T.C

MARMARA ÜNİVERSİTESİ

SOSYAL BİLİMLER ENSTİTÜSÜ

İŞLETME ANABİLİM DALI

İŞ ANALİTİĞİ BİLİM DALI

DUYARLILIĞIN HİSSE SENEDİ TAHMİNİ ÜZERİNDEKİ ETKİSİNİ DEĞERLENDİRME

Yüksek Lisans Tezi

Demet Erdoğan

İstanbul, 2022

T.C

MARMARA ÜNİVERSİTESİ

SOSYAL BİLİMLER ENSTİTÜSÜ

İŞLETME ANABİLİM DALI

İŞ ANALİTİĞİ BİLİM DALI

DUYARLILIĞIN HİSSE SENEDİ TAHMİNİ ÜZERİNDEKİ ETKİSİNİ DEĞERLENDİRME

Yüksek Lisans Tezi

Demet Erdoğan

Danışman: DR. ÖĞR. ÜYESİ NACİYE TUBA YILMAZ SOYDAN

İstanbul, 2022

GENEL BİLGİLER

İsim ve Soyadı :Demet Erdoğan

Anabilim Dalı :İşletme

Bilim Dalı :İş Analitiği

Tez Danışmanı :Naciye Tuba Yılmaz Soydan

Tez Türü ve Tarihi :Yüksek Lisans – Ocak 2022

Anahtar Kelimeler :Doğal dil işleme, duygu analizi, borsa hisse tahmini, Bert model, Neural Prophet

ÖZET

DUYARLILIĞIN HİSSE SENEDİ TAHMİNİ ÜZERİNDEKİ ETKİSİNİ DEĞERLENDİRME

Teknolojinin gelişmesiyle birlikte akıllı cihaz kullanımdaki artış ile doğru orantılı olarak sosyal medya kullanımındaki artış, kullanım şekli ve sosyal medyanın etkisi araştırma konusu olarak oldukça dikkat çekici olmuştur. Sosyal medyadaki bir hesaba fotoğraf eklemek, bir parti hakkında fikir yayınlamak vb. gibi aktivitelerin yapıldığı anda birçok takipçiye hatta o hesabı takip etmeyen kişilerin bile hemen duyabilmesine olanak sağlamıştır. Bu çalışma ile sosyal medyadaki paylaşımın gerçekten de etkisinin olup olmadığını ölçmek adına, Twitter'daki yorumlar ile Tesla hisse senedi arasındaki ilişkiye bakılmak istenmiştir. Twitter'da paylaşılan #TSLA kelimeli yorumların belli zaman aralığında alınarak BERT modeli ile duygu analizine bakılmış ve sonucunda 96% başarı elde edilmiş ve ertesi günkü TSLA market değerinin tahmin edilmesinde ise 64% başarı elde edilmiş ve sonuç olarak sosyal medyanın gerçekten de insanların düşüncelerinde etkisi olduğu kanaatine varılmıştır.

GENERAL KNOWLEDGE

Name and Surname :Demet Erdoğan

Field :Management

Programme :Business Analytics

Supervisor :Naciye Tuba Yılmaz Soydan

Degree Awarded and Date :Master– January 2022

Keywords :Natural language processing, sentiment analysis, stock market forecasting, NLP, BERT, Neural Prophet

ABSTRACT

ASSESSING THE IMPACT OF THE SENTIMENT ON STOCKS PREDICTION

With the development of technology, the increase in the use of social media in direct proportion to the increase in the use of smart devices, the way of use and the effect of social media have been quite remarkable as research subjects. Adding a photo to an account on social media, posting an opinion about a party, and similar activities allowed many followers, even those who do not follow that account, to hear immediately. In this study, it was aimed to look at the relationship between Twitter comments and Tesla stock in order to measure whether the sharing on social media really has an effect. The comments with the words #TSLA shared on Twitter were taken at a certain time interval and the sentiment analysis was checked with the BERT model, and as a result, 96% success was achieved and 64% success was achieved in estimating the TSLA market value of the next day. As a result, it has been concluded that social media really has an effect on people's thoughts.

ÖNSÖZ

Bu tez çalışmasının yürütülmesi sırasında destek ve katkılarından dolayı danışmanım Dr. Öğr. Üyesi Naciye Tuba Yılmaz Soydan'a, bu alanda yetişmemi sağlayan, üzerimde çok emeği olan ve her zaman destekleyen Prof. Dr. Şadi Evren Şeker'e teşekkür ederim.

Tez yazım aşamasında hep destekleyen arkadaşlarım Muhammad Ali Haider ve Volkan Taşköprü'ye ve,

Hayatımın her alanında olduğu gibi tez çalışmamı hazırlarken de destek verdiği, her zaman uygun koşulları hazırladığı ve hep motive ettiği için canım kardeşim Zeynep Buket Erdoğan'a, hiçbir zaman desteğini, motivasyonunu esirgemeyen, benim için her zaman en iyisini yapan canım abim Doğukan Erdoğan'a, sadece bu çalışma sürecinde değil tüm hayatım boyunca beni cesaretlendiren ve kendime inanmamı sağlayan, aldığım kararları her zaman destekleyen canım annem Şadiye Erdoğan'a ve canım babam Hüseyin Erdoğan'a tüm kalbimle teşekkür ederim.

Demet Erdoğan

İstanbul, Ocak 2022

İÇİNDEKİLER

KISALTMALAR iv

TABLOLAR DİZİNİ v

ŞEKİLLER DİZİNİ vi

GRAFİKLER DİZİNİ vii

1. GİRİŞ 1

2. SOSYAL MEDYA 1

2.1 Sosyal Medya Kavramı 1

2.2 Sosyal Medya Gelişimi 2

2.3 Sosyal Medyanın Özellikleri 3

2.4 Twitter Nedir? 5

2.5 Twitterin Analiz İçin Kullanılma Nedeni 6

2.6 Diğer Sosyal Medya Ağları 6

2.6.1 Instagram 6

2.6.2 Facebook 7

2.6.3 Linkedin 7

2.6.4 Youtube 8

2.6.5 Tiktok 8

3. YÖNTEM 9

3.1. Doğal Dil İşleme 9

3.2 Doğal Dil İşlemenin Uygulama Alanları 10

3.2.1 Soru Yanıtlama 10

3.2.2 Dil Tercüme 11

3.2.3 Bilginin Bulunması ve Getirilmesi 11
3.2.4 Kelime İşleme 12
3.2.5 Veritabanı Arayüzü 12
3.2.6 Arama Motorları 12
3.2.7 Metin Ön İşleme 13
3.2.8 Argüman Birleştirme 13
3.2.9 Konuşmadan Yazıya – Yazıdan Konuşmaya Dönüşüm 13
3.2.10 Hedeflenmiş Reklamcılık 14
3.2.11 Otomatik Arama Düzeltme ve Tamamlama 14

3.3 Makine Öğrenmesi **14**
3.3.1 Makine Öğrenmesi Kullanım Alanları 14
3.3.1.1 Sağlık 15
3.3.1.2 Siber Güvenlik 15
3.3.1.3 Lojistik ve Üretim 15
3.3.1.4 Satış ve Pazarlama 16
3.3.2 Makine Öğrenmesinde Genel Yöntemler 16
3.3.2.1 Gözetimli Öğrenme 17
3.3.2.1.1 Sınıflandırma 17
3.3.2.1.2 Regresyon 18
3.3.2.2 Gözetimsiz Öğrenme 19
3.3.2.2.1 Kümeleme 19

3.4 Derin Öğrenmede Genel Yöntemler **20**
3.4.1 Evrişimli Sinir Ağları (CNN) 21
3.4.2 Tekrarlayan Sinir Ağları (RNN) 23
3.4.3 Uzun Kısa Süreli Bellek ağları (LSTM) 24
3.4.4 Transformer 26
3.4.1 Electra 28
3.4.2 DistilBert 28
3.4.3 A Lite Bert (Albert) 28
3.4.5 BERT 29
3.4.6 Neural Prophet 35

4. DUYARLILIĞIN HİSSE SENEDİ ÜZERİNDEKİ ETKİSİNİ DEĞERLENDİRME 41
4.1 Method Yönetimi 41
4.1.1 KDD 41

4.1.2 CRISP-DM 42
4.1.3 SEMMA 43

4.2 Literatür Taraması **43**
4.2.1 Duygu Analizi 43
4.2.2 BERT 44
4.2.3 Neural Prophet 45

4.3. Problemi Tanıma (Business Understanding) **45**

4.4. Veriyi Tanıma (Data Understanding) **46**
4.4.1 Twitter 46
4.4.2 Yahoo Finans 48
4.4.3 Etiketlenmiş Duygu Analizli Veri (Labeled Data) 48

4.5. Veri Hazırlama (Data Preparation) **51**
4.5.1 Twitter Verisi Ön İşlemleri 51
4.5.2 Veri Dengesi İşlemleri 51

4.6. Model Oluşturma (Modeling) **53**
4.6.1 BERT 53
4.6.2 Neural Prophet 62

4.7. Değerlendirme (Evaluation) **65**

KAYNAKLAR 70

EKLER 75

KISALTMALAR

NLP	Natural Language Processing
DDİ	Doğal Dil İşleme
HCI	Human Computer Interaction
YSA	Yapay Sinir Ağları
LSTM	Long- Short Term Memory
CNN	Convolutional Neural Networks
BOW	Bag of Words
RNN	Recurrent Neural Network
YSA	Yinelemeli Sinir Ağları
MLM	Masked Language Modeling
MDM	Maskelenmiş Dil Modeli
NSP	Next Sentence Prediction
SCT	Sonraki Cümle Tahmini
MLP	Multi Layer Perceptron
RNN	Tekrarlayan Sinir Ağları
SVM	Support Vector Machine
KNN	K-Nearest Neighbor
VBK	Veri Tabanlarında Bilgi Keşfi
KDD	Knowledge Discovery From Database
Crisp-DM	Cross Industry Standard Process for Data Mining

TABLOLAR DİZİNİ

***Tablo 1**: Web 1.0 Web 2.0 ve Web 3.0 karşılaştırması* 3

***Tablo 2**: Tesla Tweets* 47

***Tablo 3:** Yahoo Finanse değerler* 48

***Tablo 4:** Etiketlenmiş yorumlar* 49

***Tablo 5:** Hisse senedi yorumları & duygu analizleri* 50

***Tablo 6:** Film ve hisse senedi yorumları & duygu analizleri* 50

***Tablo 7:** Dil çevirimi* 52

***Tablo 8**: Eksik değer şeması* 57

***Tablo 9:** Veri bölütleme sonucu* 58

***Tablo 10:** Tensör ve dikkat maskesi çıktısı* 59

***Tablo 11:** Bert işlem detayı* 60

***Tablo 12:** Bert model ile Tweet sentiment analizi* 61

***Tablo 13:** Tsla market değeri ve sentiment değerleri* 62

***Tablo 14:** Tsla market tahmini için hesaplanan yeni ve alınan kolonlar* 62

***Tablo 15:** Neural Prophet sonucu gelecek 50 günkü market kapanış tahmin değerleri* 64

***Tablo 16:** Gerçekleşen ve tahmin edilemeyen değer tablosu* 65

***Tablo 17:** Neural Model başarı tablosu* 66

***Tablo 18:** Confusion matrix teorik* 66

***Tablo 19:** Neural başarı değerleri* 67

ŞEKİLLER DİZİNİ

***Şekil 1**: Türkçe ve Fransıca dillerinin sözdizimsel karşılaştırması* *10*

***Şekil 2**: Sınıflandırma* *18*

***Şekil 3**: Regresyon* *19*

***Şekil 4**: Sınıflandırma ve kümleme* *20*

***Şekil 5**: Çok katmanlı algılayıcı (MLP) mimarisi* *22*

***Şekil 6**: CNN matrisleri* *22*

***Şekil 7**: RNN mimarisi* *23*

***Şekil 8**: LSTM hücre yapısı* *25*

***Şekil 9:** Transormers mimarisi* *27*

***Şekil 10**: Bert ön eğitim şeması* *30*

***Şekil 11**: Bert İnce Ayar Şeması (Fine Tunning)* *32*

***Şekil 12**: Gömme (embedding) ve bölütleme (tokenization)* *33*

***Şekil 13**: Bert maskelenmiş dil modeli* *34*

***Şekil 14**: Bert Base ve Bert Large* *35*

***Şekil 15:** Neural Prophet* *37*

***Şekil 16**: CRISP-DM döngüsü* *42*

***Şekil 17:** Github Twint API Tweet* *47*

***Şekil 18:** Hugging Face uygulama arayüzü* *54*

***Şekil 19:** Bert Base'in örnek Maskelenmiş Dil Modeli gösterimi (MLM)* *55*

***Şekil 20:** Bert Large tokanization* *55*

***Şekil 21**: Smashwords web sitesi* *56*

GRAFİKLER DİZİNİ

***Grafik 1**: Peyton Manning wikipedia sayfası erişim istatistik* *38*
***Grafik 2**: Peyton Manning wikipedia çizgi grafiği* *38*
***Grafik 3**: Peyton Manning wikipedia haftanın günlerine göre erişim* *39*
***Grafik 4**: Peyton Manning wikipedia tarihe göre etki* *39*
***Grafik 5:** Veri setindeki negatif ve pozitif etiketli satırlar* *53*
***Grafik 6:** Bölüt sayıları ve ağırlıkları* *58*
***Grafik 7:** Epoch sayısı ve doğruluk orantısı* *60*
***Grafik 8:** İşlem sayısı ve kayıp değeri* *61*
***Grafik 9:** Bert model Confusion matrix* *67*
***Grafik 10:** Neural Prophet confusion matrix* *68*

1. GİRİŞ

Akıllı cihaz kullanımındaki artış ile doğru orantılı olarak sosyal medya kullanımındaki artış, kullanım şekli ve sosyal medyanın etkisi gün geçtikçe daha da çok artmaktadır. Sosyal medyadaki bir hesaba fotoğraf eklemek, bir restorantın yemeği hakkında eleştiri yazamak, sinemada gösterimdeki bir film hakkında yorum yazmak, yeni çıkan bir hastalığın duyurulması, doğal afet sonucu hızlı bir şekilde organize olarak yardım toplanması gibi birçok aktivitenin yapıldığı anda birçok takipçiye, hatta o hesabı takip etmeyen kişilerin bile hemen duyabilmesine olanak sağlamıştır.

Sosyal bilimler gibi alanlardaki birçok araştırma sosyal medyanın insanların fikirlerini değiştirerek yaşantısındaki gidişatı da değiştirdiğini öne sürmektedir. Beşeri alanlardaki bu araştırmalardan yola çıkarak makine öğrenmesi gibi sayısal, teorik ve ölçümlenebilir araştırma yaparak bu durum test edilmek istenmiştir.

Sinema yorumları ve borsa, hisse senetleri hakkında internette yapılmış yorumların manuel olarak duygusu (negatif & pozitif) etiketlenmiş veri seti kullanılarak Bert modeli eğitimi yapılmıştır. Ardından Twitterdan Tsla kelimesi içeren yorumlar çekilerek bu yorumların duygu (sentiment) analizi yapılmış ve bu işlemin çıktısı ile borsadaki Tsla hisse senedi verileri gün bazında birleştirilmiştir. Belli bir miktar hisse senedi değerleri ile bu duygu analizindeki duygu puanları ile Neural Prophet makine öğrenmesi modelinde eğitim verilerek test verisindeki 50 gün içerisindeki değerler test edilmiştir.

2. SOSYAL MEDYA

Tez kapsamında Twitter verisi kullanılarak duygu analizi yapılmıştır ve bu yüzden öncelikle neden sosyal medya uygulamalarından Twitter seçildiğinin cevabı verilirken bazı diğer kavramların ve uygulamaların da açıklaması başlıklar halinde yapılmıştır.

2.1 Sosyal Medya Kavramı

Son yıllarda internetin ortaya çıkması ve yaygınlaşması insanlar arasında sosyal medya adı verilen bir iletişim aracının doğmasına sebep olmuştur. Akıllı telefon kullanımının da artmasıyla birlikte her yerden ve her dakika iletişim halinde olunabilmektedir. Sosyal medya; fotoğraf, video ve metin gibi birçok bilgi ve belgeyi diğer kullanıcılarla paylaşabilme imkânı sunmaktadır. Mesela seyahat için gidilen bir ülke veya şehirde yemek yemek için bir restoran araştırılırken o yerle ilgili daha önce o restoranı tercih eden müşterilerin sosyal medyada yaptıkları yorumlara bakılarak ilgili işletme hakkında bir ön bilgi sahibi olunmasına yardımcı olur. Aynı şekilde beyazperde gösterime giren bir sinema filmini izleyen seyircilerin yada satın alınacak bir eşya için daha önce alışveriş yapan müşterilerin internette yapmış oldukları yorumlar karar verme sürecinde sosyal medyanın kullanıcılarına sağladığı avantajların başında gelmektedir. Yani kısaca sosyal medya insanların düşüncelerini, görüşlerini ve fikirlerini internet üzerinden diğer insanlarla paylaştıkları bir ortamdır denilebilir.

Sosyal medya en çok katılımcının yer aldığı, son yıllarda insanlar arasında etkileşimin yüksek seviyede olduğu ve çok sayıda paylaşımın yapıldığı bir mecra olmakla birlikte kullanıcıların aktif bir şekilde yer aldığı online bir platformdur. Sosyal medya; katılımcılar, açıklık, toplum, konuşma ve bağlantılılık gibi özelliklere sahiptir (Solmaz vd. 2013: 25). Bu özellikler kısaca aşağıdaki gibi özetlenebilir;

Katılımcılar: Sosyal medyada katılımcı tarafından paylaşılan bir bilgi veya yoruma, konuyla ilgili başka bir kullanıcı tarafından geri bildirim yapılması imkanı vardır. Bu geri bildirim paylaşılan bilgi veya yorumun beğen butonuna basılarak katılımcının görüşüne katıldığını gösteren bir bildirim gönderilmesi şeklinde olabileceği gibi, yapılan yorum sonrasında kullanıcının görmesini sağlayacak bir bildirim alması şeklinde de olabilmektedir.

Açıklık: Sosyal medya servisleri oylama, bilgi paylaşımı (paylaşılan bilgi için) yorum ve anket yapılabilmesi gibi yeni özellikleriyle katılımcılara açıklık ve geri bildirim imkanı sunarak kullanıcıların aktif bir şekilde katılabilmelerine olanak sağlamaktadır. Ancak bazen içerik veya yorum yapabilme gibi erişeme yönelik durumlar çok nadirde olsa engellenebilmektedir.

Konuşma: Klasik medyada tek taraflı bir iletişim söz konusu iken, sosyal medyada karşılıklı bilgi paylaşımının daha kolay ve hızlı olduğu, katılımcı ve kullanıcının da aktif olarak konuşmaya dahil olabildiği bir ortam söz konusudur.

Toplum: Sosyal medya sayesinde kullanıcılar sevdikleri fotoğraf, manzara ya da izledikleri bir filmi o anda ve hızlı bir şekilde paylaşabilmektedirler. Bu sayede topluluklar arasında hızlı ve etkili bir iletişim imkanı sunulmaktadır.

Bağlantılılık: Sosyal medya kullanıcılara diğer internet siteleri yada insanların araştırdıkları herhangi bir konu hakkında link paylaşılmasına imkan vermektedir. Örnek olarak, instagram üzerinden satış linkleri veya reklam linklerinin verilebilmesi.

2.2 Sosyal Medya Gelişimi

1950'li yıllarda bilgisayarların kullanılmaya başlaması ile birlikte veri aktarımı ihtiyacı ortaya çıkmıştır. 1960'lı yıllarda ağ paylaşımı ile iki bilgisayar arasında veri aktarımı yapılmış olup, 1989 yılında ilk olarak Timothy Berners-Lee tarafından literatüre kazandırılan world-wide-web (www)'nun 1991 yılında kullanılmaya başlamasıyla birlikte sosyal medyanın temelleri atılmıştır (Güzel, 2016: 91).

İnternetin bulunduğu ilk andan günümüze kadar olan süreçte internet geliştirilmeye devam etmektedir. Şimdiye kadar geçen sürede internetin gelişimi 3 dönem olarak sınıflandırılmaktadır. Bu dönemlere Web 1.0, Web 2.0 ve Web 3.0 adı verilmiştir.

İnternetin en ilkel dönemi olan ve ilk toplu erişim ile başlayan dönem Web 1.0 dönemi olarak adlandırılmaktadır (Kapan, Üncel 2020: 278). Bu dönemde web sitelerinin görünümü son derece sade olmakla birlikte sadece belirli web siteleri kullanılmaktadır. Bu web sitelerinde kullanıcılar sadece paylaşılan bilgiyi alan kişiler olarak yer almaktadırlar.

Daha sonra gelişen teknolojiyle beraber kullanıcılarda içerik üretebilmeye başlamış olup aktif olarak internetin içinde yer almaya başlanılan bu yeni dönem Web 2.0 dönemi olarak adlandırılmakla birlikte bu yeni döneme kısaca yazılabilir web teknolojisi de denilmektedir (Koçyiğit, Koçyiğit 2018: 27). Web 2.0 teknolojisi ilk kez 2004 yılında O'Reilly Media tarafından 2004 yılında kullanılmaya başlayan bir sözcük olmakla birlikte kullanıcıların deneyim, bilgi ve beklentilerini paylaştıkları vikipedi, facebook, youtube gibi online uygulamaların toplamı olarak tanımlanmaktadır. Bu uygulamalar sayesinde birçok bilgi ve görsel içeriğin üretilmesi ve paylaşılması ile birlikte kullanıcılar arasında bilgi, fikir ve duygu paylaşımı sağlanmaktadır.

Günümüzde makine öğrenmesi ve yapay zekanın gelişmesiyle birlikte Web 3.0 dönemi yaşanmaktadır. Web 3.0 teknolojisi ile birlikte gelişmiş algoritmalar ve veritabanı sistemleri sayesinde kullanıcıların etkileşimini tanımlamak için kullanılan bir terimdir. Web 3.0 teknolojisi ile birlikte pazarlama ve reklam sektöründe kişiye özel olarak hazırlanan içerikler ortaya çıkmış, kullanıcıların ilgi alanlarına göre ziyaret ettiği yerler, inceledikleri ürünler ile alakalı öneriler ve reklamlar kişileştirilmiş olarak karşılarına çıkarılmaktadır.

Tablo 1'de Web 1.0,Web 2.0, Web 3.0 dönemleri aşağıdaki tabloda karşılaştırılmıştır (Aslan, Kolancı, 2019: 210).

Tablo 1: Web 1.0 Web 2.0 ve Web 3.0 karşılaştırması

	Web 1.0	**Web 2.0**	**Web 3.0**
İletişim	Yayın	İnteraktif	Çevrimiçi
Enformasyon	Statik	Dinamik	Taşınabilir ve Kişisel
Hedef	Kurum	Topluluk	Kişisel
Kullanım	Kişisel Web Siteleri	Bloglar/Wikiler	Dosya Akışları
Üretim	Eğlence	Yayıncılık	Yaratıcılık
Etkileşim	Web Biçimleri	Web Uygulamaları	Akıllı Uygulamalar
Tarama	Dizinler	Anahtar Kelimeler	Bağlam-Uygunluk
Reklam	Durağan	İnteraktif	Kişisel
Araştırma	Britannica Online	Wikipedia	Anlamsal Web
Teknoloji	Html/Ftp	Flash/Java/Xml	Rdf/Rdfs/ Owl

Makine öğrenmesi, yapay sinir ağları, doğal dil işleme gibi terimlerin genel kapsamı olan yapay zekanın gelişmesiyle birlikte Web 3.0 dönemi olarak adlandırılan dönem başlamıştır. Web 3.0 teknolojisi, gelişmiş algoritmalar ve web bazlı veritabanı sistemleri sayesinde kullanıcıların etkileşimini tanımlamak için kullanılan bir terimdir. Web 3.0 teknolojisi ile birlikte pazarlama ve reklam sektöründe kişiye özel olarak hazırlanan içerikler ortaya çıkmış, kullanıcıların ilgi alanlarına göre ziyaret ettiği yerler, inceledikleri ürünler ile alakalı öneriler ve reklamlar kişileştirilmiş olarak karşılarına çıkarılmaktadır.

2.3 Sosyal Medyanın Özellikleri

Sosyal medyadaki profil sayfaları hesap sahibini temsil etmektedir ve hesap sahibinin yaptığı bu paylaşımlardan yola çıkılarak kullanıcı hakkında bilgi sahibi olunmasını sağlamaktadır. Hesap sahibi profil sayfası üzerinde takipçilerini kendi yönetebilir ve onlarla yorum yapabilir, gönderilerini beğenebilir.

Sosyal medyanın özelliklerinin en başında hesap sahiplerinin herhangi bir mekan veya zamana bağlı kalmadan paylaşım yapabilme olanağına sahip olabilmeleri gelmektedir. Kendi çektikleri yada ürettikleri içerikleri o anda mobil uygulamalardan paylaşırlarken, uygulamalarda kendilerini takip eden yada etmeyen birçok kişi bu içerikleri görebilmektedir. Hesap sahibi paylaştığı içeriklere yapılan yorumları okuyabildiği gibi kendi de yorum yaparak karşılıklı bir sohbet ortamına katılabilme olanağına sahiptir.

Sosyal medyanın birçok alt başlığı bulunmakla birlikte en çok bilinen alt başlıklar blog, mikro-blog, forum ve sosyal ağ olarak karşımıza çıkmaktadır (Dara, 2011: 3). Bu alt başlıklar kısaca şöyle tanımlanabilir:

Blog: WeB-LOGin kelimelerinin kısaltımı olarak adlandırılan blog, web tabanlı bir paylaşım platformudur. Blogun bir çok alt türü olmakla birlikte en bilinir blog türü kişilerin günlük yaşamda yaşadıklarını olayları paylaştıkları kişisel blogtur. Kullanıcılar tarafından daha çok günlük gibi kullanılmaktadır. Kurumsal blog ise şirketlerin kendileri ile ilgili haber ve duyurularını yaptıkları yer olarak adlandırılırken, temasal blog ise sadece belli konularda yazılan gönderilerin yer aldığı platformlardır. Temasal bloglara örnek olarak yemek ve tatil konulu bloglar verilebilir ve Türkiye'de en fazla ziyaret edilen blogların başından gelmektedir.

Mikro-blog: Mikro-bloglar bloglara göre daha sınırlı paylaşım ortamı sunmaktadır. Kelime sayısı sınırlı olmakla birlikte akıllı telefonlar üzerinden takip edilebilmektedir. Twitter çok sayıda kişi tarafından kullanılan ve en çok bilinen mikro-blog platformudur. Uygulamada metin için 280 karakter uzunluğunda kelime sınırlaması vardır ve paylaşılacak bir yazının çok uzun olması halinde sadece linki paylaşılabilmektedir.

Forum: Kullanıcıların internet ortamında web sayfaları halinde herhangi bir konu hakkında sorular sorarak cevaplar aradığı, bilgi ve paylaşımlar yaptığı platformlardır. Forum alanında kategoriler yer almaktadır ve bu kategorilere bağlı alt konu grupları oluşturulur. Kullanıcılar oluşturdukları bu konu gruplarında yazılı bir şekilde ifade edebilmektedirler.

Sosyal Ağ: Kullanıcıların internet ortamında kendilerini gerçek ya da akıllarındaki sanal kimliklerini tanımlayarak diğer kullanıcılarla iletişim kurdukları ortamlara sosyal ağ denilmektedir. Kullanıcılar sosyal yaşamda kullandıkları jest ve mimik hareketlerini internet ortamında da bu ifadelere karşılık gelen sembolleri kullanarak duygu ve düşüncelerini ifade edebilmektedir. Sosyal ağlar kullanım alanlarına göre farklılık göstermektedir. Video paylaşımının yapıldığı youtube, görsel paylaşım yapılan instagram dünyada en popüler olan sosyal ağların başında gelmektedir. Sosyal ağların en önemli ortak özelliği internet ve mobil cihazların aktif olarak kullanılabildiği her an kullanıma, geliştirilmeye, paylaşım yapılmaya devam edilebilmesidir.

2.4 Twitter Nedir?

Twitter; 2006 yılında Jack Dorsey tarafından geliştirilmiş 140 karakter ile sınırlı mikroblog altyapısına sahip ve günümüzde en aktif kullanılan sosyal medya uygulamalarından biridir. Kullanıcılar üye olarak hesap oluşturur ve daha sonra hesap oluşturan diğer kullanıcıların hesaplarını takip etme imkanına sahip olurlar. Oluşturulan hesaplar gizli ve normal hesap olmak üzere iki çeşittir. Gizli hesapta kullanıcıyla sadece hesap sahibinin izin verdiği kişiler etkileşim halinde olabilirken, herkese açık olan normal bir hesabı takip eden veya etmeyen herkes görebilmektedir. Kullanıcı, takip ettiği hesap sahibi tarafından yazılan metin, paylaşılan resim veya videoları beğenebilir yada retweet ederek kendi takipçileriyle de paylaşabilmektedir. Akıllı telefonların kullanımının yaygınlaşması sayesinde uygulamaya kısa sürede ulaşılabiliyor olması nedeniyle dünyada olup biten tüm gelişmelere çok kısa sürede ulaşabilmek mümkündür.

2017 yılında 280 karaktere çıkarılan twitter'ın kelime sınırlamasının biraz daha artırılması sayesinde daha kullanışlı bir hale gelmiştir. Uygulamada o anda twitter da en çok konuşulan konuların yer aldığı trending topics adı verilen bir sekme bulunmaktadır. Bu sekmeye kısaca twitterin gündemi hatta günün gündemi denilebilir. Herhangi bir konu hakkında atılan tweetlerde kullanıcıların dikkatini çekmek için key word seçilen ve paylaşımda vurgulanmak istenilen bazı kelimelerin önüne # (hashtag) işareti konulmaktadır. Hashtag işareti sayesinde bu kelimeler ön plana çıkmakta ve diğer kullanıcılar sadece bu kelimeyi paylaşarak bile gündemde üst sıralara taşınmasını sağlamaktadır.

Cep telefonlarının yerini akıllı telefonların alması ve internetin yaygınlaşarak cep telefonlarına kadar ulaşılabilmesi insanların hayatında birçok yeniliği sebep olmuştur. Cep telefonlarında 160 karakterle sınırlı olan ve kısa mesaj (sms) adı verilen anlık olarak diğer kullanıcılarla iletişime geçilmesini sağlayan yazılı mesajların yerini akıllı telefonlardaki uygulamalar almıştır. İnternet dünyasının mesajlaşma aracı yani sms'i olarak adlandırılan Twitter; aynı mesajlaşmak gibi 280 karakter sınırlamasıyla metin yazılması olanağı sağlarken en önemli özelliği dünyadaki tüm kullanıcılarla anlık iletişime geçilebilmesini sağlamaktadır. Atılan tweet'in dünyadaki herkes tarafından görülebilmesi imkanı vardır. Twitter'ın 2006 yılında ilk çıktığında ana ekranında yazan "what are you doing?" Türkçe anlamıyla "şu an ne yapıyorsun?" yazısı 2009 da "what's happening?" yani Türkçe anlamıyla "şu anda neler oluyor?" şeklinde değiştirip bireysel kullanımdan çıkarılarak kişilerin hayatın her anı, alanı ve durumuyla ilgili gördüğü, yaşadığı, hissettiği şeyleri başkaları ile paylaşma konusunda daha teşvikçi bir yaklaşım izlemiştir. (Çalışkan, Mencik 2015:12)

2.5 Twitterin Analiz İçin Kullanılma Nedeni

Günümüzde teknolojinin ilerlemesiyle kitle iletişim aracı olarak kullanılan gazete ve dergilerin yerini akıllı telefonlarda yer alan ve istenilen her an her yerde ulaşabilme imkanı sunan sosyal medya uygulamaları almıştır. Anlık gelişen olaylardan anında haber alınmasını sağlayan uygulamaların başında da hiç şüphesiz twitter gelmektedir. Bir deprem sonrasında veya bir doğal afetten sonra insanlar anık olarak tweet atmakta birbirleriyle iletişim halinde olabilmektedir. Aynı zamanda kullanıcılar bilgi sahibi olmak istedikleri konular hakkında diğer uygulamayı kullanan kişilerin görüş ve yorumlarını arama butonunda istedikleri kelimeyi aratarak ulaşabilmektedirler. Böylelikle o kelimeyle ilgili diğer kullanıcıların görüşlerine ulaşabilirler. Kısa cümlelerle paylaşılan bu bilgiler kullanıcılara yol gösterici olabilmektedir. Dakikada binlerce tweet atılan uygulamada yapılan yorumların kullanıcıların duygularında olumlu yada olumsuz etkisi olması hiç şüphesiz beklenen bir durumdur. Twitter in popüler bir uygulama olması, erişim kolaylığı ve twitlerin en fazla 280 karakter olması nedeniyle makine öğrenmesi yöntemiyle veriyi analiz etmek için tercih edilme sebeplerinin başında gelmektedir.

2.6 Diğer Sosyal Medya Ağları

Gün geçtikçe özellikle son on yılda bir çok sosyal medya uygulaması geliştirilmiş ve dönem dönem farklı uygulamalar çok popüler olmuştur. Popülaritesi yaşanılan ülke, bölge ve kültüre göre değişkenlik gösteren uygulamaların ayrıca kullanım amaçları da değişkenlik göstermektedir. Örneğin bir grup twitter'da daha siyasi içerik takip etmek ve bu alanda paylaşım yapmak şeklinde kullanırken diğer gruptaki kişiler astroloji, şiir gibi farklı alanlarda paylaşım yapmakta ve bu alanları takip etmektedir. Örnek olarak seçilen Facebook, Youtube, Tiktok, Instagram, Linkedin ve Twitter bu dönemde en çok kullanıcı hesabı olan sosyal medya uygulamaları olarak açıklanmıştır.

2.6.1 Instagram

Instagram 2010 yılında İOS işletim sistemli telefonların kullanımına sunulduktan 2 sene sonra android işletim sistemli telefonların da kullanımına sunulmasıyla birlikte günümüzde en çok tercih edilen sosyal medya uygulamalarının arasında yer almaktadır. Uygulama tarafından kullanıcılara fotoğraf, 1 dakikalık video ve 24 saat sonra kaybolma özelliği olan hikaye paylaşımı imkanı sunulmaktadır. Bu paylaşımlarda çeşitli filtre ve efektler kullanılarak eğlenceli içeriklerin üretilebilmesi mümkündür. Kullanıcılar tarafından paylaşılan fotoğrafların açıklama kısmına # (hashtag) işareti konularak fotoğraf paylaşımı

yapılabilir. Daha sonra arama kısmından benzer kelimeler ile arama yapıldığında paylaşılan benzer fotoğraflar diğer kullanıcılar tarafından görülebilmektedir. Bu sayede farklı ülkelerdeki insanlar birbiriyle etkileşim halinde olabileceği gibi aynı zamanda reklam ve tanıtım faaliyetleri içinde en çok tercih edilen yöntemlerin başında gelmektedir. Bu sayede yapılan paylaşımlarla sadece eğlence platformu olmasının yanı sıra ayrıca önemli bir e-ticaret alanıdır.

2.6.2 Facebook

Mark Zuckerberg 2004 yılında Harvard Üniversitesi'nde öğrenci iken üniversitede okuyan öğrencilerin birbirleriyle tanışabilmesi amacıyla ilk zamanlarda ki adıyla "thefacebook" isimli sosyal paylaşım sitesini kurmuştur. Kısa sürede popüler olan site Amerika'daki diğer üniversite öğrencilerinin de kullanımına sunulmasının ardından facebook adını alarak günümüzde tüm dünyanın kullanımına sunulmuş olup halen günümüzde en çok kullanıcısı olan sosyal medya sitesidir. Kullanıcılar üye olduktan sonra kendi sayfalarını oluşturarak; fotoğraf, haber, makale, video vb gibi paylaşımlar yapabilmekte ve sayfa oluşturan diğer kullanıcılara arkadaşlık isteği gönderilerek tanışma imkanı sunmaktadır. Günümüzde marketpalace özelliği sayesinde alıcı ve satıcıların sosyal medya uygulaması üzerinden birbirini bulmasını sağlamasının yanı sıra, hesap sahiplerinin doğal afetlerde zarar gören kişilere yardım edilebilmesi için geliştirilen kriz yardım merkezi başta olmak üzere hayati konularda da insanlara destek sunuyor olması uygulamanın halen popülerliğini korumasında etkili olan özelliklerinin başında gelmektedir.

2.6.3 Linkedin

2003 yılında kullanıma açılan Linkedin işverenler ile iş arayan kullanıcılar arasında köprü kuran profesyonel bir iş ağı olmakla birlikte, diğer üye kullanıcılarla da iletişim kurulmasını sağlayan sosyal bir platformdur. Kullanıcılar üye oldukları anda deneyimler, öğrenim bilgileri, yetenekler, ilgi alanları ve başarılar alt sekmelerinin olduğu kişisel bir sayfa oluşturmaktadır ve bu sayfa kısaca kullanıcının özgeçmişidir. Kullanıcıların karşılıklı iletişim kurarak iş tecrübelerini paylaşabilmeleri, birbirlerine referans olabilmeleri ve iş bulmalarına olanak sağlamasından dolayı Linkedin, çalışanlar ve şirketler tarafından yakından takip edilen iş sektörü alanında öne çıkan bir sosyal medya uygulamasıdır. Bu sayede çalışılmak istenilen sektördeki kişilerle internet ortamında tanışılarak şirket hakkında daha detaylı bilgilere sahip olabilmek mümkündür. Kullanıcı, sayfasında yapılan çalışmaları yayımlayarak hem çalışmadaki aşamaları kaydederek aynı alanda iş yapmak isteyen diğer kişilere yol gösterebilir hem de şirketlerin dikkatini çekerek iş başvurusu yapmadan yeni iş teklifleri alabilmektedir. Borsa İstanbul'da işlem gören şirketlerin sosyal medya kullanımlarının araştırıldığı çalışmada şirketler

tarafından kullanılan en aktif uygulamanın %82 ile Linkedin olduğu sonucuna ulaşılmıştır (Özevin, 2021: 414).

2.6.4 Youtube

İnternetin yaygınlaşmasının ardından videoların e-posta yoluyla gönderilememesinden dolayı 2005 yılında video paylaşım sitesi Youtube kurulmuştur. Paylaşılan videoları izlemek için herhangi bir üyelik söz konusu değildir fakat kullanıcı tarafından içerik üretilerek siteye video yüklenmek istenirse hesap oluşturulması gerekmektedir. Hesap oluşturulduktan sonra video yüklenebilir ve belirli izlenme sayısına ulaşılması halinde videoyu üreten kullanıcı para kazanabilmektedir. Youtube da paylaşılan videoların kapasitesi 256 gb' dan küçük ve 12 saatten daha az süreli olmalıdır. İzlenen videolara yorum yapılabilir, bilgisayara indirilip kaydedilebilir ve diğer sosyal medya uygulamalarında paylaşılabilinir. Youtube'da video izlerken videonun sağ tarafında önerilen videoların yer aldığı bir liste vardır. Bu listede gösterilen videolar Youtube'un yapay zeka destekli arama motorunda aratılarak izlenilen, yorum yapılan veya beğenilen videolara en yakın olacak şekilde oluşturulmuş bir listedir. Aynı zamanda uygulamada içerik üreticilerine fikir vermek adına videoların hangi cinsiyet ve yaş grubunda kaç kez izlendiği, hangi ülkelerde popüler olduğu, kaç adet yorum ve beğeni aldığına dair bilgilerin yer aldığı youtube analytics özelliği bulunmaktadır.

2.6.5 Tiktok

Akıllı telefon kullanımının artması ile birlikte insanların arta kalan vakitlerinde eğlenceli vakit geçirmek istemesi ile birlikte 2016 yılında kısa videolar üretilerek paylaşılmasını sağlayan Tiktok uygulaması ortaya çıkmıştır. Bu uygulama ile birlikte üretilen videolara efektler ve farklı müzikler eklenebilmektedir. Paylaşılan bu videoların kısa sürede dünyadaki diğer kullanıcılar tarafından izlenebilmesi nedeniyle kullanıcı profili daha çok genç yaştaki kişilerdir.

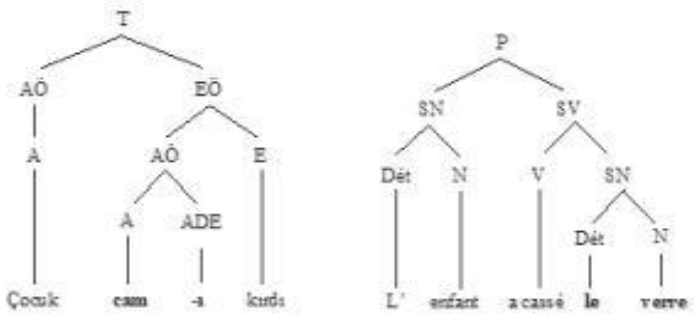
3. YÖNTEM

Yöntem bölümünde bu tez kapsamında kullanılacak modellerin teorik açıklamları ve bu modellerin dahil olduğu literatürdeki genel kavramları açıklanmıştır. Sentimen analizi (duygu analizi) doğal dil işleme modülü altında gruplandırıldığı için öncelikle doğal dil işlemenin açıklanması ardından gündelik hayatta karşılaşılabilecek alanlara göre sınıflandırması açıklanmıştır. Ardından Bert ve Neural prophet makine öğrenmesi altında yapay sinir ağları ve gözetimli öğrenme modülü altında olduğu için makine öğrenmesi ve dalları ve derin öğrenme kavramları farklı başlıklar olarak sırasıyla açıklanmıştır.

3.1. Doğal Dil İşleme

Doğal dil işleme ingilizce literatürde kısaltılmış hali ile NLP olarak yani açılımı ile natural language processing olarak geçmektedir. Dil, yapı ve köken bakımından ikiye ayrıldığı düşünülse de bu varsayım sadece insanların kendi aralarındaki iletişim ve etkileşimi için kullandığı dil için geçerli olduğu aşikardır. Oysaki teknoloji ile birlikte artık bilgisayar ve insan arasında da iletişim için bir dil doğmuştur. Bu bir yönüyle insan ve bilgisayar arasındaki etkileşim (human computer interaction HCI) olarak da görülebilir (Preece, 1994, s,70). Fakat insanlar kendi aralarındaki iletişim için doğal dil kullanırken bilgisayarların bu dili anlaması için işlemek gerekmektedir. Farklı bir açıdan, doğal dilin bilgisayar kullanılarak ya da bilgisayar aracılığıyla işlenmesine, yorumlanmasına, anlaşılmasına ya da üretilmesine kısaca doğal dil işleme denebilir.

Doğal dil işlemede veri kaynağı olarak iki grup vardır. İlk grup yazılı metin ikinci grup ise sesli metinlerdir. Doğal dil işlemede veri kaynağından alınan metin belirli ön işlemlerden geçirilir. Standart olarak kabul edilen bu işlemler; morfolojik analiz, semantik analiz (semantic analysis), sözdizimsel analiz (syntactic analysis) ve söylev analizdir (pragmatic). Morfolojik analiz, bir keliminin aldığı farklı eklerin çözümlenmesi ve kelimenin kökeninin ortaya çıkarılması için uygulanır (Yılmaz ve Yumuşak, 2021, 82) Özellikle eklemeli dillerde aynı köke sahip sözcüklerin arasındaki bağlantının çözülmesi konusunda daha başarılı sonuç verir (Bal vd. , 2017, 126). Söz dizimsel analiz, morfolojik analizin sonuçlarını kullanarak cümledeki kelimelerin sıralamalarını inceler ve yapısal tanımı oluşturur. Örneğin Fransızca'da "özne - yüklem - tümleç" şeklinde sıralama olurken Türkçe'de "özne - tümleç - yüklem" şeklinde cümle ögeleri sıralanmaktadır. Dr. Aydın İlker'in yaptığı inceleme çalışma sonuçlarından bir örnek olan Şekil 1'de (Aydın, İ., 2008: 26) Fransızca ve Türkçe'nin birer cümledeki örnekleriyle sözdizimsel analiz örneği gösterilmektedir.



Şekil 1: Türkçe ve Fransıca dillerinin sözdizimsel karşılaştırması

Semantik analiz, semantik kelime olarak anlam demektir ve cümlelerin anlamlarını analiz eder. Semantik analiz literatürde bir diğer deyişle anlam bilimsel analiz olarak da geçmektedir. Sözcük öbeklerinin ayrı ayrı ya da birleşik öbek olarak oluşturdukları anlamı ele alarak modelleme oluşturur. Genel anlamda iki tür altında incelenir; sözcüksel (lexical) ve birleşik (combination). Sözcüksel de zıt, eş anlam, bütünün parçası, çok anlamlılık gibi kavramlar incelenmektedir. Birleşik de ise sözcükselin tersine parçaların anlamındansa bütünün anlamına odaklanır. En son olarak ise söylev analizi (pragmatic), diğer analizlerden sonra gelen son bölümdür ve ekler, kelimeler, kelime öbeklerinden sonra cümlelerin ifade etmek istediğine bakan analizdir.

3.2 Doğal Dil İşlemenin Uygulama Alanları

Doğal dil işleme gün geçtikçe tüm alanlarda kullanılmaya başlamıştır. Özellikle android ve Ios alt yapılı telefon ve tablet kullanımıyla asistanlar, çeviriler, mailler gibi gündelik ve iş yaşantısında ihtiyaç duyulacak tüm alanlarda doğal dil işlemenin kullanıldığı bir bölüm bulunmaktadır. Bazı alanların açıklamaları başlıklar halinde aşağıdaki bölümde verilmiştir.

3.2.1 Soru Yanıtlama

Günümüzde yapay zekanın gelişmesi ile insan gücünün yerini alan makineler sayesinde doğal dil işleme konusunun ilk problemlerinden biri olan soru cevaplama ele alınan metinin üzerinden anlamlı sorular çıkarılarak cevaplar bulunması yada gelen cevaplardan yola çıkarak ilgili soruların bulunması işlemidir. Örneğin, bir bankanın çağrı merkezi aranıldığında çağrı merkezi çalışanıyla iletişime geçilmeden önce arayan kişinin sorununu bulabilmek için daha önce kayıt edilen sesler ile öğrenip sonrasında arayan kişinin sorduğu sorularla hangi konuda sorun yaşandığını anlayıp daha sonra çağrı merkezi çalışanına aktarmaktadır. Benzer şekilde son zamanların en büyük sağlık sorunlarından biri olan koranavirüs

salgınıyla birlikte yaşanılan pandemi sürecinde internetten alışverişin artmasıyla alışveriş yapılan sitelerde yaşanan sorunların çözümü için ilgili siteye girildiğinde karşılaşılan sorunun çözümü için çözüm merkezi sekmesi tıklandığında yazılan sorulara karşılık müşteriye siparişinin durumuyla ilgili bilgiler verebilmekte iken cevap bulamadığı sorular olduğunda daha kaliteli bir hizmet alınabilmesi için müşteriyi çağrı merkezi operatörüne yönlendirmektedir.

3.2.2 Dil Tercüme

İnternetin yaygınlaşması ve sosyal medya kullanımının artmasıyla birlikte farklı ülkelerde yaşayan insanların birbirleriyle iletişim kurmak istemesi veya bir öğrencinin farklı dille bulduğu bir kaynaktan bilgi elde etmek istemesi gibi nedenlerden dolayı makine, bir cümle yada kelimeyi başka dildeki bir metine tercüme edilebilmektedir. Bu gelişmelerden dolayı önde gelen arama motorlarından olan yandex çeviri ve google translate gibi çeviri programları gündelik hayatta önemli bir yer tutmaktadır. Son zamanlarda yapay sinir ağlarının gelişmesiyle birlikte bu çeviri programları da uzun ve karmaşık cümleleri istenilen dile çevirmekte oldukça başarılı bir konuma gelmiştir. Bu sayede araştırma yapmak isteyen araştırmacılar veya bir konu hakkında bilgi edinmek isteyen kişilerin karşılarına çıkan önemli bir makale, yazı vb materyalleri incelemek için o dili bilmese bile çeviri programları sayesinde fikir sahibi olabilmektedir.

3.2.3 Bilginin Bulunması ve Getirilmesi

Aranılan herhangi bir bilginin metin kaynakları arasında aranıp bulunması ve getirilmesi işlemidir. İnternetin yaygınlaşması sonucunda ortaya çıkmış olup, internette yapılan araştırmalar için geliştirilmiştir. İnternet bulunmadan ve yaygın olmadan önce araştırma yapılmak istenen bir konu hakkında kütüphanelere gidilerek onlarca kitap arasından konu araştırılırken bu işlem hem zaman kaybına neden olmakla birlikte hem de aranılan bilgiye ulaşılamaması gibi bir durum söz konusu olabilmektedir. İnternetin bulunmasıyla birlikte bilgi sahibi olmak istenilen konular ile ilgili internette araştırma yapıldığı zaman ilgili sözcüklerin en çok geçtiği metinlerin olduğu sayfaları bularak ekrana getirmektedir. İnternette bir konu hakkında araştırma yapmak için öncelikle anahtar sözcük yazılır ancak internette araştırma yaparken aranılan bilginin tam ve eksiksiz yazılması büyük önem arz etmektedir (Adalı, 2012: 2). Örneğin, internette Tesla kelimesi aratıldığı zaman arama sonucunda Tesla marka otomobil ekrana geleceği gibi, ünlü bilim insanı Nicola Tesla da ilgili sayfalarda araştırmacının karşısına çıkabilecektir. Sonuç olarak arama sonucunda binlerce sayfa taranıp ekrana geldiği için hepsini okuma fırsatı ve zamanı bulunmadığından dolayı aranacak kelimelerin daha net ve yalın olması araştırmacıya vakit kazandıracaktır.

3.2.4 Kelime İşleme

Kelime işleme yöntemi birçok alanda kullanılan birtakım işlemlerin birleşiminden oluşmaktadır. Metin üzerinde yazımsal hataları kontrol etmek için dilbilgisi kontrol işlemi kullanılır. Daha sonra kelimede bir yazım hatası varsa heceleme kontrol ve düzeltme işlemleri yapılır. Örneğin, internet arama motoru google da aramak istenilen bir kelime arama motoruna yazıldığında öncelikle kelimenin doğru yazımı ve benzer içerikler belirerek yönlendirme yapılır. Dil bilgisi kontrolü yapılmış, yazım yanlışı bulunmuş yani heceleme yapılarak bunun düzeltilmiş versiyonu belirtilmiş olur.

3.2.5 Veritabanı Arayüzü

Veritabanı üzerinden cevabına ulaşılmak istenen soruya ulaşmak için kullanılan bir işlem yöntemidir. Örneğin bankalar müşterilerinin tüm işlemlerini ve hareketlerini; pazarlama, satış, müşteri memnuniyeti ve güvenlik gibi nedenlerle veritabanında tutup gerekli durumlarda kullanırlar. Veritabanında tuttukları bu bilgiler sayesinde müşterilerin daha iyi bir hizmet almasını sağlamaktadır. Müşterilerin isteği olması durumunda bu bilgileri onlarla paylaşarak daha iyi bir hizmet almaları için çalışırlar. Yine aynı şekilde zincir bir market, müşterilerinin temel ihtiyacı olan malzemeleri tedarik ettiği üreticiler ile bu malzemeleri alan tüketicilerin verilerini veritabanında saklar. Böylece istenilen soruya karşılık cevap bulabilirler. Ekim ayındaki mandalina satışlarından elde edilen karın ne kadar olduğu öğrenilmek istendiğinde veritabanındaki bilgilerden cevaba ulaşılabilir.

3.2.6 Arama Motorları

Hiç kuşkusuz son yüzyılın en önemli buluşlarından biri olan internet, aranılan bilgiye en kısa sürede ve en kolay şekilde ulaşılmasını sağlamaktadır. On binlerce kaynağın yer aldığı ortamda her gün bu bilgilere yenileri eklenmektedir. Arama motorlarının içerisinde her kelimenin, hangi dokümanda ve kaç kez geçtiği bilgisinin tutulduğu veritabanı mevcuttur (Amasyalı, Diri, 2005:1). Araştırmacının araştırdığı bir bilgiyle ilgili onlarca sayfayı tek tek okuma fırsatı olmadığı için arama motorları aracılığıyla arama yaptığı zaman ilgili sitelerin url adresleri sıralanmaktadır. Arama motorları kelime grupları ile alakalı internet sitelerini en çok tıklanan sayfalardan en az tıklanan sayfalara doğru bir sıralama yaparak araştırmacıya sunmaktadır. Ulaşılmak istenen siteye girmek için adresinin tam ve doğru yazılması gerekmektedir bu yüzden internet arama motorlarının sık kullanılma nedenlerinin başında bu adresleri tam ve doğru olarak bilinmemesi yada daha kısa sürede siteye ulaşılmak istenmesi gibi nedenlerle çok sık kullanılmaktadır. Örneğin araştırmacı internet arama motoruna Marmara

Üniversitesi yazıldığı zaman karşısına ilk olarak "www.marmara.edu.tr" url adresi çıkacak ve ilgili siteye ulaşılabilecektir.

3.2.7 Metin Ön İşleme

Twitter gibi uygulamalarda kelime sınırlaması olduğu için yazım hataları ile karşı karşıya kalınmaktadır. Atılan tweet'i 280 karakterle anlatabilmek için yazılan cümlelerde bazı kelimeler için kısaltma kullanılmasından yada ingilizce karakterlerle yazı yazılmasından dolayı yazım hataları olmaktadır. Bu sebepten ötürü metnin kelimelere ayrılması, ardından kelimelerin anlamsal değeri bulunarak kelimeler köklerine ayrılırken gereksiz kelimelerin ayıklanması, gereksiz bilgilerin çıkarılırken yazım kurallarına uygunluğunun tespit edilmesi, kelimelerin köklerinin belirlenirken kelime anlamlarını bulması gibi kelimelerle ilgili sürecin tamamı metin ön işleme süreçlerine dahilinde yapılmaktadır.

3.2.8 Argüman Birleştirme

Facebook, Instagram ve Twitter gibi sosyal ağlarda paylaşılan bir görüntü veya haber için paylaşımın altına o hesabı takip etsin etmesin binlerce kişi tarafından yorum yapılabilmektedir. Yapılan bu yorumlarda birbiriyle aynı görüşe sahip olan kişilerin verdiği benzer cevaplar olabileceği gibi ilgili gönderiye katılmayıp karşı görüş olarak farklı bir bakış açısıyla başka bir tartışma konusu da yaratılabilir. Sosyal ağlar üzerinden yapılan bu tartışmalar sonucu elde edilen argümanların analiz edilerek tekrarlı olanların yada boş olanların temizlenerek, kalan argümanlardan anlamlı verilerin üretilmesi işlemine argümanların birleştirilmesi denilmektedir. Elde edilen bu rapor sonunda tartışmaya katılan kişilerden kimlerin aynı tarafta olduğu bulunabildiği gibi, farklı görüşü savunan kişilerin haritasını çıkarabilmekte mümkündür.

3.2.9 Konuşmadan Yazıya – Yazıdan Konuşmaya Dönüşüm

Literatürde speech to text - text to speech şeklinde geçen sesli komutların yazılı metine yada yazılı metinlerin sesli komutlara dönüştürülmesi işlemidir. Akıllı cihazların gündelik hayatta bu kadar yer edinmesiyle birlikte bu dönüştürme işleminin daha çok sesli asistanlar tarafından telefonlar ile ya da bilgisayarlar ile gün geçtikçe artan şekilde kullanılmaya başlamıştır. En çok bilinen sesli asistanlardan biri olan Siri üzerinden sesli komutlarla ekrana hiç dokunmadan mesaj yazabilir ve bu mesajın istenen kişiye yine yalnızca sesli komut verilerek gönderilebilir. İnternet arama motorlarında araştırılmak istenilen bir konu hakkında Google Asistana sesli komut verilerek aranılan metinlere ulaşılabilir. Benzer

şekilde farklı dilleri istenilen dile çevirmeyi sağlayan Google Translate uygulanmasında yazılan cümleyi istenilen dile çevirildikten sonra Google Asistan seçeneğine dokunularak çevirilen dildeki ilgili metinin okunuşu seslendirilebilir ve bu sayede kolayca iletişim kurulabilir.

3.2.10 Hedeflenmiş Reklamcılık

Satın almak için araştırılan bir ürün daha sonra internette gezinirken reklam olarak internet kullanıcısının karşısına çıkmaktadır. Son yıllarda gelişen bu çevrimiçi reklamcılık sayesinde şirketler hem daha çok kar etmektedir hem de reklamların ilgi alanı dahilindeki kişilere gösterilmesi sayesinde hedef kitleye ulaşan bir reklamcılık başarısı elde edilmiş olmaktadır. Satın almayı düşünen kişilerin karşılarına çıkan bu reklamlar sayesinde hem farklı bir modelle karşılaşabilme imkanı elde edebilecekleri gibi hem de daha ekonomik bir ürünü bulma ihtimali de artmaktadır. Sonuç olarak hem satın almak isteyen kişiler hem de şirketler için karşılıklı faydaya dayanan bir sistemdir.

3.2.11 Otomatik Arama Düzeltme ve Tamamlama

Araştırmacı tarafından internet arama motorunda bir şey aratılmak istendiği zaman 3-4 kelime yazıldıktan sonra olası arama terimlerin direk arama yapan kişiye önerilmesi işlemidir. Bu terimler yardımıyla arama yapıldığı zaman aratılan kelimeyle ilgili daha önce yapılan araştırmaları sıralayarak yol göstermenin yanında hem de kelimeleri tamamlama fırsatı sunmaktadır. Ayrıca yapılan yazım yanlışlarını düzeltme imkanı sağlayarak doğru verilere daha hızlı bir sürede ulaşma imkanı sunmaktadır.

3.3 Makine Öğrenmesi

Makine öğrenmesi doğal dil işlemenin üst kümesi olarak düşünülmelidir. Bu yüzden aslında doğal dil işlemenin kullanıldığı alanlarda makine öğrenmesi kullanılıyor denebilmektedir. Her doğal dil işlme problemi aslında bir makine öğrenmesi problemi olarak görülmelidir fakat her makine öğrenmesi problemi ya da projesi bir doğal dil işleme problemi değildir. Doğal dil işlemesi olarak değil de makine öğrenmesinin kullanıldığı alanlar çok genel şekilde başlıklar halinde aşağıdaki şekilde açıklanabilmektedir.

3.3.1 Makine Öğrenmesi Kullanım Alanları

Makine öğrenmesi hemen her alanda kullanıldığı için tüm alanlar açıklanmamış, bazı alanlar gruplandırılarak genel anlamda açıklama yapılmış ve örnekleri verilmiştir.

3.3.1.1 Sağlık

Sağlık alanında hastalıkların erken teşhis edilmesi ve tanımlanması büyük bir öneme sahiptir. Makine öğrenmesinin sağlık sektöründe kullanılmasıyla birlikte kanser hastalıklarına erken tanı konulması, hasta bilgilerinin dosyalar halinde değil internet ortamında saklanabilmesi, tıbbi görüntüleme cihazları sayesinde sağlık çalışanlarının gözünden kaçabilecek ufak ayrıntılarının tespiti ve hastaların geçmiş verilerini tarayarak hastalık tahmini yapılabilmektedir. Hasta tarafından tedavi süresince kullanılması gereken ilaç dozları hastanın biyolojik ve fizyolojik özelliklerine göre ayarlanabilmekte ve alınan ilaç dozları takip edilerek ne kadar süre boyunca düzenli ilaç kullanıldığı tespit edilebilmektedir. Daha önce tedavi edilen bir hastalığın tedavi süresinden yola çıkılarak benzer bir hastalığa yakalanılması durumunda ne kadar süre sonunda tedaviye yanıt alınabileceği bilgisine ulaşılabilinmektedir. Salgın hastalıklarla mücadele için geliştirilen ilaçların yan etkilerinin değerlendirilmesinde daha sağlıklı ve hızlı sonuçlar alınarak daha erken sürede müdahale edilmesini sağlamaktadır. Makine öğrenmesi hem artan maliyeti azaltmak hem de daha iyi bir sağlık sistemi oluşturulabilmesi açısından sağlık sektörüne katkı sağlamaktadır.

3.3.1.2 Siber Güvenlik

Siber güvenlik bilgisayar, sunucu, donanım ve mobil cihazların kötü amaçlı yazılımlardan korunması işlemidir. Siber saldırılar nedeniyle şirketler bilgi kaybının yanı sıra maddi kayıplarla da karşı karşıya kalabilir. Makine öğrenmesi ile siber suçun gerçekleşmeden önce tespit edilebilmesi, spam maillerin saptanarak olası dolandırıcılık faaliyetlerinin engellenmesi, kullanıcı kimlik doğrulaması yapılarak yabancı kişilerin sisteme girişinin engellenmesi mümkündür. Güvenlik önlemlerine rağmen gerçekleşen saldırılar olması halinde makine öğrenmesi ile saldırılar analiz edilerek benzer tehditlere karşı model oluşturulmaktadır. Siber saldırı yapan kişiler tarafından temel hedef sisteme yapılan sızıntılar da yakalanmaktadır. Makine öğrenmesi ile hem sisteme sızıntı yapılması engellenmektedir hem de sızıntının nereden yapılabildiği tespit edilebilmektedir. Siber güvenlik için makine öğrenmesinden faydalanılmasının getirdiği bazı dezavantajlarda mevcuttur. Model oluşturulabilmesi için kötü amaçlı yazılım kodlarına ihtiyaç duyulmaktadır. Alınan güvenlik önlemleri için şirketler bellek, veri ve bilgi işlem gücü gibi büyük miktarda kaynaklara ihtiyaç duymaktadır.

3.3.1.3 Lojistik ve Üretim

Lojistik sektöründe robot kollar ile kullanılmaya başlayan makineler daha sonra sürücüsüz otomobillerin insanların yerini almasıyla birlikte sektörde daha çok yer edinmiştir. Sürücüsüz otomobiller sayesinde yakıt tasarrufu; robot kollar sayesinde de zamandan tasarruf edilerek üretim hızında bir artış elde

edilebilmektedir. Makine öğrenmesi sayesinde teslimat aşamasında araç rota planı oluşturularak en kısa mesafe hesaplanarak sevkiyat işlemi kısa sürede yapılabilir ve aynı zamanda araç takip sistemiyle araçlar online olarak takip edilebilir. Lojistikte bir diğer önemli konu üretim planlamasıdır. Makine sayesinde geçmiş yıllardaki taleplere göre bir üretim planı oluşturularak olması muhtemel müşteri talebine göre üretim yapılabilmektedir. Hızlı ve kısa sürede üretim yapılmasını sağlayan makinelere çip takılarak verilerin makine öğrenmesi ile analiz edilmesi sonucu makinenin olası bakım zamanı ile olası bozulma süresi tahmin edilerek üretimde sürekliliğin sağlanmasında makine öğrenmesi önemli bir rol oynamaktadır. Lojistik pazarının büyümesiyle birlikte müşterilerin diğer lojistik firmaları ile çalışmaya başlama ihtimaline karşı müşteri kaybı tahmininde de makine öğrenmesinden faydalanılmaktadır.

3.3.1.4 Satış ve Pazarlama

Makine öğrenmesi hangi ürünün hangi dönemlerde satış hacminin artacağını ve potansiyel müşteri taleplerini belirleyerek satış tahmini yapılabilmesi açısından önemlidir. Müşterilerin satın alma geçmişine bağlı olarak alışveriş alışkanlıklarını öğrenerek ürün önerilerinde bulunurken, ürün fiyatlarının belirlenmesinde de etkin rol oynamaktadır. Müşterilerden gelen geri bildirimlerin raporlanarak ürünlerin geliştirilmesi için pazarlamacılara yol gösterilmesinde de makine öğrenmesinden faydalanılmaktadır. Mevcut müşterilere yeni ürün satışının daha kolay olmasından dolayı daha önce alışveriş yapan müşterilerin satın alma geçmişi analiz edilirken makine öğrenmesinden faydalanılarak müşteri tarafından satın alınabilecek ürünlerle ilgili e-postalar gönderilmektedir. Müşteriler tarafından satın alınmak istenilen bir ürünün hangi saatler arasında talep edileceğini tahmin ederek satış fırsatları yaratılması konusunda da makine öğrenmesinden yararlanılmaktadır. Belirlenen saatlere göre fiyatlar üzerinde yapılan değişiklikler rekabet avantajı yaratılmasını sağlarken aynı zamanda şirketin kar elde etmesinde de rol oynamaktadır.

3.3.2 Makine Öğrenmesinde Genel Yöntemler

Makine öğrenmesi genel anlamda, kullanılan veriye göre bazı ayarlamalar sonucu öğrenen ve öğrendiği bilgiye göre kendisi çıkarımlar yapan ve performansı gittikçe iyileştirilebilen yapay zekanın bir alt kümesi şeklinde tanımlanabilir. Yapay zeka, aynı insan zekası gibi fakat başlangıcını insanların yaptığı fakat sonrasında makinelerin öğrenerek ya da bağlam kurarak geliştirdiği ve devam ettirdiği kapsamı geniş bir kavramdır. Makine öğrenmesi ve yapay zeka çoğunlukla karıştırılarak birbirinin yerine kullanılsa da aslında her makine öğrenmesi görevi bir yapay zeka alt kümesi olmasına rağmen her yapay zeka olgusu bir makine öğrenmesi olmamaktadır. Makine öğrenmesi ya da yapay zekanın bu çağda

kullanılmadığı alan kalmamıştır. Banka, sağlık, alış veriş, ticaret, eğitim gibi tüm alanların bir bölümünde mutlaka kullanılmaktadır.

Makine öğrenmesi genel anlamda işleyiş şekli olarak iki ana gruba ayrılmaktadır. İlk ana grup gözetimli öğrenme, ikinci ana grup ise gözetimsiz makine öğrenmesidir.

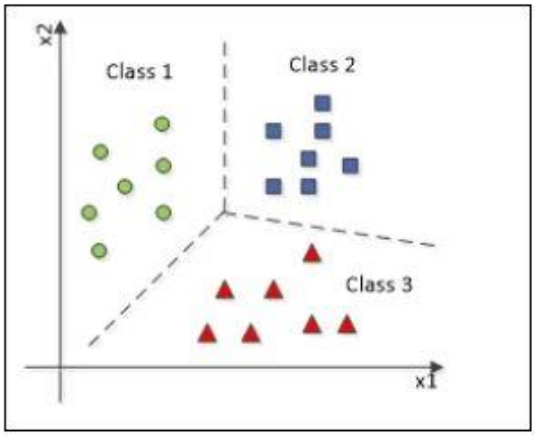
3.3.2.1 Gözetimli Öğrenme

Gözetimli makine öğrenmesi (supervised learning), makine öğrenmesinin genel mimarisi gereği öncelikle eğitim veri setindeki bilgileri öğrenerek bir sonuca varmaya çalışmaktadır. Eğitimdeki elde etmeye çalıştığı bilgi etiketteki bilgidir ve diğer tüm alanları ve bilgileri bu etiket bilgisine ulaşmak için kullanmaktadır. Gözetimli öğrenmede, sonucun önceden belli değerlerden biri olması istenilen durumlarda denetimli öğrenme kullanılmaktadır. Çıktı veri deseni sistem tarafından bilindiği için diğer tüm bilgileri bu etiket için bağ kurmaya çalışarak aradaki ilişkiyi kendisi bulur. Denetimli öğrenmede ham veri kümesinin kullanılabilir olması için öncelikle ETL, eksik veri temizleme, uç değerlerin düzenlenmesi (çıkarılması ya da sınıflandırma çalışmaları) gibi özellik mühendisliği işlemleri gerçekleştirilir. Sonrasında veri seti eğitim (training) ve test olmak üzere bölünür. Algoritme veya modele göre aşamalardan geçen eğitim seti başarısı test setinde ölçülür. Gözetimli öğrenme sınıflandırma ve regresyon olmak üzere iki ana gruba ayrılmaktadır.

3.3.2.1.1 Sınıflandırma

Sınıflandırma (classification) denetimli öğrenme alt modüllerinden birisidir. Sınıflandırma, evet & hayır, spam & spam değil, elma & armut, kadın & erkek, genç & yaşlı, satın alır & almaz gibi en az iki veya daha fazla gruptan oluşan modüldür. Gündelik hayatta karşılaşılan en meşhur sınıflandırma problemleri; e-mail spam sınıflandırması, banka müşterilerinin kredi alıp almama sınıflandırması, kanser hücrelerinin sınıflandırılması, duygu analiz sınıflandırılması, araçlarda yaya & hayvan gruplarının sınıflandırılması vb.

Sınıflandırma da kendi içerisinde ikili ve çoklu sınıf olmak üzere ikiye ayrılır. Sınıflandırma grubunda sinir ağları (neural networks), Naif Bayes (naive bayes), destek vektör makinesi (SVM), k- en yakın komşu (k-nearest neighbors KNN), rastgele orman (random forest) modelleri en çok kullanılan modellerdir.



Şekil 2: Sınıflandırma

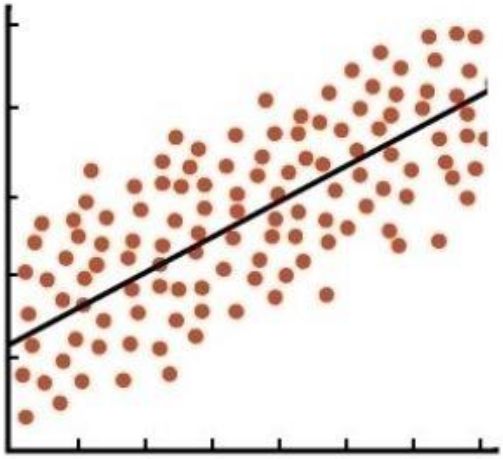
Şekil 2'de görüldüğü gibi sisteme tanıtılan veri setine göre 3 farklı sınıf olan bir modül işlenmiştir. Bu ayırım daha önce etiketlenmiş sınıflardan oluşmuştur. Daha fazla etiket grubu olursa örneğin class1, class 2 ...class n, model n tane sınıfa ayıracak demektir. Eğitim aşamsından sonra verilen test verisinde her bilgi hangi sınıfa ait olduğu düşünülürse o etiket atanır ve gerçekte doğru olup olmadığı kontrol edilir. Başarı hesaplaması tüm test veri seti bitiminde hesaplanmaktadır.

KNN

K değeri sisteme tanıtılan ve en yakınında bakması gereken komşu sayısını vermektedir. Model, yeni bir veri geldiği zaman geçmişteki verilerden öğrendiği özelliklere bakarak en yakın benzettiği grubun etiketin atamasını yapmaktadır.

3.3.2.1.2 Regresyon

Regresyon (regression), süre gelen bir durumun sınıflandırılmasında kullanılır. Örneğin bir ev satışında evin satılabilmesi ya da satılamması iki sınıflı bir karar ağacı ile çözümlenebilirken, metrekare başına değişen fiyat ile satışa sunulması modeli bir regresyon problemine dönüştürür. Sınıflandırma problemi ayrık değerlere odaklanırken regresyon genel anlamda sürekli değer tahminlerinde kullanılır. Ev fiyat tahmini, borsada hisse senedi değer tahmini, makinenin bozulabilmesi için kalan zaman gibi örnekler verilebilir. Karar ağaçları burada da kullanılabilir. Sınıflandırma modelinde düzensiz, zamansız, sırasız veriler ile çalışıbailirken regresyon modellerinde düzenli veri olmalıdır. Regresyonda, karar ağaçları (decision tree), sinir ağları (neural networks), modelleri en çok kullanılan modellerdir.



Şekil 3: Regresyon

Şekil 3 de'de görüldüğü gibi sınıflandırmanın aksine regresyon zamana & duruma göre değişen yani sürelen bir durum için kullanılır.

Karar Ağaçları

Decision tree, gerçek bir ağaç yapısından esinlenilerek oluşturulmuş, sistemin verileri ufak düğümlere ayırdığı bir yapıdır. Her düğüm birden fazla alt sınıfa ayırılır ve tüm düğümlerin, dalların bağlandığı ana düğüm (kök) en üstte olacak şekilde aşağı doğru büyüyen bir sistem oluşturulur. Hem sayısal hem de katergorik verilerin olduğu problemlerde kullanım kolaylığı sağlamaktadır.

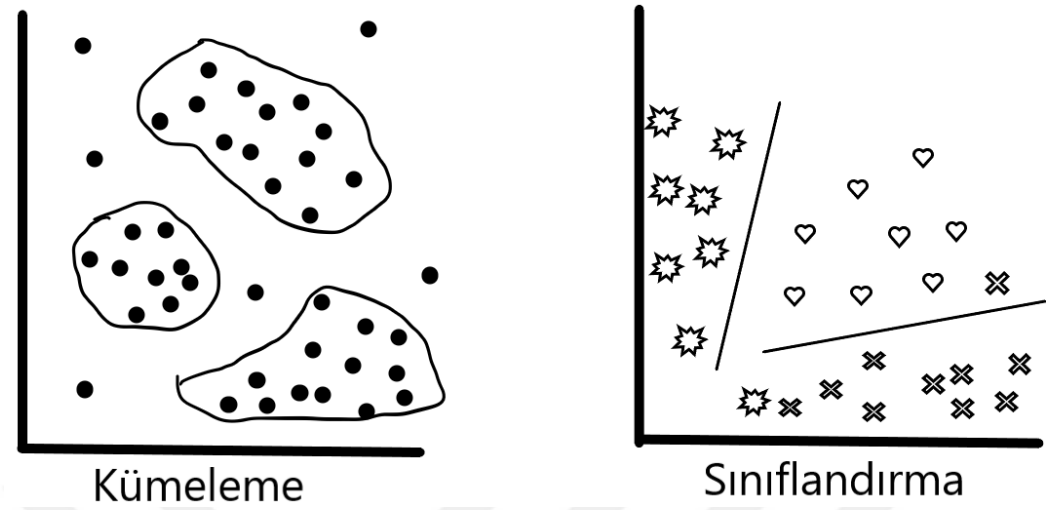
3.3.2.2 Gözetimsiz Öğrenme

Gözetimsiz öğrenme (unsupervised learning), girdi verileri etiketle belirlenmez ve öğrenme bu şekilde gerçekleştirilmez. Bu öğrenme çeşidinde denetim olmayacağı için model kendi başına bağlantılar çıkararak sonuca varmaya çalışmaktadır. Gözetimli öğrnemeye göre daha çok karmaşık ve zor problemlerin çözümünde kullanılabilir. Sistemde geçmişten gelen bir patern eğitimi olmadığı için kendisinin çözmeye çalışması daha zorlaştırır fakat her zaman veri kaynağında sonuçların tamamen belirlenmiş olması mümkün olmayacağı için bu yöntem bir çok durumda kurtarıcı niteliğindedir. En bilindik alt modeli kümelemedir.

3.3.2.2.1 Kümeleme

Kümeleme (clustering) modeli, kategorize edilmemiş veri setlerinde patern, veya ortak modeller bulmaya çalışan bir yapıya sahiptir. Kümeleme sınıflandırmanın aksine yapıyı inceleyerek veriyi anlamaya çalışır. Veri kümesinden öncesinden verilmiş hiç bir gruplayıcı ya da kategorileyici bilgi olmasa dahi model kendi başına verilerin arasındaki benzerliklere ve farklılıklara göre gruplamalar

yapmaktadır. Dezavantajlarından birisi, öncesinde bilgisi bulunmadığı için çıktıların kesin doğruluğu bilinmemekte ya da başarı ölçümü için zorluk çekilmektedir.



Şekil 4: Sınıflandırma ve kümleme

Şekil 4'te de görüldüğü gibi kümelemede kesin sınırlar ve belirlenmiş etiketler yoktur. Tüm verilerin arasındaki ilişkiye bakarak aralarında en benzer olanları gruplar. Bu benzerlikleri de yine model kendisine verilmiş tüm bilgilere bakarak ortaya çıkarır. Sınıflandırmada yapılan işlemin tam tersi işlem yapılmaktadır. Sonucu yani hangi grupların, kaç grubun olacağı baştan modele verilebilir ve o gruplara atama yapıması için özellikler test edilir.

Kümeleme daha çok tıp, biyoloji, sosyoloji, psikoloji gibi kalitatif alanlar olarak nitelendirilebilecek işler için kullanılmaktadır.

3.4 Derin Öğrenmede Genel Yöntemler

Derin öğrenme kavramı bir nevi makina öğrenmesi kavramından sonra popüler olmaya başlamıştır. Nedenlerin en başında makina öğrenmesinde kullanılacak veri boyutlarının büyümesi ve model eğitimlerinin bu boyut ile yetersiz kalması, modellerin daha fazla alanda kullanılmak istenmesi ve modelden daha fazla başarı beklentisi gibi durumlar derin öğrenme popülaritesini arttırmıştır. Video, film, müzik, grafik gibi ham veri seti ya da veri kaynağının artık bilgisayarın normal işletimcisinde işlenmesi ve makinaya bir olgu öğretilerek değerli bilgiye dönüştürülmesi çok güç ve maliyetli olmasından dolayı büyük veri kaynağı ile proje yapılmak istendiğinde akla ilk olarak derin öğrenme yöntemlerini getirmeye başlamıştır.

Yapay sinir ağları modelleri canlılardaki sinir ağlarından esinlenerek çıkarılmıştır. Veri, algoritmalar ile girdi katmanında ilk olarak modelle ilgili işleme başlattıktan sonra gizli katmana geçer ve orada probleme uygun başka işlemler yapılarak en son çıktı katmanına iletilir. Burada basitçe bir katman geçişi yerine her katmanda farklı öznitelikler eklenir ve evrilerek diğer katmana iletirilir.

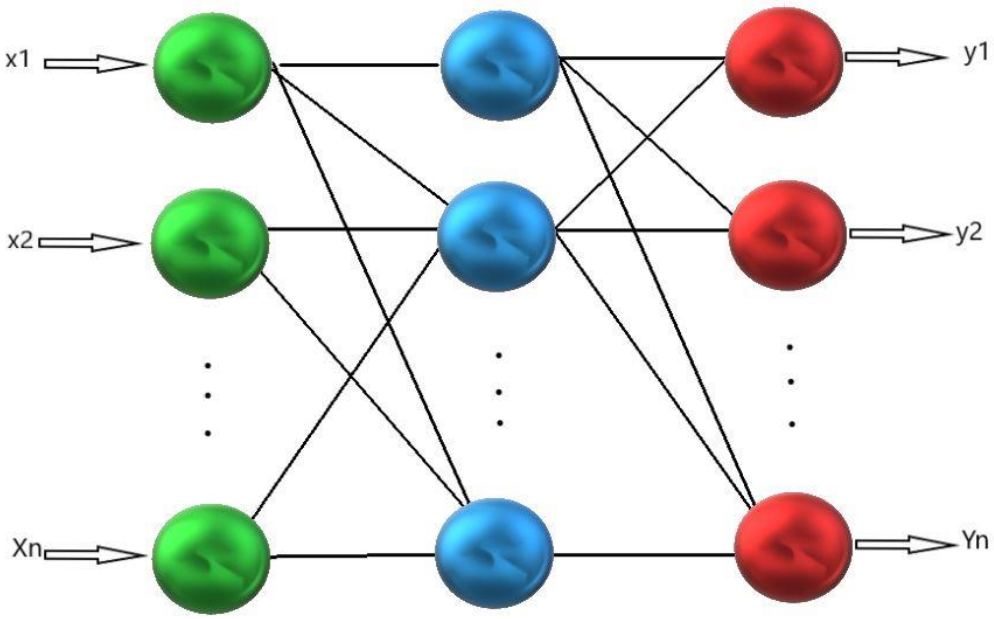
Yapay sinir ağları son dönemde popüler olmaya başlasa da aslında tarihçesi oldukça eskidir. İlk olarak literatürde geçmeye başladığı tarihler 1940'lı yıllardır (McCulloch, Pitts 1943: 115-133). Bellek yetersizliği ve bağlantılı olarak hesaplama gücünün yetersizliği sebebiyle teknoloji gelişmeye başladıktan sonra bu kavram daha çok kullanılabilme imkanı bulmuştur.

Yapay sinir ağlarındaki gizli katman sayısının artması, katmanlar arasında geçişte daha çok öznitelik kazanılmasını sağlamaktadır ve bu sayede aslında derin öğrenme modeli oluşmaktadır. Derin öğrenme ağları istenilen şekilde tasarlanabilmekte ya da önceden öğrenmesi gerçekleştirilmiş model kullanılarak yeni geliştirmeler yapılabilmektedir.

Bazı derin öğrenme genel modelleri bu tez kapsamında tanıtılacak fakat Transformatör'ün alt modeli olan Bert modeli kullanılacaktır.

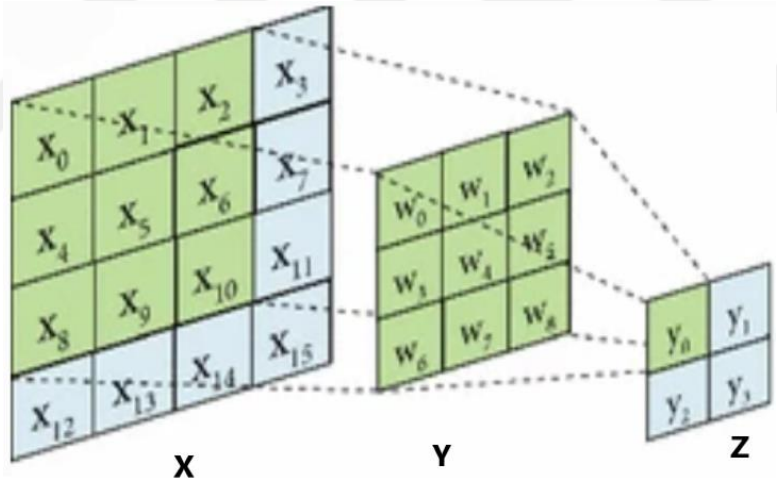
3.4.1 Evrişimli Sinir Ağları (CNN)

Literatürdeki ingilizce ismi Convolutional Neural Network (CNN) olan evrişimli sinir ağlarında yapay sinir ağına verilen girdi verisi çıktı seviyesine kadar tek yönlü iletilmektedir. Yapısında en az bir düğüm bulundurduğu ve girdi verilerinin ağırlıklar ile çarpıldığı ve sonrasında aktivasyon fonksiyonuna aktarıldığı bu tarz ileri beslemeli ağlara algılayıcı (perceptron) denilmektedir (Mikolov vd., 2013: 3). İleri beslemeli ağlarda minimum bir girdi, bir veya daha fazla gizli ve bir de çıktı katmanının olmasına çok katmanlı algılayıcı (multi layer perceptron – MLP) ismi verilmektedir (Konakoğlu, 2020: 704). Tek katmanlı yapılar daha çok doğrusal problemlerin çözümünde kullanılabilirken, çok katmanlı algılayıcılat doğrusal olmayan, karmaşık problem çözümlerinde de kullanılabilmektedir.



Şekil 5: Çok katmanlı algılayıcı (MLP) mimarisi

Ezberleme (overfitting) problemine çözüm bulunmuş yapısı olan evreşimli sinir ağlarının çok katmanlı algılayıcı yapısına sahiptir.



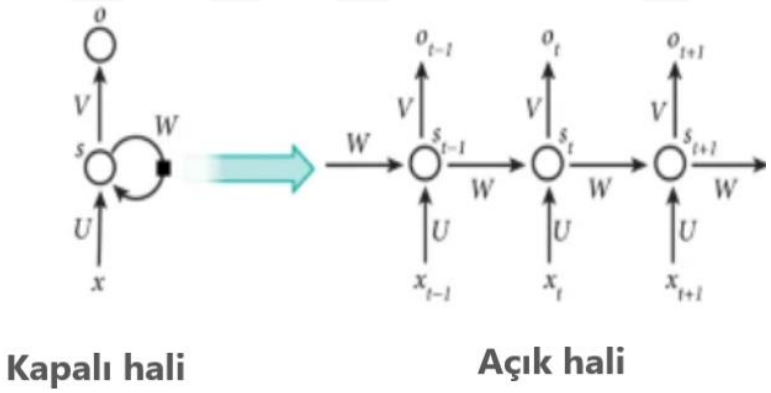
Şekil 6: CNN matrisleri

Şekil 6'da (Mosser vd., 2018) Evreşimli sinir ağının matrix sembolüyle gösterilmektedir. Girdi matrisindeki tüm değerler öznitelik haritası için kullanılır. Şekilde görünen yeşil renkli y matrisi X üzerinde kendi büyüklüğü şeklinde gezerek bu değerler Y matrisine çıktı öznitelik haritası olarak yazdırır. Makine öğrenmesi yöntemlerinde girdi matrisinin boyutu döküman sayısına, kelime sayısına, pencere boyutuna göre değişirken evrişimli sinir ağlarında bu boyut "döküman sayısı x kelime özyerleşiği uzunluğu" boyutunda olmak üzere sabitlenir (Uçan, 2020: 19).

Derin öğrenmenin alt dalı olan evreşimli sinir ağları aslında genellikle görsel ve işitsel veri analizlerinde kullanılır. Resim ve videodaki tanıma işlemleri, sınıflanadırma işlemleri, tıbbi görüntüleme işlemleri en çok kullanılan alanlardan bazırlarıdır.

3.4.2 Tekrarlayan Sinir Ağları (RNN)

Daha önce bahsedildiği gibi video, kelime dizisi, ve biribirne bağlı arka arkaya gelmiş veriler RNN (Recurrent Neural Network) mimarisini oluşturmaktadır. Geleneksel sinir ağlarında CNN'de dahil, bir önceki cümle, metin hafızasında kaydedilmediği için neden sonuç gibi ya da bağlantılı olma durumu ölçümlendirilemez. Tekrarlayan sinir ağları olarak literatüre kazandırılmış sinir ağı modeli ise bu soruna çözüm getirmiştir. Diğer klasik yapay sinir ağları her giridyi tek başına ve diğerleriyle ilişkisine bakmaksızın değerlendirir. Diğer metin, görüntü, ses vb. ile ilişkisine bakılmaması demek aslında zaman serisi problemlerinde çözüm oluşturamaması demektir. Cümle öğelerinin sıralaması önemlidir, tam ve doğru anlamı bu sırayı bir bütün olarak değerlendirildiğinde çıkmaktadır. Kelimelerin birden fazla anlamı olduğu için diğerleri ile birlikte durumları ve farklı anlamları ifade eder. Bu yüzden hafızası olan ve bağlantılılığı ölçeklendirebilen sinir ağlarına ihtiyaç doğmuştur. RNN isminden de anlaşılacağı üzere, mimarisinde metindeki her öğe için önceki çıktılara bağlantılı olarak yeniden ve yeniden çalışan bir sistem kurulmuştur.



Şekil 7: RNN mimarisi

Şekil 7'de (LeCun vd., 2015: 442) sol taraftaki RNN'in kapalı hali sağda ise kapalı halinin iç işleyiş yani açık versiyonu görülmektedir.

t – zamanı,

Xt – girdi sekansını

Ot – çıktı sekansını

W – ağırlıkları göstermektedir.

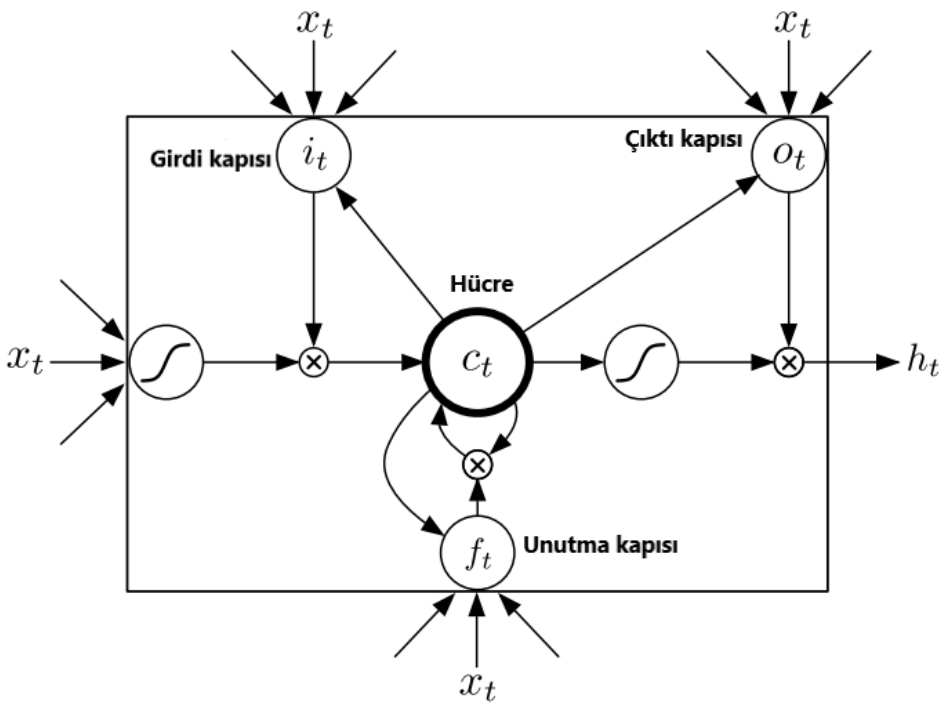
Sol taraftdaki şekilde görülen W Siyah kutucuk önceki durumlardan gelen bilgiyi bulunduran kısımdır. RNN mimarisinde giriş, gizli katmanlar ve çıkış katmanında kendi katmanına özgü yapıların ağırlıkları oluşturulur ve hepsinin eşik değeri farklıdır. Avantajı, sistem sürekli geri dönük işlem yaptığı için hata paylarını indirgemeyi hedefler. Bir önceki adımdaki bilgi ile yeni katmandaki bilgi birleştirilmesinden dolayı aradaki bağ kurulmuş olur.

RNN mimarisi gereği birçok alanda kullanım kolaylığı sağlamaktadır. Özellikle önceki katman bilgisi ve yeni katman bilgisi birleştirildiği için kelime tahmininde yüksek başarı elde edilmektedir. RNN kullanılan diğer örnek alanlar şunlardır:

- Görüntüye altyazı atama; film, dizi veya haberlerde konuşmanın altında altyazının belirmesi. Böylece farklı dildeki konuşmalar da eş zamanlı olarak alt yazı ile istenilen dile çevirilerek gösterilebilmektedir.
- Vektörü diziye çeviren sistem (vector to sequence); resim, video vektör olarak alınarak bu vektörlere metin karşılıkları atanmaktadır. Bu sadece çok büyük hafızalara ihtiyaç duyulmamakta ve maliyeti düşürmektedir.
- Bir metin alarak metin çıktısı (sequence to sequence) veren sistem; Metin çevirileri bu alanda en çok kullanılan örnek. En popüler kullanım örneği google translete'tir. Verilen metnin dilini ayırt ederek istenilen dile çeviri yapabilmektedir.
- Borsa tahmini; RNN'nin bazı eksik kalan taraflarını transformer modeli gidermektedir. Tez kapsamında Transformer başlığı altında açıklanmıştır.
- Ses ayırt edebilme, zaman serisi bazlı anormali tahmini gibi diğer alanlarda da kullanımı yaygın alanlar da diğer alanlar arasında örnek verilebilmektedir.

3.4.3 Uzun Kısa Süreli Bellek ağları (LSTM)

Long Short Term Memory (LSTM), giriş çıkış kapılarına ek olarak bünyesinde özel unutkan kapıları bulunduran özel bir Tekrarlamalı sinir ağları (RNN) türüdür (Işık, Yağcı, 2020: 35) Bu yapı RNN'de oluşan bazı dezavantajları kaldırmak için bellek hücrelerinin eklenerek oluşturulmuş bir özel mimaridir.



Şekil 8: LSTM hücre yapısı

Şekil 8'de LSTM hücre yapısı (Graves vd., 2013: 6646) görülmektedir. Yapı içerisindeki işlemler aşağıdaki gibi formülize edilmiştir.

$$i_t = \sigma\,(W_{xi}\,X_t + W_{hi}\,h_{t-1} + W_{ci}\,C_{t-1} + b_i\,)$$

$$f_t = \sigma\,(W_{xf}\,X_t + W_{hf}\,h_{t-1} + W_{cf}\,C_{t-1} + b_f\,)$$

$$c_t = f_t\,c_{t-1} + i_t \tanh\,(W_{xc}\,X_t + W_{hc}\,h_{t-1} + W_{ci}\,C_{t-1} + b_c\,)$$

$$o_t = \sigma\,(W_{xo}\,X_t + W_{ho}\,h_{t-1} + W_{co}\,C_t + b_o\,)$$

$$h_t = o_t \tanh(c_t)$$

T zamanında hücreye giren input Xt, lojistik sigmoid fonksiyonu σ, girdi İt, çıktı Ot, unutma Ft kapıları, sırasıyla aktivasyon, ağırlık, gizli, bias vektörleri Ct, W, h, b'dir.

Birbirine bağlı ve birbirinin devamı niteliğinde olan bilgiler LSTM'in bellekli mimarisi sayesinde avantajlı bir şekilde işlenebilmektedir. Örneğin bu durum, bir metinde genel bağlamın çıkarılmasına olanak sunmaktadır.

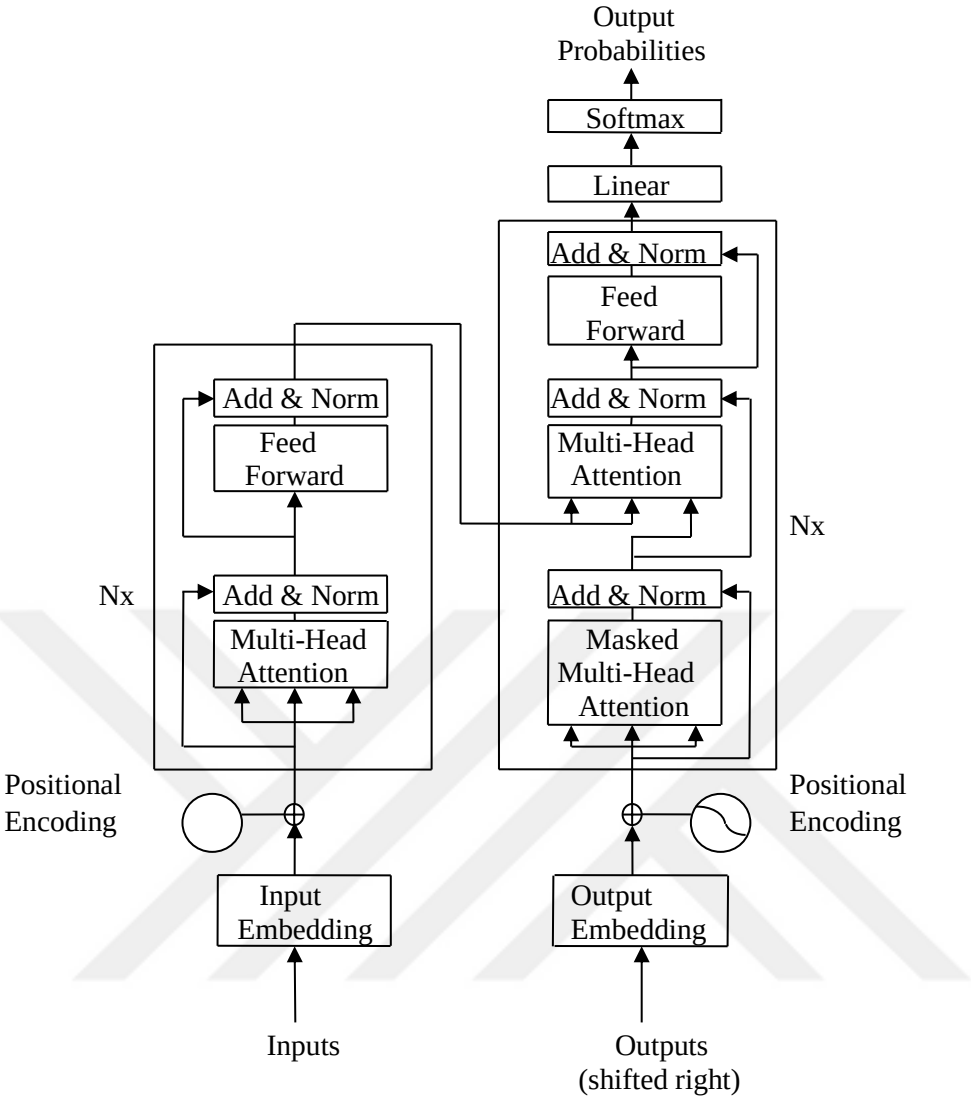
LSTM tek yönlü ve çift yönlü çeşitleri bulunmaktadır. Çift yönlü (bidirectioanal) LSTM, iki yönlü işlem yapabilmesinden dolayı metindeki anlamı çıkarma, soru cevaplama, sınıflandırma, öğrenme başarısı gibi alanlarda daha başarılıdır.

3.4.4 Transformer

Transformers'ın yapısı için ilk bahsedilmesi gereken önemli özelliği metne bir bütün olarak tek seferde bakmasıdır. Anlamı, herhangi bir yön bazlı yani soldan sağa ya da sağdol sola şeklinde incelemek yerine aynı anda çift yönlü incelemesidir. Bu yönden evrişimli sinir ağları (CNN) ve tekrarlayan sinir ağlarına (RNN) göre başarısı daha yüksektir. Transformer yapısı bilhassa büyük boyutlu girdiye (metine) sınıflandırmasına, ön eğitimine, özetlemesine oldukça uygundur ve bu alanlarda başarılıdır (Yang vd., 2019: 2). Transformer'ın gittikçe artan alan kullanımıyla birlikte komplex yapısı gelişmektedir. Bu sayede hassas problemlerin çözümünde RNN ve CNN e göre daha çok yapıtaşı olarak kullanılmaya başlanmıştır. Yine teknolojik gelişmenin bir sonucu olarak zaman, fiziksel koşullar ve işletim gücü gibi durumların tasarrufu için ön eğitimli model kullanımı artmıştır.

Transformer'ın bir çok alanda kullanılabilecek ön eğitilmiş, ayarlanmış, sıkıştırılmış mimarisi yaygın kullanımı için yardımcı olmuştur. Açık bir platform şeklinde olan transformers'ın herkesin geliştirmesine açık bırakılmış olması onun kullanımının artmasına neden olan faktörlerden biridir. Transformer mimarisi gereği her görev veya farklı problem grupları için içerisinde özel ayarlama yapılabilecek katmanları bulundurmaktadır. Bazı örnek modülleri (A lite Bert, Electra, DistilBert) başlıklar halinde açıklanmıştır. Bert modeli de bu alt modüllerden birisidir. Bu tez kapsamında Bert model kullanıldığı için daha detaylı açıklanmıştır.

Şekil 9'da (Vaswani vd., 2017: 6001) transformer mimarisi genel hatları gösterilmiştir.



Şekil 9: Transormers mimarisi

Transformer, iki ana temel olan kodlayıcı (encoder) ve kod çözücüden (decoder) oluşmaktadır. Kodlayıcıdan çıkan bilgi kod çözücüye iletilir. Yapı, göreve göre n tane giriş ve çıkış üretir. İlk iletilen metine göre ya da göreve göre girişte kaç adet olacağını kendisi belirlemektedir. Özet şekilde sırasıyla açıklanacak olursa;

- Girişte, kelimelerin gömme işlemi (embedding) yapılır ve kelimenin cümledeki pozisyonu belirlenir.
- Kelime ile ilgili ilk işlem bittikten sonra hangi kelimenin daha önemli olduğunu belirlemek için çok başlı dikkat mekanizması (attention) uygulanır.
- İşlemler sonucu her kelime için sıra, anahtar ve değer vektörleri belirlenmiş olur.
- Dikkat mekanizması aşamasından sonra giriş vektörü ile birleştirilerek normalizasyon aşamasına geçilir.

- Artık sinir ağının çıktı aşamasına geçilerek yine bu katmana da giriş ve çıkış işlemleri ve normalizasyon hesaplaması uygulanır (Sel, Hanbay, 2021: 679)

Genel aşamaları özet şeklinde yukarıdaki maddelerde açıklanmıştır. Transformer alt modelleri Electra, DistilBert, Albert ve tez çalışmasında kullanılacak Bert tanımları verilmiştir.

3.4.1 Electra

Bert ile oldukça benzer yapıya sahip olan Electra'da Maskelenmiş Dil Modelini kullanır ve girdi metnindeki bazı rastgele kelimeleri gizler (maskeler) ve metni yeniden yapılandırır. Başarısı Bert ile kıyaslanabilir fakat Bert gibi başarıya ulaşması için ondan daha fazla hesaplama işlemi yapmış olması gerekmektedir. Electra, gerçek ve sahte giriş şeklinde farklı belirteçleri ayırt edebilmektedir. Aslında modelin ana amacı da budur. Girişteki belirteçleri ayırt edebilmek için özel metin kodlayıcı eğitmektedir (Clark vd., 2020: 15) Bu bazı durumlarda maskelenmiş dil modelinden daha fazla başarı elde edilmesini sağlamaktadır.

3.4.2 DistilBert

Uzun ismi ile Distillation Bert aslında Bert modelinin damıtılmış halidir. Teknik olarak damıtma daha işlevsel, bellek gerektiren, detaylı çalışan bir modelin davranışının eğitim aşamasında kısıtlanması, bir kısmını almasıdır (taklit etmesi). BertLarge ile kıyas yapılmamakta fakat BertBase e göre parametresi oldukça azdır (yaklaşık 66M) (Sanh vd., 2019: 3). Damıtmadan dolayı eğitim süresi BertBase'e göre daha azalmış ve parametrelerdeki azalmadan dolayı da dil işleme başarısı biraz daha düşmüştür.

3.4.3 A Lite Bert (Albert)

İsminden de anlaşılabileceği gibi, bellek yükünün azaltılması amacıyla Bert'e göre daha az parametreye sahiptir. Bert gibi önceden eğitilmesi gereken modellerin ölçeklendirilmesindeki problem oluşturan iki parametreyi azaltma yöntemini bulundurur. İlki, gömme (embedding) aşamasındaki çapraz gömme şeklindeki parametre değişikliğidir. İkincisinde ise katmanlar arasındaki parametrelerin paylaşılmasıdır. Bu tekniğin avantajı ağın büyümesiyle birlikte parametrenin de büyümesini engelleyebilmektir (Guven, 2021: 630).

3.4.5 BERT

Google tarafından 2018 de bulunan BERT, Bidirectional Encoder Representations form Trasnformers kelimelerinin baş harflerinden oluşmaktadır. Bert modeli diğer klasik doğal dil işleme modellerinden ayıran özelliklerinden birisi de cümleyi öncelikle soldan sağa işlemesi ve sonrasında sağdan sola yeniden işlemesidir. Bert bulunduğunda ilk olarak ingilizce iki model oluşturulmuştur. BertBase 110 milyon parametre, 800 milyon kelime hazneli, 12 katman'dan oluşan (layer), 768 saklı büyüklüğü (hidden size) olan BookCorpus kullanılarak ve BertLarge 340 milyon parametre, 2,5 milyar kelime haznesine sahip 12 katmandan oluşan (layer) 1024 saklı büyüklüğü (hidden size) olan Wikipedia gibi kelime haznesi çok yüksek olan kaynaklarda temel model eğitimleri sağlanmıştır. Bazı farklılıklar gösteren bu iki modelin birbirleriyle karşılaştırması yapılmıştır. Bert, yeni ve klasik doğal dil işleme modellerine kıyasla başarı oranı oldukça yüksek bir modeldir. Wikipedia'nın Aralık 2021 kelime parametre sayısı 4 milyar üzerindedir.

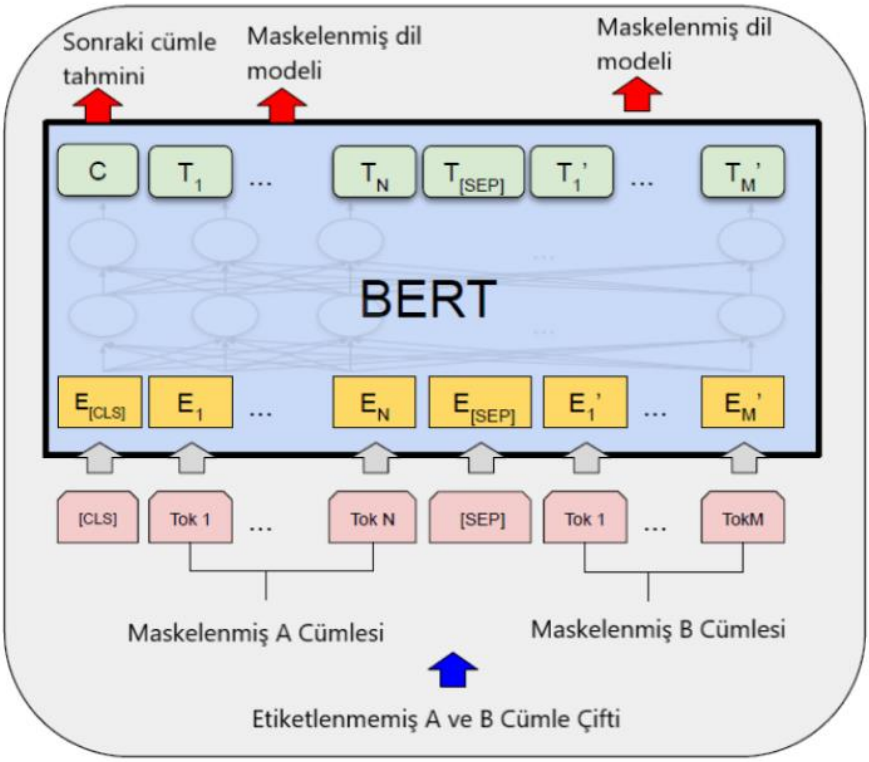
Bert sadece cümle çevirisi için değil aslında doğal dil işleminin yapılabileceği tüm alanlarda kullanılmak üzere geliştirilmiştir. Yukarı bölümlerde kısa açıklamalar ile anlatılan; soru yanıtlama, dil tercüme, bilginin bulunması ve getirilmesi, kelime işleme, arama motorları, metin ön işleme, argüman birleştirme, yazılı metinin sesli metine, sesli metinin yazılı metine çevirilmesi, hedeflenmiş reklamcılık, otomatik arama düzeltme ve tamamlama alanları, duygu analizi, yazı özetleme ve benzeri alanlar için probleme göre farklı katmanlar eklenerek Bert ile yüksek başarı ile çözümlenebilir.

Bert genel kapsamda iki ana bölümden oluştuğu söylenebilir. İlk aşamasında, yapının dilini ve bağlamı hakkında bilgi edinmeyi hedefler. İkinci aşamasında yani ince ayar aşamasında özel görevler/ problemler için model nasıl kullanılabilirin cevabını hedefler.

Diğer gözetimli öğrenen modeller gibi Bert'te de ilk aşama pre-train yani ön öğrenme aşamasıdır. Bu aşamada dili ve içeriğin ilk öğrenilmeye başladığı süreçtir. Diğer modeller gibi önce soldan sağa ya da sağdan sola şeklinde tek yönlü eğitim yaklaşımı yerine iki aşama kullanarak Bert modelinin eğitim aşaması yapılmaktadır. Bert modelinin yaratıcıları bu görevleri task 1 MLM aşaması ve task 2 NSP aşaması olarak adlandırmıştır.

Eğitim için ilk görevi MLM yaklaşımıdır. Bert, kelime dizinlerini bir bütün olarak almadan önce literatürde Masked Language Modeling (MLM) şeklinde geçen ve Türkçe karşılığı olarak Maskelenmiş Dil Modeli (MDM) olarak söylenebilecek ve 15% e karşılık gelen bir değerle cümledeki kelimeleri maskeler. Maskelenen kelimeleri dahil etmeden cümlenin geri kalanından anlam çıkararak bu boşta kalmış alanları kendisi tahmin etmeye çalışır. Daha yüksek yüzdelik maskelenirse cümlede çok kelime kaybı olmasından dolayı yerine yeni gelecek kelimenin tahmininin zor olmasından dolayı birçok deneme sonucu en iyi yüzdeliğin 15% olduğu belirlenmiştir. 15% e denk gelen rastgele seçilen

kelimelerin 80%, [Mask] kelime parçacığı (token) ile değiştirilmekte, 10% herhangi rastgele atanan bir kelime ile değiştirilmekte ve geri kalan 10%'luk kısım ise aynen orjinal kelimesi kalmaktadır (Devlin vd, 2018: 3). Buradaki kayıp değeri hesabı sadece maskelenmiş kelimeler için ölçülmüş olur. Sadece maskelenen kelimelerin tahmini cümledeki diğer kelimelere bağlı olduğu için diğer kelimelerin tahmin çalışması yapılmaz ve bu yüzden hesaplanma da yapılmaz. Bir cümledeki kelimelerle ilgili çalışılan bu aşamada kelimerin birbiri ile ilişkisi incelenmiş olur.



Şekil 10: Bert ön eğitim şeması

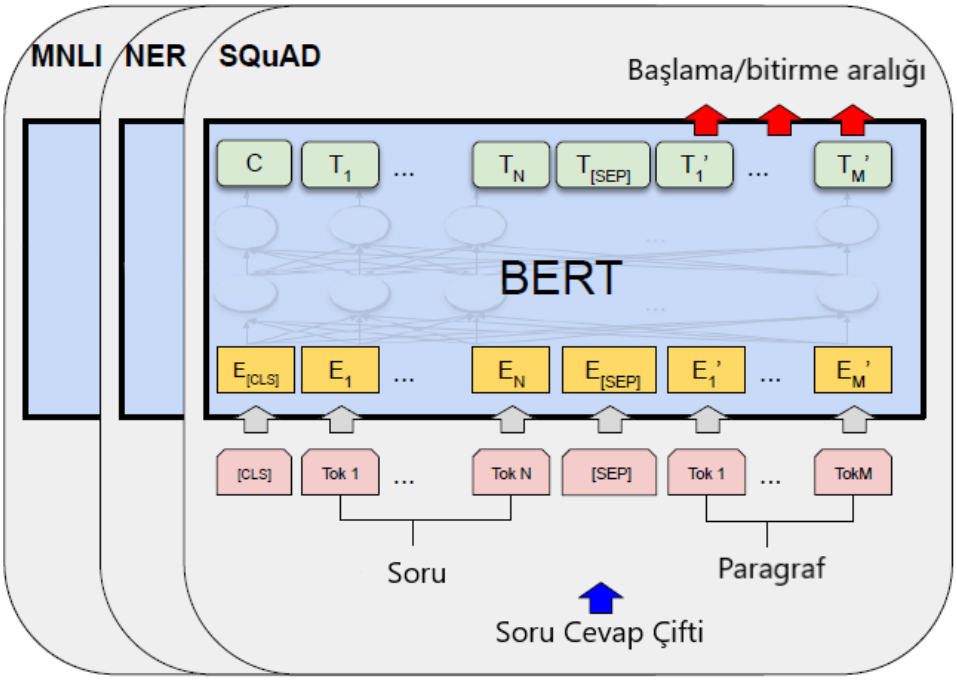
Bert modelinin yaratıcılarının makalesinden alınmış Şekil 10 da'da (Devlin vd, 2018: 3) görüldüğü gibi, etiketlenmemiş iki cümle alınır. Bu cümleler ile eğitimin ikinci aşaması olan Sonraki Cümle Tahmini (Next Sentence Prediction (NSP) için kelimelerin biribiri ile ilişkisi yerine cümlelerin birbirleriyle ilişkisine bakılır. Bu ilişkide cümlelerin birbirinin devamı niteliğinde olup olmadığı incelenir. Doğal dil işlemede klasik modellerin çözemediği bazı problemlerin çözümü Bert'te bu teknik sayesinde çözümlenmiştir. Yukarı segmelerde açıklanan soru yanıtlama, doğal dil çıkarımı, argüman birleştirme gibi problemlerde cümlelerin sırası, bağlamsal bütünlüğü, birbirleriyle ilişkisi gibi durumların yakalanması ve ona göre sonucun oluşması gerekmektedir. Bert modelindeki cümle ilişkilerini inceleyen sonraki cümle tahmini (NSP) tam olarak bu problem için bazı aşamaları içermektedir.

Ön eğitimin bu görevinde her seferinde rastgele alınan iki cümlede öncelikle birinci cümlenin 50%'si ve ikinci cümlenin 50%'si seçilir. Önceki modellerde sadece cümle yerleştirmeleri/ gömmeleri (embeddings) alt görevlere aktarılırken Bert son görev (task) parametresini başlatmak için tüm

gömülmüşleri (embeddings) aktarır. Ön eğitim çoğunlukla daha önceden de kullanılan dil modellerindeki ön eğitim literatürlerini takip eder. Buradaki fark Bert için kullanılan kitaplıklar/ öğrenme kaynakları yukarıda da detaylıca açıklanan BooksCorpus ve İngilizce Wikipedia'dır. Wikipedia için yalnızca yazılar dahil edilmişken listeler, fotoğraflar, tablolar ve başlıklar dahil edilmemiştir.

Devil ve arkadaşları (2018) modelin eğitimi sırasında uzun ve bitişik/ komşu dizilimleri çıkartabilmek için Billion Word Benchma gibi karıştırılmış cümlelerin olduğu belgeler yerine belgeler düzeyinde düzenli eğitim kaynağı kullanmanın önemini vurgulamıştır. Bu durumdan da anlaşılacağı üzere Bert gibi ön eğitim modellerine modelin ilk eğitim süreçlerinin manuel yapılması çok masraflı ve zaman alan kısıtlayıcı bir durumdur. Bu yüzden açık kaynak olan model aslında iki farklı kaynak ile bazı farklılıklarla eğitilmiştir. Problemin çeşidine göre bazı katmanlar eklenerek son eğitim aşaması son kullanıcı tarafından yapılabilir. Bu tez kapsamında ince görevi/ tekniği (fine-tunning) ile probleme uyarlamak için ön eğitim süreci sinema, finans yorumları ve duygu etiketlerini içeren veri seti ile yapılmış ve uygulama bölümünde detaylıca açıklanmıştır. Bert gibi bazı ön eğitimleri almış modeller bu şekilde probleme göre ince ayar görev (task) aşaması ile probleme yönelik daha komplex işlevlerde kullanılması sağlanmaktadır. Örneğin bir banka müşteri hizmet sisteminde kullanılması için ilk karşılamayı robotun yapması ve evet hayır sorularını sorması, kısa kelimelerle yönlendirme yapılacak birimi algılaması aslında belli bir sürelik konuşma kaydı alınarak eğitilmiş modele ince ayar yapılarak (fine tunning) yani ön eğitimde (pre-train) öğrenilmiş bölüme bazı özel katmanlar eklenmesi ile sağlanmıştır.

Gözetimli öğrenmenin ikinci aşamasını Fine-Tunning yani İnce Ayar aşaması oluşturmaktadır. Şekil 11'de (Devlin vd, 2018: 3), soru cevap çifti için aşamalar gösterilmektedir.



Şekil 11: Bert İnce Ayar Şeması (Fine Tunning)

Başlatma vektörü Bert ön eğitim şemasında E_1 olarak görülmektedir. T_n outputtaki kelime vektörleri anlamına gelmektedir. Bert ön eğitim şemasındaki gibi soru cevap için bir katman eklerken; Soru için A cümlesi gömmesi (embedding), cevap B cümlesi gömmesini (embedding) kullanmaktadır. Gömme (embedding) aşamaları bu aşamadan sonra aşağıdaki bölümde detaylıca açıklanacaktır. Soru cevap için eklenen katman şemada görüldüğü gibi ince ayar (fine tunning) görevinde eklenmektedir. Şemada görüldüğü gibi eğitim aşamasında soru ve bir paragraf veya bir metin verilir. Sorudaki anahtar kelimeler ve metindeki kelimelerden yola çıkarak cevap bu metinden bulunabilir.

İnce ayar (fine-tunning) kısmında sisteme sadece;

Başlatma vektörü $S \in \mathbb{R}^H$

Bitirme vektörü $E \in \mathbb{R}^H$ vermek yeterlidir.

i kelimesinin cevap aralığının başlangıcı olma olasılığı, T_i ve S arasındaki bir nokta hesabı ve ardından paragraftaki tüm kelimeler üzerinde bir softmax olarak hesaplanır. Softmax, gerçek değerlerinin toplamı 1 olacak şekilde vektöre çevirme işlevidir. (Softmax sisteme verilen kelimenin sınıfını bulan fonksiyon şeklinde de tanımlanabilir.)

$$P_i = \frac{e^{S \,.\, T_i}}{\sum_j e^{S \,.\, T_i}}$$

Tüm girdiler için aynı formül kullanılarak hepsi bu işlemden geçer. Konumu i ‘den j’ye kadar aday kelimenin puanı,

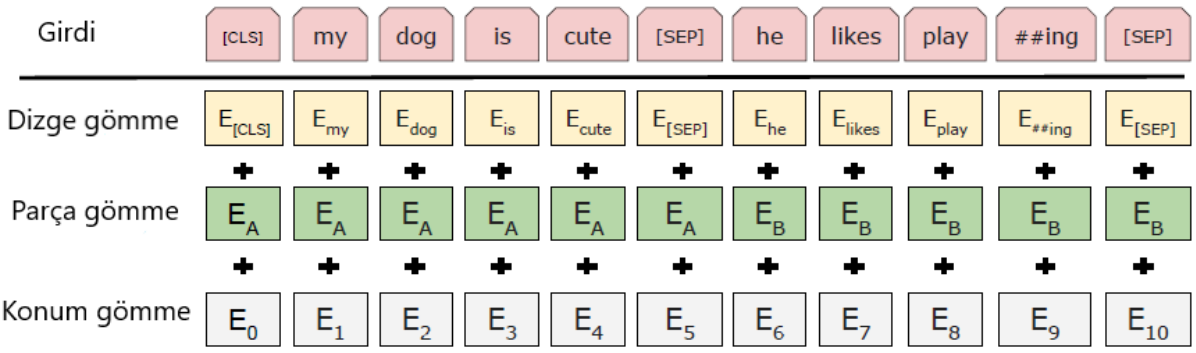
$$S \cdot T_i + E \cdot T_j$$

olarak tanımlanır. Burada j konumunun cümle içinde i konumundan sonra geldiği farzedilir (j ≥ i)

Eğitim katmanının hedefi doğru başlangıç konumu ile başlamış ve doğru bitiş konumu ile bitmiş olasılıkların log alınarak toplanmasıdır.

Şemanın üst sol köşesinde görülen C sonraki cümle tahmini için kullanılan sonraki cümle tahmininin eğitiminde elde edilmiş ince ayar aşaması (fine tunnining) haricinde anlamsız bir vektördür. İnce ayar (fine-tunning) aşamasında şemada da görüldüğü gibi input olarak alınan kelime vektör sayısı ile output olarak çıkan kelime vektör sayıları eşittir.

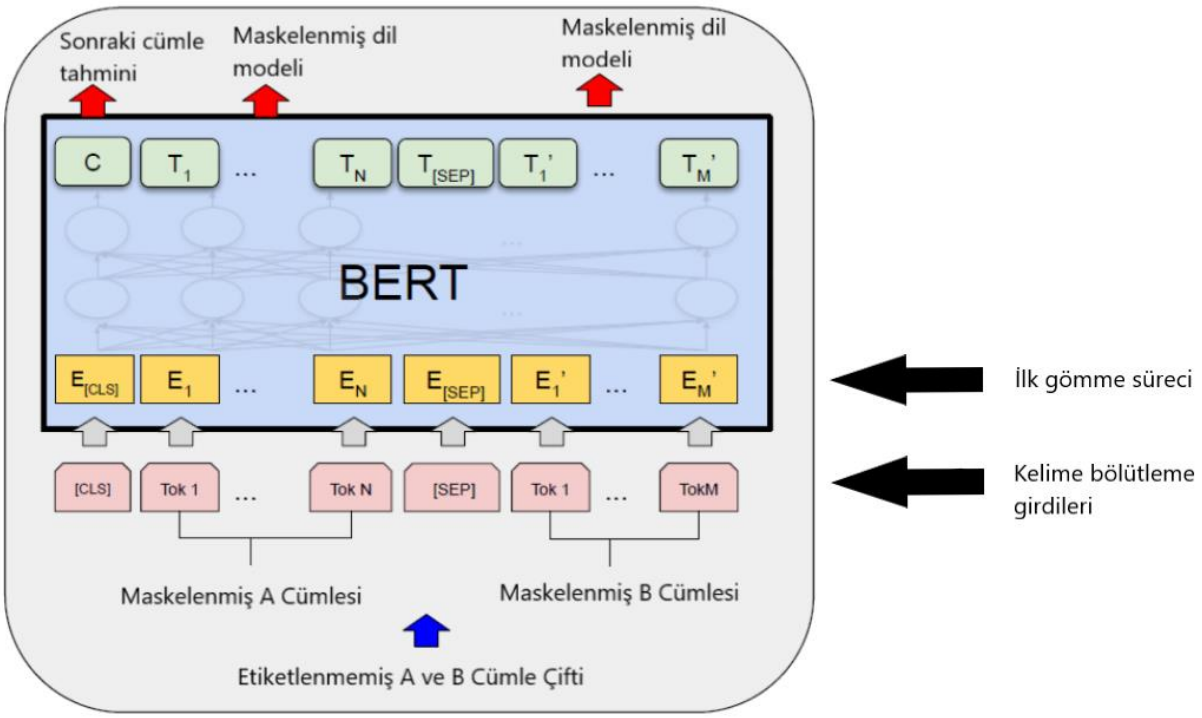
Öğrenme sürecindeki maskelenmiş dil modeli ve sonraki cümle tahmini aşamaları eş zamanlı yapılmaktadır. Bu durum birbirlerinden etkilenmesini engellediği için başarıyı arttırmakta ve zaman kaybı oluşturmamaktadır.



Şekil 12: Gömme (embedding) ve bölütleme (tokenization)

Şekil 12’de (Devlin vd, 2018: 5), gömme (embedding) ve bölütleme (tokenization) aşamaları görülmektedir.

Bert modelinin iki aşaması vardır. İlk aşaması olan ön öğrenme (pre training) aşaması kendi içerisinde maskelenmiş dil modeli ve sonraki cümle tahmini olmak üzere iki göreve ayrılmıştır. Sonrasında ince ayar (fine tunning) takip etmektedir. Gömme (embedding) şemasında da göründüğü gibi gelen girdi (input) başlatma vektörü ile dizge gömmelerine (token embeddings) çeviriler. Çıkış makalesinde belirtilen ilk Wordpieces (kelime tanıtma parçaları) 30 bin kelimeden oluşmaktadır. Gömmenin ikinci aşaması parça gömme (segment embeddings) aşamasıdır. Parça gömmede her kelime vektörlere çevirilir. Konum gömme (position embeddings) ise kelimenin cümle içerisindeki konumunun belirlendiği aşamadır.



Şekil 13: Bert maskelenmiş dil modeli

Şekil 13'de (Devlin vd, 2018: 3), Bert temel model ve gömme aşamaları gösterilmektedir. Gömme şemasında görünen dizge (token), parça (segment), konum (position) gömmelerinden (embeddings) sonra işlemin bitiminde yerleştirme vektörü elde edilmiş olur. Çıktı (output) daha önce bahsedilen C vetörü ile başlatılır. Minimum kayıp için kelime vektörleri aynı boyutta ve aynı anda oluşturulur ve bu ağlar birbirine bağlanır. Öncelikle maskelenmiş kelimelerin yerine 30 bin kelime dağarcığı olan WordPieces'teki kelimelerden tahmin edilen kelimeler alınmış olduğu için 30 bin nöronluk değer ile softmax katman değeri hesaplanır. Girdi ve çıktı sayısı eşit nöronlu olacağı için maskelenmemiş kelimelerle birlikte hesaplanmış arasındakine bakarak çapraz entropi kaybı (cross entropy loss) hesaplanır. Softmax hesaplaması kısaca formülü;

$\overrightarrow{z}$ girdi vektörü,

K sınıflandırma işlemindeki sınıf sayısı,

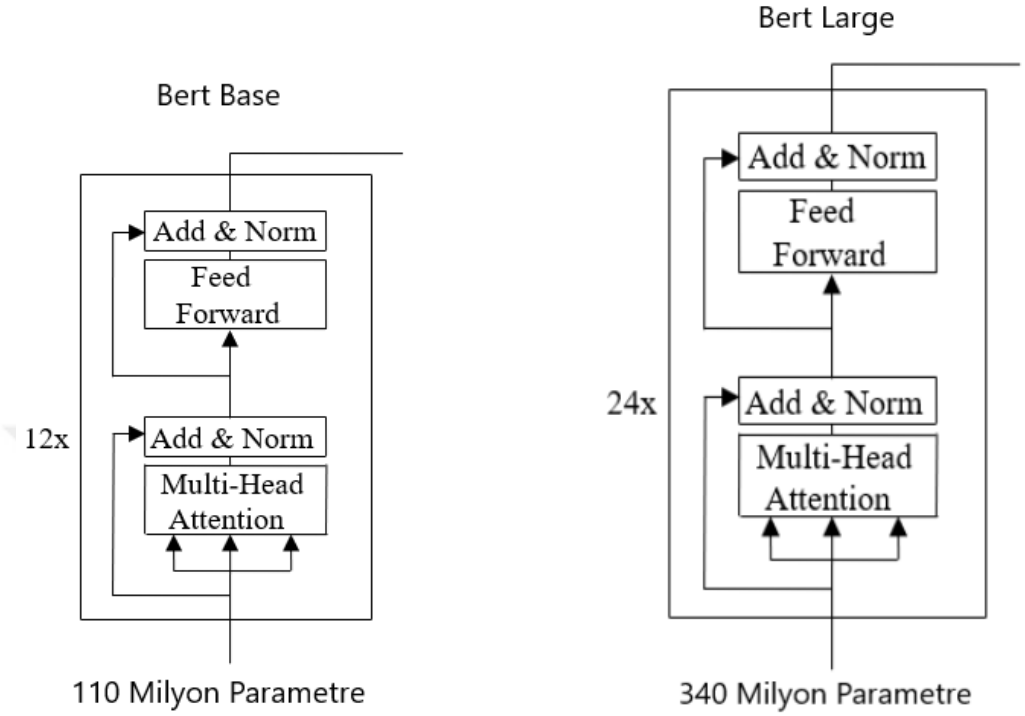
$\sum_{j=1}^{K} e^{z_j}$ değer aralığı (0-1) ve toplamı 1 olan değeri normalleştirme fonksiyonu,

e^{z_i} tüm elementler için uygulanan standart üssel fonksiyon

z_i yapay sinir ağlarında hesaplanmış değeri pozitif, sıfır veya negatif değere sahip olabilen girdi

$$\sigma(\overrightarrow{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Bert burada bahsi geçen görevleri ve aşamaları çok hızlı yapmaktadır. Örneğin, Bert ince ayar şemasında görünen SqQuAD Stanford Soru ve Cevap Veri Seti (Stanford Question & Answer Dataset) modülü bir bulutta 30 dakika ince ayar aşaması (fine tune) ve TPU kullanarak 91% F1 score elde etmiştir.



Şekil 14: Bert Base ve Bert Large

Şekil 14 'de (Devlin vd, 2018), bahsedilen 30 dakika eğitim aslında şekilde görülen Bert Base ve Bert Large gibi ne kadar parametreyle eğitildiği ve TPU gibi hızlı bir bulut sistemi kullanıp kullanmadığı gibi durumlara göre değişkenlik göstermektedir.

Bert modelinin en önemli avantajlarından birisi de kelimeyi konumununa göre değerlendirmesidir. Cümlenin hangi bölümünde hangi kelimeden önce ya da sonra geldiğine göre dikkati değişkenlik göstermektedir. Soru cevaplama katmanından anlaşılacağı gibi Bert modelinin katmanları probleme uygun özel çözümler eklenerek spesifik olarak sadece probleme yönelik ayarlanabilmektedir. Bazı örnek problemler ve ön eğitim kaynakları bu tezin ek bölümünde verilmiştir.

3.4.6 Neural Prophet

Neural prophet zaman serisi verilerini modellemek için üretilmiş sinir ağlarına dayalı python kütüphanesi modelidir. Pytorch üzerine kurgulanmış ve Facebook prophet şeklinde de literatüre işlemiş oldukça başarılı bir modeldir. Zaman serisi modelleri aslında isminden de anlaşılacağı gibi düzenli periyotlar halinde tekrar eden eylemlerin gözlemlenmesi ve gözlemler sonucu elde edilen bilginin kaydedilmesi ile elde edilir. Bu periyotlar videolardaki gibi salise boyutundan gözlemden başlayarak

saniye, dakika, saat, gün, ay, yıl şeklinde süreçler de olabilmektedir. Zaman serisi verileri çok çeşitli alanlardan olabilmektedir. Örneğin; hava durumu, hisse senedi değerleri, deniz kirliliği ölçümleri, gezegen hareketleri gibi zamana bağlı tüm gözlemler olabilmektedir.

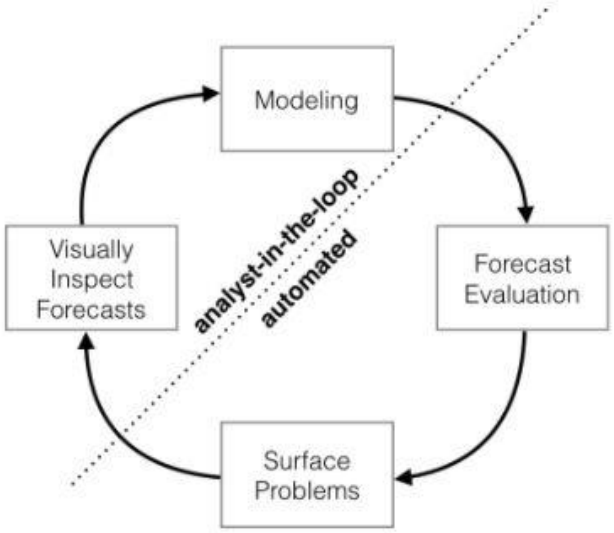
Neural prophet, Facebook, R, Python araçlarında direk açık kaynaklı olarak kullanılabilen bir model olduğu için bu tez kapsamında tercih edilmiştir. Tahmin projeleri çalışılması oldukça zor, kapsamlı ve ç0ok efor sarfetmek gereken projelerdir. Sosyal alanlarda sadece veri ile ifade edilmesi ve makine öğrenmesi yapılması başarıyı düşürmektedir. Bu gibi alanlarda tecrübe ve bilinen ve daha önce düzenli tekrarı olmamış durumların dışarıdan sisteme ek verilmesi ve sistemin yönlendirilmesi gerekmektedir. Çok esnek yapıda değillerdir ve bazen olağan üstü durumlara nadir tekrarlandığı için ayrık olay olarak algılar ve yeterince önem vermeyerek diğer durumlar gibi farz edilip sistem çalışılmaya devam ettirilir. Başarısı yüksek tahminler için deneyim, temiz ve kullanışlı veri kaynağı olmazsa olmazlardır. Prophet modelinin üretimindeki motivasyonda bu yönde olduğu için sadece üst düzey uzmanlar için değil alanda ilgisi olan ve veri seti olan herkesin kullanabileceği ve kaynağın ücretsiz olduğu bir model olmuştur.

Prophet hala Facebook tarafından kullanılan ve ürünlerin karar verilme aşamasında, ürünlerin özellikleri aşamasında, karmaşıklığı giderme konusunda hala başarılı ve geliştirilerek devam eden model olma özelliği taşımaktadır.

Tüm tahmin projeleri/ modelleri aynı yöntem ve aynı tarz veri setleri ile çözümlenemez. Prophetin kullanımında başarı elde edilebilmesi için minimumda birkaç aylık yada günlük ya da saatlik veya haftalık şeklinde tekrar eden düzenli alınmış veriler olmalıdır. Prophet ‘in diğer modellerden ayıran bazı diğer özellikleri (Letham, Taylor, 2017: 39);

- Sisteme daha öncesinde tanıtılmış fakat belli bir düzeni olmayan önemli tarihler. Örneğin seçim günleri, doğal afet günleri, tatiller, bayramlar, büyük mitingler gibi durumlar,
- Veri setinde büyük yüzdelik olmayacak şekilde aykırı (outlier) olabilecek değerler,
- Veri setinde büyük yüzdelik olmayacak şekilde eksik veriler (missing values),
- Normal gidişatı değiştiren olaylar/ eventler. Örneğin ürün tanıtım günleri, ansızın değişen hava durumuna göre kıyafet sezon değişikliği, konser & dizi gibi durumlar ile değişen ani moda akımı gibi durumlar,
- Normal akımın eğrisinin doyum noktasına ulaştığı ve artık doğrusallığını kaybettiği gelişmeler

Yukarıdaki maddeler prophet modelinin diğer modellerden daha iyi çalışmasını sağlayan ve hala geliştirilmesine devam edilen özel durumlardır. Model diğer modellere göre daha az manuel ayar yapılmasına ihtiyaç duyarak en az diğer modeller kadar başarı elde ettirebilmektedir.



Şekil 15: Neural Prophet

Şekil 15'de Neural Prophet analist şeması (Letham, Taylor, 2017: 38) yani prophet modelinin tahminleme sürecindeki çalışma prensibini göstermektedir.

Prophet tahminlemesi için sınırlı tool kullanımına açık aracı vardır; R'daki tahminleme paketi, Google CausalImpact ve Twitter'ın AnomalyDetection gibi zaman serisi bazlı paketleri. Prophet modeli ARIMA, üstel yumuşatma gibi istatistiksel model yapılarını barındırdığı ve kendi varsayılan (default) ayarlarında birçok otomatik belirlenmişlikler olduğu için uzman olmaksızın hızlı ve başarılı çalıştırılabilmektedir. Yumuşatma parametreleri (smoothing parameters) sayesinde mevsimsel geçiş için özel ayarlama, trend değişiklik süreçleri özel ayarlama, büyüme eğrileri için özel kapasite ayarlamaları, özel günlerin tanımlamalarının yapılabilmesi (seçim günleri, bayramlar gibi normal gidişatı değiştiren durumlar) gibi durumlara imkan tanıması Prophet'in diğer modellere göre bu tez için kullanılabilirliğini arttırmıştır.

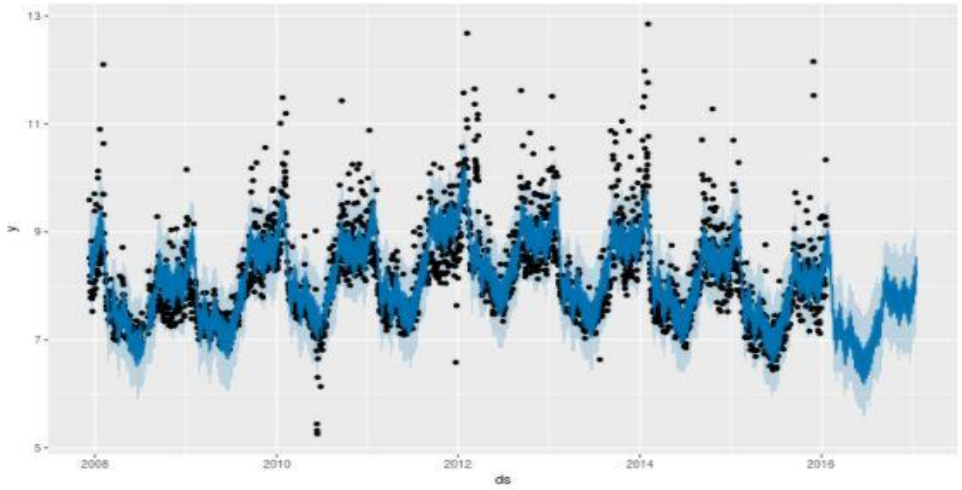
Neural prophet dört ana birleşene sahip eklemeli regresyon modeli olduğu söylenebilir (Letham, Taylor, 2017: 41);

- Doğrusal veya lojistik büyüme eğrisinden kendisine göre noktalar seçerek eğilimdeki trend değişikliklerini incelemesi ve bunları otomatikleştirmesi,
- Fourier serisi kullanarak yıllık mevsimsel birleşen modellemesi,
- Kukla (dummy) değişkeni kullanarak haftanın mevsimsel bileşen etkisini belirlemesi,
- Kullanıcı tarafından belirlenecek önemli tarihlerin sisteme girilmesine olanak sağlaması.

Bir örnek ile bu maddeler açıklanacak olursa;

Peyton Manning'in Wikipedia web sitesindeki günlük görüntüleme veri seti ile yukarıdaki maddeler şu grafiklerle açıklanabilir;

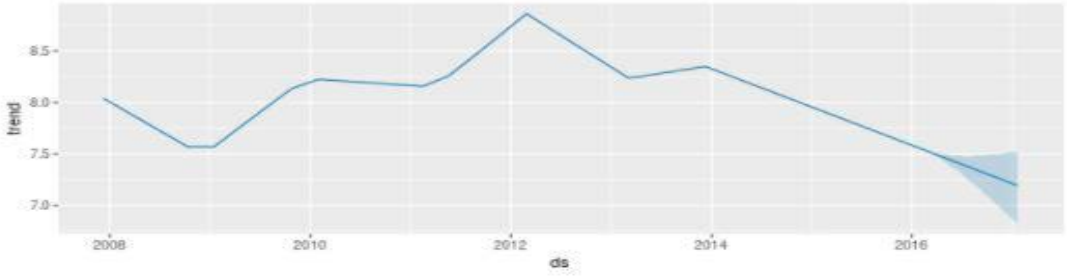
Grafik 1: Peyton Manning wikipedia sayfası erişim istatistik



Grafik 1'de Peyton Manning wikipedia sayfası erişim istatistikleri (Letham, Taylor, 2017: 42) paylaşılmıştır.

Grafikte Wikipedia'daki sayfasında görüntülenmelerin ve sonrasında da modelin tahmini görülmektedir. Futbol oyuncusu olduğu için dönemsel iniş ve çıkışları beklenmekte ve geçmiş görüntülenmelerde de bu istatistik gösterilmektedir. Prophet ile tahminleme yapılırken de bu trendlerdeki eğilimlerden örnek noktalar seçerek gelecek tahmininde bunlardan yola çıkarak tahmin oluşturulmuştur.

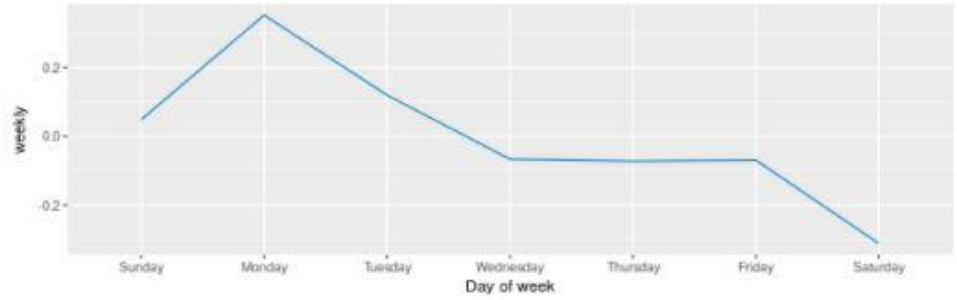
Grafik 2: Peyton Manning wikipedia çizgi grafiği



Grafik 2'de Peyton Manning wikipedia sayfası erişim istatistiklerinden çıkarılan çizgi grafiği (Letham, Taylor, 2017: 43) paylaşılmıştır.

Grafikte Wikipediadaki sayfasında görüntülenmelerin ve sonrasında da modelin tahmini çizgi grafiği şeklinde görülmektedir. Bir üst grafikteki dağınık grafik Fourier serisi kullanılarak özet bir çizgi grafiğine dönüştürülmüş ve anlaşılmasını kolaylaştırmıştır.

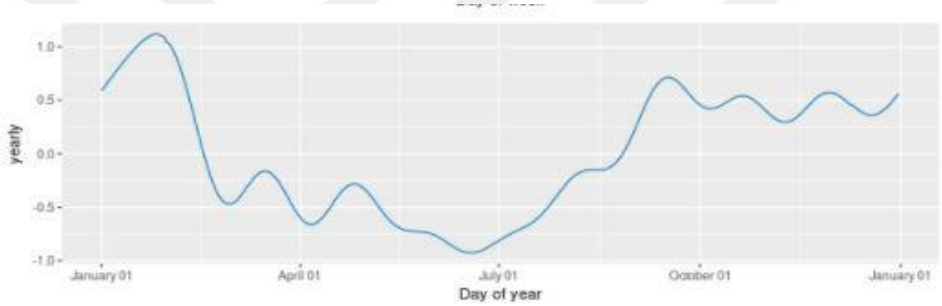
Grafik 3: Peyton Manning wikipedia haftanın günlerine göre erişim



Grafik 3'de, Peyton Manning wikipedia sayfası haftanın günlerine göre erişim istatistiği (Letham, Taylor, 2017: 43) paylaşılmıştır.

Kukla (dummy) değişkenleri kullanılarak mevsimin haftanın günleri etkisi şeklinde değerlendirilmesine dönüştürmüştür.

Grafik 4: Peyton Manning wikipedia tarihe göre etki



Grafik 4'de Peyton Manning wikipedia sayfası tarihlere göre erişim istatistiği (Letham, Taylor, 2017: 43) paylaşılmıştır.

Peyton'ın emeklilik sonrası (yaz ayları) sayfa erişimindeki azalış dikkat çekmektedir. Prophet bir web sayfasındaki erişimin öncelikle istatistiksel grafiğini, Fourier serisi kullanarak yıllık mevsimsel göstergelerini, yıllık veri setinden haftanın günlerindeki erişimlerin çıkarılmasını (maç günü ve sonraki günlerin erişim yoğunluğunun anlaşılmasını bu sayede sağlaması), ayrıca emekliliği ile birlikte (özel gün ya da trend şeklinde düşünülebilir) erişimdeki düşüşü kolayca ve özel manuel işlemler gerektirmeksizin vermektedir.

$$Y_t = g_t + s_t + h_t + \varepsilon_t$$

G(t) = Periyodik olmayan değişiklikler için trend fonksiyonunu,

S(t) = Periyodik değişiklikleri için fonksiyonu (örneğin haftalık, aylık, yıllık, mevsimsellik gibi),

H(t) = Düzeni değişitiren tatilleri temsil eden fonksiyonu,

ε (t) = Hata terimi'ni temsil etmekte ve

Prophet, temel olarak g, s, h, e bileşenleri ile açıklanabilmektedir.

Formülünün de içerdiği üzere Prophet için günlük mevsimsellik, haftalık mevsimsellik, yıllık mevsimsellik gibi trend ölçümleri modelin çıktıları arasından en çok takip edilenlerdir. ARIMA, ARMA modelleri durağan durumlar için ölçümler yaparken Prophet bu temel özellikleri dahil ederek zaman serisi yöntemleri de kapsamaktadır.

Bu tez kapsamında;

- Borsa verileri olduğu için zaman serisi modeline ihtiyaç duyulması,
- Borsa verilerinin hafta sonlarını içermemesi ve modelde zaman serisi ayarlarında bunun sorun olmaması,
- Haftasonu gibi durumların modelin direk kendisinin anlaması veya pratik şekilde ayarlanması,
- Herhangi bir uçuk değer varsa modelin bunu tolere ederek hesaba katabilmesi

Yukarıdaki açıklanan özellikler ve nedenlerden dolayı neural prophet modeli ile son aşama gerçekleştirilmiş ve duygu analizi yapılan tweetler (Bert modeli ile) ve hisse senedi değerleri birleştirilerek Neural Prophet modeline verilmiş ve gelecek tahmini için kullanılmıştır.

4. Duyarlılığın Hisse Senedi Üzerindeki Etkisini Değerlendirme

Veriye dayalı en önemli 3 proje yönetim modeli vardır. Bunlar KDD (Knowledge Discovery From Database) türkçe karşılığı ile Veri Tabanlarında Bilgi Keşfi (VBK), CRİSP-DM (Cross Industry Standard Process for Data Mining) yani Türkçe karşılığı olarak veri madenciliği için çapraz endüstri standart işlemi ya da veri madenciliği ve bilgi keşfi projeleri geliştirmek için fiili standart, SEMMA.

4.1 Method Yönetimi

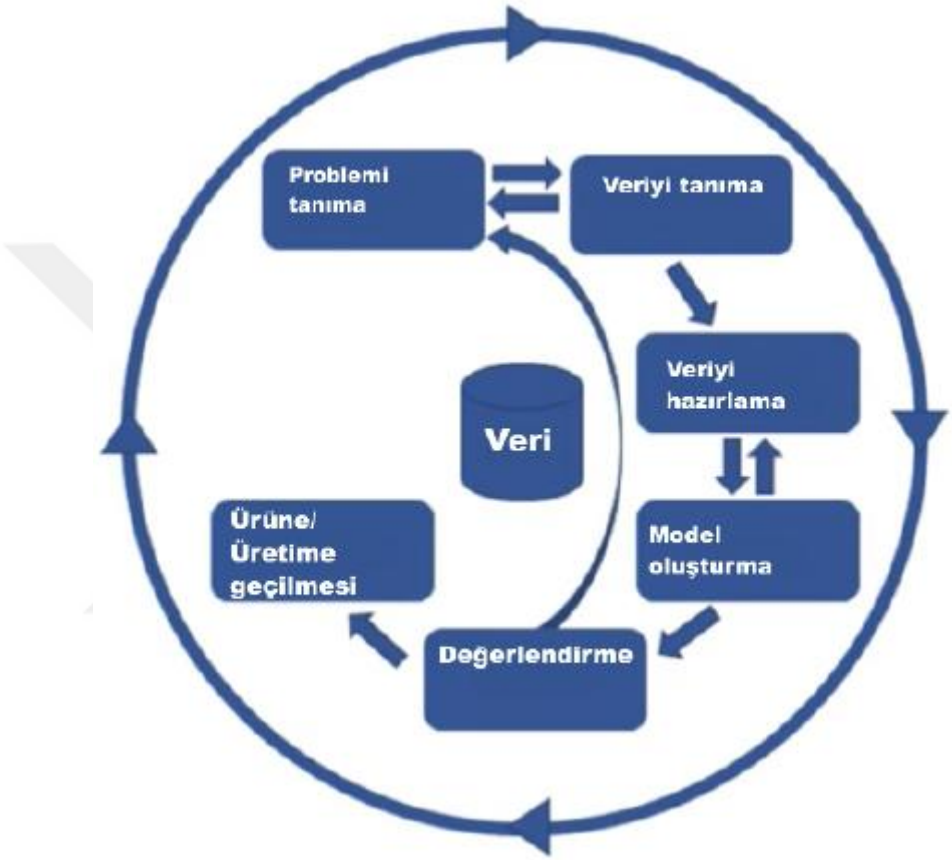
Yukarıda açıklandığı üzere, bu tez kapsamında aslında iki veri bilimi problemi üzerinde çalışldığı için bir yönetim methodu seçilmiştir. Method seçilirken en meşhur olan KDD, SEMMA ve CRISP-DM arasında proje yönetim methodları göz önünde bulundurulmuştur. Seçilen ve seçilmeyen metodlar kısa başlıklar halinde aşağıdaki bölümde açıklanmıştır.

4.1.1 KDD

KDD ingilizce knowledge and data discovery kelimelerinin baş harflerinden oluşmaktadır. Türkçe karşılığı veri tabanında bilgi keşfi şeklinde düşünülebilir. Gregory Piatte Shapiro tarafından ilk kez 1989 yılında adlandırılan bu terim aslında temeli daha öncelere dayanmaktadır. Sadece bir verinin analizini değil, veri tabanı yönetimini ve bunun bir çok yönünü, veri için ön işlemleri, modellemeyi ve bu aşamaların sonucunda da çıkarımı ve değerlendirilmesini, karmaşıklığı, görselleştirilmesini ve bunun gibi alanları içerir. Veri tabanı veya veri madenciliği gibi düşünülse de aslında bir şekilde elde edilmiş veriden yapılan işlemler sonucu yön gösterebilecek bir bilginin elde edilmesidir. KDD için göze çarpan en önemli 5 aşaması vardır. Bunlar; öncelikle problem için verinin seçimidir. Sonrasında eğer ihtiyaç varsa veri ön işlemleridir. Burada veri tipinin değiştirilmesi, eksik verilerin giderilmesi yani tamamen silinmesi ya da yerlerinin bazı yöntemle ile doldurulması gibi işlemler yapılır. Hazır halen gelen veri setinden sonra dönüşüm işlemi yapılır. Veri madenciliği aşamasında ise anomali tespiti, kümeleme, sınıflandırma, regresyon, birliktelik kural öğrenimi ya da bağımlılık modelli ve özetleme olmak üzere altı alt aşaması vardır. Bunlardan uygun olan ya da olanları kullanılır. Son olarak yorumlama ya da değerlendirme aşamasında ise üst aşamalarda yapılan çalışmaların sonucu değerlendirilir ve eğer yeterli başarıya ulaşıldı ise proje tamamlanmış sayılır. Yeterli aşamaya ulaşılamadı ise başa dönülür ve yeniden tüm aşamalar baştan yapılır.

4.1.2 CRISP-DM

Crisp-dm için 6 ana aşama vardır. Şekil 16' da (Şeker, Erdoğan 2018: 60) görüldüğü üzere aşamalar sırası ile; işi anlamak (business understanding), veriyi anlamak (data understanding), veriyi hazırlamak (data preparation), modelleme (modeling), değerlendirme (evaluation) ve dağıtım (deployment). Crisp-dm veri madenciliği için diğer kuruluşlardan bağımsız değerlendirilen bir standart sürecidir. Öncelikle problemin tanınması aşamasından sonra probleme uygun veri seçimi aşaması vardır.



Şekil 16: CRISP-DM döngüsü

Veri ön işleme süreçlerine ETL süreçleri dahil edilebilir. Burada eksik veri giderilmesi, veri temizlenmesi gibi işlemler yapılır. Sonrasında model oluşturulur ve model başarısı istenilen düzeyde ise bunun devamı olan üretime geçilme aşamasına geçilir ve bu proje genele yayılır. Eğer model değerlendirmesi istenilen seviyenin altında kaldıysa bir önceki aşama veri hazırlama aşaması veya veri yanlış seçilmişse bir önceki aşamasına geçilir oradan itibaren yeniden model için uygulamalar yapılır. Crisp-dm avantajı eğer sonuç tatmin edici değilse bir önceki aşama ya da ondan önceki aşama ya da ilk aşamadan itibaren esnek bir sistem üzerinden çalışılmasına olanak sağlamasıdır. Bu tez çalışmasında Crisp-dm görüşü ile yola çıkılarak tüm aşamaları bu yönteme uygun düşünülmüş ve uygulanmıştır. Tüm aşamalar alt kısımlarda detaylı şekilde ele alınacaktır.

4.1.3 SEMMA

Sas Enstitü'sünün yayınladığı Semma yaklaşımı aslında tam olarak bir veri madenciliği metodolojisi yerine Sas Enstitü madencilik işlevsel araç setinin mantıksal bir organizasyonu şeklinde yaklaşılabilir. SEMMA ismi sample, expolere, modify, model, assess kelimelerinin baş harflerinden oluşmuştur. Türkçe karşılıklarıyla örnekleme, keşfetme, dönüştürme, makina öğrenmesi modeli oluşturma ve değerlendirme şeklinde olan Semma yaklaşımı 5 ana aşamadan oluşmaktadır. Çıktığı döneme göre kullanımı artık çok azalmıştır. Semma'da tüm aşamaların sonunda bir değerlendirme süreci vardır. Eğer istenilen başarı seviyesi elde edilirse proje başarıyla tamamlandı demektir. Eğer istenilen başarı seviyesi elde edilemediyse o zaman örnekleme aşamasına geri dönülerek o aşamadan itibaren iyileştirme yaparak tekrar proje yapılmalıdır. Bu nedenden dolayı süreç içinde işlevsel görülmemiş ve kullanımı azalmıştır.

4.2 Literatür Taraması

Alanyazın taraması konularına göre bölümlere ayırılmış ve spesifik olarak yapılmıştır. Duygu analizine özgü alanyazı ayrı, Bert modeline alanyazın ayrı ve neural prophet'e özgün alan yazın taramaları aşağıdaki bölümde verilmiştir.

4.2.1 Duygu Analizi

Son yıllarda internet kullanımının yaygınlaşmasıyla birlikte insanlar bir mal veya hizmet konusunda karar verme süreçlerinde başka insanların yorumlarını da dikkate almaktadır. Benzer şekilde şirketler de müşterilerine daha iyi hizmet verebilmek ve kendi güçlü ve zayıf yanlarını görebilmek adına internetteki verileri kullanmaktadırlar. Bu süreçte duygu analizi bu büyük verilerin yorumlanması için son yıllarda en çok tercih edilen yöntem olmuştur.

Kaynar vd. (2016: 235-240) Yapay Sinir Ağları, Destek Vektör Makinaleri, Naive Bayes ve Merkez Tabanlı Sınıflayıcı algoritmaları kullanarak yaptıkları duygu analizi çalışmasında sinema severlerin izlemek istedikleri filme karar vermeden önce film yorumlarına ve puanına bakarak izleyecekleri filmi karar vermelerindeki etkileri duygu analizi çalışmasıyla incelemişlerdir. Başarı ölçütlerine göre en yüksek etkileri yapay sinir ağları ve Destek Vektör Makinelerinde elde etmişlerdir.

Göçgün, Onan (2021: 446-447) çalışmalarında amazon ürün yorumlarında duygu içeri içermediğini, içeriyorsa olumlu, olumsuz yada tarafsız bir yorum mu olduğunu Naive Bayes, Destek Vektör Makineleri, Karar Ağacı ve Logistik Regresyon algoritmaları ile test etmişlerdir. Yapılan duygu analizi

çalışmasından sonra en yüksek başarıyı Karar Ağacı algoritmasıyla elde ederlerken en düşük etkiyi Naive Bayes algoritmasıyla elde etmişlerdir.

Tuzcu (2020: 2-4) çevrimiçi bir kitap satış sitesinin kullanıcı yorumlarını okuyarak kitap almak isteyen müşterilerin satın almasında etkisi bulunan yorumların, negatif ve pozitifliğini Çok Katmanlı Algılayıcı, Naive Bayes, Destek Vektör Makineleri ve Lojistik Regresyon algoritmaları ile test edilmiş ve kullanarak duygu analizi çalışması yapmıştır. Negatif yorumları en iyi sınıflayan algoritma Çok katmanlı Algılayıcı algoritması iken pozitif yorumlar üzerinde en iyi algoritmanın Destek Vektör Makineleri algoritması olduğu sonucunu elde etmişlerdir.

Korkusuz ve Carus (2020: 387-396) çalışmalarında Destek Vektör Makinesi, Naive Bayes Algoritması ve K-en yakın Komşu algoritmasını kullanarak futbol karşılaşması sırasında atılan tweetlerdeki duygusal tepkilerle futbol maçı sırasındaki olaylarda gerçek zamanlı bir ilişki olduğu sonucuna ulaşırlarken en iyi sonuca Destek Vektör Makinesi algoritmasıyla ulaşmışlardır.

4.2.2 BERT

Sevli ve Kemaloğlu (2020: 37) çalışmalarında Google BERT modeli algoritmasını kullanarak twitter da felaket zamanında atılan tweetlerin gerçek veya gerçek dışı olarak etiketlenip etiketlenmediğini üzerine araştırma yapmışlardır.

Lee vd. (2019:1236) çalışmalarında biyomedikal geçmiş üzerine önceden eğitilmiş alana özgü bir dil temsil modeli olan BioBERT algoritmasını kullanmışlar ve bu algoritma ile BERT modelinden daha iyi sonuçlar elde etmişlerdir. BERT modeli sonucunda önceki modellerler karşılaştırılabilir performans elde ederlerken, BioBERT biyomedikal varlık tanımada %0,62 ,biyomedikal ilişki çıkarmada %2,80 ve biyomedikal soru yanıtlamada %12,24 daha iyi performans elde ederlerken biyomedikal geçmiş üzerine BERT ön eğitiminin karmaşık biyomedikal metinleri anlamasına yardımcı olduğunu sonucuna varmışlardır.

Sel ve Hanbay (2021: 683-684) çalışmalarında Destek Vektör Makineleri, Derin Öğrenme Yöntemleri (LSTM, CNN) ve Türkçe için ön eğitimli dil modelleri (BERT, DistilBERT, Electra) kullanarak yaptıkları çalışmalarından Türkçe twitter gönderilerinden cinsiyet tespiti yapmışlardır. Yapılan analizler sonucunda en yüksek başarımı % 80,1 ile Bert modelinden elde etmişlerdir.

4.2.3 Neural Prophet

Khayyat vd. (2021) sağlık sektöründe yapay zeka ve makine öğrenimi algoritmalarının vaka ve ölümleri tahmin etmekteki başarısından yola çıkarak Suudi Arabistan'da Covid-19'un pandemik tahminini belirleyebilmek için zaman serilerinde bağımlı Facebook Prophet modelini kullanarak verilerin gelecek tahminlerini bulmak için yaptıkları çalışmada Covid-19'dan kurtarılan vakaları bulmada düşük, ölüm oranlarını tahmin etmede yüksek bir sonuca sahip olduğu sonucuna ulaşmışlardır.

Borneklint (2021: 8-44) yaptığı çalışmasında Bitcoin ve Google fiyatlarının tahmininde ARIMA ve Facebook Prophet modelini kullanarak Bitcoin ile ilgili duyuruların fiyatta önemli bir değişikliğe sebep olduğu ve modelin performansını etkilediği sonucuna ulaşmışlardır.

Nagesh (2021) çalışmasında 01/01/2015 ile 31/12/2019 hisse senedi fiyatı, teknik göstergeler, haber makaleleri ve atılan tweetlerden yola çıkarak facebook prophet modelini kullanarak NIFTY 50 hisselerinin kapanış fiyatnı tahmin etmeye çalışmışlardır.

Akdağ ve Bozma (2021) çalışmalarında zaman serisi analiz modellerinden ARIMA ve Facebook Prophet modelini kullanarak 2008-2020 yılları arasındaki günlük bitcoin fiyatlarından hareketle bitcoin fiyat tahmini üzerine araştırma yapmışlardır. Yapılan çalışma sonrasında Facebook Prophet modelinin zaman serilerini tahmin etmede daha başarılı olduğu sonucuna varmışlardır.

4.3. Problemi Tanıma (Business Understanding)

Teknolojinin gelişmesi hatta daha detaylı bir ifade ile internetin yaygınlaşması aslında sosyal medya kullanımı da geliştirmiştir. Tablet ve telefon gibi teknolojik aletlerle birlikte daha çabuk erişim ve etkileşim sağlanmaktadır. Yukarıdaki bölümlerde detaylıca sosyal medya ve bu kapsamda teknolojik gelişim açıklanmıştır. Sosyal medyada katılımcılığa yani bir kişi tarafından paylaşılmış bilgi veya yorumun başka bir kişi tarafından oylanması, onaylanması ya da reddedilmesi yani üzerine diğer kişilerin kendi özgür düşüncelerini ekleyebilmesine olanak sağlamıştır. Açıklık ile yorum ve anket gibi geri bildirimselliğin açık halk önünde rahatça çok bütçe ayırmadan yapılabilmesine, bağlantılılık ile bir paylaşımda, bir videoda ya da yazıda, blogta link verilebilmesine ve direk farklı mecraları, platformları bağlayabilmesine olanak sağlamıştır.

Sosyal medyanın kattığı olumlu, beklenen taraflarının yanında beklenmeyen olumsuz yanları da olmaktadır. Örneğin taraflı bir yaklaşımdaki olumsuz yorum, restoranların, kitapların, filmlerin veya herhangi bir düşüncenin diğer insanlar üzerindeki etkisini de olumsuz etkileyebilmektedir. Yapılan bir çok araştırmada sosyal medyanın birçok alanda etkisinin olduğu kanıtlanmıştır. Bu etkiden yola çıkarak bu tez çalışmasında ise sosyal medyanın hisse senetleri üzerinde etkisinin olup olmadığı; Twitter

üzerinden yapılmış yorumların ertesi günkü markette hisse senedinin artış yapması veya değerinin azalmasındaki bağlantılılığına bakılacaktır.

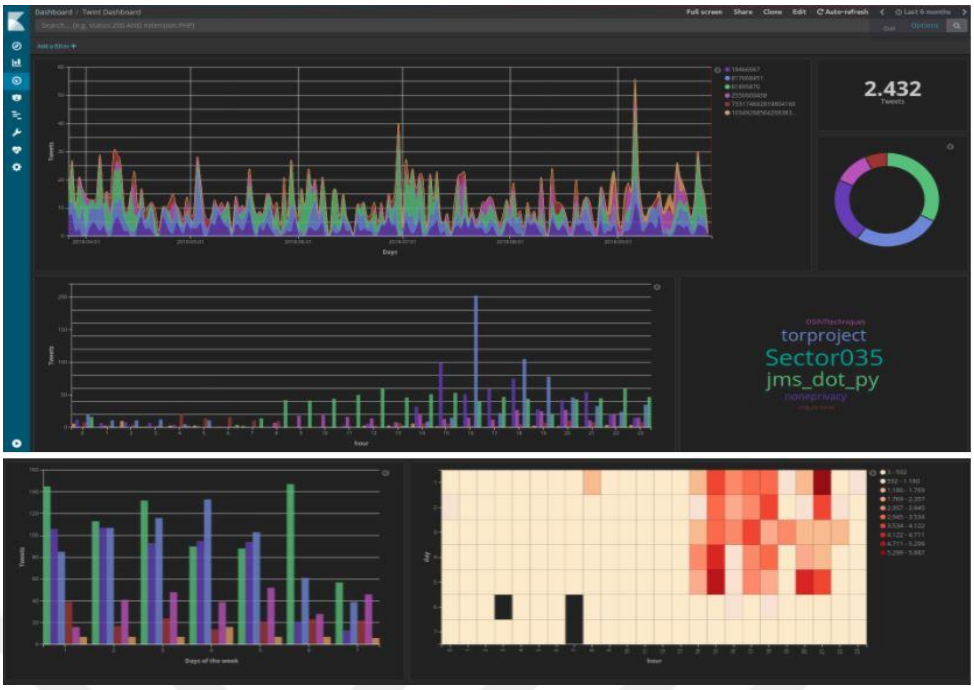
4.4. Veriyi Tanıma (Data Understanding)

Makina öğrenmesi çalışmalarında kullanılacak veri seti çok önemlidir. Genelde ham veri seti anlamsız ya da bozuk verileri içerdiğinden dolayı hata payını arttırır ve bu yüzden veri ön işleme, veri madenciliği gibi isimler alan ve birbirine çok yakın alanlar olan ön süreçler ile veri seti daha anlamlı hale getirilir.

Ham veri setleri Twitter, Yahoo Finans ve etiketlenmiş duygu analizli veri seti olmak üzere 3 farklı kaynaktan toplanmıştır.

4.4.1 Twitter

Ertesi günün market değerinde artış olması ya da azalış olmasının bağlantısını twitter üzerinde yapılan yorumlar ile ölçmek için Twitter'dan "Twint" ile https://github.com/twintproject/twint kullanılarak yorumlar alınmıştır. Modeller ve aşamalar açık kaynak olan Python ile yapıldığı için tweet alımı da teknolojisi oldukça gelişmiş ve Twitter API'sini kullanmayarak Python'da yazılmış bir araç kullanılmıştır. API'deki son 3200 tweet'i alabilme, kurulumdaki daha fazla aşaması veya zaman alımı, tweet çekme hızı, limitlendirmesi gibi eksik yanlar twint'de çözüldüğü için bu tool kullanılmıştır. Twint Twitter'a özgü çalışmasıyla Twitter'daki bir kullanıcının kimlik doğrulaması olmaksızın takipçilerini, beğendiklerini ve benzeri bilgileri alabilmektedir. Ayrıca kullanımı sadece bir kütüphane indirme süreci kadar sürmekte ve oldukça başarılıdır. Sloganları, kimlik doğrulama & API & limit yok şeklindedir. Bu nedenler ve Python 3.6 ve üzeri versiyonları için kolayca kullanılabilmesinden dolayı tercih edilmiştir.



Şekil 17: Github Twint API Tweet

Şekil 17'de Github Twint Aracı Örnek Tweet Grafiği paylaşılmıştır.

Grafikte örnek olarak Python'da bir kaç satır kod ile Twint kullanılarak çekilmiş 2432 tweet'in grafikleştirilmiş hali görülmektedir. Sol alt köşede haftanın günleri, sağ üst köşede kaç tweet çekildiği, orta kısımda tweet'lerin atıldığı saatleri ve tweet sayıları ve benzeri bilgiler çıkarttırılmıştır. Tez çalışması için 1/1/2016 ile 10/04/2021 tarihleri arasında günlük 100 farklı #TSLA kelimesini içeren tweet Twint aracı kullanılarak çekilmiştir.

Tablo 2: Tesla Tweets

date	tweet
1.01.2016	#HBRLive: Is #Tesla's Patent Sharing a Good Idea?
1.01.2016	TSLA: $240.01 - Market Cap: $31.42B 1/1/2016, 3
1.01.2016	CityFALCON score gives top rated news for FREE :
1.01.2016	Potential Tesla competitor Faraday Future is strar
1.01.2016	TSLA: $240.01 - Market Cap: $31.42B 1/1/2016, 3
1.01.2016	Faraday Future aims to make a splash at CES $TSL
1.01.2016	TSLA: $240.01 - Market Cap: $31.42B 1/1/2016, 2
1.01.2016	Panasonic to build $400m battery plant in China I
1.01.2016	A collection of testimonials from satisfied super_
1.01.2016	Tesla Motors (TSLA) to Face Threat from German
1.01.2016	TSLA: $240.01 - Market Cap: $31.42B 1/1/2016, 2
1.01.2016	We calculated 63-day Expected Return of $TSLA f
1.01.2016	3 Activity Ratios to Actively Watch https://t.co/s
1.01.2016	Looking for someone who can help guide me. I'd

Tablo 2'de TSLA kelimesini içeren Tweets , Twint Aracı ile çekilenlerin örnek satırların olduğu tablo görülmektedir.

4.4.2 Yahoo Finans

Diğer veri kaynağı ise finansal değerlerin alındığı yani hisse senedi değerlerinin olduğu Yahoo Finans. Veri kaynağına https://finance.yahoo.com/quote/TSLA/history/?guccounter=1 web sitesinden ulaşılabilir. Ayrıca bu tez çalışması kapsamında https://algotrading101.com/learn/yahoo-finance-api-guide/ Yahoo Finans API kullanılmıştır. Bu uygulamayı kullanmanın nedenlerinden bazıları; ücretsiz olması, istenilen tarih aralığının seçilerek istenilen büyüklükteki verinin çekilmesinin kolay olması, profesyonellik gerektirmeden veri çekiminin kolayca yapılabilmesi, kullanımın basit olması. Ayrıca Python dili ile çalışabilmesidir.

Tablo 3: Yahoo Finanse değerler

Date	Open	High	Low	Close	Adj Close	Volume
3/28/2016	46.321999	46.962002	45	46.051998	46.051998	19628500
3/29/2016	45.978001	46.476002	45.066002	46.026001	46.026001	20071500
3/30/2016	47.018002	47.099998	45.299999	45.377998	45.377998	20165000
3/31/2016	45.868	47.484001	45.001999	45.953999	45.953999	40064500
4/1/2016	48.966	49.580002	46.650002	47.518002	47.518002	79987500
4/4/2016	49.824001	50.424	48.728001	49.397999	49.397999	67376500
4/5/2016	48.099998	51.312	48	51.094002	51.094002	49743500
4/6/2016	50.793999	53.548	50.689999	53.084	53.084	58527500
4/7/2016	53.290001	53.868	50.902	51.439999	51.439999	44281000
4/8/2016	52.099998	52.164001	49.604	50.014	50.014	36819500
4/11/2016	50.200001	51.798	49.060001	49.984001	49.984001	45808500
4/12/2016	49.900002	50.360001	48.726002	49.563999	49.563999	28816000
4/13/2016	49.702	51.099998	49.466	50.905998	50.905998	24628000
4/14/2016	50.599998	51.368	50.209999	50.372002	50.372002	20661000

Tablo 3: Finanse API'den çekilen değerlerin örnek satırlarını göstermektedir. API kullanılarak çekilen veride tarih, hissenin açılış değeri, hissenin yüksek ve düşük değeri, gün sonu kapanış değeri ve işlem hacmi alınmıştır.

4.4.3 Etiketlenmiş Duygu Analizli Veri (Labeled Data)

Doğal dil işlemede Bert modeli gözetimli yani öğreticiyle öğrenen bir modeldir. Gözetimli öğrenmesi için daha önceden hazırlanmış ve analiz edilmiş yani duyguların etiketlenmiş halini modele vermek ve onun öğrenmesini sağlamak gerekmektedir. Bunun için açık paylaşımda olan, telif hakkı olmayan ve üzerinde çalışılmış veri kaynağı bulunması gerekmektedir.

Kaggle büyük veri kaynaklarının paylaşımı ve işlenmesi konusunda iyi olduğu için Bert modelinin ilk adımdaki eğitiminde kullanılmak üzere tez kapsamında kaggle'dan etiketlenmiş veri setleri alınmıştır. İlk veri seti

https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data kaynağından alınmıştır ve veri setinin ilk halinde negatif, aşağı yukarı negatif, nötr, aşağı yukarı pozitif ve pozitif duygu etiketleri bulunmaktadır. Bu veri setinde filmler hakkında yazılmış yorumlar vardır ve bunlar manuel olarak 6 farklı duygu şeklinde etiketlenmiştir fakat tez kapsamında negatif yani olumsuz ve pozitif yani olumlu olmak üzere iki ana gruba düşürülmüş ve nötr yorumlar dahil edilmemiş ve sadece negatif ve pozitif etiketlerin olması gerektiği için aşağı yukarı nagatif ve negatif değerler ve aşağı yukarı pozitif ve pozitif değerler de kendi aralarında birleştirilmiştir.

Veri setinin önemli bazı özellikleri; stanford tarafından hazırlanmış ve cümlenin tüm halinin duygu etiketi belirlenmiş daha sonra sekmelere ayrılmış halinin de her sekmesi için duygu etiketi belirlenmiştir. Sekmeler başka cümledeki sekmelerle aynı ise bir tanesi alınmış ve tekrarlanmasına izin verilmemiştir.

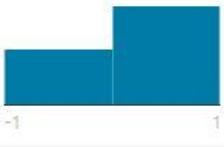
Bir diğer etiketlenmiş veri seti de yine kaggle'dan https://www.kaggle.com/yash612/stockmarket-sentiment-dataset adresinden alınmıştır. Sadece sinema ya da film yorumları kullanımı modeldeki başarıyı daha az etkilemesinden dolayı hisse senedi marketi hakkında etiketlenmiş veri seti ihtiyacı doğurmuştur.

Tablo 4: Etiketlenmiş yorumlar

Text	Sentiment
NG - nice PNF BY - breakout - need follow thru	1
Won't believe AAP uptrend is back until it crosses above MA(50)	-1
X swing still on	1
SWY - 30% of float short and breaking out - ouch	1
BIOF wants 4.90's comin!!!	1
VS inverted head and shoulder play out well. Wasn't able to catch the entry. E	1
red, not ready for break out.	-1
EI close to breaking out now. My trigger is at 30.40.	1
user BAC For a quick Trade to late..But for investing ~11.98 is a good entry po	1
CHDN - ong 49.02. Trailing Stop 56.66 from 6 prior Stops of 54.17, 49.88, 4	1
AAP VOME today is impressive. At this rate and well probably get to 30M shar	1
user: been adding VXY long off the bottom today for trade, also got WPI near l	-1
I repeat, if global economy is going to get better this year go with C instead of	1
GOOG go ONG on close above 725.	1
NKD looking like a good short. Failed to break price level resistance at 116 tod	-1
GS like the price action. So far holding above 130. Still deciding whether to go	1
SBX buy when it clears resistance at 54.5	1
here is NEW up target for AAP and notice shakeout reading at -8.49 ! and we	1
Too early to short into this move. Stock Market needs a few days to settle dow	-1

Etiketlenmiş hisse senedi yorumları yine aynı nedenlerden dolayı https://www.kaggle.com/yash612/stockmarket-sentiment-dataset adresinden alınmıştır. Bu veri setinde bir öncekinden farklı olarak duygu etiketi sadece iki tanedir. Olumlu ya da olumsuz yani negatif / olumsuz yorumda (-1) ve pozitif/ olumlu için (+1) kullanılmıştır.

Tablo 5: Hisse senedi yorumları & duygu analizleri

5791 unique values	
AAP - user if so then the current downtrend will break. Otherwise just a short-term correction in me...	-1
Monday's relative weakness. NYX WIN TIE TAP ICE INT BMC AON C CHK BIIB	-1
GOOG - ower trend line channel test & volume support.	1
AAP will watch tomorrow for ONG entry.	1
i'm assuming FCX opens tomorrow above the 34.25 trigger buy. still very much like this setup.	1

Veri setinde, örnek tabloda göründüğü gibi 5791 cümle ve bunların duygu analiz etiketleri vardır. Bu veri seti ile sinema yorumları veri seti birleştirilmiştir.

Tablo 6: Film ve hisse senedi yorumları & duygu analizleri

Text	Sentiment
EI close to breaking out now. My trigger is at 30.40.	1
user BAC For a quick Trade to late..But for investing ~11.98 is a	1
CHDN - ong 49.02. Trailing Stop 56.66 from 6 prior Stops of 5	1
AAP VOME today is impressive. At this rate and well probably g	1
user: been adding VXY long off the bottom today for trade, also	-1
I repeat, if global economy is going to get better this year go witl	1
GOOG go ONG on close above 725.	1
NKD looking like a good short. Failed to break price level resista	-1
GS like the price action. So far holding above 130. Still deciding	1
SBX buy when it clears resistance at 54.5	1
here is NEW up target for AAP and notice shakeout reading at -	1
Too early to short into this move. Stock Market needs a few days	-1
AAP obbers hit Apple Store in Paris. Prefer to take merchendis	1
AXA Here is my tip of the day. Buy now! You do not even have t	1
X out the 1/2 due to the move today and the Kirby on the daily a	1
AAP 10 min short won now longs have fully added in, look for a	1
NG stellar move - still long swing - full position - I believe user &	1
ATICE & TADE SETP: KO AN CNX	1
IDCC 50SMA acting as a good support. Short-term momentum i	1
NEW POST: BAC Is At 19 Month Highs, But Showing Signs of S	1
c two daily bull flags on top of each other=bullish still	1
NG - traditional and AT point and figure - bullish BY signals mult	1
jpm closed>10sma on Monday(u know my rules)...large cup n h	1

Örnek satırların olduğu yukarıdaki tablo bu birleştirme sonucu toplamda iki kolon olan ve bir kolon yorum ve bir kolon da -1 negatif yorumu +1 ise pozitif yorumu içeren üç veri setinin toplamı toplamda 14 bin satır olmuştur.

4.5. Veri Hazırlama (Data Preparation)

Toplanan veya oluşturulan ham veri çoğunlukla veri ön işleme & ETL süreçleri gerektirmektedir. ETL ingilizce açılımı ile extract, transform, load kelimelerinin baş harflerinden oluşmaktadır. Türkçe karşılığı ise ham verinin kaynağından alınması ya da oluşturulması aşamasından sonra bazı ön işlemler geçirerek değer kazandırılması ve yeniden belirlenen bir sisteme yüklenmesidir. Bu tez kapsamında son aşaması olan bir sisteme yüklenmesi hariç diğer tüm aşamaları uygulanmıştır. Ön işlemde veri ile bazı hesaplamaların yapılması, veri içinde aykırı değerlerin bulunması ve temizlenmesi yani ayıklanması, ve verinin bazı formatlarının dönüştürülmesi gibi adımlar sayılabilir.

4.5.1 Twitter Verisi Ön İşlemleri

Twitter'da saniyede bile binlerce yorum / fikir paylaşılması aslında aranılan konu ile ilgili bilgi erişimini hem zorlaştırmaktadır hem de konu ile ilgili fikir almak isteyen kişi için çok sayıda yoruma ulaşabilmesi açısından fikir edinmesini kolaylaştırmaktadır. # işareti kullanılarak yazılan anahtar kelime ile istenilen konuda bilgi bulunabilir. Bu tez kapsamında Twitter'dan nasıl veri çekildiği ve hedef twitler yukarıdaki bölümde açıklanmıştır. #Tsla kullanılarak çekilen twitler model başarısı için eğitim veri setinin ingilizce olmasından dolayı ingilizce olmalıdır.

Bunun için spacy kütüphanesi kullanılarak ingilizce yorumlar alınmıştır. Spacy langdetect ve spacy python kütüphanaleri yüklenerek 'en' şeklinde ingilizce istenildiği belirtilerek veriler temizlenebilmektedir. Bu kütüphanenin önemli özelliklerinden birisi uzun cümle guruplarında başarılı çalışmasıdır. Alternatif olarak pycld2, textblob, googletrans vb. kütüphaneler de kullanılabilir. Textblob birden fazla dil seçimi için başarısı spacy'den daha yüksek fakat tek dil seçimi yapılacağı için yeterli bulunmuştur. İngilizce dışında bulunmuş Çince, Rusca, Japanca gibi dillerin olduğu satırlar silinmiştir.

4.5.2 Veri Dengesi İşlemleri

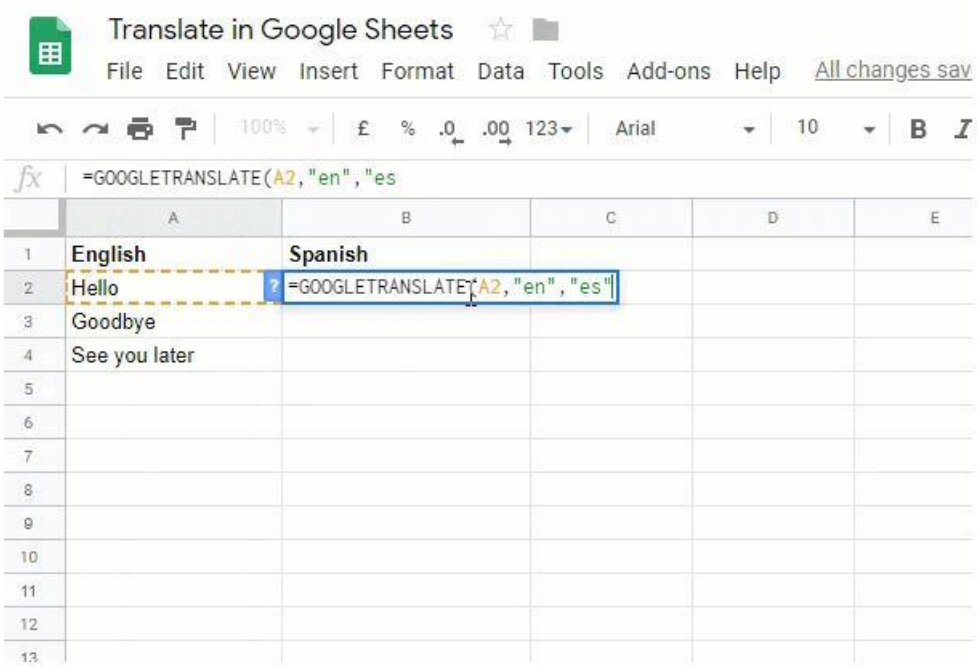
Birleştirilen sinema yorum verileri ve hisse senedi yorum verilerinde negatif ve pozitif etiketlerde sayıca dengesizlik tespit edilmiştir. Aslında öznitelik mühendisliği işlemi sayılan bu verideki dengenin sağlanması sınıflama başarısını oldukça çok etkilemektedir. Klasik veri öğrenmesi yapılarında bu

dengesizlik model üzerinde daha çok etki yaratırken derin öğrenmedeki modellerde etkisi azalmaktadır fakat doğal dil işleme tam anlamıyla sayısal ifadelerle ilgili olmadığı için her etki değerlendirilmiş ve iyileştirme yoluna gidilmiştir. Veri kümesinde etiketli kümeler arasında yeterli denge olmaz ise 90% veya 99% oranındaki doğruluk (accuracy) elde ederken bir diğer taraftan hata matrisi (confusion matrix) incelenince yüksek başarının nedeni çok yüksek oranda etikete sahip veri kümesindeki doğru bilme oranından kaynaklandığı görülebilmektedir. Örneğin veri setinde 80% ve 20% ağırlıkları olacak şekilde etiketlenmiş iki kümeli veri setinde test verileri rastgele 80% oranlı küme için verilmiş etiket atansa bile doğruluğu yüksek çıkacaktır. Bu durum aslında modelin doğru işlemediği anlamına gelmesine rağmen doğruluk testlerine bakılarak yanıltıcı kararlar aldırabilmektedir.

Birleştirilen veri setindeki dengesizlik (imbalance data) bazı manuel işlemlerle giderilmiştir. İlk aşamada incelenen iki etiketteki negatif yorumların pozitif yorumlara göre sayısı az olduğu için negatif yorumların arttırılması gerektiği düşünülmüştür.

Bu dengesizlik için ingilizce yorumların ispanyolcaya çevirilmesi ve yeniden ingilizceye çevirilmesi sağlanmıştır.

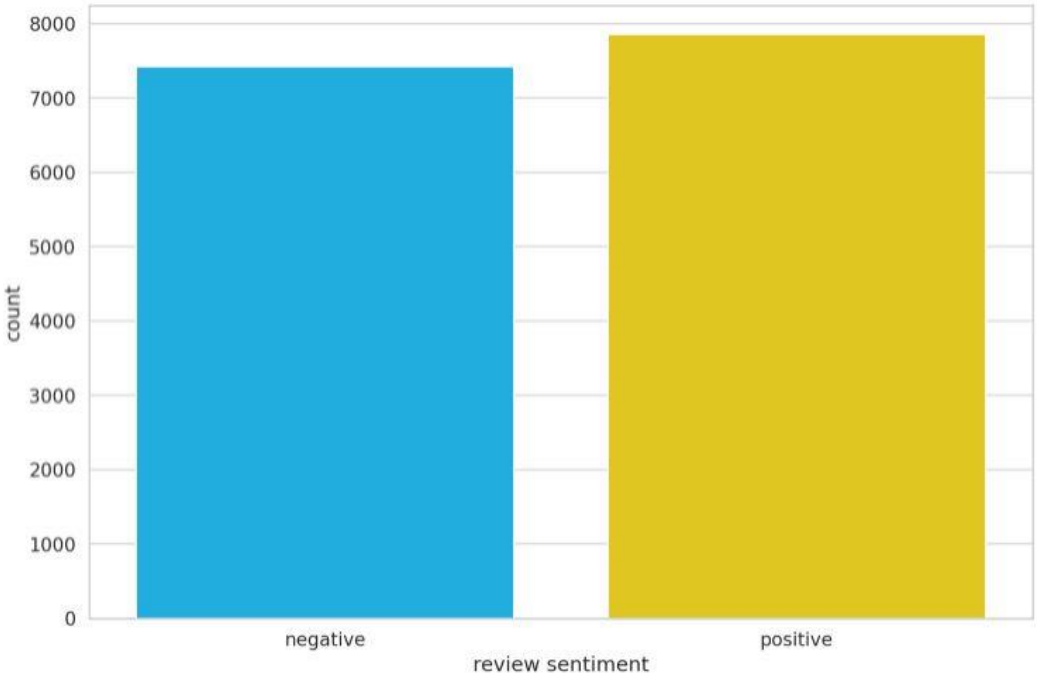
Tablo 7: Dil çevirimi



Şekilde basit bir örnekte göründüğü gibi Google translate'in excel verisyonu kullanılarak negatif etikete sahip yaklaşık bin satırlık cümle önce İspanyolca'ya çevirilmiş sonrasında ingilizceye çevirilmiştir. Eğer tekrar çevirilmiş cümle eskisi ile aynı ise bazı farklı kelimeler eklenmiş ya da bazı kelimelerin eş anlamlıları ile değiştirilmiş ve öyle İspanyolca'ya çevirilmiş sonrasında ingilizceye çevirilmiştir. Eski

değiştirilmemiş hali ve yeni çevirilmiş halleri farklı olan tüm satırlar alınmış ve veri setinde daha çok denge sağlanmıştır.

Grafik 5: Veri setindeki negatif ve pozitif etiketli satırlar



Bu çevirme işleminden sonra pozitif ve negatif etiketli verilerin toplamı 15,270 satır olmuştur.

4.6. Model Oluşturma (Modeling)

Problemi anlama aşamasından sonra ham verinin temizlenmesi ve diğer veri ön işlemlerinin olduğu veri hazırlama aşaması yapılmış ve model için veri hazır hale getirilmiştir. Bu tez kapsamında birbirinden farklı iki model uygulanacaktır. Birinci model doğal dil işlemede duygu analizi modellerinin alt kümesinde olan BERT modelidir. İkinci model ise Bert modelinin sonucunu kullanan ve makina öğrenmesinde gözetimli öğrenme modellerinin bir alt kümesi olan Neural Prophet modelidir.

4.6.1 BERT

Öncelikle Bert model gibi önceden eğitilen modeller ile çalışmak için eğer imkan varsa önceden eğitilmiş model direk kullanılmalıdır. Bu imkan yoksa problemin durumuna göre model kullanılmadan önce de özel veri seti ile eğitilebilmektedir.

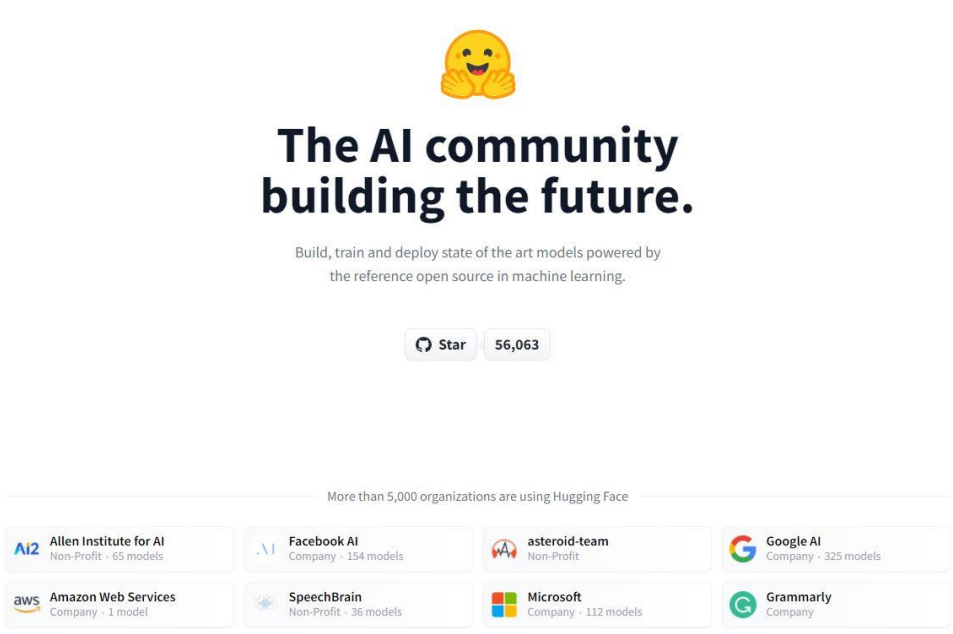
Bu tez aşamasında model eğitimi yapılırken büyük veri seti kullanılacağı ve birden fazla model kullanılacağı için hem hızlı çalışılmalı hem de iyi bir donanıma yani GPU'ya sahip olunmalıdır. Veri

seti öğrenmesi ve sonucundaki başarıyı etkileyeceğinden dolayı veri seti rastgele karıştırılarak her seferinde sıfırdan başlayarak tüm model baştan çalıştırılmaktadır.

Uzun istatistiksel işlemler içeren model ve problemler için çözümlerinde kullanılan GPU geçmişte daha çok görüntü işleme için kullanılmaktaydı. Bert model gibi çalıştırıldığında hız gerektirecek ve komplex model için Google ücretsiz olarak COLAB platformu sunmuştur. Bu tez kapsamında Colab platformu ile çalışılmış ve BERT model bu platformda işlenmiştir.

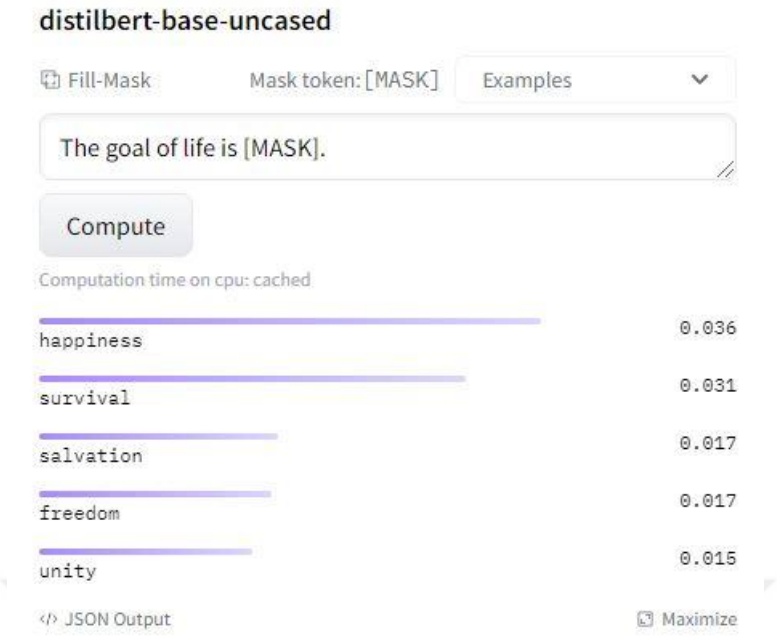
Zamana bağlı t-1, t, t+1 gibi beraber değerlendirilecek durumların yani bağlı düşünülmesi gereken indexlerin sıralaması önemlidir. Bu önemin nedeni cümledeki kelimelerin sırasına, videodaki fotoğrafların sırasına (ki bu saniye 24 ayrı fotoğraftan oluşmaktadır), bir şarkıda notalara bağlı olduğu içindir. Örneğin atasözündeki iki kelimin yeri değiştirilirse anlam eksikliği oluşmakta ve vurgulanmak istenilen durumun ağırlığı değişmektedir.

Bert modeli yukarıda detaylı açıklandığı gibi ön eğitilmiş ve bu aşamada ince ayar aşamasında/ tekniği ile (fine-tunning) probleme özel katman eklenerek prolemle ilgili veri setinden sonraki etiketlenmiş öğrenme süreci çalışılmıştır.



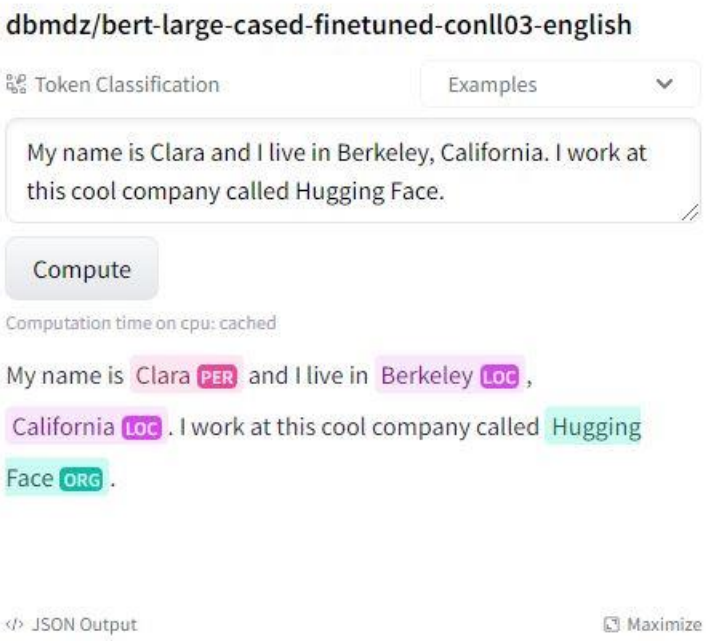
Şekil 18: Hugging Face uygulama arayüzü

Şekil 18'de Hugging Face uygulama arayüzü (https://huggingface.co/) gösterilmektedir. Bahsedilen önceden eğitilmiş model Hugging face API'sinden çekilmiştir. Bert'in de bunlardan birisi olduğu 5 binden fazla büyük organizasyonun paylaştığı açık kaynaklı modeller bulunmaktadır.



Şekil 19: Bert Base'in örnek Maskelenmiş Dil Modeli gösterimi (MLM)

Şekil 19'da Bert Base'in örnek Maskelenmiş Dil Modeli gösterimi (MLM) (Hugging Face API) gösterilmektedir. Yukarıdaki bölümlerde açıklanan maskelenmiş dil modeli (MLM) görevine göre örnek bir cümle seçilerek sonuçları gösterilmiştir.



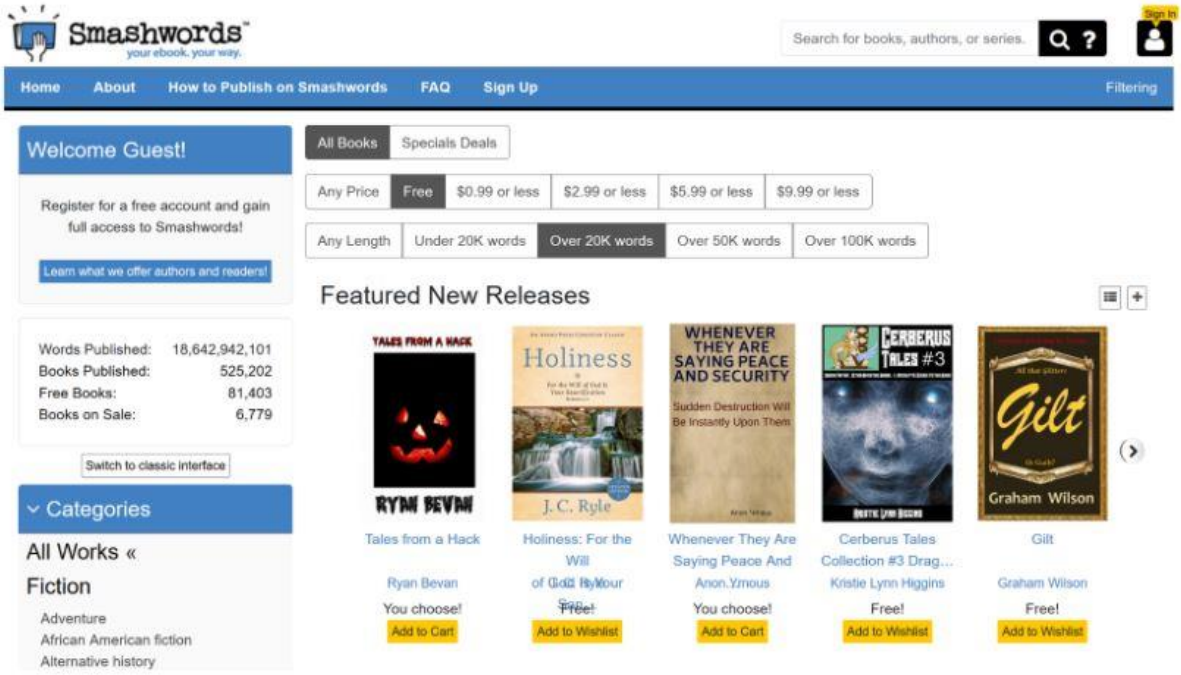
Şekil 20: Bert Large tokanization

Şekil 20' de, Bert Large'ın örnek kelimelere ayırma (tokanization) işlemi (Hugging Face API) gösterilmektedir.

Hugging Face'deki modeller 160 dilden fazla dili içerdiği için oradan alınan modellerin dil ayrımını model kendisi yapmaktadır.

Ön işlemlerden geçirilmiş 15 bin satırlık, 2 etiket sınıfından oluşan veri ile öğretilmiştir.

Yukarıdaki bölümde Bert modeli anlatılırken ön eğitim sürecinden bahsedilmiştir. Bu tez kapsamındaki modelin ön eğitimi 800 milyon kelime haznesi olan Toronto Book Corpus ve 2,5 milyon kelimeden oluşan ingilizce Wikipedia kaynakları ile yapılmıştır. Transformer'ın yapısında olan kodlayıcı (encoder) ve kod çözücü (decoder) sürecinin encoder süreci yani kelimenin kökeni, yapısı, modelin dili algılama süreci bu ön öğretim (pre-train) ile modele verilmiştir. Bert modeli bu süreçte kod çözücü sürecini işleyebilecek duruma gelmiştir.



Şekil 21: Smashwords web sitesi

Şekil 21'de, Toronto Book Corpus için Smashwords web sitesi kullanıcı arayüzü gösterilmektedir.

Toronto Book Corpus artık herkese açık bir kaynak olmadığı için smashwords aracı olarak kullanılarak oradaki kitaplar çekilebilmektedir. Zhu ve arkadaşları araştırmalarında (2015) kitap siteleri, kitaplar, kullanıcı yorumları, kitapların içerisindeki tüm karakterler, nesneler, sahneler, anlambilim, kişilerin (okuyucuların) düşünceleri, hisleri, hikaye, olgu, hikayenin gelişmesi vb bilgiler modelin ön eğitimi için çok zengin kaynak olduğunu belirtmiştir.

Bert transformers'ın bir alt modeli olduğu için yukarıda da ekran ara yüzü verilen Hugging face kütüphanesi eklenerek model için ilk başlangıç yapılmıştır.

Veri setinin kullanılabilir olması sonucu direk etkileyeceği için büyük önem taşımaktadır. Bu yüzden veri seti hakkında bazı sorgulamalar yapılmış ve eksik değer olup olmadığı kontrol edilmiştir.

Tablo 8: Eksik değer şeması

```
[ ]  df.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 15270 entries, 0 to 15269
     Data columns (total 2 columns):
      #   Column     Non-Null Count  Dtype
     ---  ------     --------------  -----
      0   Text       15270 non-null  object
      1   Sentiment  15270 non-null  int64
     dtypes: int64(1), object(1)
     memory usage: 238.7+ KB
```

Eksik değer (missing value) bölümündeki değerin non-null olduğu görüntülendiği için burada bir işlem yapılmamıştır. Eksik değer çıksaydı sistemden eksik değerin olduğu satırlar silinmeli ya da çeşitli yöntemler ile eksik satır ve sütunlar tamamlanmalıydı.

Veri setinde kontrol edilmesi gereken ikinci önemli aşama veri dengesidir (data imbalance). İlk aşamada veri setinde denge kurulamamış bu yüzden bazı ön işlemler uygulanmış ve bu işlemler veri dengesi işlemleri başlığında detaylıca açıklanmıştır.

Makine öğrenmesi modellerinde genellikle ham metin ile işlem yapılması çok zordur ve bu yüzden metin sayısal ifadelere çevirilir. Bert modeli açıklamasında bu çevirim yani bölütleme (tokenization) teorik olarak tezin yöntem kısmında açıklanmıştır. Bu tez kapsamında değerlendirilecek olursa bölütlemenin aşamaları;

- Cümleler birbirinden ayırılır ve sınıflandırma yapılabilmesi için özel belirteçler eklenir

- Dolgu tanıtılır ve sabit uzunluktaki dizilerin geçirilmesine izin verilir (padding)

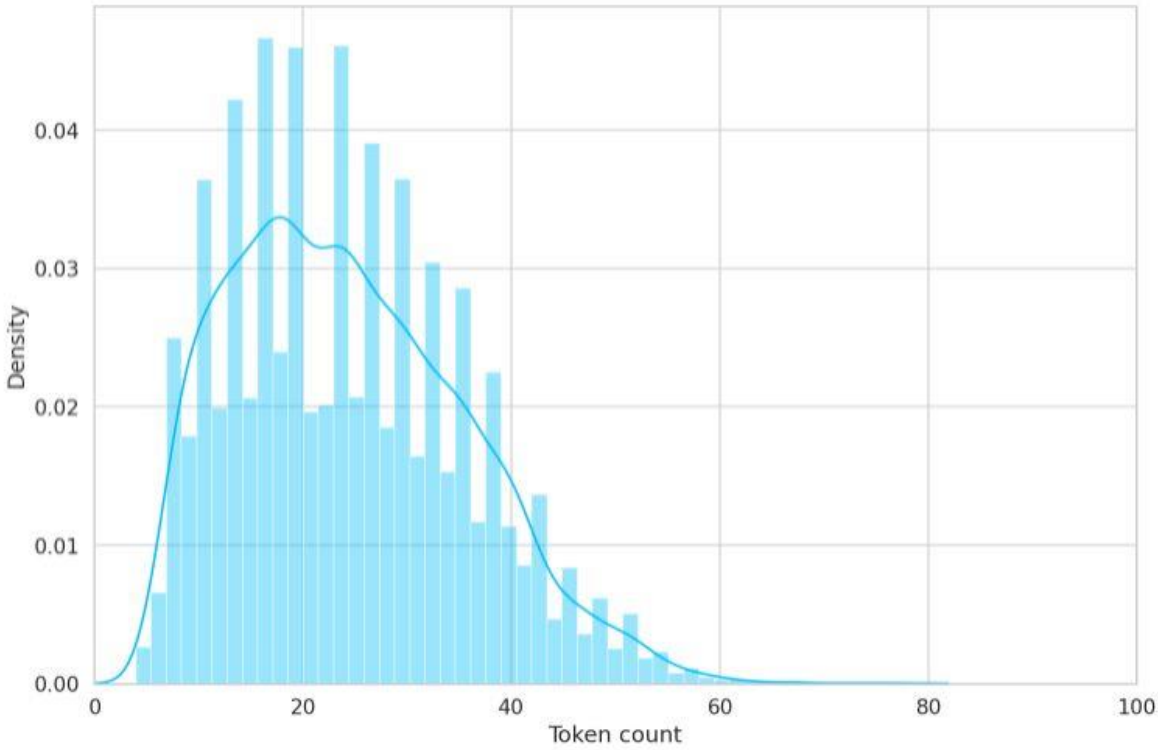
- Dikkat maskesi olarak adlandırılan 0 ped belirteçleri (pad token) ve 1 gerçek belirteçler (real token) dizisi oluşturulur

Transformatör kitaplığı Bert modeli gibi çok çeşit üst model sağlayıcısıdır ve bu yüzden önceden oluşturulmuş belirteçleri içerir. Bu tez kapsamında Bert modeli kullanılarak duygu analizi yapılacaktır. Bert modelinin kasalı (cased) ve kasasız (uncased) sürümlerinden kasalı (cased) sürümü duygu analizinde daha iyi çalıştığı için kasalı kullanılmıştır. Örneğin duygu anlatırken metindeki yazılışlarında "iğrenç" ve "İĞRENÇ" arasında oldukça fark vardır ve case bu farkı daha çok kapsamaktadır.

Bölütleme (tokenization) işlemi metni bölütlere ve bölütleri de benzersiz tam sayılara (id) dönüştürür. Cümle sonlarında bitirildiğine dair işaret kullanılır (cümle bitiş işareti). Bu işlemler encode_plus yöntemi ile yapılmaktadır. Belirteçler yani bölütlenmiş özel id edinmiş benzersiz sayılar Tensör'lerde

(Tensor) saklanır. Dikkat maskesi (attention mask) ismi verilen daha önce açıklanmış sabit uzunluk ise bu tez kapsamında 32 seçilmiştir. Bert modeli sabit uzunluk belirlenen metinler ile çalışmaktadır. Metin uzunluğu bu başta belirlenen uzunluğu geçse dahi sorun olmamaktadır fakat işlemci için yavaşlatacağı için genellikle 250 kelime gibi değerler seçilir. Bu tez kapsamında max_length 512 seçilmiş ve sonrasında bu uzunluğun yeterli mi yoksa fazla mı olduğu kontrolü sağlanarak değiştirilmiştir. Bu seçimden sonra eğitim için verilen 15 bin satırlık verinin bölütleri hakkında bilgi sorgulaması sonucu:

Grafik 6: Bölüt sayıları ve ağırlıkları



Grafikte 6' da görüldüğü gibi geneli 80 kelimeden az içeren cümleler olduğu için max_length uzunluğu 80 şeklinde düşürülmüştür.

Sonraki aşamada makinanın öğrenmesi ve veriminin hesaplanabilmesi için metin veri seti 90% eğitim (training), 5% doğrulama (validation) ve 5% test (testing) şeklinde bölünmüştür. Doğrulama kısmı modelde metin ile eğitilmeyen ayırılan test kısmından da ayırılan bölümdür.

Tablo 9: Veri bölütleme sonucu

```
[ ]  df_train.shape, df_val.shape, df_test.shape

     ((13743, 3), (763, 3), (764, 3))
```

Tüm veri setini bölme işlemi sonucu eğitim için 13743 satır metin, doğrulama için 763 satır metin ve test için 764 satır metin elde edilmiştir.

Artık metin Bert modelinde kullanılmak üzere işlenmeye başlamaya hazır hale getirilmiştir. Açıklandığı gibi öncelikle bölütleme (tokenization) ile bölütlerine ayrılır sonrasında sınıflandırma işlemi için olması gereken cümlenin en başına ve cümlenin en sonuna eklenir. Metin değeri belirtilen maksimum uzunluktan kısa ise (sabit değer belirtildiği için) boşluklar doldurulur fakat eğer metin daha uzun ise o zaman belirtilen sabit değer kadar kelime ile ifade edilir. Bu işlem ile dikkat maskeleri oluşturulmuş olur (attention mask) ve cümleler bu işlemler ile birlikte artık tensör objesi şeklinde geri döndürülür.

Tablo 10: *Tensör ve dikkat maskesi çıktısı*

```
[ ]  F.softmax(model(input_ids, attention_mask), dim=1)

     tensor([[0.3311, 0.6689],
             [0.4530, 0.5470],
             [0.5488, 0.4512],
             [0.3166, 0.6834],
             [0.7348, 0.2652],
             [0.3665, 0.6335],
             [0.4566, 0.5434],
             [0.6227, 0.3773],
             [0.6790, 0.3210],
             [0.5921, 0.4079],
             [0.4571, 0.5429],
             [0.6397, 0.3603],
             [0.3888, 0.6112],
             [0.5508, 0.4492],
```

Bölütlerine ayırılmış kelimerin id numaraları almış halleri ve dikkat maskesi değerleri ve yukarı bölümde açıklandığı gibi olasılıkların toplam değeri 1 olduğu örnek bir kaç satır çıktı ile gösterilmiştir. Bu çıktı aslında maskelenmiş kelimelerin olasılıklarının tahminini ve üzerine softmax fonksiyonunun uygulanmasının sonucudur.

Modelle ilgili tüm aşamalar yapıldığı için artık toplanan sinema yorumları, borsa yorumları ve onların duygu puanları (-1 & 1) ile modelin eğitiminin yapılması başlayabilir.

Bu aşamada model kaç kez eğitim (train) edilmesi gerektiği sisteme belirtilmelidir. Tez kapsamında bir çok deneme yapılmış ve en iyi sonucu yani hiperparametrelerin en iyi olduğu değerler bulunup onlar uygulanmıştır.

Örneğin denenen hiperparametre değerleri:

Batch size için: 16, 32

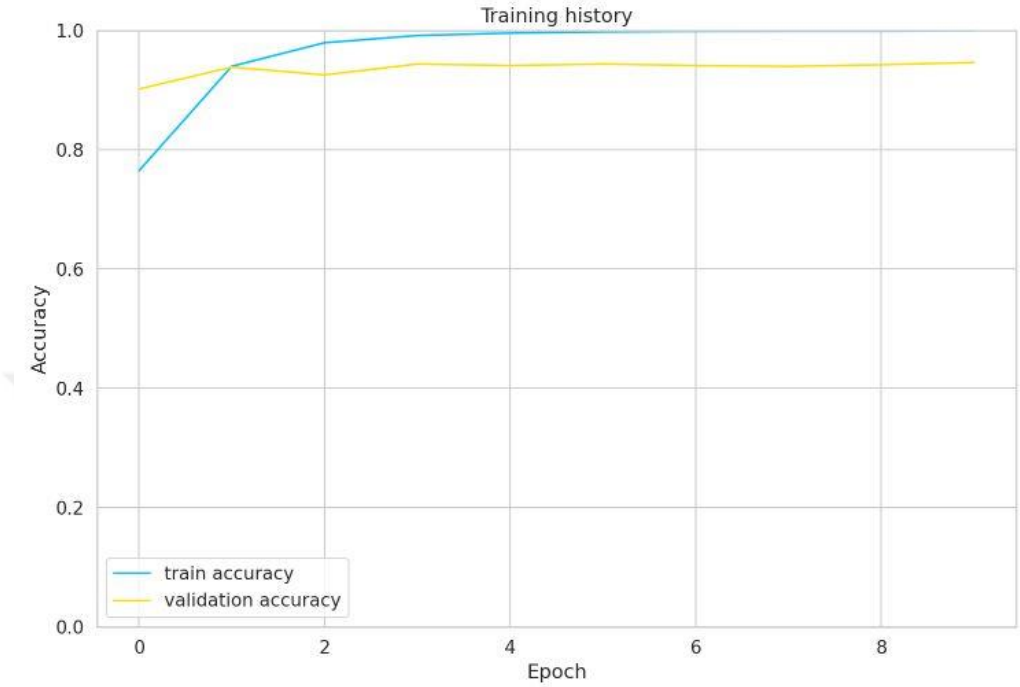
Öğrenme oranı (learning rate) (Adam): 5e-5, 3e-5, 2e-5

Epoch (devir) sayısı: 2,3,4, 10

Bahsedildiği üzere girdiler (batch size) 32 şerli alınarak bu işlem 10 kez tüm set bitene kadar yapılmıştır. 10 kez bu işlem sıfırdan tüm eğitim veri seti boyunca yapıldığı için oldukça zaman almıştır. Eğer epoch değeri 2 ya da 3 gibi bir değer seçilseydi işlem kısa fakat loss değeri yani öğrenmede eksikler,

maskelenmiş kelimelerdeki doğru tahmin ve sonucu olarak test verisindeki doğruluk başarısı düşük elde edilmiş olurdu. Modelin devir tekrar sayısı ile öğrenme sonucunda elde edilen doğruluk oranının grafiği aşağıda verilmiştir.

Grafik 7: Epoch sayısı ve doğruluk orantısı



Tablo 12'de ise bir epoch çalışma ortalama süresi gösterilmiştir.

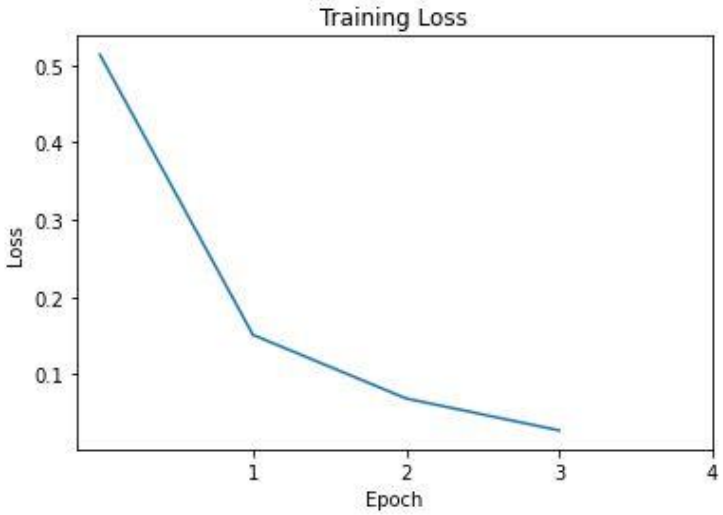
Tablo 11: Bert işlem detayı

```
Val   loss 0.4275357335014481 accuracy 0.9462647444298821

CPU times: user 58min 35s, sys: 32.9 s, total: 59min 8s
Wall time: 59min 13s
```

Aşağıdaki grafikte bu tablodaki kayıp değeri ve işlem tekrar sayısı arasındaki bağlantı verilmiştir.

Grafik 8: İşlem sayısı ve kayıp değeri



Şemada işlem sayısı arttıkça loss değerlerinin düştüğü görülmektedir.

Her tekrar (epoch) süresi CPU ile yaklaşık bir saat sürmüştür. GPU kullanım durumunda yaklaşık 1-2 dakika işlem süresi öngörülmektedir.

Modelin eğitim ve test sonucu çıktıları değerlendirme başlığında açıklanmıştır. Sinema, film, borsa, finans gibi alanlar için yapılan yorumlar ile öğrenimi gerçekleştirilmiş Bert modeli artık Twitter'dan #Tsla kelimesi ile toplanan günlük 100 tweetlerin duygu analizini yapması için hazırlanmış durumdadır. Modele bu tüm tweet veri seti verilerek her yorum için 0 ile 1 arasında duygu puanlaması yaptırılmıştır.

Tablo 12: Bert model ile Tweet sentiment analizi

date	Text	Sentiments
1/1/2016	#HBRLive: Is #Tesla's	0,9999610185623160
1/1/2016	TSLA: $240.01 - Mark	0,9999639987945550
1/1/2016	CityFALCON score giv	0,9999347925186150
1/1/2016	Potential Tesla compet	0,288075170828,96
1/1/2016	TSLA: $240.01 - Mark	0,9999639987945550
1/1/2016	Faraday Future aims t	0,354088188032,619
1/1/2016	TSLA: $240.01 - Mark	0,9999639987945550
1/1/2016	Panasonic to build $40	0,9999576807022090
1/1/2016	A collection of testimor	0,9997859597206110
1/1/2016	Tesla Motors (TSLA) to	0,43861677113,4547
1/1/2016	TSLA: $240.01 - Mark	0,9999639987945550
1/1/2016	We calculated 63-day	0,9722539186477660
1/1/2016	3 Activity Ratios to Act	0,9998416900634760
1/1/2016	Looking for someone v	0,9999247789382930
1/1/2016	Will Faraday Future Re	0,74341471190,5643
1/1/2016	Will Faraday Future Re	0,545435868843,924
1/1/2016	$TSLA https://t.co/jVy	0,12456854165066000
1/1/2016	TSLA: $240.01 - Mark	0,9999635219573970
1/1/2016	Will Faraday Future Re	0,49723406846,1693

Tablo 12'de de görüldüğü gibi, 0 negatif yorumları 1 ise pozitif yorumları anlatacak şekilde model tüm yorumlar için bu aralıkta puanlama yapmıştır. Model çalıştırılması sonucu her yorum için 0 ile 1 aralığında duygu analiz değeri çıkarmıştır. Günde 100 tweet alınmıştır ve her satırın farklı sentiment analiz değeri vardır fakat hisse senedi için günde bir kez kapanış değeri vardır ve bu yüzden günlük 100 tweet değerlerinin bir ortalaması alınarak hisse senedi değer tablosu ile birleştirilerek neural prophet için kullanıma hazır hale getirilmiştir.

4.6.2 Neural Prophet

Neural prophet zaman serisi verilerini modellemek için üretilmiş sinir ağlarına dayalı python kütüphanesi modelidir. Neural Prophet aşamaları yöntem başlığında açıklanmıştır. Modelin çıktıları olarak;

Tablo 13: Tsla market değeri ve sentiment değerleri

	Date	Open	High	Low	Close	Volume	Sentiments
0	1/4/2016	46.144001	46.276001	43.799999	44.681999	34135500	0.622939
1	1/5/2016	45.271999	45.377998	44.000000	44.686001	15934000	0.371077
2	1/6/2016	44.000000	44.009998	43.195999	43.807999	18895500	0.704350
3	1/7/2016	42.838001	43.688000	42.734001	43.130001	17771500	0.687406
4	1/8/2016	43.571999	44.088001	42.153999	42.200001	18140500	0.758833
5	1/11/2016	42.801998	42.889999	40.599998	41.570000	20457000	0.900371
6	1/12/2016	42.320000	42.748001	41.062000	41.993999	15459500	0.751501
7	1/13/2016	42.402000	42.529999	40.000000	40.062000	20632000	0.705132
8	1/14/2016	40.442001	42.000000	38.675999	41.236000	32453500	0.800704
9	1/15/2016	39.793999	41.014000	39.450001	40.998001	27893000	0.689339

Yukarıdaki tabloda TSLA market için Yahoo Finance'den çekilmiş değerler ile Bert modelinin sonucu günlük 100 adet tweet'in sentiment analizinin ortalaması birleştirilmiştir.

Tablo 14: Tsla market tahmini için hesaplanan yeni ve alınan kolonlar

	ds	y	add1	...	add7	add8	add9
1401	7/28/2021	649.789978	647.000000	...	48.705501	3.749044e+09	0.644466
1402	7/29/2021	700.000000	649.789978	...	58.563120	3.779439e+09	0.621174
1403	8/2/2021	719.000000	700.000000	...	66.042256	3.813054e+09	0.712506
1404	8/3/2021	711.000000	719.000000	...	66.056548	3.834675e+09	0.656869
1405	8/4/2021	716.000000	711.000000	...	66.313889	3.851677e+09	0.642823
1406	8/5/2021	711.900024	716.000000	...	67.156988	3.864597e+09	0.536683
1407	8/6/2021	710.169983	711.900024	...	60.348163	3.849021e+09	0.596637
1408	8/9/2021	713.989990	710.169983	...	64.053169	3.863736e+09	0.617222
1409	8/10/2021	712.710022	713.989990	...	62.437456	3.850483e+09	0.699707
1410	8/11/2021	706.340027	712.710022	...	61.476225	3.840744e+09	0.615540

ds – tarihi

y – ertesi günün market kapanış değerini

add1 – açılış değerini

add2 – en yüksek değeri

add3 – en düşük değeri

add4 – kapanış değerini

add5 – işlem hacmini

add6 – basit hareketli ortalama (SMA) hesaplaması sonucunu

add7 – (RSI) hesaplaması sonucunu

add8 – (OBV) hesaplaması sonucunu

add9- sentiment sonucunu göstermektedir.

Tablo 15: Neural Prophet sonucu gelecek 50 günkü market kapanış tahmin değerleri

0	626.517508
1	606.082947
2	575.260401
3	602.863023
4	607.866821
5	603.175207
6	603.233533
7	613.689447
8	610.822226
9	622.121210
10	600.752835
11	607.577271
12	622.273623
13	628.494972
14	621.699381
15	628.002743
16	661.537898
17	687.510812
18	674.535884
19	695.553332
20	682.493594
21	686.881281
22	679.769734
23	688.591159
24	661.085058
25	645.815533
26	656.012266
27	657.872027
28	690.813951
29	671.961222
30	658.040807
31	651.664428
32	646.108644
33	647.995677
34	660.790587
35	656.936867
36	651.280743
37	644.040383
38	661.900063
39	642.073113
40	649.074863
41	683.722755
42	718.701709
43	709.933796
44	716.653789
45	715.937343
46	702.022179
47	716.075318
48	710.610738
49	708.802991

Yukarıdaki tablo TSLA hisselerinin neural prophet modelinin sonucunda elde edilen 50 günlük tahmin değerlerini göstermektedir.

Tablo 16: Gerçekleşen ve tahmin edilemeyen değer tablosu

```
Actual y [0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0]
Predicted [1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1]
```

Yukarıdaki tabloda 1 market değerinin bir sonraki gün arttığı 0 ise azaldığını göstermektedir. Actual gerçekte olan artış ve azalışı predicted ise tahmin edilen artış ya da azalıştaki değerleri vermektedir.

4.7. Değerlendirme (Evaluation)

Sadece doğruluk (accuracy) sonucuna bakarak modelin değerlendirmesi yanlış yönlendirmektedir. Bu yüzden birden fazla metriğe bakmak model başarısı anlamında daha doğru yorum yapılmasını sağlamaktadır. Doğruluk değeri, doğru tahmin edilen değerlerin veri setindeki tüm değerlere oranı şeklinde açıklanabilir. Doğruluk değerinin tek başına yeterli olmaması örneğin; suçlu ve masum insanların olduğu 100 kişilik bir grupta 10 kişi suçlu geri kalan 90 kişi masum şeklinde dağılım vardır. Herhangi 10 kişinin suçlu olduğu söylenirse ve aslında suçlu olup ama masum insanların arasında olan olursa oldukça yanıltıcı bir durum açığa çıkmış olur.

$$Doğruluk = \frac{True\ Pozitif + True\ Negatif}{True\ Pozitif + Falsa\ Pozitif + \ True\ Negatif + False\ Negatif}$$

Verilen örneğe göre suçlu olan fakat masum etiketi alan kişi (false negatif) yanlış şekilde değerlendirilmiş olur.

Kesinlik (precision), etiket verilen yani seçilen durumların gerçekte de doğru etiketi aldığı durunlardır.

$$Kesinlik = \frac{True\ Pozitif}{True\ Pozitif + Falsa\ Pozitif}$$

Örneğin, kredi veren kurum olarak kredi verilmesi gereken kişiye riskli etiketi vererek kredi verilmemesine neden olursa bu kurumun kaybıyla sonuçlanıp maliyeti yüksek bir durum çıkarmış olur. Bu yüzden false pozitif değerinin yüksek olması model verimliliği açısından önemlidir.

Duyarlılık (recall) değeri False negatiflerin değeridir. Anlamı, bir üstteki örneğin tersi yani bankada kredi isteyen bir kişinin aslında almaması gerekirken güvenilir şekilde etiket çıkmasından dolayı o kişiye bankalar tarafından kredi verilmesi durumunu oluşturur.

F1 score değeri, kesinlik ve duyarlılık test değerlerinin harmonik ortalamasıdır. Harmonik ortalama normal ortalamadan farklıdır ve değerlerin direk toplanıp ikiye bölünmesi yerine durumlardaki uç noktaların değerlendirilmesi ve ona göre ortalama alınmasına neden olmaktadır.

$$F_1 = \ 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık}$$

F1 score değerinin doğruluk yerine daha çok kullanılmasının nedenlerinden birisi de veri kümesindeki eşit olmayan dağılım hakkında model başarısını daha doğru göstermesidir.

Bu değerler ve anlamları göz önünde bulundurulduğunda model başarısı için hepsi incelenmiştir.

Tablo 17: Bert Model sonuç tablosu

	precision	recall	f1-score	support
negative	0.95	0.97	0.96	387
positive	0.97	0.95	0.96	377
accuracy			0.96	764
macro avg	0.96	0.96	0.96	764
weighted avg	0.96	0.96	0.96	764

Bert modelinin kesinlik (precision) değeri yani pozitif olanların gerçekte de ne kadar pozitif olduğu, duyarlılık (recall) değerleri tablodan da görülebileceği gibi oldukça yüksek başarı elde edilmiştir.

Tablo 18: Confusion matrix teorik

		Tahmin Edilen	
		Negatif (N) -	Pozitif (P) +
Gerçekte Olan	Negatif -	True Negatif (TN)	False Pozitif (FP)
	Pozitif +	False Nagatif (FN)	True Pozitif (TP)

Tablo örnek şeklinde açıklanacak olursa,

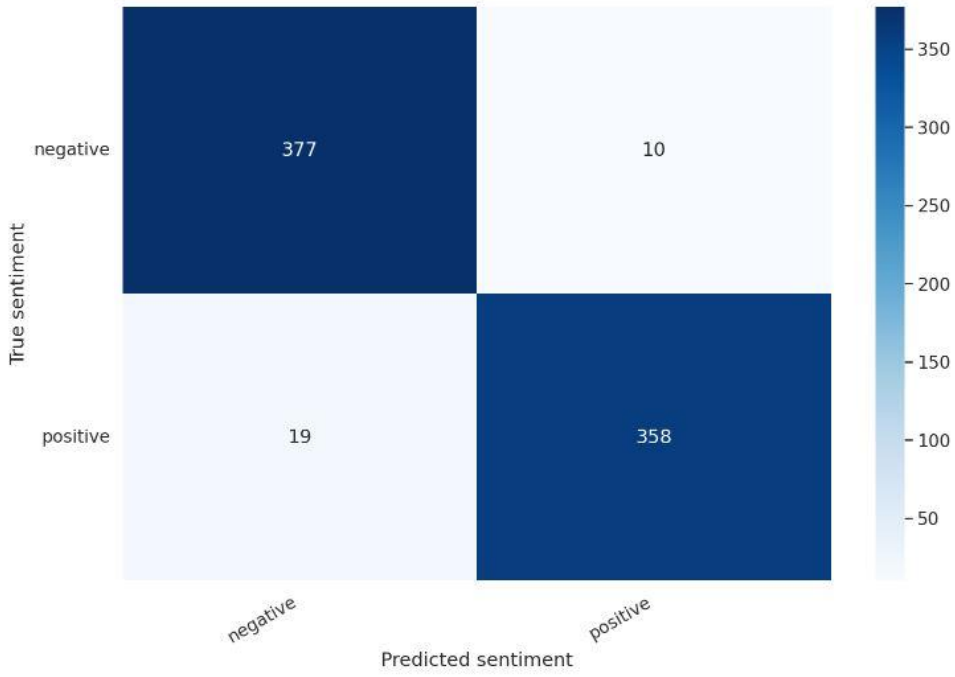
True pozitif (TP) değeri, hırsız olarak etiketlenen kişilerin gerçekten de hırsız olduğu durumu,

True negatif (TN) değeri, hırsız değil şeklinde etiketlenen kişilerin gerçekten de hırsız olmama durumu,

False pozitif (FP) değeri, hırsız olmayan bir kişiye hırsız etiketinin verildiği durumu,

False negatif (FN) değeri, gerçekte hırsız olan bir kişiye hırsız değil etiketinin verildiği durumu göstermektedir.

Grafik 9: Bert model Confusion matrix



Bert modelinde tüm veri setini bölme işlemi sonucu eğitim için 13743 satır metin, doğrulama için 763 satır metin ve test için 764 satır metin elde edilmişti. 764 satırın yani 764 yorumun test sonucuna göre 377 tanesi true negatif, 368 yorumun ise true pozitif olarak doğru etiketleri aldığı görülmüştür.19 yorumun false negatif (type 2 error) ve 10 yorumun false pozitif (type 1 error) etiketlenerek modelin yaptığı yanlışlar da tabloda gösterilmektedir.

Bu tez kapsamında Bert modelinin başarısı yeterli bulunmuş olup Twitter'dan alınan tweetlerin duygu puanlamaları model ile yapılmıştır. Çıktı tablosu model başlığı altında paylaşılmıştır.

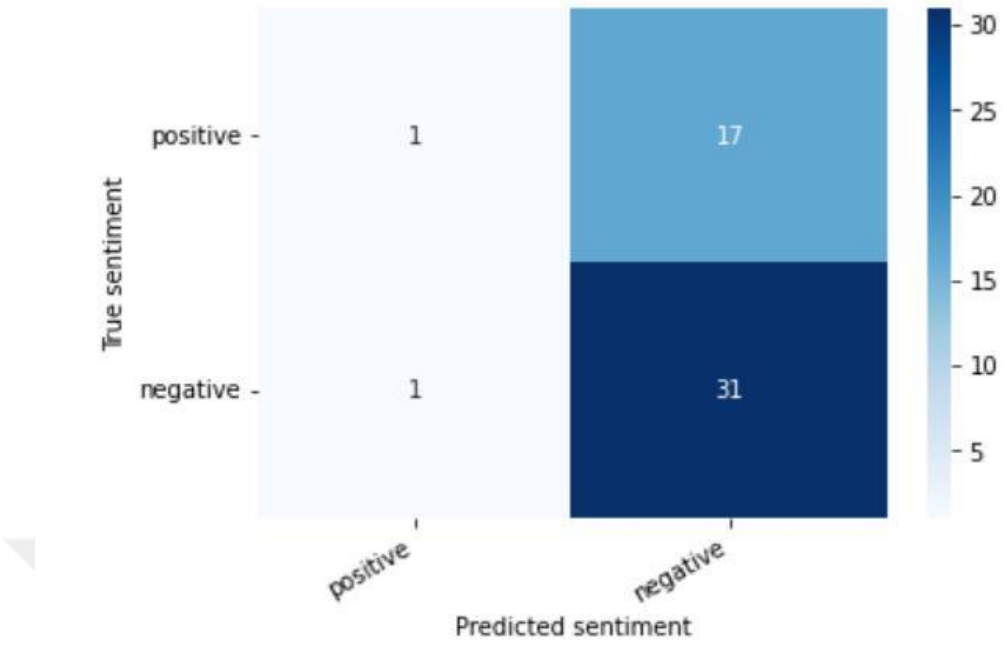
50 günlük tahmin için model başarısı 64 olarak hesaplanmıştır. TP, TN, FP, FN değerleri aşağıdaki confusion matrixte gösterilmektedir.

Neural Prophet modelinden elde edilen model başarısı;

Tablo 19: Neural Prophet sonuç değerleri

```
Accuracy is 64.0
True positve 31
True negative 1
False positive 17
False negative 1
50
```

Grafik 10: Neural Prophet confusion matrix



Teknolojik gelişmenin sonucu olarak internet kullanımındaki kolay erişim sayesinde sosyal medyanın da önemi her geçen gün artmaktadır. Sosyal medyadaki kulllanım artışı etkileşimi, hatta anlık etkileşimi arttırmaktadır. Bu etkileşimin olumlu ve olumsuz yanları bulunmaktadır. Olumlu yanlarından bazıları; doğal afet gibi durumlarda hızlı duyurularak kampanya düzenlenmesi ve büyük yardımların toplanabilmesi, daha önce hiç gidilmemiş yerler, restoranlar hakkında fikir edinerek gidilebilmesi vb. Olumsuz yanlarından bazıları ise; hızlı etki ve tepki oluşturabilecek ve halkın huzurunu bozabilecek propagandalar, rakiplerin bilinçli yaptırdığı kötü yorumlar vb.

Sosyal medya kullanımının artması ve etkileşimin de artmasından dolayı bu tez çalışması kapsamında sosyal medyadaki yorumların borsa üzerine etkisi ölçülmüştür.

Öncelikle sinema, film, borsa, hisse senedi yorumları ve bu yorumların pozitif ve negatif olacak şekilde sentiment analizleri ile Bert modeli eğitilmiştir. Twitter'dan belirli tarih aralıklarında olmak üzere günlük 100 #TSLA market hakkında yazılmış tweet alınarak bu yorumların Bert modeli ile duygu analizi yapılarak günlük bir değer ortalaması tespit edilmiştir. Bu duygu etiketinin olumlu ise ertesi gün hisse değerinin artış göstermesi, negatif ise ertesi gün hisse senedi değerinin azalması beklenmiş ve bu test edilmiştir.

Sonuç olarak Bert model başarısının 96% civarında olmasına rağmen Neural Prophet başarısı 64% civarında olmuştur. Başarı tatmin edici ve başta ön görülen sosyal medyadaki yorumun ertesi gün hisse senedi değerlerine etkisi vardır tezi doğrulanmıştır.

Günlük 100 tweet'in o günkü duyguyu çıkarmak için eksik kalabileceği ve bu yüzden çok daha fazla tweet alınmasıyla başarının arttırılabileceği öngörülmektedir. Ayrıca daha büyük veri seti ve uzun süreli veri toplama işleminin yapılması, modeli etkileyebilecek dış veri setlerinin kullanılması (örneğin siyasi bazı toplantı günleri, döviz değerleri vb), daha büyük bir işletim sistemine sahip modellerin işlevini de arttıracağı için başarıyı arttırabilecek faktörler olarak düşünülmektedir.

KAYNAKLAR

Adalı, E. (2012). Doğal Dil İşleme. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, *5*(2).

Akdağ, M., Bozma, G., (2021). Stok Akış Modeli ve Facebook Prophet Algoritması İle Bitcoin Fiyatı Tahmini. *Uluslararası Ekonomi İşletme ve Politika Dergisi, 5(1)*, 16-30.

Amasyalı, M. F., & Diri, Banu (2005). Bir soru cevaplama sistemi: Baybilmiş. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, *1*(1).

Aslan, E. Ş., & Kolancı, D. (2019). Semantik Web'in Marka İtibarına Etkisi Üzerine Bir Araştırma. *The Journal of International Scientific Researches*, *3*(4), 208-220.

Aydin, İ. (2008). Türkçe Ve Fransızca'nın Sözdizimi Üzerine Birkaç Söz 1. *Synergies Turquie*, (1), 25-35.

Bal, G., Gültunca, C., Karakaş, A., & Güngör, T. (2017). Özgeçmişlerden Ve İş İlanlarından Doğal Dil İşleme Yöntemleri İle Bilgi Çıkarımı, *TBD 31. Ulusal Bilişim Kurultayı Bildiriler Kitabı,* 125-131

Borneklint, N. (2021). Forecasting prices of Bitcoin and Google stock with ARIMA vs Facebook Prophet.

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Çalışkan, M., & Mencik, Y. (2015). Değişen Dünyanın Yeni Yüzü: Sosyal Medya. *Akademik Bakış Uluslararası Hakemli Sosyal Bilimler Dergisi*, (50), 254-277.

Dara, A. Y. (2011). Kamu yönetiminde halkla ilişkiler ve sosyal medya. *Selçuk İletişim*, *7*(1), 5-15.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Göçgün, Ö. F., & Aytuğ, O. N. A. N. (2021). Amazon Ürün Değerlendirmeleri Üzerinde Derin Öğrenme/Makine Öğrenmesi Tabanlı Duygu Analizi Yapılması. *Avrupa Bilim ve Teknoloji Dergisi*, (24), 445-448.

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649). Ieee.

Guven, Z. A. (2021, September). The Effect of BERT, ELECTRA and ALBERT Language Models on Sentiment Analysis for Turkish Product Reviews. In *2021 6th International Conference on Computer Science and Engineering (UBMK)* (pp. 629-632). IEEE.

Güzel, E. (2016). Dijital kültür ve çevrimiçi sosyal ağlarda rekabetin aktörü:"Dijital habitus. *Gümüşhane Üniversitesi İletişim Fakültesi Elektronik Dergisi*, *4*(1).

Sel, İ., & Hanbay, D. (2021). Ön Eğitimli Dil Modelleri Kullanarak Türkçe Tweetlerden Cinsiyet Tespiti. *Firat University Journal of Engineering*, *33*(2), 675-684

Hugging Face API, https://huggingface.co/ adresinden alındı.

Işık, A. H., & Yağcı, A. ş. (2020). Sıradan Sıraya LSTM Modeli ile Telegram Bot Uygulaması. *Gazi Mühendislik Bilimleri Dergisi*, *6*(1), 32–39. https://web.archive.org/web/20201212071004id_/https://dergipark.org.tr/tr/download/article-file/1081622

Kapan, K., & Üncel, R. Gelişen Web Teknolojilerinin (WEB 1.0-WEB 2.0-WEB 3.0) Türkiye Turizmine Etkisi. *Safran Kültür ve Turizm Araştırmaları Dergisi*, *3*(3), 276-289.

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

Khayyat, M., Laabidi, K., Almalki, N., & Al-zahrani, M. (2021). Time Series Facebook Prophet Model and Python for COVID-19 Outbreak Prediction. *CMC-COMPUTERS MATERIALS & CONTINUA*, *67*(3), 3781-3793.

Koçyiğit, M., & Koçyiğit, A. (2018). Değişen ve Gelişen Dijital İletişim: Yazılabilir Web Teknolojisi (Web 2.0). *Dijital Kültür ve İletişim*, *1*, 19-48.

Konakoğlu, B. (2020). Çok Katmanlı Algılayıcı Yapay Sinir Ağı ile Jeodezik Elipsoidal Koordinatların (φ, λ, h) 3 Boyutlu Global Kartezyen Koordinatlara (X, Y, Z) Dönüşümü. *Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *10*(3), 702–710.

Korkusuz, R., & CARUS, A. (2020). Futbol Müsabakaları ile İlgili Tweetlerin Anlık Duygu Analizi. *Avrupa Bilim ve Teknoloji Dergisi*, 386-396.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.

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

Mosser, L., Dubrule, O., & Blunt, M. J. (2018). Stochastic reconstruction of an oolitic limestone by generative adversarial networks. *Transport in Porous Media*, *125*(1), 81-103.

Nagesh, P. (2021). Combination of Facebook Prophet and Attention-Based LSTM with Multi-Source data for Indian Stock Market Prediction.

Özevin, O. Firmaların Sosyal Medya Kullanımlarının Ağ, Sektör ve Finansal Performans Açısından Araştırılması: BIST Uygulaması. *Finans Ekonomi ve Sosyal Araştırmalar Dergisi*, *6*(3), 408-420.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, H., and Carey, T. (1994). *Human-Computer Interaction*. Pearson Education Limited: Essex, England.

Tuzcu, S. (2020). Çevrimiçi Kullanıcı Yorumlarının Duygu Analizi ile Sınıflandırılması. *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, *1*(2), 1-5.

Uçan, A.(2020). Türkçe His Analizinde Optimizasyon Ve Ön-Eğitimli Modellerin Kullanımı. (Doktora Tezi)

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sevli, O., & Kemaloğlu, N. Olağandışı Olaylar Hakkındaki Tweet'lerin Gerçek ve Gerçek Dışı Olarak Google BERT Modeli ile Sınıflandırılması. *Veri Bilimi*, *4*(1), 31-37.

Solmaz, B., Tekin, G., Herzem, Z., & Demir, M. (2013). İnternet ve sosyal medya kullanimi üzerine bir uygulama. *Selçuk İletişim*, *7*(4), 23-32.

Şeker, Ş. E., & Erdoğan, D. (2018). *Knime ile Uçtan Uca Veri Bilimi* (First ed., Vol. 1). Demet ERDOĞAN. https://www.dr.com.tr/ekitap/knime-ile-uctan-uca-veri-bilimi

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37-45.

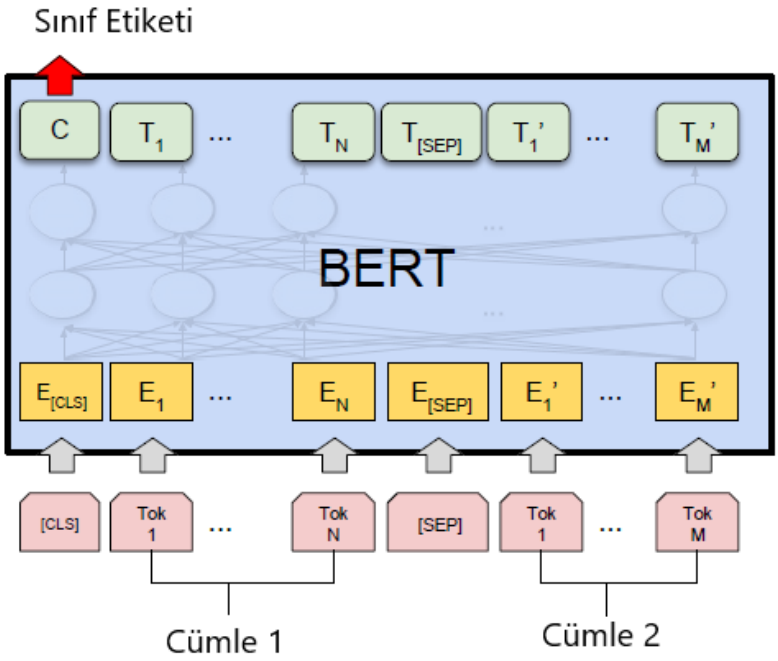
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*.

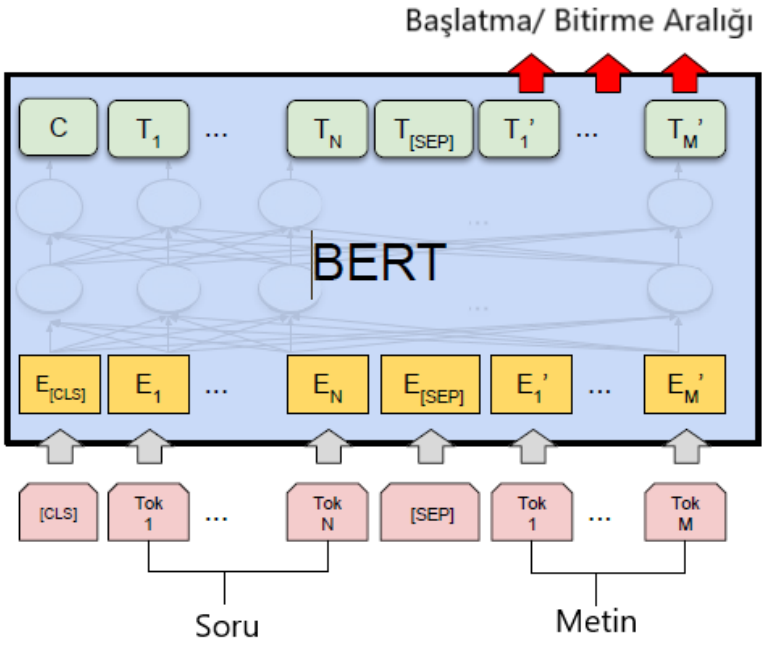
Yılmaz, H. & Yumuşak, S. (2021). Açık Kaynak Doğal Dil İşleme Kütüphaneleri . İstanbul Sabahattin Zaim Üniversitesi Fen Bilimleri Enstitüsü Dergisi , UAKK2020 , 81-85

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision* (pp. 19-27).

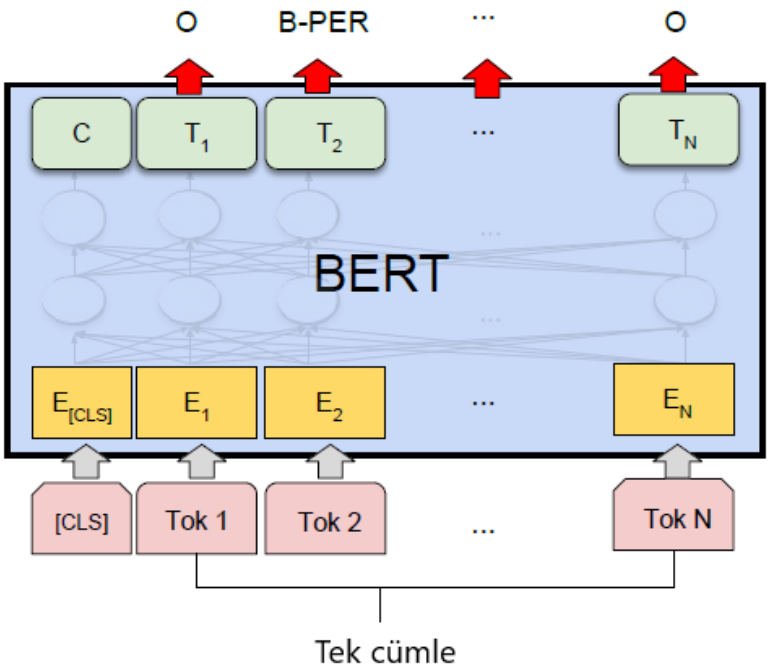
EKLER



Bert modeli cümle cümlelerin bağlantılıp olup olmadığının belirlenmesi için eklenen katman



Bert modelinde soru cevap katmanı



Bert modeli tek cümle analiz katmanı