T.C. DOĞUŞ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

HİSSE SENETLERİ VE SOSYAL MEDYA ARASINDAKİ İLİŞKİNİN MAKİNE ÖĞRENMESİ TEKNİKLERİ İLE BELİRLENMESİ

YÜKSEK LİSANS TEZİ

EMİNE ATEŞ

201695001

TEZ DANIŞMANI

DR. ÖĞR. ÜYESİ AYSUN GÜRAN

İSTANBUL, 2019

	YÜKSEK LİSANS TEZ SINAV TUTANAĞI	Doküman No	FR.1.26
		Yürürlük Tarihi	1.11.2017
		Revizyon Tarihi	1.11.2017
		Revizyon No	1
		Sayfa	1 / 1

SOSYAL BİLİMLER / FEN BİLİMLERİ ENSTİTÜSÜ

Tarih : 13/09/2019

Anabilim/Anasanat Dalı : Bilgi.sayar...Mühendisliği.....

Öğrencinin Adı Soyadı : Emine...ATEŞ.....

Öğrenci No : 2016.95.001.....

Tez Danışmanının Adı Soyadı : Dr..Öğr..Üyesi..Aysun..GÜRAN.....

İkinci Tez Danışmanının Adı Soyadı : Prof.Dr..Selim..Akyokuş.....

Tezin Başlığı : Hitit Sözlükleri ve Sözel Medya Arşivlerinin Makine Öğrenmesi Yöntemleri ile Belirlenmesi

Doğuş Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin 32.Maddesi uyarınca yapılan değerlendirmeler sonunda;

☑ **tezin kabul edilmesine** ☐ **tezde düzeltme verilmesine** ☐ **tezin reddedilmesine**

(oy birliği) / oy çokluğu ile karar verilmiştir.Gereği için arz olunur.

Danışman Üye

Dr.Öğr.Üyesi Aysun GÜRAN

Üye

Dr. Öğr. Üyesi
Nilgün Güler Bayazıt

Üye

Dr Öğr Üyesi
M. Zahid Gürbüz

Üye

Üye

Anabilim/Anasanat Dalı Başkanı Onayı: Dr. Öğr. Üyesi Yasemin Koçak Uçarer

(Form No: FR-001, Revizyon Tarihi:30.09.2016, Revizyon No:00)

YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum "Hisse Senetleri ve Sosyal Medya Arasındaki İlişkinin Makine Öğrenmesi Teknikleri İle Belirlenmesi" adlı çalışmanın, tarafımdan, akademik kurallara ve etik değerlere uygun olarak yazıldığını ve yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

Emine ATEŞ

11.09.2015

ÖNSÖZ

Tez çalışmam boyunca her anlamda emek ve katkılarından dolayı danışman hocam Sayın Dr. Aysun Güran'a, fikir ve yönlendirmeleri için sayın Prof. Dr. Selim Akyokuş'a, arkadaşlarıma ve tabii ki bu süreçte beni destekleyen aileme teşekkür ederim.

İstanbul, 2019 **Emine ATEŞ**

ÖZET

Bu tez çalışmasının amacı Bist30 endeksinde bulunan hisseler hakkında Twitter mikroblog sitesi üzerinden yapılan yorumlar ile günlük, haftalık ve aylık periyotlarda Bist30 değer değişimleri arasındaki korelasyon ve nedensellik ilişkisinin araştırmak ve elde edilen sonuçları yorumlamaktır. Bu amaçla tez çalışmasında veri kümesi olarak 07.05.2018-30.04.2019 tarihleri arasında Twitter mikroblog sitesinde herkese açık profile sahip kullanıcıların Bist30 hisseleri hakkında paylaştığı 57.933 etiketli kısa ileti, 20.276 etiketsiz kısa ileti ve Bist30 endeksi hisselerinin açılış kapanış fiyat farkları kullanılmıştır. Eğitim amaçlı kullanılan 57.933 adet twitter yorumu bu 12 aylık periyotta finans ile ilgili kişilerce manuel olarak iletilerin içerdiği duygu durumlarına göre pozitif, negatif ve nötr olmak üzere üç kategoride etiketlenmiştir. Etiketlenen bu veriler makine öğrenmesi algoritmaları ile eğitilerek sınıflandırılmıştır ve algoritmalar arasından başarı oranı en yüksek olanı tespit edilerek, bu algoritma eğitim esnasında kullanılmayan yeni bir test veri setinin sınıflandırılması için kullanılmıştır. Test veri setinin sınıflandırılması ile pozitif, negatif sınıflardaki kısa ileti sayıları elde edilmiş ve bu sayılar ile literatürde kullanılan 4 farklı duygu skoru oluşturulmuştur. Bölüm 5'te belirtilen bu duygu skorları S_1, S_2, S_3 ve S_4 olarak isimlendirilmiştir. Tezin kapsamında bu 4 çeşit duygu skoru ile günlük, haftalık ve aylık bazda Bist30 hisseleri değer değişimleri arasındaki Pearson Korelasyon analizi gerçekleştirilmiştir. Pearson korelasyonu değişkenler arasındaki ilişkinin nedenselliğini analiz etmemektedir. Değişkenler arasındaki nedensellik ilişkisi için Granger nedensellik analizi uygulanmıştır ve nihayetinde elde edilen sonuçlar yorumlanmıştır.

Kısa iletilerin bilgisayar tarafından algılanabilmesi adına sayısallaştırılması ve vektörler ile ifade edilebilmesi için kelime tabanlı N-gramlara dayalı geleneksel kelime çantası modeli dışında, Yapay Sinir Ağlarına (YSA) dayalı Doc2vec mimarisi de kullanılmıştır. Eğitim veri kümesi sınıflandırılırken Lojistik Regresyon (LR), Destek Vektör Makineleri (DVM), Naive Bayes (NB), Karar Ağaçları (KA), K-EnYakın Komşu (KYK) sınıflandırıcıları ve Rastgele Orman (RO), Gradyan Artırma (GA) ve Maksimum Oylama (MO) topluluk öğrenmesi algoritmaları kullanılmıştır. Deneyler sonunda en iyi sınıflandırma algoritmasının LR olduğu ortaya çıkmıştır.

LR ile etiketsiz veri setindeki kısa iletilerin sınıfları tahmin edilmiştir.Yeni oluşan veri seti üzerinden pozitif ve negatif duygu içeren tweetlerin polarite değerleri hesaplanmış ve bu duygu skorları ile Bist30 hisseleri arasındaki, ilişkilerin istatistiksel analizlerinin yapılması sağlanmıştır.Bist30 endeksinin açılış ve kapanış fiyatları ile duygu skorları ele alındığında günlük ve haftalık periyotta orta kuvvette ilişki bulunurken, aylık dönemde 0,74 gibi kuvvetli bir ilişkiye sahip olduğu saptanmıştır.Hisse bazında günlük ve haftalık zaman diliminde zayıf ilişkiye sahipken, aylık periyotta örneğin Ereğli Demir ve Çelik Fabrikaları T.A.Ş. (EREGL), Türkiye İş Bankası A.Ş. (ISCTR), Tofaş Türk Otomobil Fabrikası A.Ş. (TOASO) gibi hisselerin kuvvetli ilişkiye sahip olmaları dikkat çekmiştir. Ay bazında incelendiğinde 2018 yılının Ağustos ve 2019 yılının Şubat ayında kuvvetli ilişki bulunduğu gözlemlenmiştir. Granger nedensellik analizi yapıldığında günlük ve haftalık periyotta duygu skorları ve Bist30 endeksinin birbirlerinin geçmiş değerlerinden etkilendiği; 9 hissede tek veya çift yönlü, 11 tane hissede ise günlük atılan tweet sayısı ile hisse fiyatı arasında nedensellik ilişkisi tespit edilmiştir. Tüm sonuçlar ayrıntılarıyla Bölüm 6'da açıklanmıştır.

Anahtar Kelimeler: Makine Öğrenmesi, Bist30, Twitter Duygu Skorları, Pearson Korelasyonu, Granger Nedensellik Analizi

ABSTRACT

The aim of this study is searching the effect of the comments on the Twitter microblog site about opened-closed of the Bist30 and stock prices rates in daily, weekly and monthly periods.For this purpose, the data set of 57,933 supervised short messages, 20,276 unsupervised short messages and the opening and closing price differences of Bist30 index shared by users have public profile on Twitter microblog site between 05.07.2018 and 04.30.2019 were used as data set in the thesis study.57,933 twitter comments were classified into three categories as positive, negative and neutral by examining the sensitives of the sentences manually during this 12-month period.These classified data were tested by training with machine learning algorithms and among the algorithms, the highest success rate was selected and the classes were estimated on the test data set with random monthly and stock samples.These sensitives scores mentioned in Chapter 5 are called S_1, S_2, S_3 and S_4. Within the scope of the thesis, Pearson Correlation analysis was conducted between these 4 type of sensitives scores and Bist30 rates on daily, weekly and monthly basis. Pearson correlation does not analyze the causality of the relationship between variables.For the causality relationship between the variables, Granger causality analysis was applied and the results obtained were interpreted.

For short messages to be detected and to be expressed with vectors by the computer, in except the traditional word bag model based on word-based N-grams, used Doc2Vec model based on Neural Networks. Logistic Regression, Support Vector Machine, Naive Bayesian, Decision Trees, K-Nearest Neighbor, Random Forest, Gradient Boosting and Maximum Voting Algorithms were used. As a result of the experiments, the best classification algorithm is Logistic Regression.

Logistic Regression was used to estimate the classes of short messages in the unsupervised data set. Sensitives scores of positive and negative tweets were calculated for the new data set and statistical analyzes of the relationships between sensitives scores and Bist30 stock prices were provided.Between the opening and closing prices of the Bist30 index and the sensitives scores, it was found that there was a medium correlation on the daily and weekly periods, while it had a strong correlation such as 0.74 in the monthly period.On the basis of stocks, while it has a weak relationship in

daily and weekly time periods, it is noteworthy that the stocks such as EREGL(Ereğli Iron and Steel Factories), ISCTR(Turkey Business Bank), TOASO(Tofaş Turk Automobile Factory) have strong relations in the monthly period.On a monthly basis, it was observed that the month of 2018 had a strong relationship in August and 2019 of February.When Granger Causality Analysis was performed, it was found that sensitives scores and Bist30 index were effects by each other's past values in daily and weekly periods; One or two-way stock prices and sensitives scores were determined in 9 stocks, and causality was determined between the number of tweets per day and stocks in 11 stocks. All results are described by detail in Chapter 6.

Keywords: Machine Learning, Bist30 Stock Market, Twitter Sensitivity Scores, Pearson Corelation, Granger Causality Analysis

İÇİNDEKİLER

Sayfa No.

TABLO LİSTESİ

Sayfa No.

ŞEKİL LİSTESİ

Sayfa No.

KISALTMALAR

ARA : Artırılmış Regresyon Ağacı (Boosted Regression Tree – BRT)

BIST : Borsa İstanbul

CTS : Centroid Tabanlı Sınıflandırıcı (Centroid-based Classifier)

ÇKA : Çok Katmanlı Algılayıcılar (Multi-Layer Perceptron - MLP)

ÇR : Çoklu Regresyon (Multi Regression – MR)

DBH : Dağıtılmış Bileşik Hafıza (Distributed Memory Concatenated - DMC)

DJIA : Dow Jones Endeksi (Dow Jones Industrial Average)

DKT : Dağıtılmış Kelime Torbası (Distributed Bag of Words - DBOW)

DOB : Dengeli Ortak Bilgi (Balanced Mutual Information)

DOH : Dağıtılmış Ortalama Hafıza (Distributed Memory Mean - DMM)

DVM : Destek Vektör Makinesi (Support Vektör Machine – SVM)

DVRTÖ : Destek Vektör Regresyon Topluluk Öğrenmesi (Support Vektor Regression Ensemble - SVRE)

GA : Gradyan Artırma (Gradient Boosting)

GN : Granger Nedenselliği (Granger Causality - GC)

KA : Karar Ağaçları (Decision Trees - DT)

KAP : Kamu Aydınlatma Platformu

KDBNN : Kendi Kendini Düzenleyen Bulanık Sinir Ağları (Self Organizing Fuzzy Neural Networks - SOFTNN)

KYK : K-En Yakın Komşu Algoritması (K Nearest neighborhood - KNN)

LDA : Lineer Diskriminant Analizi (Linear Discriminant Analysis – LDA)

LNR : Linear Regression (Lineer Regresyon)

LR : Lojistik Regresyon (Logistic Regression)

NB : Naive Bayes Algoritması

OB : Ortak Bilgi (Mutual information – MI)

RO : Rassal Orman (Random Forest - RF)

ROR : Rassal Orman Regresyon (Random Forest Regression – RFR)

TBA : Temel Bileşenler Analizi (Principal Component Analysis - PCA)

TDK : Türk Dil Kurumu

TO : Topluluk Ortalaması (Ensemble Averaging - EA)

VO : Vektör Otoregresyon (Vector Autoregression - VAR)

YSARTÖ : Yapay Sinir Ağları Regresyon Topluluk Öğrenmesi (Neural Networks Regression Ensemble – NNRE)

1. GİRİŞ

Veri madenciliği ortaya çıkarılan bilgilerin bilgisayar programları sayesinde sınıflandırılmasını veya geçmiş verilere dayanarak gelecek verilerin tahmin edilmesini sağlar. Veri madenciliğinin bir alt araştırma alanı olan metin madenciliği, özellikle metinsel verilerin sayısallaştırılarak analize uygun hale getirilmesi bakımından önem taşımaktadır. Dijital medya, şirketlerin müşteri ilişkisi yönetimi uygulamaları (kampanya, kullanıcı yorumlarının değerlendirilmesi), e-ticaret sitelerinde yapılan yorumlar, kullanıcı deneyimlerinin paylaşıldığı platformlar ve bloglar zengin bir veri kaynağı oluşturmaktadır. Bu zengin veri kaynaklarının yorumlanması değerli bilgilerin elde edilmesi hedefiyle metin madenciliği teknikleri ile yapılabilmektedir.

Metin madenciliği kapsamında araştırılan konulardan biri olan duygu analizi metinlerin okuyucuda hissettirdiği duyguyu ortaya çıkararak sınıflandırılması işlemidir ve belirli bir konu hakkında yapılan yorumların olumlu,olumsuz, nötr gibi kategorilere ayrılmasını amaçlamaktadır. Bir ürün hakkında yapılan kullanıcı yorumlarının duygu analizi yöntemleriyle olumlu, olumsuz olarak belirlenmesi üretici açısından değerli bir bilgidir. Artık sosyal medya televizyondan daha çok büyük bir kitleye çok kısa sürede ulaşabilmekte ve gerek kullanıcılar gerekse üreticilerden hızlı geri bildirim alabilmeyi sağlamaktadır. Bu yüzden özellikle duygu analizinde Twitter gibi mikro blog sitelerinin değerli bir veri kaynağı olabileceği ön görülmektedir.

Bu tez çalışmasında Bist30 endeksinde bulunan hisseler hakkında Twitter mikroblog sitesi üzerinden yapılan yorumların günlük, haftalık ve aylık periyotlarda Bist30 değer değişimlerine olan etkisi araştırılmıştır. Bu etki analizi gerçekleştirilirken korelasyon analizi ve Granger nedensellik analizi uygulanmış ve elde edilen sonuçlar yorumlanmıştır.

Tez çalışmasında 2018 Mayıs itibariyle 12 aylık zaman diliminde, Twitter kullanıcılarının herkese açık profillerinden Bist30 hisseleri hakkında paylaşmış olduğu Twitter yorumları manuel olarak etiketlenmiş sonrasında bu veri seti üzerinde makine öğrenmesi teknikleri ile sınıflandırma çalışması yapılmış ve en iyi sınıflandırıcı tespit edildikten sonra hazırlanan test tweetleri sınıflandırılarak olumlu, olumsuz ve nötr tweet sayısı tespit edilmiştir. Bu tespit ile Bist30 hakkındaki kısa iletilerin duygu skorları elde edilmiş ve sonrasında bu duygu skorları ile Bist30 hisse değer değişimleri arasındaki

ilişki ve nedensellik ilişkisi sonuçları yorumlanmıştır. Tez çalışması hisse senetleri değerleri ile sosyal medya arasındaki ilişkiyi inceleyen en kapsamlı Türkçe çalışmalar arasındadır.

1.1. Tezin Katkısı

Sosyal medya platformları borsa gibi dinamik, sürekli değişen ve bir çok faktörden etkilenen alanlarda analiz için iyi bir veri kaynağı oluşturmaktadır. Bu yüzden her geçen gün de popüleritesi artmaktadır. Borsada işlem gören hisse senetlerinin mikro blog sitelerindeki yorumlardan etkilenip etkilenmediği iş dünyası kadar akademik anlamda da merak edilen ve geliştirilmek istenen bir konu olmuştur. Bu çalışmada da Twitter üzerinden yapılan yorumların makine öğrenmesi teknikleri ile tahmin edilerek, Bist30 endeksini oluşturan firmaların hisse senetlerinin günlük, haftalık ve aylık periyotlarda değer değişimlerine istatistiksel olarak araştırılması hedeflenmiştir.

Tez çalışması ile bu alandaki en geniş veri kümesi oluşturulmuştur. Finans ile ilgili kişiler tarafından 57.933 adet tweetin cümle içindeki duygu durumunu ifade eden sınıfı belirlenerek manuel etiketleme yapılmıştır. Oluşan veri kümesi ile pozitif, negatif, nötr sınıflardaki kısa iletiler ile herbir kategori için borsaya özgü sözcükler çıkarılmıştır. Özellikle güncel tarihli geniş bir literatür çalışması yapılarak duygu analizinde en çok kullanılan makine öğrenmesi teknikleri tespit edilerek açıklanmıştır. Günlük, haftalık ve aylık periyotlarda 4 çeşit duygu skoru hesaplaması yapılarak, bu duygu skorları ile Bist30 değer değişimleri arasındaki ilişkilerinin hem korelasyon, hem de nedensellik ilişkisi araştırılmıştır. Bildiğimiz kadarıyla bu çalışma Türkçe veri kümeleri üzerinde yapılan en kapsamlı çalışmalar arasındadır.

2. LİTERATÜR TARAMASI

Bu bölümde duygu analizi ve duygu analizinin finans alanına uygulanması ile ilgili olan literatür çalışmaları sunulmuştur.

2.1. Literatürdeki Duygu Analizi Çalışmaları

(Can & Alataş, 2017) çalışmasında duygu analizi yapılırken temelde veri sözlüğü, makine öğrenmesi veya her iki yöntemin birleşimi olan hibrid yaklaşımı benimsenmiştir. Sözlük yönteminde, kelimelerin olumlu-olumsuz-nötr olarak sınıflandırılması için Türk Dil Kurumu Türkçe Sözlüğü'ndeki sıfatlardan faydalanılarak oluşturulmuştur. Kelime çiftleri ve jargon sözlüğü gibi sözlük türleri genişletilmiştir. Kelimelerin frekanslarına göre N-gram skorlama yöntemiyle cümlenin sınıfı belirlenmiştir.

(Akgül, Ertano & Diri, 2016) çalışmasında belirli bir Twitter kullanıcısının yorumları sınıflandırılmış ve anahtar kelimeler kategori bazında oluşturulmuştur. Test tweet iletilerindeki kelimeler ile anahtar kelimeler karşılaştırılmıştır. Navie Bayes (NB) algoritması kullanılarak nötr kelimelerin ağırlıklı olduğu saptanmıştır.

(Baykara & Göktürk, 2017) çalışmasında duygu sınıflandırması için (NB), Destek Vektör Makinesi (DVM) ve Lojistik Regresyon (Logistic Regression – LR) arasında karşılaştırma yapılmıştır. 1-gram ve 2-gram yöntemlerinin birleştirilmesiyle elde edilen veri kümesi kullanıldığında NB sınıflandırılmasının diğerlerine göre başarılı olduğu görülmüştür.

(Onan, 2017) çalışmasında İnternet Film Veri Tabanı'ndan (Internet Movie Database - IMDb) yararlanılarak elde edilen film yorumlarını sınıflandırmıştır. Çalışmada NB, Centroid Tabanlı Sınıflandırıcı (Centroid-based Classifier – CB), Çok Katmanlı Algılayıcılar (ÇKA) ve DVM yöntemleri kullanılmıştır. Eğitim veri setinde en iyi sonucun ÇKA ile elde edilirken, test verisinde DVM ile aynı oranda başarı göstermiştir.

(Kaynar vd., 2016) çalışmasında Twitter yorumları üzerinde K-En Yakın Komşu Algoritması (KYK) ve DVM karşılaştırması gerçekleştirmiş ve özellik sayısı arttıkça DVM'in daha iyi sonuç verdiği bulunmuştur.

(Hug, Ali & Rahman, 2017) çalışmasında DVM ve KYK ile birlikte kullanıldığı hibrid bir algoritma geliştirilmiştir. Özellik seçimi yapılırken DVM ve KYK'den hangisi başarılı ise o özellik seçilerek başarı oranı arttırılması hedeflenmiştir.

(Gupta, Pruthi & Sahu, 2017) çalışmasında duygu sınıflandırılması için tweetlerdeki sıfatların tweet içindeki ağırlıklarına göre negatif ve pozitif olarak ayrılmasını sağlayan bir formül geliştirilmiştir. Bu yöntemin sözlüğe göre daha başarılı olduğu tespit edilmiştir.

(Rout vd., 2018) çalışmasında duygu analizi gerçekleştirirken unigram gösterimi ile NB algoritmasının en iyi sonuca ulaştırdığını belirtmişlerdir.

2.2. Duygu Analizi ile Hisse Senetleri Arasındaki İlişki

(Bollen, Mao & Zeng , 2011), 6 sınıflı (sakin, uyarı, dikkat, hayati, kibar, mutlu) duygu analizi çalışmasında 10 aylık bir periyotta 9 milyon tweet için Kendi Kendini Düzenleyen Bulanık Sinir Ağları (Self Organizing Fuzzy Neural Networks - SOFTNN) ile yapılan sınıflandırmanın Dow Jones Endeksi (Dow Jones Industrial Average) ile Granger Nedensellik (Granger Causality - GC) analizi 7 günlük gecikme uzunluğunda çalıştırıldığında "sakin" sınıfının etkili olduğu görülmüştür.

(Mittal & Goel, 2012), veri kaynağı olarak 7 aylık bir periyottaki Twitter metinlerini kullanarak 4 sınıflı bir duygu analizi gerçekleştirmiş ve tweet duygu skorları ile DJIA arasındaki ilişkiyi incelemiştir.

(Kim, Jeong, & Ghani, 2014), M ve H medyaları tarafından yayınlanan yaklaşık 70 bin ekonomi ile ilgili makalelerin kelimeleri önişlemeden geçirelerek pozitif ve negatif olmak üzere 2 sınıflı kelime sözlüğü oluşturmuştur. Kelimelerin duygu skorunu ise hisse senetlerinin artış ve azalışlarına göre belirlemiştir. Çalışma ile Kore Borsası ile duygu analizi arasında anlamlı bir ilişki olduğu bulunmuştur.

(Eliaçık & Erdoğan, 2015), finans ile alakalı yorum yapan kullanıcının konuya ilgi düzeyi ve topluluk içerisindeki inandırıcılığını baz alan bir yaklaşım önermiştir. Bu özgün yöntem de tahmin edilen duygu polaritesi ile BIST100 arasında anlamlı bir ilişki olduğu tespit edilmiştir. Ayrıca veri setinden olağan dışı olayların olduğu haftalar çıkarıldığında ilişki kuvvetini belirleyen Pearson katsasıyısının arttığı gözlemlenmiştir.

(Gündüz & Çataltepe, 2015), finansal makale, internet haberleri ve Kamu Aydınlatma Platformu (KAP) bildirimlerinden hazırlanan 18 aylık veri setinde özellik

seçimi yöntemlerinden Ortak Bilgi (Mutual information – MI), Dengeli Ortak Bilgi (Balanced Mutual Information-BMI), SMOTE, Ki-kare (Chi-Square) karşılaştırılması yapılmış ve BMI yönteminin ve NB sınıflandırılması ile eğitilen internet haberlerinin BIST100 tahmininde daha başarılı olduğunu belirtmiştir.

(Nguyen, Shirai & Velcin, 2015), sosyal medya datası kullandıkları çalışmalarında SentiWordNet sözlüğü yardımıyla konu ve duygu veren kelimelerin yakınlıklarını hesaplayarak duygu skoru belirlemişlerdir. Bu yöntem ile hisse değeri tahmini %2.07 oranında arttığını belirtmişlerdir.

(Ranco vd., 2015), anormallik gösteren hisse değerlerinin Twitter duygu skoru arasında anlamlı bir ilişki bulunmuştur. 15 aylık bir periyotta 1.5 milyondan daha fazla tweet üzerinde çalışılmış; Granger Causality Testi sonucunda polarite skorunun hisse bazında çok etkili olmadığı görülmüştür. Ancak Olay Çalışması (Event Study) çalışması yapıldığında; firmaların çeyrek dönem kazançlarının açıklandığı tarihlerde hisse değerlerinin anormal artış/azalış olup olmadığı incelenmiştir. Olay tarihinden +/- 10 gün sonraki hisse değerleri takip edildiğinde olay sonrası 10 günlük süreçte hisse değerlerinin duygu skorundan etkilendiği açıkça görülmüştür.

(Yang, Mo & Liu, 2015), twitterda en çok finansal yoruma sahip 50 kullanıcıyı tespit etmiş ve onların takipçilerinden oluşan 3 çeşit finansal topluluk oluşturmuştur. SentiWordNet sözlüğünü kullanarak bu 3 finansal topluluğun sınıflarını belirlemeye çalışmıştır. Bu toplulukların borsadaki hareketlerle olan ilişkisini lineer regresyon kullanarak incelemiştir.

(Pagolu vd., 2016) çalışmasında 2.5 milyon tweet arasından 3.216 tanesi manuel etiketleme yapılarak pozitif, negatif, nötr olmak üzere 3 sınıfa ayrılmıştır. LibSVM ile eğitilen datanın başarı oranının lojistik regresyona göre daha iyi olduğu bulunmuştur.

(Oliveira, Cortez & Areal, 2017), hisse değeri, dalgalanma, ticaret hacmi kategorilerinde borsaların mikroblog sitelerinden etkilenmesi konusu üzerinde yaptıkları bir çalışmada; günlük hesaplanan 4 çeşit duygu skoru ile hisse değeri, dalgalanma ve ticaret hacmi için ayrı ayrı Autoregressive modelleri (AR(5)) uygulanmıştır. Bu AR modelleri de Çoklu Regresyon (ÇR), RO, DVM, YSA, Topluluk Ortalaması (Ensemble Averaging - EA) sınıflandırma modelleri ile karşılaştırılmıştır. Tahminlerde en başarılı modelin DVM olduğu ortaya çıkmıştır.

(Snir, 2017) çalışmasında 2000'den fazla şirketin çeyreklik finansal raporlarından oluşan bir veri seti üzerinde L&M (Finansal sözlük) sözcükleri kullanılarak NB modeli uygulandığında diğer yöntemlere göre daha başarılı olduğu bulunmuştur.

(Yıldırım & Yüksel, 2017) çalışmasında 500 adet tweet iletisi pozitif ve negatif olarak sınıflandırılmış ve NB algoritmasının diğerlerine göre daha başarılı olduğu görülmüştür. Aynı zamanda Spearman korelasyonu ile duygu skoru arasında anlamlı bir ilişki olduğu vurgulanmıştır.

(Deng vd., 2018) kaynak olarak Twitter ve Reuters finansal haberlerinden yararlandıkları çalışmalarında 4 yıllık zaman diliminde yaklaşık 18 milyon veri kullanmıştır. SentiStrenght methoduyla kısa iletileri pozitif, negatif ve nötr sınıflarına ayırmıştır. DJIA borsasındaki günlük-saatlik hisse fiyatlarındaki değişimler ile duygu skorları arasında Granger Nedensellik analizi yapmışlardır. Günlük nedensellik ilişkisi bulunmazken, saatlik analizde negatif tweetlerin daha etkili olduğunu tespit etmişlerdir.

(Ruan, Durresi & Alfantoukh, 2018), 8 aylık Twitter veri seti üzerinde yaptıkları bir çalışmada kullanıcının topluluk üzerinde etkisini de duygu skoruna dahil edilerek firmaların anormal hisse artış/azalış gösteren değerleri arasında Pearson ilişkisi incelemişlerdir. Uzun vadede (167 gün) sadece duygu polaritesinin, 40'ar tweetten 8 firma için örneklem alındığında ise kullanıcı güvenirliliğinin daha etkili olduğunu saptamışlardır.

(Weng vd., 2018) tarafından Citi Grup hisse senetleri üzerinde yapılan bir çalışmada Yapay Sinir Ağları Regresyon Topluluk Öğrenmesi (Neural Networks Regression Ensemble – NNRE), Destek Vekör Regresyon Topluluk Öğrenmesi (Support Vektor Regression Ensemble - SVRE), Artırılmış Regresyon Ağacı (Boosted Regression Tree – BRT), Rassal Orman Regresyon (Random Forest Regression – RFR) olmak üzere 4 çeşit topluluk öğrenme yöntemleri uygulanmıştır. Temel Bileşenler Analizi (Principal Component Analysis – PCA) özellik seçimi kullanılarak 4 topluluk öğrenme yönteminin hata oranları karşılaştırıldığında BRT ve RFR'nin daha iyi sonuç verdiği tespit edilmiştir.

Literatürde yer alan bu makale ve yayınlar Tablo 2.1'de özetlenmiştir.

Tablo 2.1 Literatür Araştırması

Makale	Kaynak	Periyod	Data Sayısı (Yaklaşık)	Sınıf	Yaklaşım	Method	Sonuç
(Bollen, Mao & Zeng, 2011)	M (Twitter)	10 ay	9.8 milyon	6	MÖ	LNR, KDBNN	Duygu skoru ve DJIA arasında GN ile anlamlı ilişki bulunmuştur.
(Mittal & Goel, 2012)	M (Twitter)	7 ay	476 milyon	4	MÖ	LNR, LR,DVM, KDBNN	Duygu skoru ve DJIA arasında GN ile anlamlı ilişki bulunmuştur.
(Kim, Jeong, & Ghani, 2014)	FM (M ve H medya)	1 yıl	78 bin	2	S	Sözlük	Kore borsası ile duyu analizi arasında anlamlı ilişki bulunmuştur.
(Eliaçik & Erdogan, 2015)	M (Twitter)	6 ay	619 bin	2	MÖ	DVM	Bist100 haftalık değişimde olağan dışı olaylar çıkarıldığında duyu analizi ile anlamlı ilişki bulunmuştur.
(Gunduz & Cataltepe, 2015)	FM, K	24 ay	111 bin	3	MÖ, I	BOW, NB, DOB, Kikare, SMOTE	DOB kullanılarak NB ile BIST100 tahmini %74 başarılı bulunmuştur.
(Nguyen, Shirai & Velcin 2015)	FM (Yahoo)	1 yıl	249 milyon	5, 2	MÖ	DVM, Latent Dirichlet Tahsisi, JST (joint sentiment/topic model)	Aspect-based sentiment yaklaşımı %54 başarılı bulunmuştur.

Tablo 2.1 (Devamı) Literatür Araştırması

Makale	Kaynak	Periyod	Data Sayısı (Yaklaşık)	Sınıf	Yaklaşım	Method	Sonuç
(Ranco vd., 2015)	M (Twitter)	15 ay	1.5 milyon	3	MÖ	DVM	GN yöntemiyle DJIA30 ile duygu skoru arasında Olay Çalışması incelemesinde anlamlı bulunmuştur.
(Yang, Mo & Liu, 2015)	M (Twitter)	5 ay	1.6 milyon	2	S,MÖ	LNR, Sözlük	-
(Akgül, Ertano & Diri, 2016)	M (Twitter)	4 ay	(500-5000) bin	3	S, MÖ	Sözlük (TDK), n-gram	Sözlük tabanlı yaklaşım 3-gram dan %70 ile daha başarılı bulunmuştur.
(Kaynar vd., 2016)	IMDb	-	2 bin	2	MÖ	NB, YSA, DVM	YSA %89.73 ile DVM'ne göre daha başarılı bulunmuştur.
(Pagolu vd., 2016)	M (Twitter)	12 ay	2.5 milyon	3	MÖ	Word2Vec, n-gram, LR, LibSVM	LibSVM %71.82 ile LR'dan daha başarılı bulunmuştur.
(Baykara & Gürtürk, 2017)	M (Twitter)	-	100	3	S, MÖ	NB ve sözlük	Belli bir kullanıcı için duyu skoru hesaplanarak NB sınıflandırması yapılmıştır.
(Can & Alataş 2017)	Birden fazla kaynak ile araştırma yapılmıştır.	-	-	-	-	-	NB ve DVM en çok kullanılan algoritmalardır.

Tablo 2.1 (Devamı) Literatür Araştırması

Makale	Kaynak	Periyod	Data Sayısı (Yaklaşık)	Sınıf	Yaklaşım	Method	Sonuç
(Gupta, Pruthi, & Sahu, 2017)	M (Twitter)	-	900 bin	3	MÖ	DVM, KYK	DVM + KYK hibrid model yaklaşımı uygulanmıştır.
(Huq, Ali & Rahman, 2017)	M (Twitter)	-	1000	2	MÖ	KYK, DVM	Özellik sayısı arttıkça DVM daha iyi sonuç vermiştir.
(Oliveira, Cortez & Areal, 2017)	M (Twitter)	3 yıl	31 milyon	2	MÖ,I	ÇR, RO, DVM, YSA, TO, AR Modeli	DVM diğerlerine göre daha başarılı bulunmuştur.
(Onan, 2017)	M (Twitter)	1 ay	10 bin	2	MÖ	n-gram, NB, DVM, LR	NB (1g+2gr hibrid) %77.78 oranında başarılı bulunmuştur.
(Snir, 2017)	FR (Q4)	-	50 bin	2	S, MÖ	Sözlük, NB, RO, DVM	Finansal sözlük ile NB %55.42 oranında daha başarılı bulunmuştur.
(Yıldırım & Yüksel, 2017)	M (Twitter)	3 ay	9.5 bin	2	MÖ	Zeror, J48, KYK, NB	NB %70 ile daha başarılı ve Spearman korelasyonu ile duygu skoru arasında ilişki bulunmuştur.
(Deng vd., 2018)	M (Twitter), FM (Reuters)	4 yıl	18 milyon ve 3 milyon (FM)	2	ZS, I	VAR, GC	Günlük-saatlik incelemede DIJA ile GC arasında anlamlı ilişki, negatif duygu skorunun daha etkili olduğu bulunmuştur.

Tablo 2.1 (Devamı) Literatür Araştırması

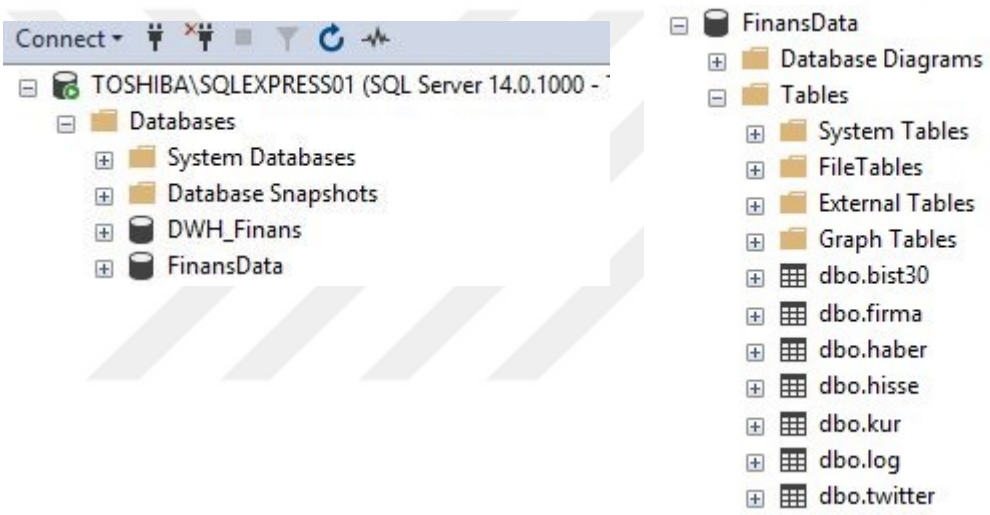
Makale	Kaynak	Periyod	Data Sayısı (Yaklaşık)	Sınıf	Yaklaşım	Method	Sonuç
(Rout vd., 2018)	M (Twitter)	-	60 bin	3	MÖ	n-gram, NB, DVM	SentiWordNet kullanılarak önerilen yaklaşım %80.68 ile daha başarılı bulunmuştur.
(Ruan, Durresi & Alfantoukh 2018)	M (Twitter)	8 ay	8 firma (3 bin) toplam 700 milyon	2	MÖ	YSARTÖ, DVRTÖ, ARA, ROR, HİBRİD, TBA	Topluluk üzerinde güvenirliliği olan kullanıcıların Pearson ilişkisi daha etkili bulunmuştur.
(Weng vd., 2018)	M (Yahoo, Google vs..)	3 yıl	-	2	MÖ	NNRE, SVRE, BRT,RFR	TBA ile özellik seçimi yapıldığında BRT ve RFR'nin daha iyi sonuç verdiği görülmüştür.

Kaynak: M - Microblog siteleri, A - Anket Çalışması, FM - Finansal Makale/Haber, FR - Finansal Rapor, K- KAP

Yaklaşım: S - Sözlük tabanlı yaklaşım, MÖ - Makine öğrenmesi, I – İstatistik, ZS - Zaman Serisi

3. VERİ SETİNİN HAZIRLANMASI

Twitter adlı sosyal medya platformu, standart Tweepy api uygulaması hizmeti ile herkese açık profillerin tweetlerine geriye yönelik en fazla 7 gün olmak üzere erişim imkanı sağlamaktadır. Bu çalışmada kullanılacak veri setinin oluşturulması için öncelikle her gün çalışacak biçimde bir crawler(arama robotu) geliştirilmiştir. 07.05.2018 – 31.04.2019 tarihleri arasında python dilinde hazırlanmış exe uygulamaları ile her iş günü için BIST30 firmaları hakkında gün içinde yazılmış tweetler SQLEXPRESS yardımıyla FinansData adlı veri tabanına tablolara kaydedilmiştir. DWH_Finans adlı ayrı bir veri tabanı oluşturularak burada analiz için kullanılacak özet bilgiler ve data temizliği için gerekli prosedürler hazırlanmıştır.



Şekil 3.1 Veri Tabanı Yapısı ve Tablolar

Şekil 3.1'deki "twitter" tablosunda tutulan bilgiler:

tag: Hisse kodu

tweet_id: Kullanıcıya ait benzersiz tweet numarası

screen_name: Kullanıcı adı

created_dt: Tweetin yazıldığı tarih

created_tm: Tweetin yazıldığı saat

tweet: Kullanıcı yorumu

followers_count: Kullanıcının takipçi sayısı

- “firma” tablosunda BIST30 firmalarının hisse kodlarının listesi,
- “bist30” tablosunda her iş gününe ait BIST30 açılış ve kapanış değerleri,
- “hisse” tablosunda her firmaya ait hisse kodlarının gün sonunda açılış ve kapanış fiyatları ile fark bilgileri,
- “kur” tablosunda merkez bankasından alınan gün sonu dolar, euro ve değişim bilgileri tutulmaktadır.

Kullanıcılar yorum yaparken birden fazla firmanın hisse kodunu hashtag yapabilmektedir. Bu sebeple tweetler üzerinde çalışılırken; tag ve tweet_id bilgilerinin birlikte kullanılarak benzersiz kayıtlar elde edilmiştir.

Veri tabanında biriken datalardan 07.05.2019 ile 01.03.2019 tarihleri arasında rastgele seçilerek toplamda 57.933 tweet üzerinde, içlerinde finansal bilgiye de sahip farklı kişiler tarafından manuel etiketleme yapılarak eğitim verisi elde edilmiştir. Etiketleme yapılırken pozitif, negatif ve nötr olmak üzere 3 sınıf seçilmiştir. Cümle bütünlüğüne göre sezgisel yaklaşımla tweetlerin sınıflarının belirlenmesi sağlanmıştır. Etiketleme işlemi tamamlandıktan sonra rastgele 500 adet tweet seçilerek tekrar etiketleme yapılmıştır. Bu işlemin sonucunda; kullanıcıların %80 oranında ortak görüşe sahip olduğu görülmüş olup etiketleme güvenilirliliği test edilmiştir.

3.1. Bist30 Hisseleri

Tablo 3.1 Bist30 hisseleri ile ilgili yazılan kısa ileti sayılarını göstermektedir.

Tablo 3.1 Bist30 Firmaları ve Tweet Sayıları

Hisse Kodu	Firma Adı	Tweet Sayısı
AKBNK	Akbank T.A.Ş.	3.872
ARCLK	Arçelik A.Ş.	1.616
ASELS	Aselsan Elektronik Sanayi ve Ticaret A.Ş.	6.640
BIMAS	BİM Birleşik Mağazalar A.Ş.	344
DOHOL	Doğan Şirketler Grubu Holding A.Ş.	1.476
EKGYO	Emlak Konut Gayrimenkul Yatırım Ortaklığı A.Ş.	2.043
ENJSA	Enerjisa Başkent Elektrik Perakende Satış A.Ş.	749
EREGL	Ereğli Demir ve Çelik Fabrikaları T.A.Ş.	4.106
FROTO	Ford Otomotiv Sanayi A.Ş.	257
GARAN	Türkiye Garanti Bankası A.Ş.	5.256
ISCTR	Türkiye İş Bankası A.Ş.	2.340
KRDMD	Kardemir Karabük Demir Çelik Sanayi ve Ticaret A.Ş.	4.052

Tablo 3.1 (Devamı) Bist30 Firmaları ve Tweet Sayıları

Hisse Kodu	Firma Adı	Tweet Sayısı
KCHOL	Koç Holding A.Ş.	1.667
KOZAL	Koza Altın İşletmeleri A.Ş.	3.761
KOZAA	Koza Anadolu Metal Madencilik İşletmeleri A.Ş.	5.022
PGSUS	Pegasus Hava Taşımacılığı A.Ş.	1.147
PETKM	Petkim Petrokimya Holding A.Ş.	4.742
SAHOL	Hacı Ömer Sabancı Holding A.Ş.	1.310
SODA	Soda Sanayii A.Ş.	1.283
SISE	Türkiye Şişe ve Cam Fabrikaları A.Ş.	870
HALKB	Türkiye Halk Bankası A.Ş.	4.529
TAVHL	TAV Havalimanları Holding A.Ş.	1.231
TKFEN	Tekfen Holding A.Ş.	1.215
TOASO	Tofaş Türk Otomobil Fabrikası A.Ş.	2.145
TCELL	Turkcell İletişim Hizmetleri A.Ş.	1.216
TUPRS	Türkiye Petrol Rafinerileri A.Ş.	2.210
THYAO	Türk Hava Yolları A.O.	4.699
TTKOM	Türk Telekomünikasyon A.Ş.	3.993
VAKBN	Türkiye Vakıflar Bankası T.A.O.	1.816
YKBNK	Yapı ve Kredi Bankası A.Ş.	2.602

Tablo 3.2 ise manuel etiketleme neticesinde 3 farklı kategorideki toplam kısa ileti sayılarını belirtmektedir.

Tablo 3.2 Tweetlerin Sınıflara Göre Dağılımı

Sınıf	Tweet Sayısı
Pozitif	14.691
Negatif	10.802
Nötr	32.440
Toplam	57.933

3.2. Veri Ön İşleme

Veri temizleme işlemi veri tabanında prosedür yardımıyla analiz için hazır hale getirilmiştir. Crawlerların çalışması sırasında oluşan aksaklıktan dolayı tekrar eden kayıtlar tekilleştirilmiş, sadece iş günleri ve borsa kapanış saatinden önceki tweetler dikkate alınmıştır. Cümle içindeki '@','$' karakterleri ile başlayan kelimeler, 'https://tco','pic.twitter.com/',' pictwittercom/' ifadeler, noktalama işaretleri ve sayılar kaldırılmıştır. ('%+', '+%', '-%','%-', ':)', ':(' vb. ifadeler artan/azalan veya olumlu/olumsuz anlam içermeye yardımcı olduklarından hariç tutulmuştur.)

Cümle içerisinde cümlenin sınıfına etkisi olmayan “durak kelimeler” olarak adlandırılan ve, şu, bu, de, da vb. kelimeler için tablo oluşturulmuştur. Bu kelimelerin, “#”, sayı ve sayıların dahil edilip edilmemesi tahmin oranını etkileyip etkilemediği Bölüm 6’da incelenmiştir.

Tweetin sınıfı üzerinde belirleyici bir etkisi olmayan sıkça tekrar eden aşağıdaki gibi kelime öbekleri kaldırıldığında ve sadece ilgilenilen hissenin artış/azalışına yönelik bilgi içeren bölümler tespit edilip ayıklandığında başarı oranına etkisinin olup olmadığı araştırılmıştır. Bu tip kontrollerin başarı oranında anlamlı bir artışı olmamıştır.

Sıkça tekrar edilen kelime öbekleri:

“ytd değildir”, “YTD”, “al sat tut tavsiyesi değildir.”, “Kesinlikle al sat tut tavsiyesi değildir.”, “al sat tut anlamı içermez”, “ANALİZLERİM YATIRIM TAVSİYESİ DEĞİLDİR.”, “en çok para girişi yaşayan hisseler”, “en çok yorum alan hisseleri”,” Para Giriş ve Çıkışı Olanlar”

3.3. Veri Setinin Sayısallaştırılması

Tweetlerin makine öğrenmesi algoritmalarıyla sınıflandırılabilmeleri için cümle içerisinde bulunan kelime veya kelime öbeklerinin sayısallaştırılmaları gerekmektedir. Her bir kelimenin veya kelime öbeklerinin bütün tweetler ve belge içerisindeki sıklıklarına göre hesaplanabilir. Yada belge bazında kelime sıklıklarına göre vektörler oluşturulabilir. Terim frekansı-ters doküman sıklığı (Tf-Idf), N-gram ve Kelime Vektörü (Word2Vec) yöntemleri örnek verilebilir.

3.3.1. Terim frekansı-ters doküman sıklığı (Tf-Idf)

Terim Frekansı doküman içinde geçen kelimelerin kullanılma sıklıklarına göre sayısal değerlere çevirilerek matris oluşturulması işlemidir.

Kelimenin belgedeki sıklığına ve tüm belgeler içindeki önemine göre aşağıdaki formüle ile hesaplanır.

$$\text{tf-idf}\,(t,d,D) = \text{tf}(t,d) \times \text{idf}(t,D) \tag{3.1}$$

t : terim sayısı, d: doküman sayısı D: tüm dokümanların sayısı

$$W_{x,y} = tf_{x,y} * \log(\frac{N}{df_x}) \quad (3.2)$$

$tf_{x,y}$ = y dokümanı içinde geçen x terimin sayısı

df_x = x terimini içeren doküman sayısı

N = toplam doküman sayısı

3.3.2. N-Gram

N-gram yöntemi metin içerisindeki kelimelerin veya kelime içindeki karakterlerin n sayısı kadar ayrılarak vektörlere dönüştürülmesi işlemidir. Aşağıda en sık kullanılan çeşitleri ve örnekleri gösterilmektedir.

Örnek : "bist100 kararsız açıldı, artı-eksi gidip geliyor, önemli tahtalarda, ayı-boğa kavgası var."

a) 1-Gram (Unigram):

Her kelimenin tek başına öz nitelik olarak değerlendirilmesidir.

"bist100", "kararsız", "açıldı", "artı-eksi", "gidip", "geliyor", "önemli", "tahtalarda", "ayı-boğa", "kavgası", "var."

b) 2-Gram (Bigram):

Kelimelerin ikişerli kombinasyonlarının kullanılmasıdır.

"bist100 kararsız", "kararsız açıldı", "açıldı artı-eski", "artı-eksi gidip", "gidip geliyor", "geliyor önemli", "önemli tahtalarda", "tahtalarda ayı-boğa", "ayı-boğa kavgası", "kavgası var"

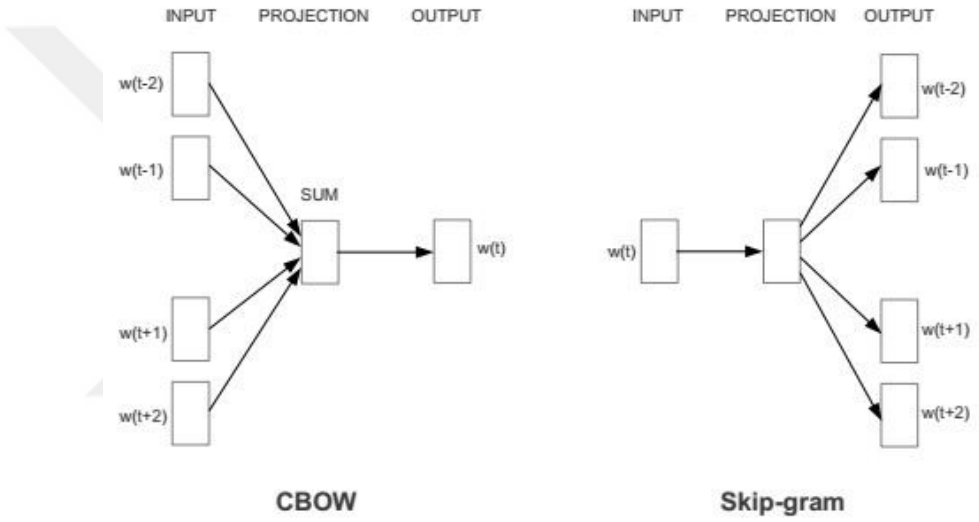
c) 3-Gram (Trigram)

Kelimelerin üçerli kombinasyonların kullanılmasıdır.

"bist100 kararsız açıldı", "kararsız açıldı artı-eski", "açıldı artı-eksi gidip", "artı-eksi gidip geliyor", "gidip geliyor önemli", "geliyor önemli tahtalarda", "önemli tahtalarda ayı-boğa", "tahtalarda ayı-boğa kavgası", "ayı-boğa kavgası var"

3.3.3. Doküman vektörlerinin (Doc2Vec) kullanılması

Kelime vektörünün (Word2Vec) doküman bazında geliştirilmiş halidir. Kelimelerin vektörel gösterilmesiyle birlikte çalışma mantığı yapay sinir ağlarına benzemektedirler. Sürekli Kelime Torbası (Continuous Bag of Words - CBOW) yönteminde her kelime özellik vektörü olarak eğitilerek kelime vektörlerini oluşturur. Bu kelimelerle ilişkili olabilecek kelimeler tahmin edilmeye çalışılır. Skip Gram yönteminde ise tek kelime vektörü ile o kelime ile ilişkili olabilecek kelimeler tahmin edilir.

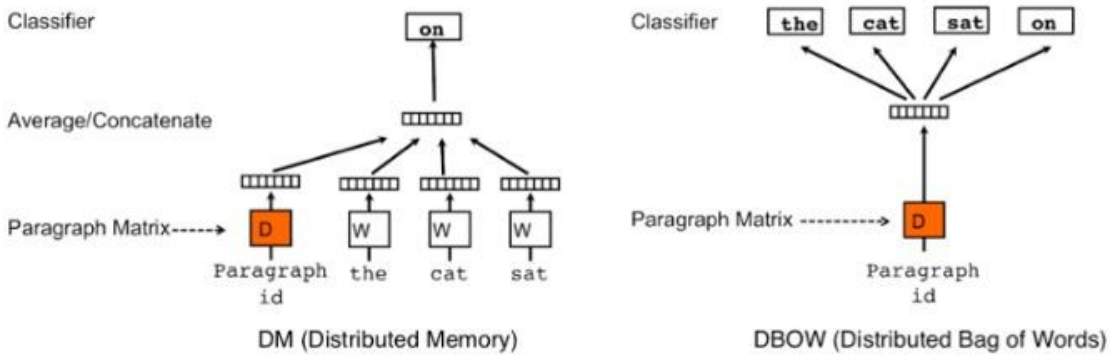


Şekil 3.2 Word2Vec Uygulama Şeması

Kaynak: https://towardsdatascience.com/another-twitter-sentiment-analysis-with-python-part-6-doc2vec-603f11832504 Erişim Tarihi:18 Ocak 2018

Doc2Vec yönteminde ise kelimeler doküman bazında ele alınır. Kelime vektörleri eğitilirken doküman vektörü de eğitilmiş olur. Dağıtılmış Hafıza (Distributed Memory - DM) yönteminde her bir doküman için kelime vektörlerinin input parametleri olarak verilerek bu kelimelerden bir kelime tahmin edilmeye çalışılmıştır. (Le & Mikolov, 2014) Dağıtılmış Bileşik Hafıza (Distributed Memory Concatenated - DMC) ve Dağıtılmış Ortalama Hafıza (Distributed Memory Mean - DMM) olarak çeşitleri vardır. DBH ve DOH arasındaki fark, ilkinin kelime vektörlerinin birleşiminden oluşması, ikincisinin ise bunların ortalamalarını almasıdır. DBH eğitim sırasında daha fazla hafıza tüketir ve daha büyük bir model üretir. (Tao, Chen & Lee, 2016)

Dağıtılmış Kelime Torbası (Distributed Bag of Words - DBOW) yönteminde dokümandan rastgele örneklenen kelimeleri tahmin etmeye çalışır. Yani doküman vektörüne sınıflandırma görevi verilir. (Le & Mikolov, 2014)



Şekil 3.3 Doc2Vec Uygulama Şeması

Kaynak: https://towardsdatascience.com/another-twitter-sentiment-analysis-with-python-part-6-doc2vec-603f11832504 Erişim Tarihi:18 Ocak 2018

3.3.4. Eğitim ve test verisinin bölünmesi

Doğru sınıflandırma yöntemini bulabilmek için veri setinin iki parçaya ayrılarak; bir parçası ile datalar eğitilerek diğer ayrılan parçanın doğruluğunun test edilmesi sağlanır. Bu bölümlendirmenin nasıl yapılacağı konusunda farklı yaklaşımlar geliştirilmiştir.

a) K-Katlı çapraz doğrulama (kFold):

Veri seti k tane eşit parçaya ayrılır. K-1 adedi eğitim verisi olacak biçimde model çalıştırılır. K defa her seferinde daha önce kullanılmayan parça test verisine bölünür. Hepsi çalıştırıldıktan sonra ortalamaları alınarak doğrulama ve performans ölçütleri elde edilir. Bu çalışmada literatürde sıkça kullanılan K-Katlı çapraz doğrulama yöntemi kullanılmış olup sonuçları Bölüm 6'da paylaşılmıştır.

b) Dışarıda tutma yöntemi (Hold-out):

Veri setinin 2/3 eğitim, diğer kalanı ise test verisi olmak üzere rastgele seçilir. K defa tekrarlanarak doğruluk ölçütlerinin ortalamaları alınarak bulunur.

c) Birini dışarıda bırakan çapraz doğrulama (Leaving-one out cross validation LOOCV):

Bir tane gözlemin test verisi olarak kullnıldığı yöntemdir. K defa tekrarlanarak elde edilen doğruluk ölçütlerinin ortalamasının hesaplanması ile elde edilir.

d) Yeniden Örnekleme (Bootstrap):

Seçilen örneklemin iade edilerek yeni veri kümeleri oluşturması yöntemidir. Bu sebeple veri kümeleri içerisinde tekrar eden gözlemler olabilir. Her bir veri setinden oluşan doğruluk ölçütlerinin ortalaması alınır.

4. KULLANILAN SINIFLANDIRMA YÖNTEMLERİ

4.1. Lojistik Regresyon

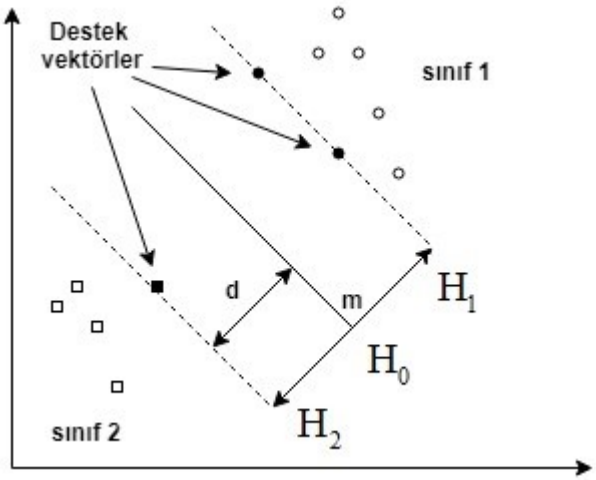
Lojistik regresyon bağımlı değişkenin nominal bir değer olduğu ve bu sebeple varsayımların sağlanamadığı durumlarda kullanılır. En küçük kareler yöntemi, kategorik değişkenlerin normal dağılım varsayımına uymaması sebebiyle bu değerlerin tahmininde kullanılamaz. Bu yüzden lojistik regresyon modeli maximum olabilirlik yöntemiyle tahmin edilir. (Kalaycı,2010,116)

$$L = \ln[\frac{p_i}{1-p_i}] \rightarrow \hat{Y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{4.1}$$

Makine öğrenmesi alanının sınıflandırma algoritmaları arasında da sıkça kullanılmaktadır. Bu analizde; tf-idf yöntemiyle hesaplanan kelime değerleri bağımsız değişkenleri (x değerlerini), duygu durumuna göre tweetlere verilen pozitif, negatif ve nötr olmak üzere 3 kategorili nominal bağımlı değişkeni (Y değişkenini) temsil etmektedir.

4.2. Destek Vektör Makinesi

Bir düzlemde veri kümesinin doğrusal veya doğrusal olmayan fonksiyon ile sınıflara ayrılması yöntemidir. Sınıflar birden fazla doğru ile ayırılabilir. Bu doğrulardan birbirine en uzak olanı bulmak en idealidir. En optimum uzaklığa sahip doğruya H_0 düzlemi denir.



Şekil 4.1 Destek Vektör Makinesi

$$H_0 = W^T x + b = 0 \tag{4.2}$$

W ağırlık vektörü, x öznitelik sayısını, b sabit katsayıyı ifade eder.

H_1 ve H_2 düzlemleri üzerindeki gözlemlere "destek vektör" adı verilir. x_1 destek vektörü ile H_2 düzlemi arasındaki uzaklık d ve H_1 ve H_2 arasındaki uzaklık ise m ile ifade edilirse aşağıdaki formüller ile hesaplanır. (Özkan,2016,170)

$$d = \frac{\left|W^T x_1 + b\right|}{w} \quad \text{ve} \quad m = 2d = \frac{2}{w} \tag{4.3}$$

Bu çalışmada lineer destek vektörü kullanılmıştır.

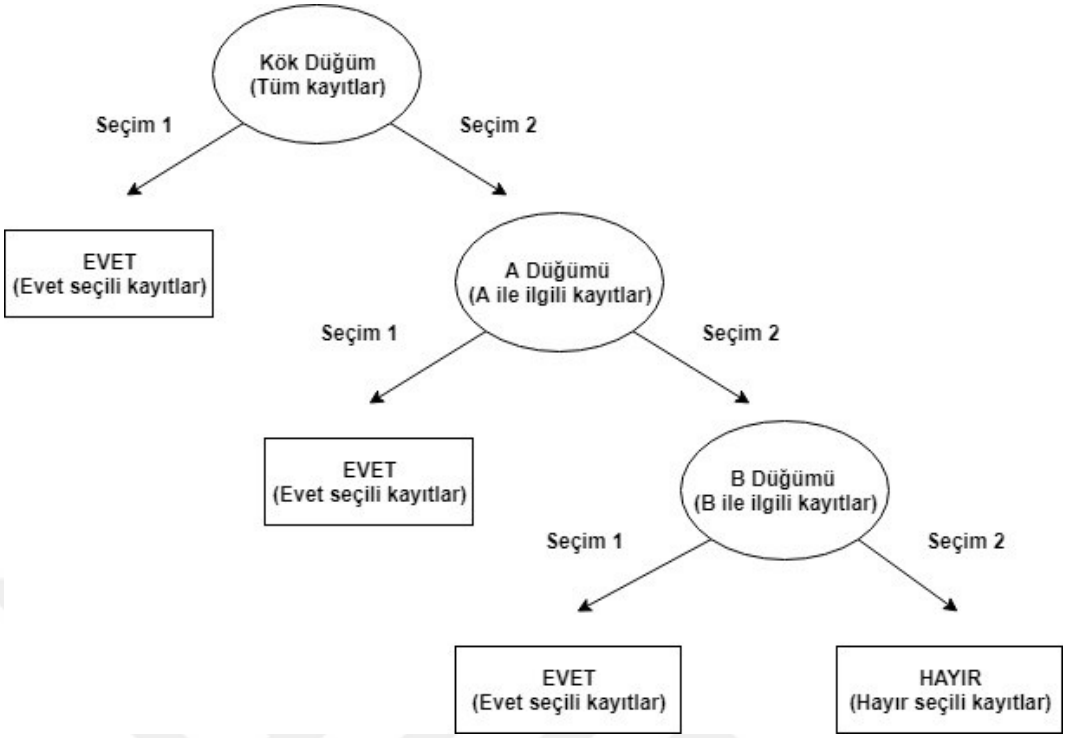
4.3. Naive Bayes Algoritması

Bayes Teoremi'ne dayanan istatistiksel bir sınıflandırma yöntemidir. İki tane A ve B olayında, birinin gerçekleşme olasılığı diğerine bağlı ise koşullu olasılık söz konusudur. M sınıfı olan C dğerlerinden oluşan örnek veri setinden sınıfı bilinmeyen X veri setinin tahmin edilmesi aşağıdaki formül ile ifade edilir. En büyük olasılığa sahip değer o verinin sınıfını belirler. (Özkan,2016,157)

$$P(C_i \setminus x) = \frac{P(X \setminus C_i)P(C_i)}{P(X)} \tag{4.4}$$

4.4. Karar Ağaçları

Karar Ağaçları akış şemaları ağaçlara benzeyen kök, dal ve yaprak adı verilen yapılardan oluşan algoritmalardır.En üst düğüm kök olup; seçilen kriterlere göre yapraklara doğru dallanma söz konusudur.Hangi kriter ile başlanacağı konusunda da çeşitli algoritmalar bulunmaktadır.Entropi kullanan ID3 ve C4.5 yanı sıra, sınıflandırma ve regresyon için Twing, Gini ve Regresyon Ağaçları Algoritmaları da kullanılmaktadır. (Özkan,2016,94) Bu çalışmada sınıflandırma temelli algoritmalar çalıştırılmış ve en başarılı olan 6. Bölüm'de paylaşılmıştır.



Şekil 4.2 Örnek Bir Karar Ağacı Akış Şeması

4.5. K-En Yakın Komşu Algoritması

Başlangıçta sınıfları belli olan bir veri setine dahil edilen yeni gözlemlerin tahmin edilmesinde kullanılan bir sınıflandırma yöntemidir. Yeni gözlem değerine tüm gözlemlerin uzaklıkları hesaplanır.İlk başta belirlenen bir k değeri kadar, bu uzaklıklar küçükten büyüğe sıralandığında; en fazla hangi sınıf varsa o gözlemin sınıfı olarak kabul edilmiş olur. Parametre olarak öklid uzaklığına dayanan aşağıdaki 3 çeşit formül uygulanır. (Özkan,2016,193) Bu çalışmada uzaklıklar ve k değerinin en başarılı olanı seçilerek sonuçları 6.Bölüm de paylaşılmıştır.

Öklit Uzaklığı:

$$d(i,j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{4.5}$$

Manhattan Uzaklığı:

$$d(i,j) = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| \tag{4.6}$$

Minkowski Uzaklığı:

$$d(i,j) = \sqrt[m]{\sum_{k=1}^{p} \left(\left| x_{ik} - x_{jk} \right|^{m} \right)} \tag{4.7}$$

4.6. Topluluk Öğrenmesi

Topluluk Öğrenmesi farklı sınıflandırma algoritmalarının ağırlıklarına göre puanlama yaparak en doğru değerlere ulaşmayı amaçlamaktadır. En önemli avantajı kullandığı algoritmaların verilerinin kombinasyonunu kullanıyor olmasıdır. (Kılınç, 2016)

4.6.1. Torbalama

Torbalama eğitim veri setinden rastgele n adet yeni eğitim veri setleri üretilerek tekrar eğitmeyi amaçlayan topluluk öğrenmesi yöntemidir. Bazı gözlemler dahil edilmezken, bazıları birkaç kez eklenebilir. Her bir sınıflandırıcı, farklı örnekler içeren eğitim setleri tarafından eğitilir ve sonuçları çoğunluk kuralı ile birleştirilir. (Kılınç, 2016) Bu çalışmada Random Forest kullanılmış olup Bölüm 6 da sonuçları paylaşılmıştır.

4.6.2. Gradyan Artırma

Artırma yönteminde önceki sınıflandırıcı tarafından belirlenemeyen veriler kullanılır. Her örnek, veri setinde bir ağırlığa sahiptir. Her öğrenme sürecinden sonra, her bir sınıflandırıcının sınıflandırma hatası dikkate alınarak örneklerin ağırlıkları güncellenir. Her bir sınıflandırıcının doğruluğuna dayanan ağırlıklı ortalama, yeni bir

örneği sınıflandırmak için seçilir ve sınıflandırma işlemi gerçekleştirilir. (Kılınç, 2016) Artırma art arda eğitim verisinin sınıflandırmada zayıf kalmış örneklerine uygulanarak çalışır. Her artırma ile yanlış sınıflandırılmış örneklerin ağırlıkları artarken, doğru sınıflandırılan örneklerin ağırlıkları azalır. Böylece önceki adımlarda sınıflandırılması zor olan örneklere odaklanır, başarı oranı artmış olur. (Krauss & Huck, 2017). Bu çalışmada Gradyan Artırma kullanılmış olup Bölüm 6 da sonuçları paylaşılmıştır.

4.6.3. Oylama

Oylama (voting) yöntemi makine öğrenmesi algoritmalarından gelen tahminleri birleştirerek en iyi sınıflandırıcıyı bulma yöntemidir. Sınıflandırıcılar arasında en çok oyu alma yöntemi "hard voting", sınıflandırıcıların olasılıklarının ortalamasını alarak tahminde bulunma yöntemi ise "soft voting" olarak adlandırılmaktadır. Bu çalışmada bölüm 6'da en iyi sonucu veren voting yöntemi paylaşılmıştır.

4.7. Doğrulama Süreci

Modelin doğruluğunu ve diğer modellere göre performansını karşılaştırabilmek için bazı ölçütler kullanılır.Eğitim veri setinden faydanılarak elde edilen tahminlerden bir matris elde edilir. Buna karışıklık matrisi (confusion matrix) adı verilir. (Özkan,2016,87)

Tablo 4.1 Üç sınıflı Karışıklık Matris Örneği

		Tahmini Sınıf		
		pozitif	**negatif**	**nötr**
Gerçek Sınıf	**pozitif**	A	B	C
	negatif	D	E	F
	nötr	G	H	I

Bu matris yardımıyla aşağıdaki ölçütler hesaplanır.

4.7.1. Doğruluk oranı (Accuracy)

Doğru tahmin edilen sınıf sayısının toplam gözlem sayısına oranıdır.Diğer bir deyişle gerçek değere yakınlık yüzdesini verir. N toplam gözlem sayısı olmak üzere aşağıdaki formül ile hesaplanır.

$$\text{Doğruluk Oranı} = \frac{A + E + I}{N} \tag{4.8}$$

4.7.2. Kesinlik ölçütü (Precision)

Bir sınıfın doğru tahmin sayısının o sınıfın tüm tahmin sayısına oranı olup sınıf içindeki başarısını belirler.

$$P_{pozitif} = \frac{A}{(A+D+G)} \quad P_{negatif} = \frac{E}{(E+B+H)} \quad P_{nötr} = \frac{I}{(I+C+F)} \tag{4.9}$$

4.7.3. Duyarlılık ölçütü (Recall)

Bir sınıfın doğru tahmin sayısının gerçekte o sınıfın gözlem sayısına oranı olup tahmin başarısını belirler.

$$R_{pozitif} = \frac{A}{(A+B+C)} \quad R_{negatif} = \frac{E}{(E+D+F)} \quad R_{nötr} = \frac{I}{(I+G+H)} \tag{4.10}$$

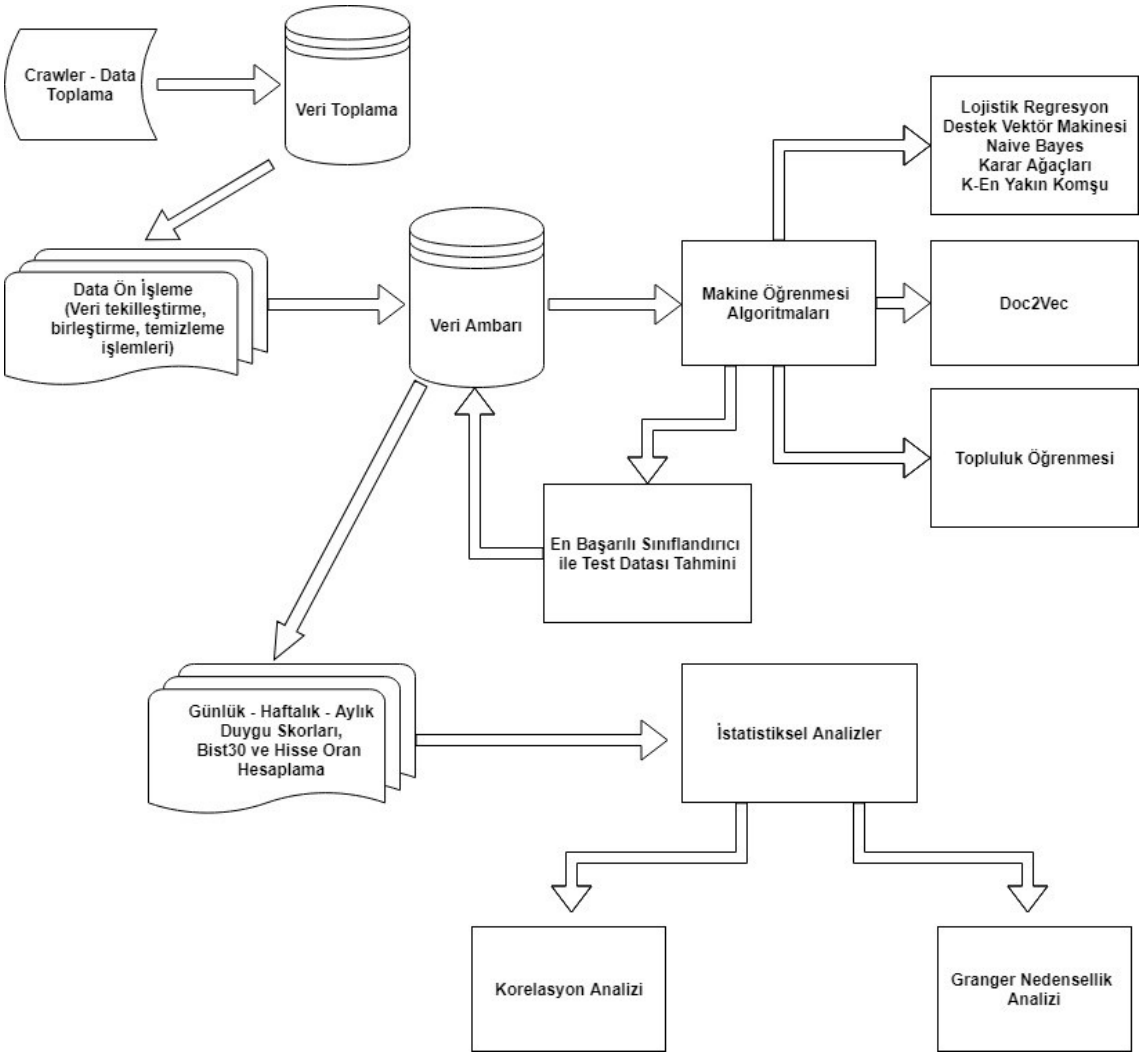
4.7.4. F ölçütü

Duyarlılık ve kesinlik oranlarının harmonik ortalaması hesaplanarak aşağıdaki formül ile elde edilir.

$$F = 2\frac{(Precision)(Recall)}{(Precision)+(Recall)} \tag{4.11}$$

5. UYGULAMA ADIMLARI VE ANALİZ

Makine öğrenmesi yapay zekanın bir alt kümesi olup, verilen bir veri setinin istatistiksel yöntemler yardımıyla öğretilmesini sağlayan ve işledikleri bilgiler sonucunda algoritmaların geliştirilmesini sağlayan bir bilgisayar bilimidir. Bu çalışmada; manuel olarak etiketlenen 57.933 adet tweeti, regresyon ve sınıflandırma algoritmaları yardımıyla en iyi modeli bulmak amaçlanmıştır. Sınıflandırma algoritmalarının çalışabilmesi için de metin verilerinin Bölüm 3.3 de anlatıldığı gibi sayısal değerlere çevrilmesi gerekir. (Kaynar vd., 2016)



Şekil 5.1 Çalışmanın Uygulama Şeması

5.1. Tweetlerin Sınıflandırılması

Bölüm 3'te belirlenen tüm mikro blog sayısallaştırma yöntemleri Bölüm 4 ile belirtilen tüm sınıflandırma algoritmalarına uygulanmış ve en başarılı olan yöntem ile tweetlerin sınıflandırması sağlanmıştır. Detaylar Bölüm 6'da paylaşılmıştır.

5.2. MikroBlog Polarite Hesaplanması

Literatürde yaygın olarak kullanılan duygu skoru pozitif ve negatif yorum sayılarının farklarının toplamlarına oranıdır. Ancak bu duygu skoru analiz edilecek veriler ile kullanılacağından her dağılım için yeterli olmayabilir. Alternatif hesaplar için araştırma yapılmış ve aşağıdaki formüllerin de kullanılabildiği bulunmuştur. Nötr yorumların duygu durumuna herhangi bir katkısı olmadığı için hesaplamaya sadece pozitif ve negatif sınıflı tweetler dahil edilmiştir.

p_t : t zamandaki pozitif yorum yapılmış tweetlerin sayısı

n_t : t zamandaki negatif yorum yapılmış tweetlerin sayısı

$$S_1 = \frac{p_t - n_t}{p_t + n_t} \qquad \text{(Ranco vd., 2015)} \qquad (5.1)$$

$$S_2 = \ln \frac{p_t + 1}{n_t + 1} \qquad \text{(Oliveira, Cortez \& Areal 2017)} \qquad (5.2)$$

$$S_3 = \frac{p_t}{p_t + n_t} - \frac{p_{t-1}}{p_{t-1} + n_{t-1}} \qquad \text{(Oliveira, Cortez \& Areal 2017)} \qquad (5.3)$$

$$S_4 = 1 - \sqrt{1 - \left(\frac{p_t - n_t}{p_t + n_t}\right)^2} \qquad \text{(Oliveira, Cortez \& Areal 2017)} \qquad (5.4)$$

5.3. İstatistiksel Analiz

5.3.1. Korelasyon analizi

Korelasyon analizi iki sürekli değişken veya birden fazla değişkenin bağımlı bir değişken ile arasında anlamlı doğrusal bir ilişki olup olmadığını incelemek için kullanılan bir analizdir. Korelasyon katsayısı ise bu ilişkinin derecesini gösterir. Değişkenlerden biri artarken diğeri azalıyorsa negatif, her ikisi de de aynı yönde değişiyorsa pozitif korelasyon ilişkisi vardır. Korelasyon analizi iki değişken arasındaki neden-sonuç ilişkisi hakkında bilgi vermez. Analiz öncesi normallik ve doğrusal ilişkinin varlığı grafik ve testler yardımıyla kontrol edilmelidir. (Kalaycı,2010,116)

Bu çalışmada SPSS yardımıyla Pearson ve Spearman Sıra Korelasyon analizleri test edilmiş ve sonuçları Bölüm 6'da paylaşılmıştır.

5.3.2. Granger nedensellik analizi

Granger zamana bağlı gecikmeli ilişkisi bulunan iki değişkenin nedenselliğinin yönünü istatistiksel açıdan belirmeyi amaçlayan bir analizdir. Aşağıdaki eşitliklerden; $\mathbf{y_t}$'deki değişiklikler $\mathbf{x_t}$'deki değişikliklere neden oluyorsa $\mathbf{y_t}$'den $\mathbf{x_t}$'ye doğru Granger nedenselliği vardır. Nedensellik ilişkisi tek yönlü olabileceği gibi çift yönlü de olabilmektedir. (Öner, İçellioğlu & Öner 2018)

$$y_t = \alpha_1 + \sum_{i=1}^{m} \beta_i x_{t-i} + \sum_{j=1}^{m} \gamma_j y_{t-j} + e_{1t} \tag{5.5}$$

$$x_t = \alpha_2 + \sum_{i=1}^{n} \theta_i x_{t-i} + \sum_{j=1}^{n} \delta_j y_{t-j} + e_{2t} \tag{5.6}$$

a) Durağanlık ve Birim Kök Testi

Granger testinin varsayımlarından biri serilerin durağan olmasıdır. Zaman içerisinde ortalama, varyans ve ortak varyansı değişmeyen serilere durağan denir. Durağanlık kontrolü birim kök testleri ile yapılır. En yaygın kullanılan Dickey Fuller (ADF) testi olup, aşağıdaki regresyon modellerine sabit veya trend eklenerek hesaplanmasıyla elde edilir. (Pekkaya & Bayramoğlu, 2008)

Sabitsiz ve trendsiz model: $$\Delta Y_t = \delta Y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta Y_{t-1} + \varepsilon_t \quad (5.7)$$

Sabitli ve trendsiz model: $$\Delta Y_t = \mu + \delta Y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta Y_{t-1} + \varepsilon_t \quad (5.8)$$

Sabitli ve trendli model: $$\Delta Y_t = \mu + \beta T + \delta Y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta Y_{t-1} + \varepsilon_t \quad (5.9)$$

Birim kök sınaması regresyon modellerine aşağıdaki hipotezin kurulması ile test edilir. Zaman serisinin durağanlaştırılrması sırasında birinci dereceden farkı alındıysa seri 1. dereceden bütünleşik olup, I(1) ile gösterilir. (Öner, İçellioğlu & Öner 2018)

$H_0 : \delta = 0$ ise, Y_t birim köke sahiptir ve durağan değildir.

$H_1 : \delta < 0$ ise, Y_t birim köke sahip değildir ve durağandır.

b) Gecikme Uzunluğunun Hesaplanması

Vektör Otoregresyon (Vector Autoregression - VAR) modeliyle gecikme uzunluğu hesaplanır. Akaike (AIC), Schwarz (SC), Hannan-Quinn (HQ) gibi kriterlerin değerlerini en küçük yapan gecikme sayısı Granger modelinin gecikme uzunluğunu verir. (Pekkaya & Bayramoğlu, 2008)

Bu çalışmada EViews programı yardmıyla duygu skorları ile Bist30 veya hisse oranları, USD ile Bist30, tweet sayısı ile hisse oranları arasında Granger analizleri yapılmıştır. Sonuçları Bölüm 6 da paylaşılmıştır.

6. SONUÇLAR

6.1. Eğitim Veri Setinde Sıkça Kullanılan Durak Kelimeler

Durak kelime, kısaltma, ek , bağlaç gibi kelimeleri birbirine bağlamaya yarayan ancak kendi başlarına cümleye herhangi bir anlam katmayan kelimelere denir. Twetlerin içinde frekanslarına göre sıralandığında aşağıdaki kelimelerin sıkça olduğu görülmüştür.

Tablo 6.1 En sık kullanılan 50 durak kelime

ve	mi	ye	zaten	ancak
bu	den	dk	biz	işte
bir	mı	bazı	rt	sanki
çok	ya	mu	acaba	nin
için	bi	sen	beni	veya
en	ki	hala	yi	yada
de	şu	bile	size	sizce
da	dan	hiç	siz	kendi
ne	ise	tek	hep	bizi
ile	an	bunu	yani	kimse

6.2. Eğitim Veri Setinden Sözlük Oluşturulması

Eğitim veri setindeki cümlelerde kullanılan kelimeler ile sözlük hazırlanmıştır. Zemberek programı ile python dilinde kelime çözümlemesi yapılarak kökleri ve köklerin tipi (sıfat, isim, fiil) bulunmuştur. Pozitif, negatif ve nötr sınıfında en sık kullanılan sıfat ve isimler Tablo 6.2 de örnekler gösterilmiştir.

Tablo 6.2 Eğitim Sözlüğünden Örnekler

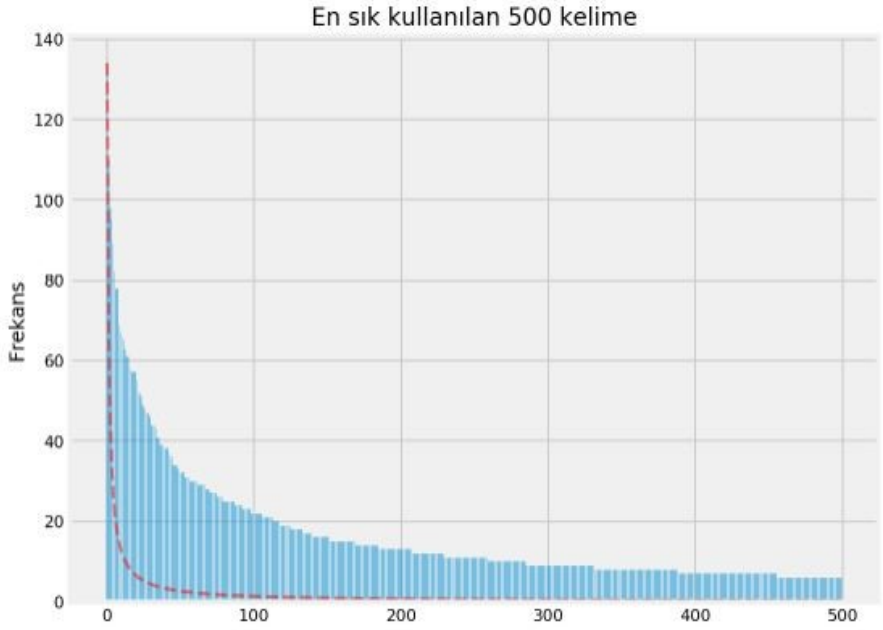
pozitif		**negatif**		**nötr**	
sıfat	**isim**	**sıfat**	**isim**	**sıfat**	**isim**
güzel	yukarı	sert	düş	iyi	hisse
üstün	bereket	kötü	dip	uzun	yatırım
yüksek	olumlu	kırmızı	eksi	yakın	analiz
yeşil	inşallah	fena	yazık	büyük	teknik
sakin	pozitif	rezil	sıkıntı	fazla	para
muhteşem	fırsat	perişan	negatif	doğru	yabancı
harika	maşallah	gergin	olumsuz	canlı	tavsiye
muazzam	şükür	berbat	bela	yeter	aracı
şahane	artı	tuhaf	kriz	belli	borsa
güçlü	getiri	çürük	kayıp	zor	kurum

Tablo 6.2 (Devamı) Eğitim Sözlüğünden Örnekler

pozitif		negatif		nötr	
sıfat	**isim**	**sıfat**	**isim**	**sıfat**	**isim**
	boğa				ayı
	tobo				obo
	tepe				direnç
	fincan (kulp)				destek
	rsi				flama
					gap
					doji
					devre (kesmek)

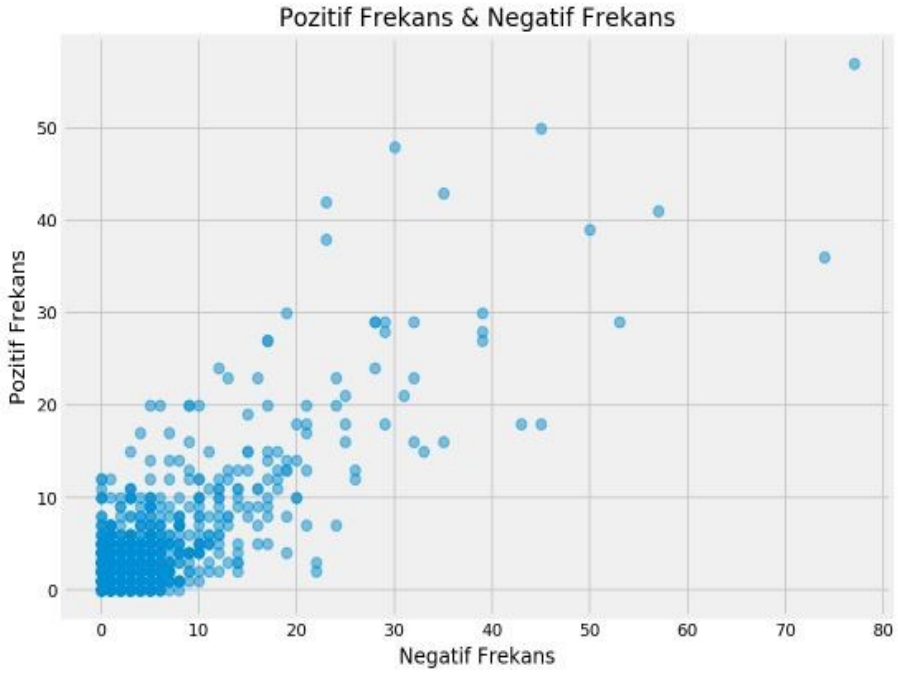
6.3. Zipf Yasası

Zipf's Yasası veri setinde en sık kullanılan sözcüklerin büyükten küçüğe doğru sıralanarak histogram üzerinde gösterilmesidir. Grafik üzerinde ortaya çıkan eğriye Zipf's Eğrisi denir. Log-log ölçeğinde ise doğrunun doğrusal olması beklenir. (Zhang, Dong & Mu, 2018) Her sözcüğün gözlem sayısına göre sıra numarası oranlandığında sözcüklerin sıklıkları bulunmuş olur.



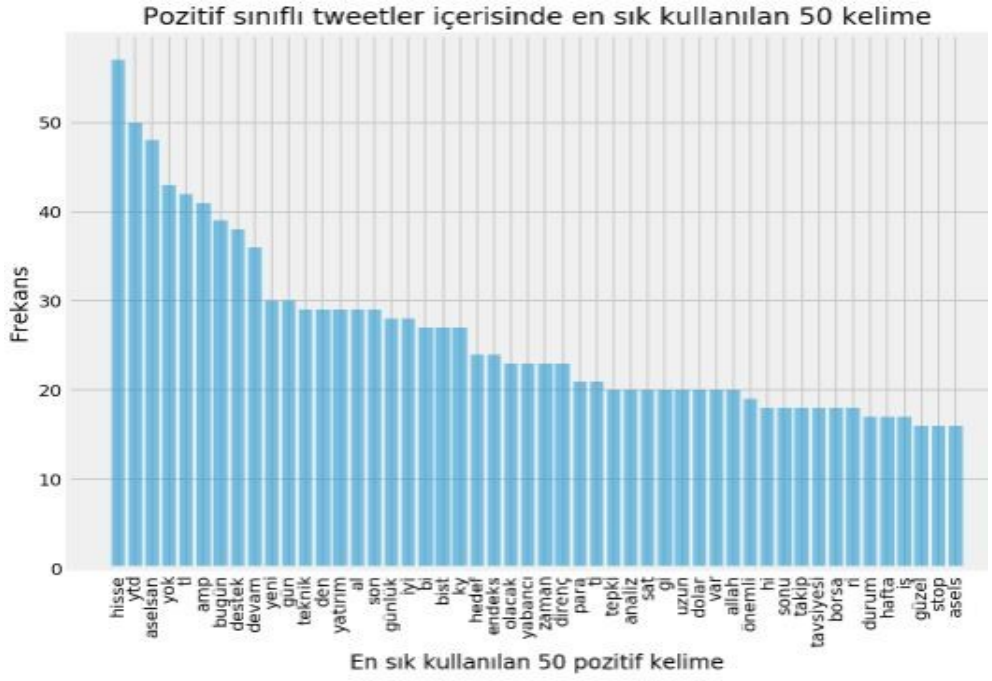
Şekil 6.1 Zipf Eğrisi

Dağılım Zipf eğrisini takip ediyor olsa da eğrinin üzerinde kalan yüksek frekanslı kelimeler bulunmaktadır.

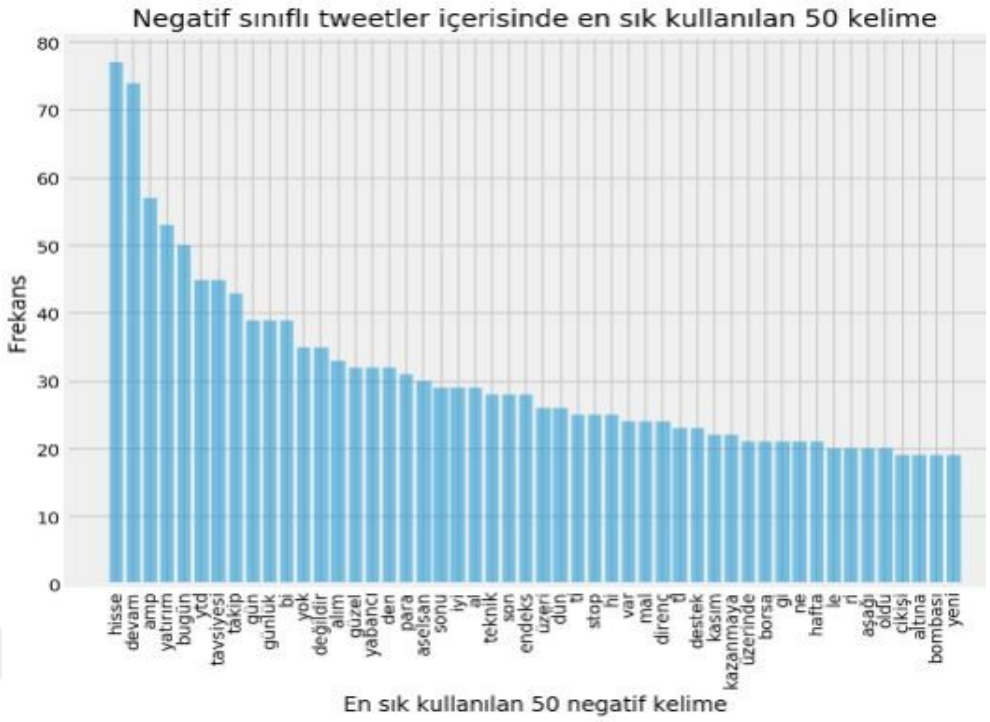


Şekil 6.2 Pozitif ve Negatif Frekans Saçılım Grafiği

20 frekansın altında kalan pozitif ve negatif kelimelerin arasında anlamlı bir ilişki olmadığı grefikten görülmektedir. Sınıfların aralarında ilişki olmaması sınıflandırma tahmininde başarıyı arttımayı sağlar.



Şekil 6.3 En Sık Kullanılan 50 Pozitif Kelime Frekans Dağılımı



Şekil 6.4 En Sık Kullanılan 50 Negatif Kelime Frekans Dağılımı

Grafiklerden “hisse”, “ytd”, “ tavsiye” gibi kelimelerin her iki sınıfta da sıkça kullanıldığı ortaya çıkmıştır. Sözlükten bu kelimelerin nötr sınıfına ait olduğunu görebiliriz.

6.4. Tf-Idf Sonuçları

Tablo 6.3 kFold ve Tf-Idf Uygulanmış Algoritma Doğruluk Oranları

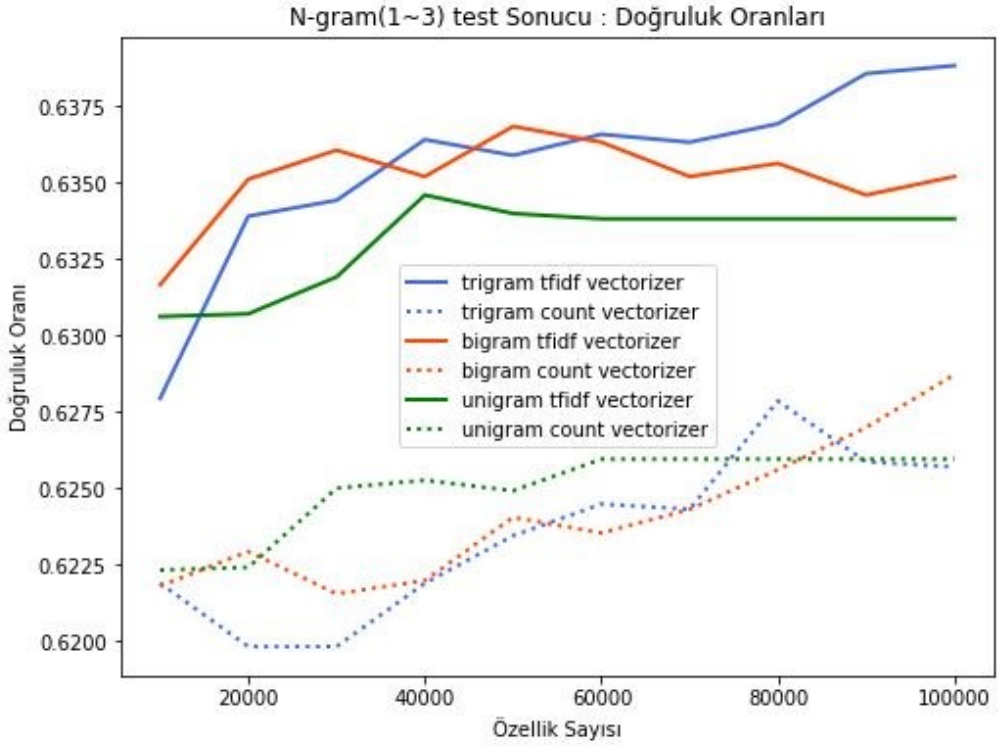
	1-Gram	**2-Gram**	**3-Gram**
LR	%63,41 (+/- 0,28)	%63,60 (+/- 0,29)	**%63,47 (+/- 0,24)**
DVM	%62,53 (+/- 0,48)	%62,32 (+/- 0,54)	%62,32 (+/- 0,47)
NB	%59,55 (+/- 0,32)	%61,58 (+/- 0,43)	%61,40 (+/- 0,56)
KA	%58,01 (+/- 0,30)	%57,60 (+/- 0,36)	%57,47 (+/- 0,41)
KNN	%56,16 (+/- 0,45)	%55,43 (+/- 0,45)	%55,34 (+/- 0,49)

Sınıflandırma algoritmalarının en başarılı olanının Lojistik Regresyon olduğu görülmektedir. En yüksek tahminin bigram olmasına rağmen standart hata oranı daha fazla olduğu için trigram ile aralarında anlamlı bir fark yoktur. Lojistik regresyon ile istenilen n-gram kullanılarak tahmin yapılabilir.

Tablo 6.4 kFold ve Tf-Idf Uygulanmış Algoritma F1 Skor Sonuçları

	1-Gram	2-Gram	3-Gram
LR	%53,68	%53,54	%53,24
DVM	%55,34	%55,20	%54,96
NB	%41,98	%51,77	%53,67
KA	%46,86	%46,07	%45,99
KNN	%49,01	%48,65	%48,54

Lojistik Regresyon ve Destek Vektör Makinesi'nde n-gram bazında F ölçütleri birbirine yakın görünmektedir. Algoritmalar arasında incelendiğinde ise Destek Vektör Makinesi'nin daha başarılı olduğu görülmüştür. Ancak aralarında önemli bir ölçüde farklılık görülmemektedir.

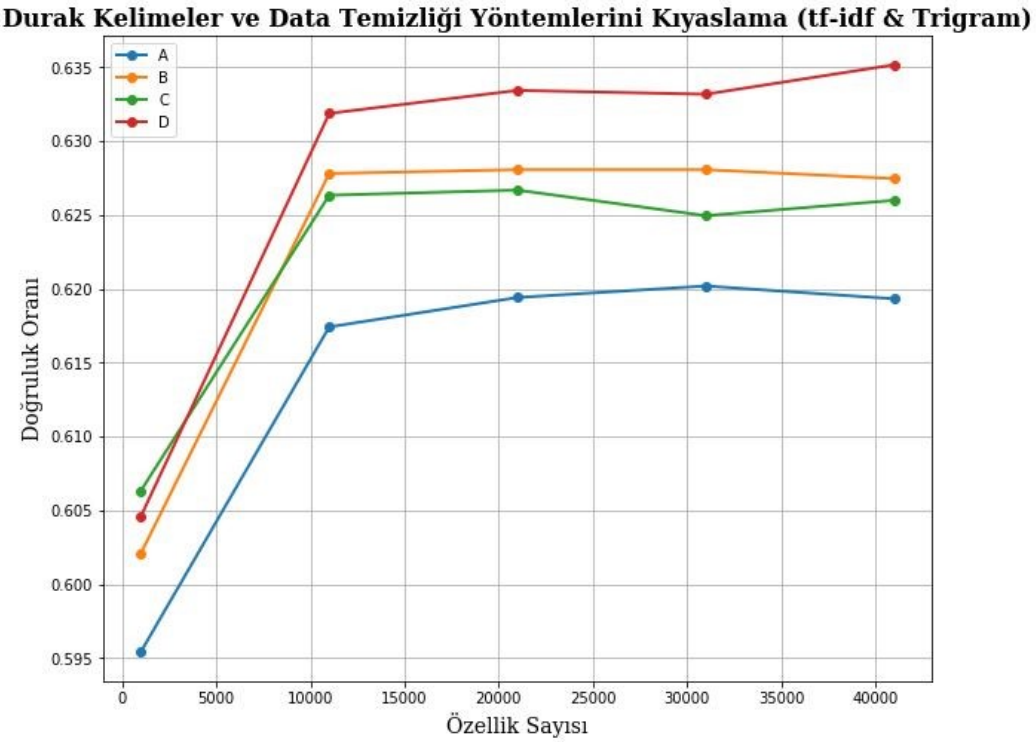


Şekil 6.5 Tf ve Tf-idf Doğruluk Oranlarının Karşılaştırılması

Grafikten özellik sayısının artıkça tahmin başarısının arttığını ve trigramın bigrama göre daha başarılı olduğu söylenebilir. Ancak k katlı çapraz doğrulama ile ortalamada aralarında anlamlı bir fark olmadığı tablodan anlaşılmaktadır. Aynı zamanda tf-idf yönteminin terim sıklığına göre daha iyi tahmin ettiği görülmüştür.

Veri seti üzerinde durak kelimelerin dahil edilip, edilmemesi veya benzer tweetler üzerinde filtreler uygulandığında başarı oranını anlamlı ölçüde değiştirebilme olasılığına karşı farklı veri setleri oluşturularak kıyaslaması yapılmıştır.

Şekil 6.6 incelendiğinde farklı veri setleri üzerinde lojistik regresyon ile tahmin yapıldığında aralarında en fazla %1.6 kadar fark oluştuğu görülebilir. 6.4 Başlığında paylaşılan sonuçlar D veri setine aittir. Diğer veri setlerinde daha iyi performans elde edilmediği için bu veri setiyle çalışmalara devam edilmiştir.



Şekil 6.6 Farklı Veri Setleri Üzerinde Lojistik Regresyon Sonuçları

A: Durak kelimeler hariç filtre uygulandığında doğruluk oranı: 61.93%

B: Sadece filtre uygulandığında doğruluk oranı: 62.75%

C: Kelime köklerine göre filtre uygulandığında doğruluk oranı: 62.60%

D: Genel data temizliği yapılmış datanın doğruluk oranı: 63.52%

6.5. Doc2Vec Sonuçları

Tablo 6.5 Doc2Vec Sonuçlarının Doğruluk Oranları

	1-Gram	2-Gram	3-Gram	Sonuç
DKT	%57,93 (+/- 0,014)	%57,39 (+/- 0,013)	%58,05 (+/- 0,019)	3-gram
DBH	%55,47 (+/- 0,021)	%55,54 (+/- 0,021)	%55,58 (+/- 0,021)	3-gram
DOH	%56,53 (+/- 0,017)	%56,63 (+/- 0,020)	%56,41 (+/- 0,015)	2-gram
DKT+DBH	%58,51 (+/ 0,016)	%57,82 (+/- 0,015)	%57,98 (+/- 0,017)	1-gram
DKT+DOH	%58,55 (+/- 0,016)	%58,34 (+/- 0,019)	**%58,96 (+/- 0,013)**	**3-gram**

Tablo 6.5 incelendiğinde en iyi sonucun DKT ve DOH birleşiminin trigram ile çalıştırılması ile elde edilmiş olup; Lojistik Regresyon'dan daha başarılı olmadığı görülmektedir.

6.6. Topluluk Öğrenme Sonuçları

6.6.1. Torbalama ve Artırma Sonuçları

Tablo 6.6 Topluluk Öğrenme Sonuçları

	1-Gram	2-Gram	3-Gram
Rastgele Orman	**%59,69 (+/- 0,48)**	%59,66 (+/- 0,37)	%59,47 (+/- 0,43)
Gradyan Artırma	%59,47 (+/- 0,33)	%59,55 (+/- 0,32)	%59,62 (+/- 0,32)

Tablo 6.6 incelendiğinde; 10 katmanlı çapraz doğrulama yöntemiyle elde edilen sonuçlara göre unigram ile eğitilmiş Rastgele Orman topluluk öğrenme yönteminin en başarılı olduğu ortaya çıkmıştır. Ancak doğruluk oranı Lojistik resgresyon sonucundan daha düşük olduğu görülmektedir.

6.6.2. Oylama sonucu

Sınıflandırıcılar arasından en başarılı ilk 3 tanesi olan Lojistik Regresyon, Destek Vektör Makinesi ve Naive Bayes seçilip oylama yapıldığında aralarında yine en iyi tahminin Tablo 6.7'deki doğruluk oranlarına sahip olan Lojistik Resgresyon olduğu

görülmektedir. Ancak doğruluk oranları Lojistik Regresyon'un tek başına çalıştırıldığındaki doğruluk oranından daha düşüktür.

Tablo 6.7 Topluluk Öğrenme Sonuçları

	1-Gram	2-Gram	3-Gram
Oylama Sonucu	%62,97 (+/-0,23)	%63,30(+/-0,18)	%63,09(+/-0,21)

6.7. İki Sınıflı Tf-Idf Sonuçları

Tablo 6.8 İki Sınıflı Doğruluk Oranları Sonuçları

	1-Gram	2-Gram	3-Gram
LR	73,23% (+/- 0,55)	72,99% (+/- 0,56)	72,92% (+/- 0,77)
DVM	72,97% (+/- 0,55)	72,84% (+/- 0,94)	72,52% (+/- 0,65)
NB	69,34% (+/- 0,77)	70,94% (+/- 0,81)	71,41% (+/- 0,80)
KA	64,28% (+/- 0,88)	63,62% (+/- 0,76)	63,85% (+/- 1,13)
KYK	66,81% (+/- 0,52)	66,36% (+/- 0,87)	65,92% (+/- 0,93)

Tablo 6.8'den anlaşıldığı üzere iki sınıflı tweet sınıflandırılmasında başarı oranı artmıştır. Buradan nötr tweetlerin başarı oranını azaltığı sonucu çıkarılabilir. Ancak pozitif ve negatif tweetleri ayırabilmemiz için nötr sınıfınında da sınıflandırmaya dahil edilerek tahmin edilmesi gerekmektedir.

6.8. Değerlendirme ve Duygu Skorlarının Hesaplanması

57.933 eğitim veri setiyle çalışılarak sınıflandırma algoritmaları, Topluluk Öğrenmesi ve Doc2vec sonuçlarını da kıyasladıktan sonra en iyi modelin halen Lojistik Regresyon olduğu tespit edilmiştir. 2018 Mayıs-2019 Nisan periyodundaki tweet verilerinden her ayı kapsayacak şekilde eğitim datasının %35'ini oluşturacak biçimde test datası seçilmiştir. Böylece 78.209 adet yeni veri seti oluşmuş; yaklaşık %25'i test datası geri kalanı da eğitim datası olmak üzere dağıtılmıştır. Sonuç itibariyle, 57.933 manuel etiketleme yapılmış tweetler lojistik regresyon yöntemiyle 20.276 adet test verisinin sınıfı tahmin edilmiştir.

Yeni oluşturulan veri setinden günlük, haftalık, aylık periyotlarda Bölüm 5'te belirtilen 4 çeşit duygu skorları hesaplanmıştır. Her üç periyot içinde; bu duygu skorları ile Bist30 açılış-kapanış fark değişimleri ve Bist30 içerisinde yer alan hisselerin hisse oranları arasında ilişkileri incelenmek amaçlanmıştır.

6.9. Korelasyon Analizi Sonuçları

Tüm istatistik analizlerinde olduğu gibi korealsyon testlerine başlamadan önce verilerin varsayımları sağlayıp sağlamadığı kontrol edilmelidir. Öncelikle serilerin normal dağılıma uyup uymadıkları kontrol edilmesi için aşağıdaki hipotez kurulur.

$\boldsymbol{H_0}$: Verilerin gösterdiği dağılım ile normal dağılım arasında fark yoktur.

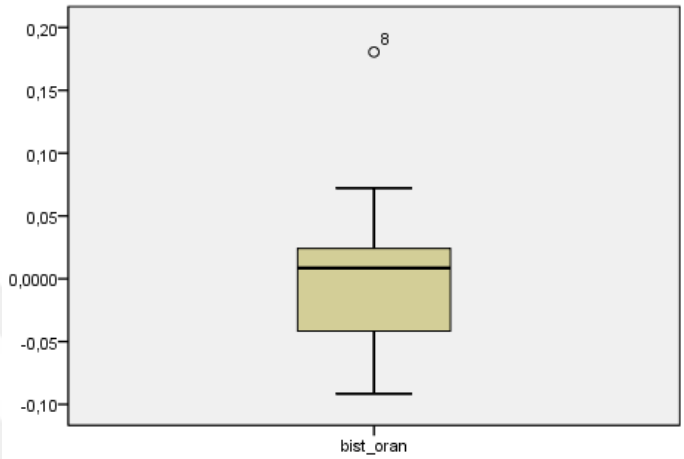
$\boldsymbol{H_1}$: Verilerin gösterdiği dağılım ile normal dağılım arasında fark vardır.

Tablo 6.9 Verilerin Normal Dağılım Testleri

		Günlük	**Haftalık**	**Aylık**
S_1	Ortalama	0.145754	0.147410	0.150700
	Çarpıklık	-0.411	-0.608	-0.836
	Basıklık	-0.231	0.091	-0.428
	Kolmogorov-Smirnov[a]	**0.008**	0.200*	0.200*
	Shapiro-Wilk	**0.002**	0.104	0.183
S_2	Ortalama	0.296271	0.298828	0.304345
	Çarpıklık	-0.285	-0.537	-0.815
	Basıklık	-0.311	0.029	-0.456
	Kolmogorov-Smirnov[a]	**0.029**	0.200*	0.200*
	Shapiro-Wilk	**0.034**	0.176	0.199
S_3	Ortalama	0.000250	0.001504	0.006591
	Çarpıklık	-0.284	-0.663	-0.064
	Basıklık	0.114	0.487	-0.669
	Kolmogorov-Smirnov[a]	0.200*	0.200*	0.200*
	Shapiro-Wilk	0.272	0.180	0.759
S_4	Ortalama	0.028053	0.017138	0.013091
	Çarpıklık	1.316	1.053	-0.258
	Basıklık	1.308	0.323	-1.015
	Kolmogorov-Smirnov[a]	**0.000**	**0.013**	0.200*
	Shapiro-Wilk	**0.000**	**0.000**	0.627
Bist30 Oran	Ortalama	-0.000153	-0.000928	0.003555
	Çarpıklık	-0.398	-0.371	1.171
	Basıklık	0.966	0.541	**2.005**
	Kolmogorov-Smirnov[a]	0.200*	0.200*	0.200*
	Shapiro-Wilk	**0.028**	0.837	0.279

Normallik testlerinden Kolmogorov-Smirnov gözlem sayısı büyük olduğunda, Shapiro-Wilk ise gözlem sayısı 29 dan az olduğunda tercih edilir. Bu sebeple günlük ve haftalık veri setleri için Kolmogorov-Smirnov, aylık veriler için Shapiro-Wilk test

sonuçları değerlendirilmesi uygun olur. Testlerin sig değerleri 0.05'ten büyük olduğu zaman H_0 hipotezi kabul edilir ve verilerin normal dağıldığı sonucuna varılır. Tablo 6.9'a bakıldığında günlük periyotta S_1, S_2 ve S_4 skorlarının, haftalık periyotta ise S_4 skorunun normal dağılmadığı görülebilir.

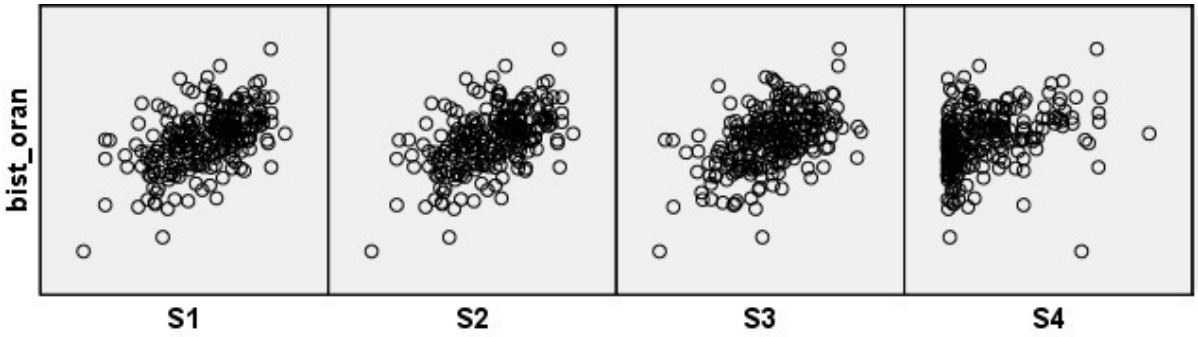


Şekil 6.7 Aylık Bist30 Oranının Kutu Grafiği

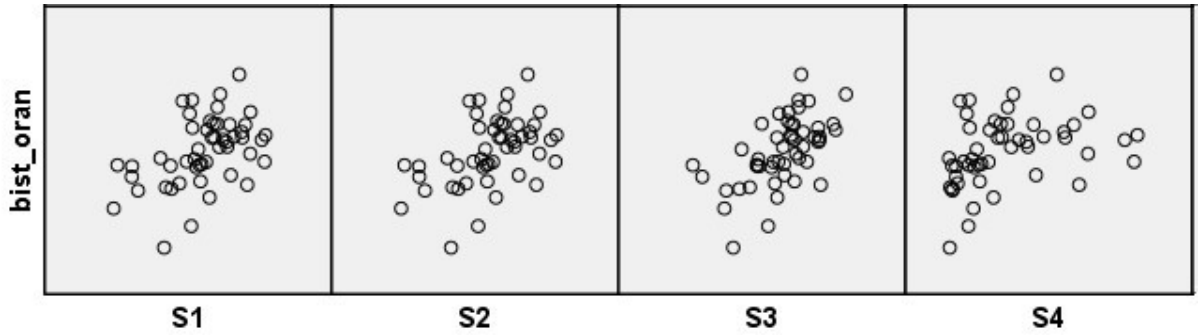
Bist oran dağılımı sola çarpık olup, 8. Gözlem değeri aykırı değerdir. Veri setinden 8. Gözlem değeri incelendiğinde 2019 yılının ocak ayına ait olduğu ve önceki aya göre değişiminin fazla olduğu ortaya çıkmıştır.

Tablo 6.10 Aylara Duygu Skorları ve Bist30 Dağılımı

	Aylar	**Tweet Sayısı**	S_1	S_2	S_3	S_4	**Bist30 Oranı**
1	2018-06	3.307	0,1393	0,2800	0,0404	0,0097	-0,0302
2	2018-07	6.925	0,1616	0,3258	0,0111	0,0131	0,0085
3	2018-08	3.998	0,1806	0,3649	0,0095	0,0165	-0,0458
4	2018-09	5.664	0,1601	0,3227	-0,0103	0,0129	0,0721
5	2018-10	7.128	0,1034	0,2073	-0,0284	0,0054	-0,0784
6	2018-11	5.174	0,1847	0,3734	0,0407	0,0172	0,0343
7	2018-12	7.934	0,0445	0,0891	-0,0701	0,0010	-0,0374
8	2019-01	11.509	0,2223	0,4520	0,0889	0,0250	**0,1805**
9	2019-02	15.020	0,2027	0,4109	-0,0098	0,0208	0,0141
10	2019-03	1.938	0,0549	0,1094	-0,0739	0,0015	-0,0916
11	2019-04	6.513	0,2036	0,4123	0,0744	0,0209	0,0130

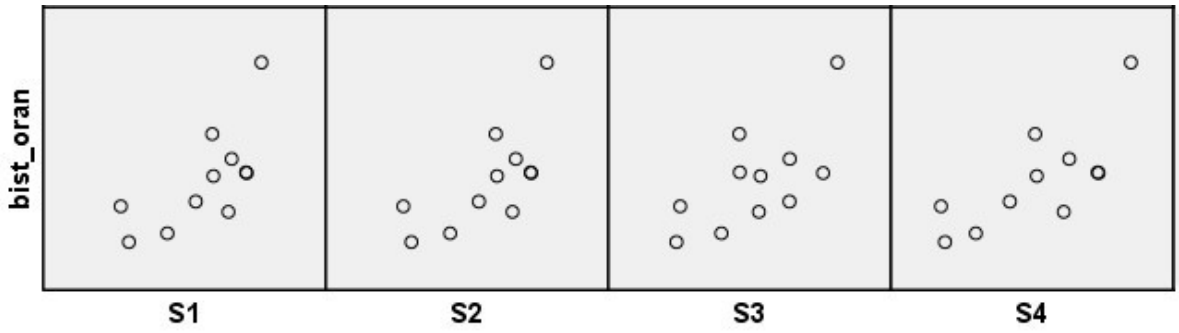


Şekil 6.8 Duygu Skorları ile Bist30 Arasındaki Günlük Saçılım Grafikleri



Şekil 6.9 Duygu Skorları ile Bist30 Arasındaki Haftalık Saçılım Grafikleri

Günlük ve haftalık bist oran değerleri ile duyu skorlarının saçılım grafikleri incelendiğinde S_1, S_2, S_3 duygu skorları ile Bist30 oranlarının arasında pozitif korelasyon olduğu, ancak S_4 skorunda korelasyonun olup olmadığı anlaşılamamaktadır.



Şekil 6.10 Duygu Skorları ile Bist30 Arasındaki Aylık Saçılım Grafikleri

Aylık Bist30 oranı ile duygu skorları arasında pozitif kolerasyonun olduğu, S_4 skorunun ise periyot artıkça Bist30 oranıyla olan ilişkisinin arttığı öngörülebilir.

Tablo 6.11 Bist30 ve Hisse Bazında Duygu Skorlarının Korelasyon İlişkisi

	Günlük	**Haftalık**	**Aylık**	**En iyi Polarite Skoru**
BİST30				
S_1	0,51	0,47	**0,70**	
S_2	0,51	0,46	**0,70**	S_4
S_3	0,49	0,58	**0,68**	
S_4	0,39	0,45	**0,74**	
AKBNK				
S_1	0,27	0,37	-	
S_2	0,28	0,38	-	S_3
S_3	-	0,35	**0,66**	
S_4	-	-	-	
ARCLK				
S_1	0,23	0,29	-	
S_2	0,26	0,34	**0,61**	S_2
S_3	-	-	-	
S_4	-	-	-	
ASELS				
S_1	0,42	0,40	-	
S_2	0,47	0,46	-	
S_3	0,33	0,37	-	
S_4	0,16	-	-	
BIMAS				
S_1	0,25	-	-	
S_2	0,28	-	-	
S_3	-	0,44	-	
S_4	-	-	-	
DOHOL				
S_1	0,26	-	-	
S_2	0,30	-	-	
S_3	0,18	-	-	
S_4	-	-	-	

Tablo 6.11 (Devamı) Bist30 ve Hisse Bazında Duygu Skorlarının Korelasyon İlişkisi

	Günlük	Haftalık	Aylık	En iyi Polarite Skoru
EKGYO				
S_1	0,27	-	-	
S_2	0,29	-	-	
S_3	-	-	-	
S_4	-	-	-	
ENJSA				
S_1	0,18	-	-	
S_2	0,23	-	-	
S_3	-	-	-	
S_4	-	-	-	
EREGL				
S_1	0,23	-	**0,76**	
S_2	0,25	-	**0,75**	S_4
S_3	-	-	**-**	
S_4	-	0,40	**0,78**	
FROTO				
S_1	-	-	-	
S_2	-	-	-	
S_3	-	-	-	
S_4	-	-	-	
GARAN				
S_1	0,30	0,31	**0,61**	
S_2	0,32	0,32	**0,61**	S_1 - S_2
S_3	0,24	0,42	-	
S_4	-	-	-	
HALKB				
S_1	0,27	0,31	-	
S_2	0,30	0,33	-	S_3
S_3	0,26	0,35	**-**	
S_4	-	-	-	

Tablo 6.11 (Devamı) Bist30 ve Hisse Bazında Duygu Skorlarının Korelasyon İlişkisi

	Günlük	Haftalık	Aylık	En iyi Polarite Skoru
ISCTR				
S_1	0,37	-	**0,86**	
S_2	0,41	0,34	**0,92**	S_2
S_3	-	-	-	
S_4	-	-	-	
KCHOL				
S_1	0,15	-	-	
S_2	0,18	-	-	
S_3	-	-	-	
S_4	-	-	-	
KOZAA				
S_1	0,29	0,36	-	
S_2	0,31	0,34	-	
S_3	0,25	0,46	-	
S_4	-	0,36	-	
KOZAL				
S_1	0,32	-	-	
S_2	0,34	-	-	
S_3	-	0,33	-	
S_4	-	-	-	
KRDMD				
S_1	0,36	0,37	-	
S_2	0,40	0,38	-	
S_3	0,26	0,43	-	
S_4	-	-	-	
PETKM				
S_1	0,37	0,36	-	
S_2	0,40	0,36	-	
S_3	0,26	-	-	
S_4	0,16	0,28	-	

Tablo 6.11 (Devamı) Bist30 ve Hisse Bazında Duygu Skorlarının Korelasyon İlişkisi

	Günlük	Haftalık	Aylık	En iyi Polarite Skoru
PGSUS				
S_1	0,24	-	-	
S_2	0,27	-	-	
S_3	-	-	-	
S_4	-	-	-	
SAHOL				
S_1	-	-	-	
S_2	0,17	-	-	S_3
S_3	-	-	**0,63**	
S_4	-	-	-	
SISE				
S_1	0,25	-	-	
S_2	0,28	-	-	
S_3	-	-	-	
S_4	-	-	-	
SODA				
S_1	0,26	-	-	
S_2	0,30	-	-	
S_3	-	-	-	
S_4	-	-	-	
TAVHL				
S_1	0,15	-	-	
S_2	0,24	-	**0,64**	S_2
S_3	-	-	-	
S_4	-	-	-	
TCELL				
S_1	0,31	-	-	
S_2	0,35	-	-	
S_3	-	-	-	
S_4	-	-	-	

Tablo 6.11 (Devamı) Bist30 ve Hisse Bazında Duygu Skorlarının Korelasyon İlişkisi

	Günlük	Haftalık	Aylık	En iyi Polarite Skoru
THYAO				
S_1	0,29	0,38	-	
S_2	0,34	0,41	-	
S_3	0,23	-	-	
S_4	0,14	-	-	
TKFEN				
S_1	0,30	0,32	-	
S_2	0,35	0,41	-	
S_3	0,19	0,42	-	
S_4	0,19	-	-	
TOASO				
S_1	0,20	-	-	
S_2	0,21	0,37	-	S_3
S_3	0,17	0,41	**0,82**	
S_4	-	-	-	
TTKOM				
S_1	0,35	-	**0,64**	
S_2	0,37	-	**0,64**	S_1 - S_2
S_3	0,15	-	**-**	
S_4	-	-	**0,63**	
TUPRS				
S_1	-	-	-	
S_2	-	-	-	
S_3	-	-	-	
S_4	-	-	-	
VAKBN				
S_1	0,38	0,46	-	
S_2	0,41	0,50	-	S_3
S_3	-	**0,57**	-	
S_4	-	-	-	

Tablo 6.11 (Devamı) Bist30 ve Hisse Bazında Duygu Skorlarının Korelasyon İlişkisi

	Günlük	Haftalık	Aylık	En iyi Polarite Skoru
YKBNK				
S_1	0,23	0,32	-	
S_2	0,30	0,31	-	
S_3	-	-	-	
S_4	0,14	-	-	

Tablo 6.11'de normal dağılıma uyan seriler için Pearson, uymayanlar için Spearman Sıra Korelasyonu hesaplanmış olup, sadece ilişkinin anlamlı olduğu değerler gösterilmiştir. Koyu renkli değerler duygu skorları arasında yüksek korelasyonun olduğunu ifade etmektedir. Tablo incelendiğinde Bist30 bazında S_4 duygu sokurunun, hisse bazında ise S_2 duygu skorunun diğerlerine göre daha kuvvetli ilişkili olduğu, aynı zamanda aylık verilerde diğer periyotlara göre tüm duygu skorları ile Bist30 endeksi arasında daha kuvvetli ilişkinin olduğu görülmektedir. Aylık periyotta duygu skorlarıyla; GARAN, SAHOL, TAVHL, TTKOM, VAKBN hisselerinin orta kuvvetli, Bist30, EREGL, TOASO hisselerinin kuvvetli, ISTCR'nin ise çok kuvvetli ilişkisi vardır. FROTO, TUPRS hisselerinin ile duygu skorları arasında anlamlı bir ilişki olmadığı tespit edilmiştir. FROTO Bist30 listesine sonradan dahil olmasından dolayı ilişkinin var olmamasına sebep olmuş olabileceği düşünülebilir.

Tablo 6.12 Orta Kuvvette İlişkili Olan Hisseler ile Duygu Skorları

	Günlük	Haftalık	Aylık
BİST30	S_1 - S_2	S_3	S_3
AKBNK			S_3
ARCLK			S_2
GARAN			S_1 - S_2
SAHOL			S_3
TAVHL			S_2
TTKOM			S_1 - S_2 - S_4
VAKBN		S_2 - S_3	

Tablo 6.12'deki hisselerin çoğunlukla S_2 skoru ile orta şiddetli ilişkisinin olduğunu söyleyebiliriz.

Tablo 6.13 Kuvvetli İlişki Bulunan Aylar

	S_1	S_2	S_4
2018-08	0,70	0,71	
2019-02			0,72

Günlük olarak incelendiğinde 2018 yılının ağustos ayında S_1 ve S_2 duygu skorları 2019 şubat yında ise S_4 duygu skoru ile Bist30 arasında kuvvetli bir korelasyon ilişkisi olduğu görülmektedir.

6.10. Granger Nedensellik Analiz Sonuçları

Null Hypothesis: DBIST_ORAN has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=1)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-5.590152	0.0023
Test critical values:	1% level	-4.420595	
	5% level	-3.259808	
	10% level	-2.771129	

*MacKinnon (1996) one-sided p-values.
Warning: Probabilities and critical values calculated for 20 observations
and may not be accurate for a sample size of 9

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(DBIST_ORAN)
Method: Least Squares

Sample (adjusted): 3 11
Included observations: 9 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DBIST_ORAN(-1)	-1.667360	0.298267	-5.590152	0.0008
C	-0.004053	0.037883	-0.106985	0.9178

R-squared	0.816992	Mean dependent var	0.007322
Adjusted R-squared	0.790849	S.D. dependent var	0.248143
S.E. of regression	0.113484	Akaike info criterion	-1.321188
Sum squared resid	0.090150	Schwarz criterion	-1.277360
Log likelihood	7.945344	Hannan-Quinn criter.	-1.415768
F-statistic	31.24980	Durbin-Watson stat	2.310497
Prob(F-statistic)	0.000824		

Şekil 6.11 1.Farkı Alınmış Bist30 Oranı Durağanlık Test Sonuçları

Bist30 aylık oranı durağan olmadığı için birinci farkı alınarak Dickey-Fuller testi yapıldığında Şekil 6.11 de görüldüğü üzere durağan hale getirilir. Bu kontrole Prob. Değerinin 0.05'ten küçük ve t-Statistic değerinin de test kriter değerlerinden küçük olana kadar devam edilir.

VAR Lag Order Selection Criteria
Endogenous variables: S2 BIST_ORAN
Exogenous variables: C

Sample: 1 242
Included observations: 237

Lag	LogL	LR	FPE	AIC	SC	HQ
0	595.0226	NA	2.30e-05	-5.004410	-4.975144	-4.992614
1	618.0113	45.39537	1.96e-05	-5.164652	-5.076853*	-5.129264
2	625.8946	15.43395*	1.90e-05*	-5.197423*	-5.051091	-5.138442*
3	627.4995	3.114993	1.93e-05	-5.177211	-4.972346	-5.094637
4	630.1895	5.175674	1.96e-05	-5.166156	-4.902759	-5.059990
5	633.7149	6.723698	1.96e-05	-5.162151	-4.840222	-5.032393

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Şekil 6.12 S_2 ve Bist Oranının Gecikme Sayısının Belirlenmesi.

VO modeliyle Şekil 6.12'deki S_2 ve Bist30 oranı için gecikme uzunluklarının en çok yıldıza sahip satırdan anlaşıldığı üzere 2 olduğu görülmektedir. Bu işlemler tüm ilişkileri incelenen seriler için ayrı ayrı belirlenir.

Tablo 6.14 Duygu Skorları ile Bist30 Granger Nedensellik Yönleri

	$S_1 - BO$	**$S_2 - BO$**	**$S_3 - BO$**	**$S_4 - BO$**
Günlük	$S_1 \leftrightarrow BO$	$S_2 \leftrightarrow BO$	$S_3 \leftarrow BO$	$S_4 \leftarrow BO$
Haftalık	$S_1 \leftrightarrow BO$	$S_2 \leftrightarrow BO$	$S_3 \rightarrow BO$	$S_4 \leftarrow BO$
Aylık	-	-	-	-

Tablo 6.14 incelendiğinde; günlük ve haftalık zaman dilimlerinde S_1 ve S_2 duygu skorları ile Bist30 oranı arasında çift yönlü nedensellik ilişkisi bulunmaktadır. Haftalık

periyotta Bist30 oranlarındaki değişimin nedeninin S_3 olduğu dikkat çekmektedir. Diğerlerinde ise Bist30 değişiminin S_3 ve S_4 duygu skorlarını etkilediği ortaya çıkmıştır. Aylık için korelasyon ilişkisinin anlamlı olmasına rağmen skorlar arasında nedensellik ilişkisi bulunamamıştır. Buradan Twitter yorumlarının Bist30 değişimlerinden uzun vadede birlikte aynı yönde hareket etkilerini ama birbirlerinden etkilenmediklerini söyleyebiliriz. Kısa vadede ise en az bir yönde birbirlerinden etkilenmektedirler.

VAR Granger Causality/Block Exogeneity Wald Tests

Dependent variable: S1

Excluded	Chi-sq	df	Prob.
D1_USD	1.910360	3	0.5912
BIST_ORAN	15.95026	3	0.0012
All	18.14571	6	0.0059

Dependent variable: D1_USD

Excluded	Chi-sq	df	Prob.
S1	8.010413	3	0.0458
BIST_ORAN	2.501661	3	0.4750
All	15.73352	6	0.0153

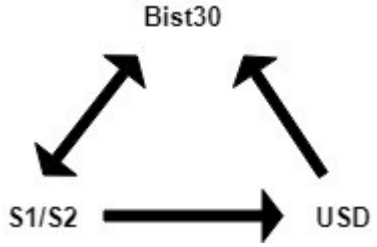
Dependent variable: BIST_ORAN

Excluded	Chi-sq	df	Prob.
S1	8.240727	3	0.0413
D1_USD	9.780691	3	0.0205
All	17.76251	6	0.0069

Şekil 6.13 Günlük S_1, Bist30 ve Usd Arasındaki Nedensellik İlişkisi

Dolar kurunun hareketli olduğu dönemlerde Twitter yorumlarına konu olması Bist30 değerlerini etkilediği öngörülmektedir. Şekil 6.13 teki VO modeliyle S_1 duygu skoru, Bist30 ve USD arasında nedensellik ilişkileri çıkarılmıştır. Prob. Değeri 0.05'ten küçük olan bağıntılarda H_0 hipotezi reddedilir. Bu durumda, S_1 duygu skorunun

bağımlı değişken olduğu 1. bağıntıda; S_1 Bist30'un, USD bağımlı değişken olduğu 2.bağıntıda; S_1 USD'nin, Bist30'un bağımlı değişken olduğu 3. bağıntıda; hem S_1 hem de USD Bist30'un Granger nedeni olduğu sonucuna varılmıştır. Şekil 6.14'te nedensellik yönleri gösterilmiştir.



Şekil 6.14 S_1-S_2 , Bist30 ve USD Arasındaki Nedensellik Yönleri

S_1 ve duygu skorları ile Bist30 arasında çift yönlü etkilenme söz konusuyken diğerlerinde tek yönlü ilişki bulunmuştur. USD kurunun duygu skorundan etkileniyor olması da dikkat çekmektedir. S_3 ve S_4 duygu skorları ile USD arasında anlamlı bir ilişki bulunmamaktadır. Sonuç olarak günlük periyotta Bist30 oranındaki değişim USD değişiminden ve dolaylı olarak duygu skorlarından etkilendiğini söyleyebiliriz.

Tablo 6.15 Hisse Bazında Granger Nedensellik Yönleri

	$\mathbf{S_2 - HO}$	$\mathbf{TS - \lvert HO \rvert}$
AKBNK	$S_2 \leftrightarrow HO$	-
ARCLK	$S_2 \leftrightarrow HO$	$TS \leftarrow HO$
ASELS	$S_2 \leftarrow HO$	-
BIMAS	-	$TS \leftarrow HO$
DOHOL	-	-
EKGYO	$S_2 \leftarrow HO$	-
ENJSA	-	-
EREGL	$S_2 \leftarrow HO$	$TS \leftarrow HO$
FROTO	$S_2 \leftarrow HO$	-
GARAN	-	-
ISCTR	-	-
KRDMD	-	$TS \leftarrow HO$
KCHOL	-	-
KOZAL	-	-
KOZAA	-	$TS \rightarrow HO$
PGSUS	$S_2 \leftarrow HO$	$TS \leftarrow HO$
PETKM	-	$TS \leftarrow HO$

Tablo 6.15 Hisse Bazında Granger Nedensellik Yönleri (Devamı)

	$S_2 - HO$	$TS - \lvert HO \rvert$
SAHOL	-	-
SODA	-	$TS \leftarrow HO$
SISE	-	-
HALKB	-	-
TAVHL	-	-
TKFEN	-	-
TOASO	-	-
TCELL	-	-
TUPRS	$S_2 \leftarrow HO$	$TS \leftarrow HO$
THYAO	-	$TS \leftarrow HO$
TTKOM	-	-
VAKBN	$S_2 \rightarrow HO$	-
YKBNK	-	$TS \rightarrow HO$
Toplam	9	11

TS: Tweet sayısı, HO: Hisse oran

Tablo 6.15'te günlük periyotta korelasyon ilişkisi diğerlerine göre yüksek çıkan S_2 duygu skoru için hisse oranı ve tweet sayılarının hisse oranlarının nedensellik ilişkisi araştırılmıştır. Bist30 firmaları içerisinden hisse oranları arasında 9 hissede nedensellik ilişkisi bulunmuştur. AKBNK ve ARCLK hisselerinde çift yönlü ilişkisi bulunurken, VAKBN haricindeki hisselerin S_2 skorlarındaki değişimin hisse oranlarındaki değişimden kaynaklandığı görülmektedir. Hisse başına düşen veri setindeki toplam tweet sayıları ile hisse oran değişimleri incelenmiştir. 11 tanesinde nedensellik ilişkisi tespit edilmiştir. KOZAA ve YKBNK hisse oranındaki değişimin nedeni tweet sayısı olurken, diğerlerinde tweet sayılarının hisse oranlarındaki değişimden kaynaklandığı anlaşılmaktadır.

YKBNK, THYAO, SODA, PETKM, KRDMR, KOZAA, BIMAS gibi hisselerin S_2 skoru ile nedensellik ilişkisi bulunmazken, tweet sayıları ile hisse oranları arasında nedensellik ilişkisi olduğu dikkat çekmektedir. Bu duruma diğer duygu skoru neden olmuş olabilir. Örneğin PETKM hissesi üzerinde S_3 skoru üzerinden nedensellik analizi yapıldığında hisse oranından S_3'e doğru tek yönlü ilişki olduğu bulunmuştur. O halde tweet sayısındaki değişimi destekleyen S_3 skorudur.

Tablo 6.16 Önemli Olayların Olduğu Aylardaki Granger Nedensellik Yönleri

	Granger Nedenseliğinin Yönü		
2018 Haziran (Başkanlık seçimi) 3.307 tweet	Bist Oran	→	S_2 / S_3
2018 Temmuz (Kabine değişimi) 6.925 tweet	USD	→	Bist Oran
2018 Ağustos (Trump konuşması) 3.998 tweet	USD	→	$S_2 / S_3 / S_4$
2019 Mart-Nisan (Yerel seçimler) 8.551 tweet	USD	→	Bist Oran
	USD	→	S_3

Tablo 6.16'da belirlenen önemli olayların olduğu günlerde; Bist30, duygu skorları ve USD kuru arasında Granger Nedenselliği araştırıldığında, 2018 yılının haziran ayında Bist30 değişiminin twitter yorumlarını etkilediği, diğer olaylarda ise doların tweet yorumlarına ve bist oranına etkisi olduğu görülmektedir.

7. TARTIŞMA ve GELECEK ÇALIŞMALAR

Çağımızda sosyal medyanın yaygın kullanılıyor olması hayatımızın kültürel, ekonomi gibi birçok alanını da etkilemektedir. İnsanların duygu ve düşüncelerini açıkca paylaşmaları ve birbirlerinin yorumlarından etkileniyor olmaları markaların satışlarını olumlu veya olumsuz yönde etkileyebilme gücüne sahiptir. Bu çalışmada sosyal medya ve borsa arasında böyle bir etkinin olup olmadığı veya ne oranda etkilendiği araştırılmak istenmiştir.

Mikroblog sitesi olan Twitter kullanılarak 1 yıllık süre boyunca insanların günlük Bist30 endekslerindeki hisseler için paylaştıkları yorumlar incelenmiş ve bu yorumların duygu durumlarına göre finans alanı ile ilgili kişilerce pozitif, negatif ve nötr olmak üzere 3 kategoride sınıflandırılmıştır. Makine öğrenmesi teknikleriyle bu veriler eğitilmiş, en iyi sınıflandırma yöntemi belirlendikten sonra test verileri tahmin edilmiştir. Birçok çalışmada DVM, KNN gibi algoritmalar en iyi sonuca sahipken, bu çalışmada LR diğerlerine göre daha iyi sonuç vermiştir. LR ile tahmin yapıldıktan sonra pozitif ve negatif tweetler kullanılarak 4 çeşit duygu skoru elde edilmiş ve bu duygu skorları ile Bist30 endeksinin açılış/kapanış fiyatları ile endekse bağlı hisselerin fiyatları arasındaki korelasyon ilişkisi incelenmiştir. Farklı duygu skorlarının kullanılması Bist30 endeks değerleri ile korelasyonu açısından uyumluluğu sağlayarak çeşitliliği arttırmıştır. Aynı zamanda verilerin bir zaman serisi olduğu düşünülürse geçmiş değerlerinden de etkilenme durumuna karşı Granger Nedensellik Analizi yapılmıştır. Korelasyon ilişkisi ve Granger Nedensellik sonuçları (Ranco vd.,2015) çalışmasına benzer sonuçlar ortaya koymuştur. Önceki çalışmalara yakın sonuçlar bulunmuş; duygu skorları yani sosyal medya yorumlarının Bist30 endeksinin çoğu zaman tek, bazen çift yönlü etkilendiği, hisse bazında aralarında zayıf korelasyon olmasına rağmen birbirlerinin geçmiş değerlerinden etkilenebildiği görülmüştür. Aylık incelemede Granger Nedenselliği için daha fazla veriye ihtiyaç duyulmaktadır. Borsanın günlük olarak bir çok faktörden etkilendiği düşünülürse kısa vadeli (günlük) nedensellik ilişkisinin daha etkili olduğu mantıklı görünmektedir. Korelasyon ilişkisinde ise günlük dalgalanma ilişkinin kuvvetini değiştirmektedir. Bu bağlamda aylık dönemde daha istikrarlı ve kuvvetli bir ilişki gözlemlenmiştir.

Sonraki çalışmalarda hisse bazında veya bir marka grubu üzerinden araştırma yapılabilir. Sosyal medyanın yanı sıra, hisse fiyatlarını etkilediği düşünülen başka

parametreler de eklenerek çoklu regresyon ve zaman serileri analizlerinden AR modelleri üzerinde çalışılabilir.

Sonuç itibariyle istatistik ve bilgisayar bilimleri sosyal medya gibi büyük veri kaynaklarını kullanarak ihtiyaç duyulan tüm bilgi ve çıkarımları insanlara sunabilmektedir. Bu imkanlardan faydanılarak istenilen esneklikte yeni algoritmalar geliştirilebilir.

KAYNAKÇA

Akgül, E. S., Ertano, C., & Diri, B. (2016). Twitter verileri ile duygu analizi.

Aytuğ, O. N. A. N. (2017). Twitter Mesajları Üzerinde Makine Öğrenmesi Yöntemlerine Dayalı Duygu Analizi. *Yönetim Bilişim Sistemleri Dergisi*, 3(2), 1-14.

Baykara, M., Gürtürk, U., & Teknolojileri, N. B. (2017). "Sosyal Medya Paylaşımlarının Duygu Analizi Yöntemiyle Sınıflandırılması", *2. International Conferance on Computer Science and Engineering*, 911-916.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1), 1-8.

Can, Ü., & Alataş, B. (2017). Duygu Analizi ve Fikir Madenciliği Algoritmalarının İncelenmesi. *International Journal of Pure and Applied Sciences*, 3(1), 75-111.

Deng, S., Huang, Z. J., Sinha, A. P., & Zhao, H. (2018). The Interaction between Microblog Sentiment and Stock Return: An Empirical Examination. *MIS quarterly*, 42(3), 895-918.

Eliaçik, A. B., & Erdogan, N. (2015). "Mikro Bloglardaki Finans Toplulukları için Kullanıcı Ağırlıklandırılmış Duygu Analizi Yöntemi". *In UYMS.*

Gunduz, H., & Cataltepe, Z. (2015). Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection. *Expert Systems with Applications*, 42(22), 9001-9011.

Gupta, A., Pruthi, J., & Sahu, N. (2017). Sentiment Analysis of Tweets using Machine Learning Approach. Ankita Gupta et al, *International Journal of Computer Science and Mobile Computing*, 6(4), 444-458.

Huq, M. R., Ali, A., & Rahman, A. (2017). Sentiment analysis on Twitter data using KNN and SVM. *Int J Adv Comput Sci Appl*, 8(6), 19-25.

https://towardsdatascience.com/another-twitter-sentiment-analysis-with-python-part-6-doc2vec-603f11832504 Erişim Tarihi:18 Ocak 2018

Kalaycı, Ş.(2010). *SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri*. Ankara: Asil.

Kalyani, J., Bharathi, P., & Jyothi, P. (2016). Stock trend prediction using news sentiment analysis. *International Journal of Computer Science and Information Technology*. 8 (3), 67-76.

Kaynar, O., Görmez, Y., Yıldız, M., & Albayrak, A. (2016). "Makine öğrenmesi yöntemleri ile Duygu Analizi". *In International Artificial Intelligence and Data Processing Symposium* (IDAP'16), September (pp. 17-18).

Kılınç, D. (2016). The Effect of Ensemble Learning Models on Turkish Text Classification. *Celal Bayar Üniversitesi Fen Bilimleri Dergisi*, 12(2).

Kim, Y., Jeong, S. R., & Ghani, I. (2014). Text opinion mining to analyze news for stock market prediction. *Int. J. Advance. Soft Comput. Appl*, 6(1), 2074-8523.

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2), 689-702.

Le, Q., & Mikolov, T. (2014, January). "Distributed representations of sentences and documents". *In International conference on machine learning* (pp. 1188-1196).

Liu, J., Gong, Q., Wang, T., Zhu, W., & Li, Q. (2013). Looking for Gold in the Sands: Stock Prediction Using Financial News and Social Media. *In PACIS* (p. 26).

Mittal, A., & Goel, A. (2012). Stock prediction using twitter sentiment analysis. Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf), 15.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24), 9603-9611.

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, 73, 125-144.

Onan, A. (2017). Twitter Mesajları Üzerinde Makine Öğrenmesi Yöntemlerine Dayalı Duygu Analizi. *Yönetim Bilişim Sistemleri Dergisi*, 3(2), 1-14.

Öner, H., İçellioğlu, C. Ş., & Öner, S. (2018). Volatilite Endeksi (VIX) ile Gelişmekte Olan Ülke Hisse Senedi Piyasası Endeksleri Arasındaki Engel-Granger Eş-Bütünleşme ve Granger Nedensellik Analizi. *Finansal Araştırmalar ve Çalışmalar Dergisi*, 10(18), 110-124.

Özkan, Y.(2016). *Veri Madenciliği Yöntemleri*. İstanbul: PapatyaBilim.

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016, October). "Sentiment analysis of Twitter data for predicting stock market movements." *In 2016 international conference on signal processing, communication, power and embedded system* (SCOPES) (pp. 1345-1350). IEEE.

Pekkaya, M., & Bayramoğlu, M. F. (2008). Hisse senedi fiyatları ve döviz kuru arasındaki nedensellik ilişkisi: YTL/USD, İMKB 100 ve S&P 500 Üzerine Bir Uygulama. *Muhasebe ve Finansman Dergisi*, (38), 163-176.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. *PloS one*, 10(9), e0138441.

Rout, J. K., Choo, K. K. R., Dash, A. K., Bakshi, S., Jena, S. K., & Williams, K. L. (2018). A model for sentiment and emotion analysis of unstructured social media text. *Electronic Commerce Research*, 18(1), 181-199.

Ruan, Y., Durresi, A., & Alfantoukh, L. (2018). Using twitter trust network for stock market analysis. *Knowledge-Based Systems*, 145, 207-218.

Snir, I. (2017). Using Text Mining and Machine Learning to Predict the Impact of Quarterly Financial Results on Next Day Stock Performance.

Tao, J., Chen, L., & Lee, C. M. (2016). "DNN Online with iVectors Acoustic Modeling and Doc2Vec Distributed Representations for Improving Automated Speech Scoring". *In INTERSPEECH* (pp. 3117-3121).

Yang, S. Y., Mo, S. Y. K., & Liu, A. (2015). Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance*, 15(10), 1637-1656.

Yıldırım, M., Yüksel, C. A.(2017). Sosyal Medya ile Hisse Senedi Fiyatının Günlük Hareket Yönü Arasındaki İlişkinin İncelenmesi: *Duygu Analizi Uygulaması. Uluslararası İktisadi ve İdari İncelemeler Dergisi*, 33-44.

Weng, B., Lu, L., Wang, X., Megahed, F. M., & Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. *Expert Systems with Applications*, 112, 258-273.

Zhang, L., Dong, W., & Mu, X. (2018). Analysing the features of negative sentiment tweets. *The Electronic Library*, 36(5), 782-799.

ÖZGEÇMİŞ

1983 İzmir doğumluyum. Üsküdar Lisesi'nden (Yabancı Dil Ağırlıklı Lise) mezun olduktan sonra Süleyman Demirel Üniversitesi Bilgisayar Programcılığı Bölümü'nü bitirdim. Sonrasında lisans eğitimimi Mimar Sinan Güzel Sanatlar Üniversitesi İstatistik Bölümü'de tamamladım.

Yazılım üzerine başladığım iş hayatıma farklı sektörlerde iş zekası danışmanı olarak devam ettim. Son 4 yıldır ise finans sektöründe olup; şu anda uluslararası bir bankanın IT departmanında Kıdemli İş Zekası ve Veri Ambarı Uzmanı olarak çalışmaktayım.