) SVM (c) RF algoritması kullanılarak etkisiz-kelimelerden elde edilen sonuçlar.

(a)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	446	186	290
	Negatif	203	423	133
	Nötr	351	391	577

(b)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	118	1	3
	Negatif	42	201	79
	Nötr	90	48	168

(c)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	179	3	6
	Negatif	7	220	45
	Nötr	64	27	199

Tablo 4.5a, birinci veri kümesinin etkisiz-kelime verisinde PL kullanılarak elde edilen sonucu göstermektedir. Bu tabloda PL'nin hem olumlu hem olumsuz verilere göre nötr verilerde daha iyi olduğu görülmektedir. Aynı zamanda PL'nin, nötr verilerin sınıflandırılması söz konusu olduğunda %50'den fazla bir doğrulukla gerçekleştirdiği görülmektedir.

Tablo 4.5b, ilk veri kümesinin etkisiz-kelime verisinde SVM kullanılarak elde edilen sonucu göstermektedir. Bu tabloda SVM'nin negatif veriler üzerinde, hem olumlu hem de nötr verilere kıyasla daha iyi performans gösterdiği görülmektedir. Nötr verilerin %67'sinden fazlası SVM tarafından doğru şekilde sınıflandırılırken, pozitif verilerin sadece %47.2'sinin doğru şekilde sınıflandırıldığı görülmektedir.

Tablo 4.5c, ilk veri kümesinin etkisiz-kelime verisinde RF kullanılarak elde edilen sonucu göstermektedir. Bu tabloda RF'nin negatif verilere %88'lik bir doğrulukla sahip olduğu görülmektedir. Hem pozitif hem de nötr verilerden ziyade negatif veriler üzerinde daha iyi performans gösterdiği görülmektedir. Ayrıca Tablo 4.5c'den, her sınıfın RF kullanılarak sınıflandırılmasında %71'den fazla bir doğruluk performansının elde edildiği söylenebilir.

Table 4.6: İlk veri kümesindeki etkisiz-kelimeleri kullanılarak elde edilen sonucun performansı

	Yanlış sınıflandırma oranı	Kesinlik	Hassasiyet	F1 skoru
PL	0.518	0.493	0.482	0.487
SVM	0.351	0.713	0.649	0.679
RF	0.203	0.816	0.797	0.806

Tablo 4.6, ilk veri kümesinin etkisiz-kelime verisinde elde edilen tüm sonuçlarını göstermektedir. Tablodan RF'nin en iyi performansa sahip yöntem olduğu açıkça görülebilir. RF genel durumda %79.7'lik bir hassasiyetla ile daha yüksek bir performans elde etmiştir. Ardından %64.9 ile SVM ve ardından da %48.2 ile PL gelmektedir.

Tablo 4.7a, birinci veri setinin gövdelenmiş verileri üzerinde PL kullanılarak elde edilen sonucu göstermektedir. Bu tabloda PL'nin hem pozitif hem de negatif verilerden daha fazla nötr verilerde iyi olduğu gözlemlenmiştir. Aynı zamanda, PL'nin, sadece nötr verilerin sınıflandırılmasında %66'dan daha fazla bir doğruluk payı olduğu gözlemlenmiş ve pozitif ile negatif verileri sınıflandırılması söz konusu olduğunda %53'ten az olduğu görülmüştür.

Tablo 4.7b, ilk veri kümesinin gövdelenmiş verileri üzerinde SVM kullanılarak elde edilen sonucu göstermektedir. Bu tabloda SVM'nin negatif veriler üzerinde hem pozitif hem de nötr verilerden daha iyi performans gösterdiği görülmektedir. Nötr verilerin %80'i SVM tarafından doğru şekilde sınıflandırılırken, bu durumda pozitif verinin sadece %63.6'sının doğru şekilde sınıflandırıldığı görülmektedir.

Table 4.7a: İlk veri kümesindeki (a) PL (b) SVM (c) RF algoritması kullanılarak gövdelenmiş verilerden elde edilen sonuçlar

(a)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	520	137	193
	Negatif	147	525	142
	Nötr	333	338	665

(b)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	159	11	18
	Negatif	40	214	32
	Nötr	51	25	200

(c)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	173	8	13
	Negatif	15	212	53
	Nötr	62	30	184

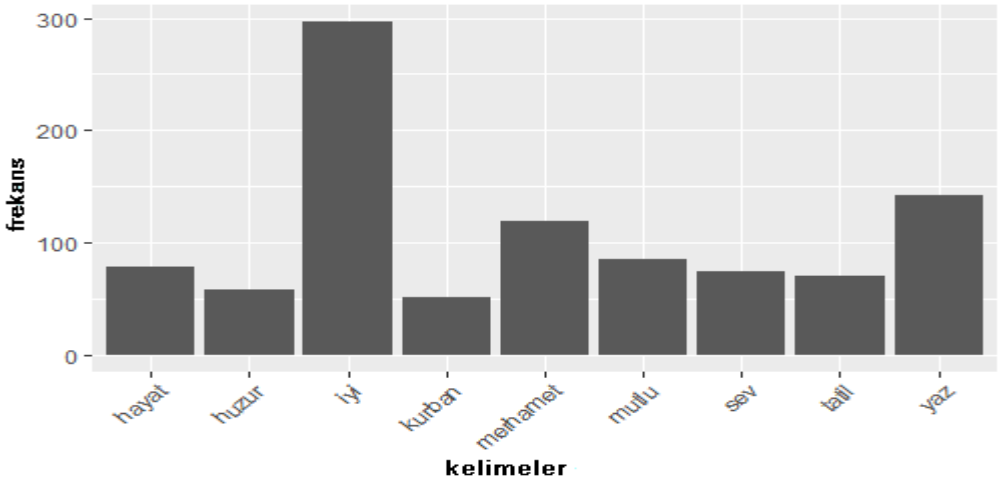
Tablo 4.7c, birinci veri kümesinin gövdelenmiş verileri üzerinde RF kullanılarak elde edilen sonucu göstermektedir. Bu tabloda RF'nin negatif verilerin sınıflandırılmasında %84'ten fazla bir doğruluk payı olduğu görülmektedir. Ayrıca bu tabloda RF kullanılarak sınıflandırmada %69 doğrulukta bir performans elde edildiği söylenebilir.

Tablo 4.8, ilk veri setinin gövdelenmiş verileri üzerinde elde edilen tüm sonuçların performansını göstermektedir. Tablodan açıkça görüleceği gibi SVM, en iyi performansa sahip yöntemdir. Genel durumda %76.4'lük bir hassasiyetla daha yüksek bir performans elde etti, ardından %75.9 ile RF ve %57 ile PL gelmektedir.

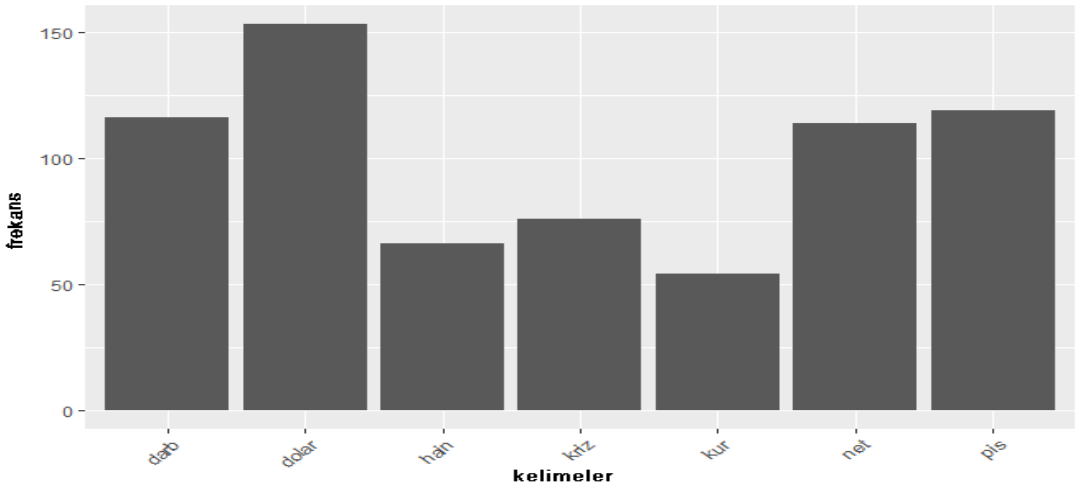
Tablo 4.8: İlk veri kümesindeki gövdelenmiş veriler kullanılarak elde edilen sonucun performansı

	Yanlış sınıflandırma oranı	Kesinlik	Hassasiyet	F1 skoru
PL	0.43	0.585	0.57	0.577
SVM	0.236	0.773	0.764	0.768
RF	0.241	0.772	0.759	0.765

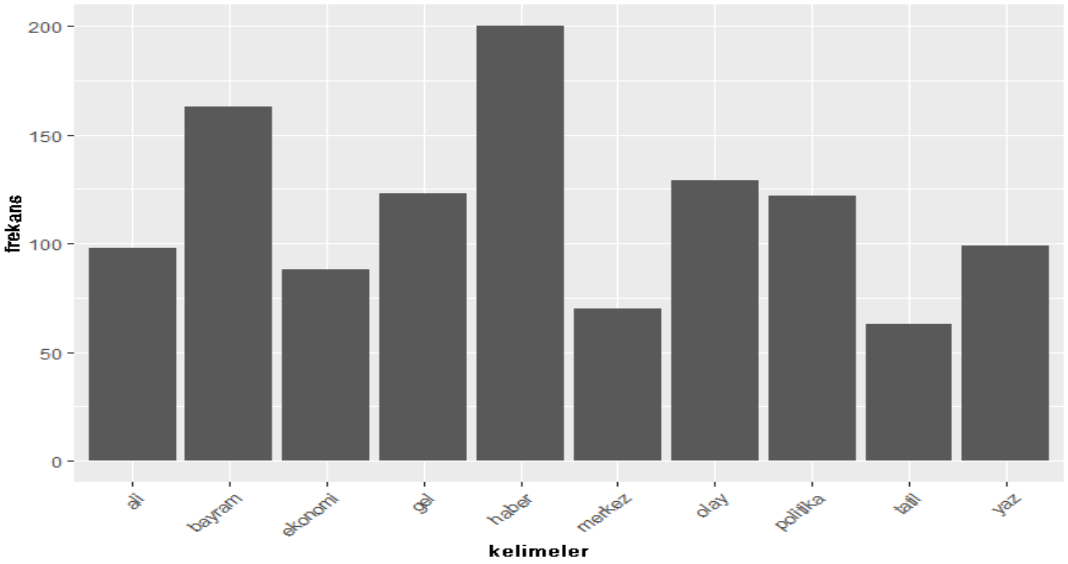
Aşağıdaki şekil 4.1, 4.2 ve 4.3 çubuk grafikler, ilk durumda hem pozitif hem negatif hem de nötr tivitlerde en fazla ortaya çıkan kelimeleri (50'den fazla zaman gösteren kelimeler) temsil eder ve şekil 4.4, 4.5 ve 4.6 kelime bulutu, her pozitif, negatif ve nötr tivitin en çok kullanılan ilk 50 kelimesinden oluşur.



Şekil 4.1: İlk veri kümesindeki pozitif tivitlerde en sık kullanılan sözcükler



Şekil 4.2: İlk veri kümesindeki negatif tivitlerde en sık kullanılan sözcükler

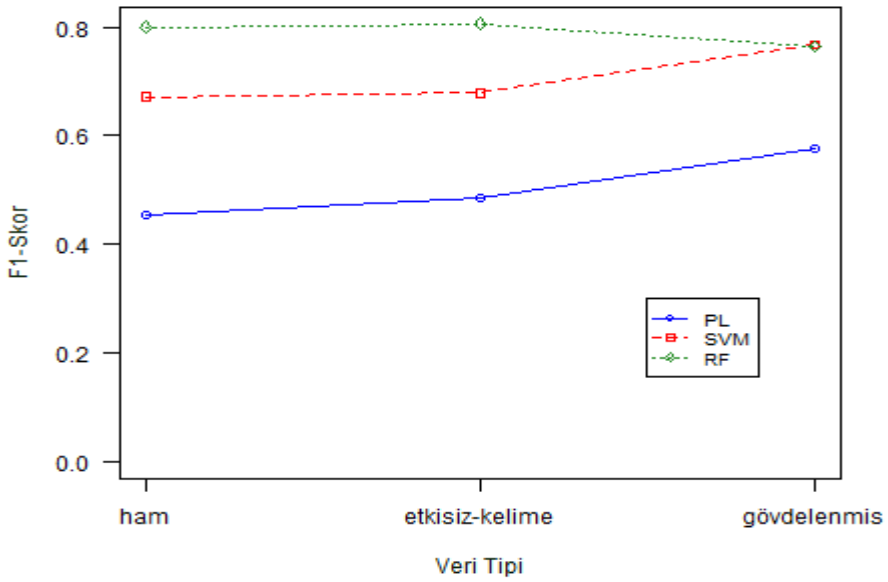


Şekil 4.3: İlk veri kümesindeki nötr tivitlerde en sık kullanılan sözcükler

Şekil 4.4: İlk veri kümesindeki pozitif kelimelerin kelime bulutu

Şekil 4.5: İlk veri kümesindeki negatif kelimelerin kelime bulutu

Şekil 4.6: İlk veri kümesindeki nötr kelimelerin kelime bulutu



Şekil 4.7: İlk veri kümesinde kullanılan her bir yöntemde elde edilen performansı gösteren grafik.

Tablo 4.9: İlk veri kümesindeki sonucu hesaplamak için her yöntemi aldığı süresi

	Ham	Etkisiz-kelime	Gövdelenmiş
PL	1.69 dk	1.48 dk	1.62 dk
SVM	0.65 dk	0.68 dk	0.3 dk
RF	69 dk	89.4 dk	11.52 dk

Şekil 4.7 ve Tablo 4.9'deki elde edilen sonuçtan yola çıkarak, RF'nin performansı ham verilerden etkisiz-kelimelere ve daha sonra elde edilen verilere doğru gididiğinde azalır. SVM'nin yanı sıra PL'nin performansı, verilerin ham veriden etkisiz-kelime verisine dönüşümü ve daha sonra da verilerin gövde bulunduktan sonra sürekli olarak artar. SVM, veri gövdelendiktan sonra küçük bir farkla RF'ye geçer. SVM ayrıca, kullanılan diğer yöntemlerle karşılaştırıldığında en kısa sürede (bir dakikadan az) yürütülen yöntem olarak göze çarpmaktadır.

Sınıflandırma yönteminin ikinci durumunda, eğitim veri seti olarak 2500 veri (her biri pozitif, negatif ve nötr), test verileri olarak 1000 (her biri pozitif ve negatif) veri kullanırken, tüm veri 10.500 olarak (her biri 3.500 olacak şekilde) kullanılmıştır. PL kullanan yöntemde sonucu bulmak için pozitif, negatif ve nötr her biri kullanılır. Yukarıdaki örnekte olduğu gibi, orijinal veri kümesine ham veri kümesi denir, etkisiz-kelimeler kaldırıldıktan sonra ve dizge yapıldıktan sonra veri kümesi etkisiz-kelime veri kümesi olarak adlandırılır ve daha sonra gövde bulunduktan sonra veri kümesi gövdelenmiş veri kümesi olarak adlandırılmıştır. Elde edilen sonuçlar aşağıdaki tablolarda sunulmuştur.

Tablo 4.10a, ikinci veri kümesinin ham verileri üzerinde PL kullanılarak elde edilen sonucu temsil etmektedir. Bu tabloda, PL'nin nötr veriler üzerinde hem pozitif hem de negatif verilerden daha iyi performans gösterdiği görülmektedir. Aynı zamanda, PL'nin, nötr verilerin sınıflandırılması söz konusu olduğunda, %50'den fazla bir doğrulukla gerçekleştirdiği görülmektedir.

Tablo 4.10b, ikinci veri kümesinin ham verileri üzerinde SVM kullanılarak elde edilen sonucu temsil etmektedir. Bu tabloda SVM'nin, negatif verileri sınıflandırmada pozitif ve nötr verilere göre %62.2'lik bir doğrulukla daha iyi performans göstermektedir. SVM'nin pozitif sınıflandırmada ise %59.8 ve negatif sınıflandırmada sadece %58'lik bir doğruluk elde ettiği görülmektedir.

Tablo 4.10c, ikinci veri kümesinin ham verileri üzerinde RF kullanılarak elde edilen sonuç gösterilmektedir. Bu tabloda, RF'nin negatif verilere göre olumlu sonuç verdiği, sırasıyla %97.2 ve %95.6'lık bir doğrulukla negatif ve pozitif verilere sahip olduğu düşünülmektedir. RF, bu durumda nötr verilerin sınıflandırılmasında en düşük doğruluğa sahip olmasına rağmen, diğer yöntemlerin sınıflardırmasında elde edilen sonuçlardan %72.8'lik doğruluk elde etmeyi başarmıştır.

Tablo 4.10: İkinci veri kümesindeki (a) PL (b) SVM (c) RF algoritması kullanılarak ham verilerden elde edilen sonuçlar

(a)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	1427	401	795
	Negatif	780	1631	691
	Nötr	1293	1468	2014

(b)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	598	222	133
	Negatif	174	622	287
	Nötr	228	156	580

(c)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	956	0	24
	Negatif	16	972	248
	Nötr	28	28	728

Tablo 4.11: İkinci veri kümesindeki ham verileri kullanarak elde edilen sonucun performansı

	Yanlış sınıflandırma oranı	Kesinlik	Hassasiyet	F1 skoru
PL	0.517	0.497	0.483	0.49
SVM	0.4	0.601	0.6	0.6
RF	0.115	0.897	0.885	0.891

Tablo 4.11, ikinci veri kümesinin ham verileri üzerinde elde edilen tüm sonuçların performansını temsil eder. Tabloda rasgele orman'ın en iyi performansa sahip yöntem olduğu açıkça görülebilir. Genel durumda, rasgele orman %88.5'lik bir hassasiyetla daha yüksek bir performans elde eder ve ardından %60 SVM ve daha sonra %48.3 ile PL izlemektedir.

Tablo 4.12: İkinci veri kümesindeki (a) PL (b) SVM (c) RF algoritması kullanılarak etkisiz-kelime verilerinden elde edilen sonuçlar

(a)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	1760	569	954
	Negatif	566	1520	532
	Nötr	1174	1411	2014

(b)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	382	228	64
	Negatif	154	637	305
	Nötr	464	135	631

(c)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	740	20	48
	Negatif	14	956	220
	Nötr	246	24	732

Tablo 4.12a, ikinci veri kümesinden elde edilen etkisiz-kelime verisine PL kullanılarak elde edilen sonucu göstermektedir. Bu tabloda, PL'nin nötr veriler üzerinde hem pozitif hem de negatif verilerden %57'sinin üzerinde daha iyi performans gösterdiği görülmektedir. PL'nin yalnızca nötr veriler ile pozitif ve negatif verilerin sınıflandırılması söz konusu olduğunda sırasıyla %50.2 ve %43.4 doğrulukla çalıştığı görülmektedir.

Tablo 4.12b, ikinci veri kümesinin etkisiz-kelime verilerinde SVM kullanılarak elde edilen sonucu temsil eder. Bu tabloda, SVM'nin negatif verilerde daha iyi performans gösterdiği ve ardından nötr verileri elde edilen %63.7 ve %63.1'lik bir doğrulukla gösterdiği görülmektedir. Tablodan pozitif verilerin sadece %38.2'sinin doğru bir şekilde sınıflandırıldığı da görülebilir.

Tablo 4.12c, ikinci veri kümesinin etkisiz-kelime verilerinde RF kullanılarak elde edilen sonucu temsil eder. Bu tabloda, RF'nin hem pozitif hem de nötr verilerden ziyade negatif verilerde daha iyi performans gösterdiği ve negatif verileri %95'den fazla bir doğrulukla sınıflandırdığı görülmektedir. Ayrıca Tablo 4.12c'de, RF kullanarak her sınıfın sınıflandırılmasında %73'den fazla doğrulukla bir performans elde edildiği söylenebilir.

Tablo 4.13: İkinci veri kümesindeki etkisiz-kelime verileri kullanılarak elde edilen sonucun performansı

	Yanlış sınıflandırma oranı	Kesinlik	Hassasiyet	F1 skoru
PL	0.496	0.518	0.504	0.511
SVM	0.45	0.554	0.55	0.552
RF	0.191	0.817	0.809	0.813

Tablo 4.13, ikinci veri kümesinin etkisiz-kelime verisinden elde edilen tüm sonuçların performansını göstermektedir. Tablodan RF'nin en iyi performansa sahip yöntem olduğu açıkça görülebilir. Genel durumda, RF %80,9'luk bir hassasiyetla daha yüksek bir performansa sahip olduğu görülmektedir, ardından %55 ile SVM ve daha sonra %50.4 ile PL gelmektedir.

Tablo 4.14a, ikinci veri kümesinin gövdelenmiş verileri üzerinde PL kullanılarak elde edilen sonucu göstermektedir. Bu tabloda, PL'nin nötr verilere göre daha iyi performans gösterdiği görülmektedir. PL'nin nötr veriler üzerinde %60'ın üzerinde bir doğruluk, pozitif verilerde %52.3, negatif verilerde ise %51.1 doğruluk söz konusudur.

Tablo 4.14b, ikinci veri setinin gövdelenmiş verileri üzerinde SVM kullanılarak elde edilen sonucu temsil etmektedir. Bu tabloda, SVM'nin nötr veriler üzerinde daha iyi performans gösterdiği, ardından pozitif verilerde %72.7 ve %69'luk bir doğrulukla olduğu görülmektedir. Tabloda, negatif verinin sadece %61'inin doğru şekilde sınıflandırıldığı da görülebilir.

Tablo 4.14c, ikinci veri kümesinin gövdelenmiş verileri üzerinde RF kullanılarak elde edilen sonucu göstermektedir. Bu tabloda, aynı zamanda, %83'ün üzerinde bir doğrulukla pozitif veriyi sınıflandırması açısından, hem negatif hem de

pozitif veriyi sınıflandırmasında daha iyi bir performans gösterdiği görülmektedir. Ayrıca, Tablo 4.14c'de, negatif verileri sınıflandırırken %79.2'lik bir performans elde edildiği ve nötr verileri sınıflandırırken ise sadece %50.8'lik bir performans elde edildiği görülmektedir.

Tablo 4.14: İkinci veri kümesindeki (a) PL (b) SVM (c) RF algoritması kullanılarak gövdelenmiş verilerden elde edilen sonuçlar

(a)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	1830	595	720
	Negatif	525	1789	660
	Nötr	1145	1116	2120

(b)

		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	690	81	129
	Negatif	33	610	144
	Nötr	277	309	727

(c)

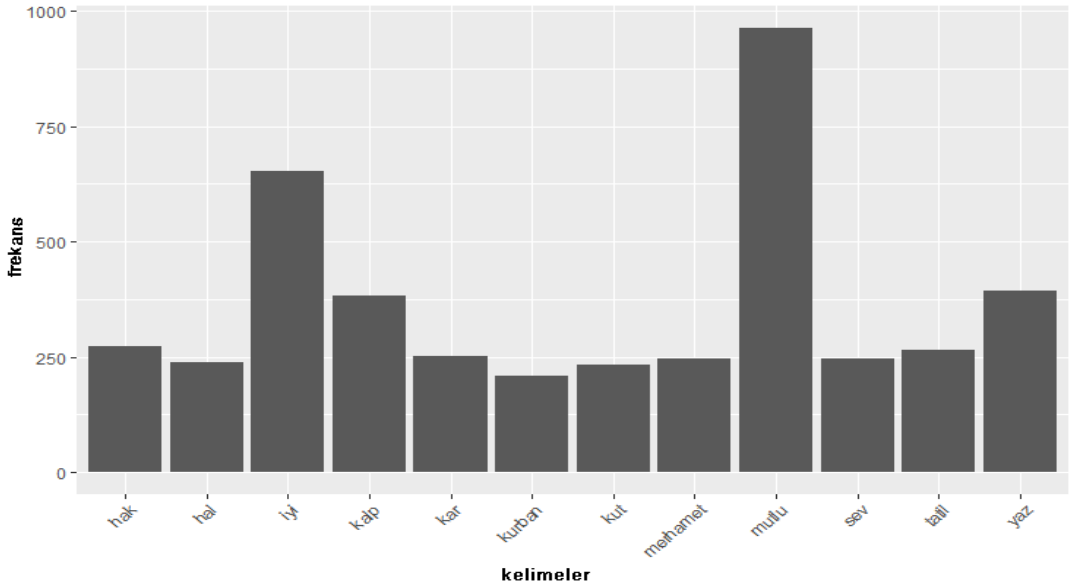
		Doğru Sınıf		
		Pozitif	Negatif	Nötr
Öngörülen Sınıf	Pozitif	836	112	204
	Negatif	20	792	288
	Nötr	144	96	508

Tablo 4.15, ikinci veri kümesinin gövdelenmiş verileri üzerinde elde edilen tüm sonuçların performansını göstermektedir. Genel durumda, RF %71,2'lik bir hassasiyetla daha yüksek bir performansa sahiptir, ardından ise %67.6 ile SVM ve daha sonra %54,7 ile PL gelmektedir. Genel durumdan da anlaşıldığı üzere RF'nin en iyi performansa sahip yöntem olduğu açıkça görülebilir.

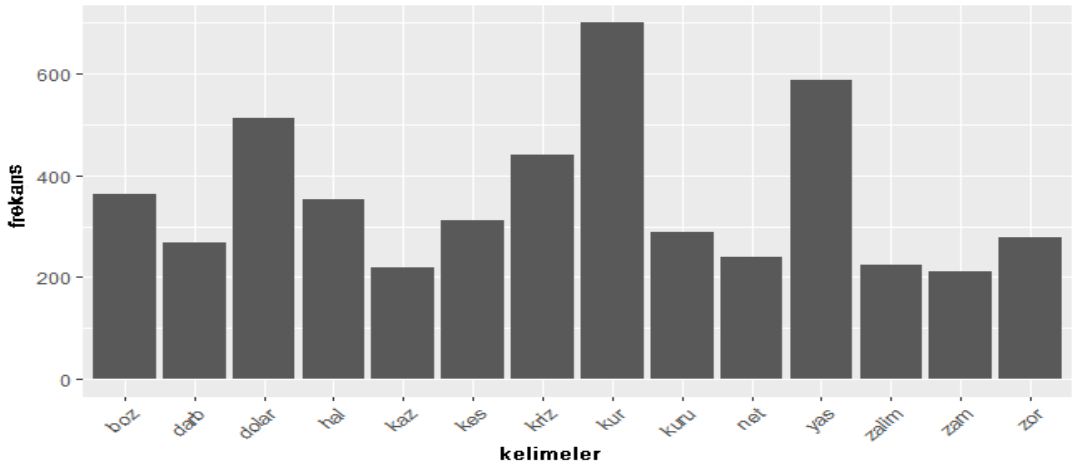
Tablo 4.15: İkinci veri kümesindeki gövdelenmiş veriler kullanılarak elde edilen sonucun performansı

	Yanlış sınıflandırma oranı	Kesinlik	Hassasiyet	F1 skoru
PL	0.453	0.556	0.547	0.551
SVM	0.324	0.699	0.676	0.687
RF	0.288	0.708	0.712	0.71

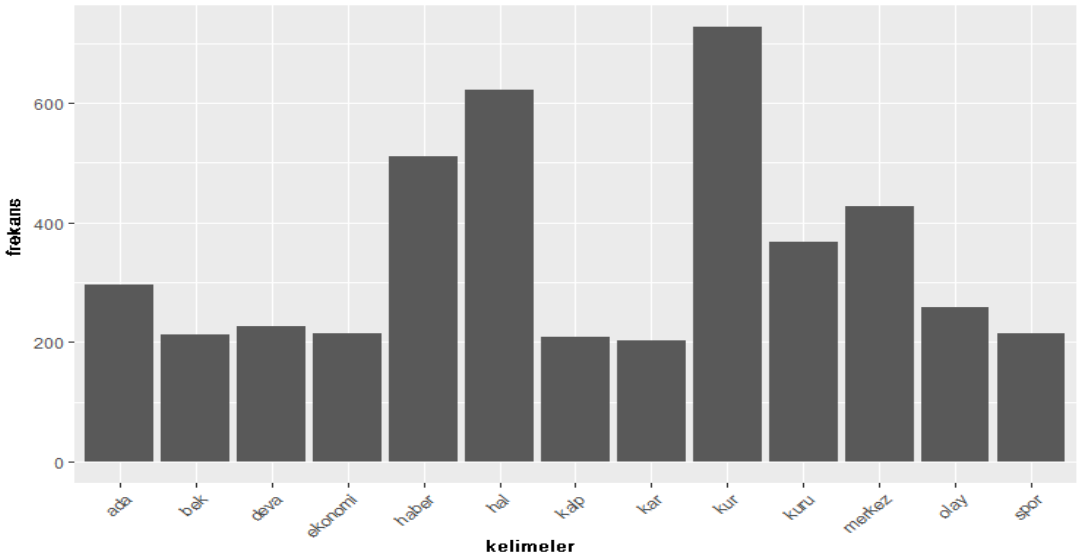
Şekil 4.8, 4.9 ve 4.10'daki çubuk grafikler ikinci durumda hem pozitif, negatif ve nötr verilerde en çok meydana gelen kelimeleri (200'den fazla kez görünen kelimeler) temsil eder ve şekil 4.11, 4.12 ve 4.13'teki kelime bulutu, her pozitif, negatif ve nötr tivitin ilk 100 en çok kullanılan kelimesinden oluşur.



Şekil 4.8: İkinci veri kümesindeki pozitif tivitlerde en sık kullanılan sözcükler.



Şekil 4.9: İkinci veri kümesindeki negatif tivitlerde en sık kullanılan sözcükler

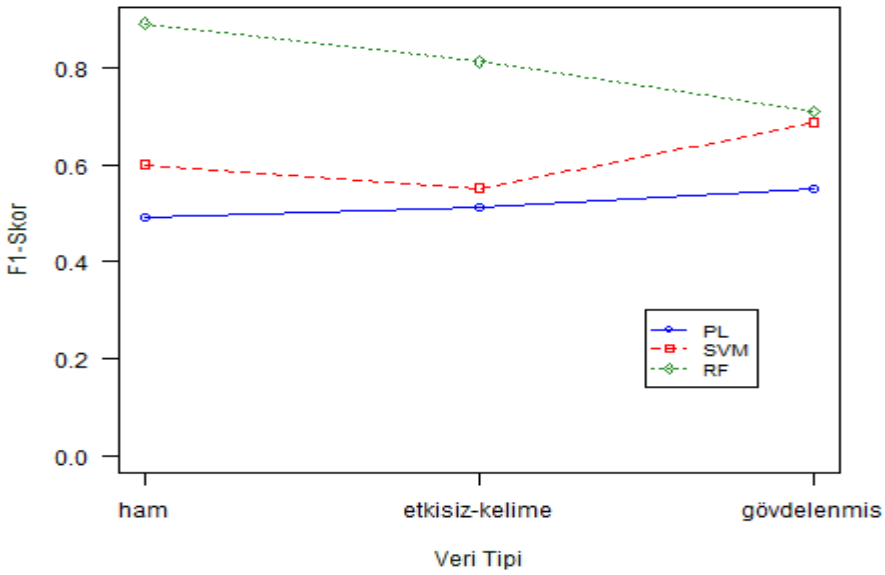


Şekil 4.10: İkinci veri kümesindeki nötr tivitlerde en sık kullanılan sözcükler

Şekil 4.11: İkinci veri kümesindeki pozitif kelimelerin kelime bulutu

Şekil 4.12: İkinci veri kümesindeki negatif kelimelerin kelime bulutu

Şekil 4.13: İkinci veri kümesindeki nötr kelimelerin kelime bulutu



Şekil 4.14: İkinci veri kümesinde kullanılan her bir yöntemde elde edilen performansı gösteren grafik.

Tablo 4.16: İkinci veri kümesindeki sonucu hesaplamak için her yöntemi aldığı süresi

	Ham	Etkisiz-kelime	Gövdelenmiş
PL	6.36 dk	5.33 dk	6.12 dk
SVM	3.14 dk	3.63 dk	2.91 dk
RF	7,200 dk	5,760 dk	4,320 dk

Şekil 4.14 ve Tablo 4.16'da elde edilen sonuçtan, rasgele orman'ın (RF) performansı, veriler ham verilerden etkisiz-kelime verilerine ve daha sonra da gövdelenmiş verilere dönüştürüldüğünde azalmaktadır. Yine de, verilerin ham olduğu zaman, gereksiz kelimeler çıkarıldıktan sonra ve verilerin kökü bulunduktan sonra, RF'nin diğer iki yöntemden yine de daha iyi performans gösterdiği görülebilir. Destek vektör makinesi (SVM) performansı, verilerden gereksiz kelimeler kaldırıldığında azalır ve daha sonra verilerin kökü bulunduktan sonra tekrar artar. İlk durumda olduğu gibi, SVM diğer tüm yöntemler arasında en hızlı işlem süresine sahiptir. Ayrıca PL kullanan yöntemin performansı, her aşamada gereksiz kelimeler çıkarıldıktan sonra ve dizgecikleme yapıldıktan sonra artar. Verilerin kökü bulunduktan sonra performans daha da artar.

5. SONUÇ VE İLERİYE DÖNÜK ÇALIŞMALAR

Tez çalışmasının kısa bir özeti ve katkısı, yapılan ve ileride yapılması düşünülen çalışmalar aşağıdaki alt bölümlerde verilmiştir.

5.1 Yapılanlar

Bu tezde Twitter API kullanılarak Twitter'dan toplam 13K tivit toplanmıştır ve toplanan tivitler, içeriklerine göre üç farklı sınıfa (pozitif, negatif ve nötr) ayrılmıştır. Toplanan tivitler öncelikle bağlantıları, sayıları, noktalama işaretleri ve anlamlı olmayan karakteri kaldırılarak temizlenmiştir. Bundan sonra tivitler dizgeye dönüştürülmüş, onlardan gereksiz kelimeler çıkarılmış ve ayrıca kökleri bulunmuştur. Ayrıca, verilerin bir kısmı alınmıştır; pozitif, negatif ve nötr her sınıftan 1,000 veri içeren toplam 3,000 veri alınmış ve bu ilk veri kümesi olarak tanımlanmıştır. Her sınıftan 3,500 veri içeren toplam 10,5000 veri ikinci veri kümesi olarak kullanılmıştır. Makine öğrenmesi sınıflandırıcıları ve ayrıca PL tivitlerin duygularını belirlemek için kullanılmıştır. Tezde kullanılan Türkçe sözlüğü dizimi, yıllarca derlenen yaklaşık 6800 pozitif ve negatif kelimeyi içeren karşılaştırmalı İngilizce veri kümesini elle Türkçeye çevirerek geliştirilmiştir (Hu ve Liu, 2004). SVM ve RF, duyguları belirlemek için uygulanan iki makine öğrenim sınıflandırıcısıdır. Sonucu analiz ettikten sonra, tivitlerin her bir sınıfındaki en yaygın kelimeler de tanımlanmıştır ve her bir yöntemin hem birinci hem de ikinci veri kümesinde yürütülmesi için gereken zaman da belirlenmiştir.

5.2 İleriye Dönük Çalışmalar ve Öneriler

Bu araştırmadan elde edilen sonuçlara göre işlenmiş veri yerine ham veri üzerinde daha iyi gerçekleştiği için RF algoritmasının ham veri için tavsiye edildiği söylenebilir. Eğer veri üzerinde PL uygulandıysa kelimelerin kökeninin bulunması fayda sağlamaktadır. Çoğu durumda RF'nin daha iyi performans göstermesine ve üç durumun hepsinde PL'den daha iyi olmasına rağmen sonuçları hesaplamak RF için

çok zaman almaktadır ve bu yüzden zaman maliyetinin önemli olduğu durumlarda sırasıyla SVM ya da PL kullanılabilir.

Bu tez çalışmasında kullanılan Zemberek yazılımı, Türkçe metinlerdeki kısaltılmış kelimeleri tam olarak yazmak, doğru yazılmayan kelimeleri otomatik olarak düzeltmek için kullanılan fakat henüz performans olarak eksikleri bulunan önemli bir girişimdir. İlerideki araştırmalar için, bütün kısaltmaları düzeltmek amacı ile kısaltma sözlükleri kullanılabilecek, belirli bir tivitle bağdaşlaştırılan resimlerin bağlantısını almak yerine üzerinde yazılan kelimeleri metne dönüştürülebilecek ve daha fazla kelimeyi var olan kelimeler listesine ekleyebilecek çalışmalar yapılabilir. Böylece daha fazla ilişkili kelimenin eklenmesi ile daha başarılı sonuçların elde edilmesi sağlanabilecektir.

RF algoritmasının, her üç durumda da pozitif verileri sınıflandırırken kullanılan diğer yöntemlerden daha iyi performans gösterdiği görülmüştür ancak veriler gövdelendikten sonra SVM, negatif ve nötr verilerde gözlemlenen 4 vakanın 3'ünde RF'den daha iyi sonuç vermektedir. Bu nedenle, pozitif gövdelenmiş verileri sınıflandırmak için RF kullanıp, diğerleri için daha sonra ayrıca bir SVM yapısı kullanarak hibrid ve hiyerarşik bir algoritmanın geliştirilmesi ile duygu analizi performansının iyileştirilmesi ileriye dönük bir çalışma olarak düşünülmektedir.

5.3 Sonuç

Bu tezde, üç farklı duygu halinin; örneğin mutlu, üzgün ve nötr gibi halleri iki farklı yöntem (kutupsallık sözlüğü ve sınıflandırma) kullanarak analizleri yapılmış ve her iki yöntem iki farklı veri seti ile test edilmiştir. Sınıflandırma aşamasında, SVM ve RF algoritmaları kullanılmıştır. İkinci veri seti toplamda 10.500 tivit içerirken birinci veri seti 3.000 tivitten oluşmaktadır. Elde edilen sonuçlara göre, dizgecikleme kullanımı ve etkisiz-kelimelerin çıkarılması ile PL yönteminin kullanımında doğruluğun arttığı açıkça görülmektedir. Ayrıca, dizgecikleme yapılması ve etkisiz-kelimelerin çıkarılmasından sonra verideki kelimenin kökenini bulmak için Zemberek kullanımı doğruluğu daha da arttırmaktadır. Veri sınıflandırması için makine öğrenmesi algoritmalarının kullanımı ile iyi bir doğruluk sağlandığı görülmüştür. Kullanılan sınıflandırıcılar arasında RF'nin çoğu durumda SVM'den

daha iyi performans gösterdiği gözlenmiştir. Tüm sınıflandırma problemleri için en uygun algoritma olmamasına rağmen, bu durumda RF'nin daha iyi performans gösterdiği söylenebilir.

6. **KAYNAKÇA**

Akın, A. A., & Akın, M. D., "Zemberek, an open source NLP framework for Turkic languages", *Available at: https://github.com/ahmetaa/zemberek-nlp (2007, accessed 1 March 2016)*.

Anastasia, S. and Budi, I. "Twitter sentiment analysis of online transportation service provider", *Proceedings of the International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 359 – 365, (2016).

Anjaria M., Guddeti R. M., "Influence factor based opinion mining of Twitter data using supervised learning", *Proceedings of the* 6th *International Conference on Communication Systems and Networks (COMSNETS), 1-8*, (2014).

Ay Karakuş B., Talo M., Hallaç İ. R., Aydin G., "Evaluating deep learning models for sentiment classification", *Concurrency and Computation: Practice and Experience*, (2018).

Aytekin C. "An opinion mining task in Turkish language: A model for assigning opinions in Turkish blogs to the polarities*", Journalism and Mass Communication, 3(3), 179–198*, (2013).

Baccianella S., Esuli, A., & Sebastiani F., "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining", *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC), Valletta, Malta, 10:2200–2204*, (2010).

Bassel G. W., Glaab E., Marquez J., Holdsworth M. J., Bacardit J., "Functional network construction in Arabidopsis using rule-based machine learning on large-scale data sets". *The Plant Cell, 23(9):3101–3116*, (2011).

Behdenna S., Barigou F., Belalem G., "Document level sentiment analysis: a survey", *Enterprise Application Integration Endorsed Transactions on Context-aware Systems and Application, 1 – 8*, (2018).

Bilgin Ö. Çetinoglu Ö., & Oflazer, K., "Building a WordNet for Turkish", *Romanian Journal of Information Science and Technology, 7(1–2):163–172*, (2004).

Cambria E., Olsher D., & Rajagopal D.,"SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis", *Proceedings of the 28th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence, 1515–1521,* (2014).

Cambria E., Schuller B., Liu B., Wang H., Havasi C.,"Knowledge-based approaches to concept-level sentiment analysis", *Institute of Electrical and Electronics Engineers Intelligent Systems 28(2):12–14,* (2013)

Cambria, E., Schuller, B., Liu, B., Wang, H., Havasi, C., "Statistical approaches to concept-level sentiment analysis", *Institute of Electrical and Electronics Engineers Intelligent Systems 28(3):6–9*, (2013).

Can F., Kocberber S., Balcik E., Kaynak C., Ocalan H.C., Vursavas O. M., "Information retrieval on Turkish texts", *Journal of the American Society for Information Science and Technology.* 59(3):*407-421*, *(2008).*

Can U., Alatas B., "Duygu analizi ve fikir madenciliği algoritmalarının incelenmesi" *International Journal of Pure and Applied Sciences, 75 – 111*, (2017).

Cao Q., Duan W., Gan Q., "Exploring determinants of voting for the ''helpfulness'' of online user reviews: a text mining approach", *Decision Support System, 50:511–21,* (2011).

Cesarano C., Dorr B., Picariello A., Reforgiato D., Sagoff A., "OASYS: An opinion analysis system", *Proceedings of Association for the Advancement of Artificial Intelligence Spring Symposium on Computational Approach to Analyzing Weblogs*, (2006).

Chaovalit P. and Zhou L., "Movie review mining: A comparison between supervised and unsupervised classification approaches", *Proceedings of the 38th Institute of Electrical and Electronics Engineers Annual Hawaii International Conference on System Sciences, 112c-112c,* (2005).

Chief, Astathropoulos, Greenonion, Cmantas, "Introduction to Turkish language morphology", (22th January), *https://github.com/skroutz/ turkish_stemmer*, (2014).

Coban O., Ozyildirim B. M., Ozel S. A., "An empirical study of the extreme learning machine for Twitter sentiment analysis", *International Journal of Intelligent Systems and Applications in Engineering, 6(3)*, (2018).

Çoban Ö., Özyer B., Özyer G. T., "Sentiment analysis for Turkish Twitter feeds*", 23th Signal Processing and Communications Applications Conference (SIU),* (2015).

Cortes C, Vapnik V., "Support-vector networks", *Machine Learning*, *20: 273-297,* (1995).

Çoşkun N., "Türkçe Tümcelerin Öğelerinin Bulunması" Yüksek Lisans Tezi, *İstanbul Teknik Üniversitesi Fen Bilimler Enstitüsü,* Bilgisayar Mühendisliği Anabilim Dali, İstanbul (2013).

Cruz F. L., Troyano J. A., Enriquez F., Ortega J. F., Vallejo C. G., "Long autonomy or long delay?' The importance of domain in opinion mining", *Expert Systems with Applications; 40:3174–3184,* (2013).

Cummins N., Amiriparian S., Ottl S., Gerczuk M., Schmitt M., Schuller B., "Multimodal Bag of Words for Cross Domains Sentiment Analysis", *Institute of Electrical and Electronics Engineers International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* (2018).

Da Silva, N.F., Hruschka, E.R. and Hruschka JR. E.R., "Tweet sentiment analysis with classifier ensembles", *Decision Support Systems, 66: 170-179*, (2014).

Deerwester S, Dumais S, Landauer T, Furnas G, Harshman R., "Indexing by latent semantic analysis", *Japan Analytical & Scientific Instruments Show;41:391–407*, (1990).

Dehkharghani R., Yanikoglu B., Saygin Y., Oflazer K., "SentiTurkNet: a Turkish polarity lexicon for sentiment analysis", *Language Resources and Evaluation, 50 (3). 667-685*, (2015).

Eroğul U., "Sentiment analysis in Turkish", *(Master's thesis, Middle East Technical University, Ankara, Turkey), retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp*, (2009).

Etter M., Colleoni E., Illia L., Meggiorin K., & D'Eugenio A., "Measuring organizational legitimacy in social media: assessing citizens' Judgments With Sentiment Analysis", *Business & Society, 57(1), 60–97*, (2016).

Fahrni A., Klenner M., "Old wine or warm beer: target-specific sentiment analysis of adjectives", *In: Proceedings of the Artificial Intelligence and Simulation of Behaviour symposium on affective language in human and machine, 60–3,* (2008).

Fu X., Liu G., Guo Y., Wang. Z., "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon". *Knowledge-Based Systems 37:186–95*, (2013).

Ghag K., and Shah K., "Comparative analysis of the technique of Sentiment Analysis", *International Conference on Advances in Technology and Engineering (ICATE), 1-7*, (2013).

Go A., Bhayani R. and Huang L., "Twitter sentiment classification using distant supervision", *CS224N Project Report, Stanford, 1:12*, (2009).

Güven Z. A., Diri B., Çakaloğlu T., "Classification of Turkish Tivit by n-Stage Latent Dirichlet Allocation",*Electric Electronics, Computer Science, Biomedical Engineerings' Meeting* , (2018).

Han J., Kamber M., Pei J., "Data Mining Concepts and Techniques", 3rd Edition, *the Morgan Kaufmann Series in Data Management Systems (Selected Titles),* (2012).

Hatzivassiloglou V, McKeown K., "Predicting the semantic orientation of adjectives*", Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL'97)*, (1997).

Hayran A. and Sert M., "Sentiment analysis on microblog data based on word embedding and fusion techniques," *25th Signal Processing and Communication Applications Conference, Antalya,* 1-4, (2017).

Hu and Liu, **https://www.cs.uic.edu/liub/fbs/sentiment-analysis.html #datasets, KDD*, (2004).

Hu N., Bose I., Koh N. S., Liu L. ''Manipulation of online reviews: an analysis of ratings, readability, and sentiments'', *Decision Support Systems, 52:674–784*, (2012).

Islam T., Bappy A. R., Rahman T., Uddin M. S., "Filtering political sentiment in social media from textual information", *Proceedings 5th International Conference on Informatics, Electronics and Vision (ICIEV), 663 – 666,* (2016).

Jackson, D., "What is social listening & why is it important?", (20th September), *https://sproutsocial.com/insights/social-listening/* , (2017).

Jain A. P., Katkar Mr. V. D., “Sentiment analysis of Twitter data using data mining”, *International Conference on Information Processing (ICIP) Vishwakarma Institute of Technology, 807 - 810* (2015).

Jürgens P., Jungheer A. and Shoen H., “Small worlds with a difference: New Gatekeepers and the filtering of political information on Twitter” *Association for Computing Machinery Web Science, Koblenz Germany*, (2011).

Karabulut Y. E., Küçüksille E. U., “Twitter profesyonel izleme ve analiz aracı”, *Journal of Technical Sciences, 8:17-24,*(2018).

Kaya M., Fidan G. and Toroslu I. H., "Sentiment analysis of Turkish political news," *International Joint Conferences on Web Intelligence and Intelligent Agent Technology, vol.01, Institute of Electrical and Electronics Engineers Computer Society*, *174 – 180,* (2012).

Kaya M., Fidan G. and Toroslu I. H., "Transfer Learning Using Twitter Data for Improving Sentiment Classification of Turkish Political News", *Springer International Publishing Switzerland, 139 – 148*, (2013).

Khan F. H., Qamar U. and Javed M.Y., “Sentiview: A visual sentiment analysis framework”, *In Information Society (i-Society), International Conference on Institute of Electrical and Electronics Engineers, 291 – 296,* (2014).

Kim S., Hovy E., “Determining the sentiment of opinions”, *In: Proceedings of international conference on Computational Linguistics (COLING’04);* (2004).

Kiprono K. W., Abade E. O., “Comparative Twitter sentiment analysis based on linear and probabilistic models”, *International Journal on Data Science and Technology, 2:41-45*, (2016).

Kristin N. S., “Social media and political campaign thesis”, *University of Tennessee, Honors Thesis Projects,* Knoxville, USA (2011).

Kubat M., “An Introduction to Machine learning”, *Springer International Publishing Switzerland, pg. 85*, (2015).

Kumara T., “Supervised learning based approach to aspect based sentiment analysis”, (4th October), *https://www.slideshare.net/tharinduzonic/supervised-learning-based-approach-to-aspect-based-sentiment-analysis*, (2015).

Li Y., Li T.,"Deriving market intelligence from microblogs". *Decision Support Systems*, *55(1):206 – 217,* (2013).

Lin C. & He Y., "Joint sentiment/topic model for sentiment analysis", *Proceeding of the 18th Association for Computing Machinery : Conference on Information and Knowledge Management (CIKM), 375 – 384,* (2009)

Liu B., "Sentiment Analysis: Mining Opinions, Sentiments and Emotions", *Cambridge University Press, pg. 3*, (2015).

Liu Y., Yu X., Liu B., Chen Z., "Sentence-level sentiment analysis in the presence of modalities", *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing, 8404:1-16,* (2014).

Maks I., Vossen P., "A lexicon model for deep sentiment analysis and opinion mining applications", *Decision Support System, 53:680–688*, (2012).

Maynard D., Funk A., "Automatic detection of political opinions in tweets", *In: Proceedings of the 8th international conference on the semantic web, 88–99.*, (2011).

Medhat W., Hassan A., and Korashy H., "Sentiment analysis algorithm and applications: A survey*", Ain Shams Engineering Journal, 5(4):1093-1113,* (2014).

Miller G. A., "WordNet: A lexical database for English", *Communications of the Association of the Computing Machinery, 38(11):39–41*. (1995).

Miller G. A., Beckwith R., Fellbaum C., Gross D., Miller K., "WordNet: an on-line lexical database", *Oxford University Press, 3(4):235–244,* (1990).

Mohammad S., Dunne C., Dorr B., "Generating high-coverage semantic orientation lexicons from overly marked words and a thesaurus", *In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'09),* (2009).

Mulki H., Haddad H., Ali C. B., Babaoğlu İ., "Preprocessing impact on Turkish sentiment analysis", *26th Signal Processing and Communications Applications Conference (SIU),* (2018),

Oğul B. B., Ercan G., "Sentiment classification on Turkish hotel reviews", *24th Signal Processing and Communication Application Conference (SIU)*, (2016).

Pak, A., and Paroubek, P. "Twitter as a corpus for sentiment analysis and opinion mining", *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC' 10),* 1320-1326, (2010).

Pang B. and Lee L., "Opinion Mining and Sentiment Analysis", *Foundations and Trends in Information Retrieval, 2(1–2):1-135*, (2008).

Pang B., Lee L. and Vaithyanathan S., "Thumbs up? sentiment classification using machine learning techniques", *In Proceedings of the Association for Computational Linguistic 2th Conference on Empirical Methods in Natural Language Processing,10:79 – 86,* (2002).

Patel S., "SVM (Support Vector Machine) - Theory", (3th May), *https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72,* (2017).

Qiu G., He X., Zhang F., Shi Y., Bu J., Chen C., "DASA: dissatisfaction-oriented advertising based on sentiment analysis", *Expert Systems with Applications, 37(9):6182–6191*, (2010).

Quinlan J.R., "Induction of decision trees", *Machine Learning, 1:81–106*, (1986).

Sağlam, F., Sever, H. and Genç, B. "Developing Turkish sentiment lexicon for sentiment analysis using online news media", *13th International Conference of Computer Systems and Applications (AICCSA) Institute of Electrical and Electronics Engineers / American Chemical Society* , (2016).

Shahana P.H. and Omman B., "Evaluation of features on sentimental analysis", *Procedia Computer Science, 46:1585-1592*, (2015).

Sleator D., Temperley D., "Parsing English with a link grammar", *Carnegie Mellon University Computer Science technical report, CMU-CS-91-196*, (1991).

Subrahmanian V.S., Reforgiato D., "AVA: Adjective-verb-adverb combinations for sentiment analysis", *Institute of Electrical and Electronics Engineers Intelligent Systems, 23:43-50*, (2008).

Symeonidis S., "How to classify sentiment?", *https://www.kdnuggets.com/2018/03/5-things-sentiment-analysis-classification.html,* (2018).

Thelwall M., Buckley K., & Paltoglou G., "Sentiment strength detection for the social web", *Journal of the American Society for Information Science and Technology, 63(1):163–173*, (2012).

Tocoglu M. A., Alpkocak A., "TREMO: A dataset for emotion analysis in Turkish", *Journal of Information Science, 1–13*, (2018).

Tripathy A., Agrawal A. and Rath, S.K., "Classification of sentimental reviews using machine learning techniques", *Procedia Computer Science, 57:821-829*, (2015).

Türkmenoğlu C. and Tantuğ C. A., "Sentiment analysis in Turkish media", *International Conference on Machine Learning,* (2014).

Turney P. D., "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 417-424*, (2002).

Tyagi A., Chandra N., "A proposed approach with analysis of speech signals for sentiment detection", *Proceedings of the 5th International Conference on Communication Systems and Network Technologies, 339 – 344,* (2015).

Ucan A., "Automatic sentiment dictionary translation and using in sentiment analysis," *(Master's thesis, Hacettepe University, Ankara, Turkey), Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp*, (2014).

Vaghela V. B., Jadav B. M., "Analysis of various sentiment classification techniques", *International Journal of Computer Applications, 140(3):22 – 27*, (2016).

Vapnik V., "The nature of statistical learning theory", *Springer-Verlag Berlin, Heidelberg,* (2013).

Vural A. G., B.B Cambazoglu, B. B., Senkul P. and Tokgoz Z. O., "A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish," *In Computer and Information Sciences III, Springer London, 437-445,* (2013).

Weiss M. S., Indurkhya N., "Rule-based machine learning methods for functional prediction", *Journal of Artificial Intelligence Research*, *3:383 – 403,* (1995).

Wenhao Z., Xu H., Wei W., “Weakness finder: find product weakness from Chinese reviews by using aspects based sentiment analysis”, *Expert System Application, 39:10283–91*, (2012).

Yuan G.; Ho C.; Lin C. “Recent advances of large-scale linear classification”, *Processings of the Institute of Electrical and Electronics Engineers. 100(9):2584-2603,* (2012).

7. **EK**

Agulutif (bitişken): Bir aglutinatif dil, kelimelerin anlamlarını belirlemek için farklı türlerde morfemlerden oluştuğu bir dil türüdür.

GDELT: 100'den fazla dilde dünyanın her yerinden gelen yayın, basılı ve web haberlerini izleyen küresel bir veritabanı.

Duygu: Görüş veya fikir.

SentiTurkNet: İlk Türkçe kutupsallık sözlüğü.

SWNetTR: Ucan (2014) tarafından geliştirilen 27 bin kelime içeren bir Türkçe SentiWordNet.

SWNetTr-GDELT: GDELT'den indirilen Türkçe haber metninin kökü atama polaritesiyle.

SWNetTR-PLUS: SWNetTr uzantısı SWNetTr-GDELT içinde bulunan ancak SWNetTr'de bulunmayan 10.000 benzersiz sözcük ekleyerek. oluşturmuştır.

WordNet: İngilizce dilsel veritabanı.

Zemberek: Turkic dilleri için açık kütüphaneleri kaynaklı bir NLP çerçevesi.

8. **ÖZGEÇMİŞ**

Adı Soyadı	:	Harisu Abdullahi SHEHU
Doğum Yeri ve Tarihi	:	Potiskum, 15 Nisan 1995
Lisans Üniversite	:	Gediz Üniversitesi
E posta adresi	:	harisushehu@gmail.com
İletişim Adresi	:	Pamukkale Üniversitesi

Yayınlar Listesi :

• Sharif Md. H.,Uyaver S.,Sharif Md. H.U., Ince F.I., Shehu H. A., Galip F., "Particle Filter to Track Movers From Laser Scanned Datasets", *International Conference on Emerging Technologies in Data Mining And Information Security*, India, (2018).

Poster Sunumu :

• Shehu H. A.,Bomai A., "Tracking Movers From Laser Scanned Dataset", *Genç Beyinler Yeni Fikirler Proje Pazarı ve Bitirme Projeleri Ortak, 2016* (Doküz Eylul Üniversitesi İzmir).