TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DİNAMİK SOSYAL AĞLARDA AKAN VE ÇOK BOYUTLU VERİ ÜZERİNDEN ANALİZ VE TAHMİN YAPILMASI

DOKTORA TEZİ

Onur Can SERT

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Doç. Dr. Tansel ÖZYER

NİSAN 2020

Fen Bilimleri Enstitüsü Onayı

.....

Prof. Dr. Osman EROĞUL
Müdür

Bu tezin Doktora derecesinin tüm gereksinimlerini sağladığını onaylarım.

.....

Prof. Dr. Oğuz ERGİN
Anabilim Dalı Başkanı

TOBB ETÜ, Fen Bilimleri Enstitüsü'nün **121117703** numaralı Doktora Öğrencisi **Onur Can SERT**'in ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"DİNAMİK SOSYAL AĞLARDA AKAN VE ÇOK BOYUTLU VERİ ÜZERİNDEN ANALİZ VE TAHMİN YAPILMASI"** başlıklı tezi **20.04.2020** tarihinde aşağıda imzaları olan jüri tarafından kabul edilmiştir.

Tez Danışmanı: **Doç. Dr. Tansel ÖZYER** .....
TOBB Ekonomi ve Teknoloji Üniversitesi

Jüri Üyeleri: **Prof. Dr. Faruk POLAT (Başkan)** .....
Orta Doğu Teknik Üniversitesi

Doç. Dr. Fatma Betül ATALAY SATOĞLU .....
TOBB Ekonomi ve Teknoloji Üniversitesi

Prof. Dr. Ali Aydın SELÇUK .....
TOBB Ekonomi ve Teknoloji Üniversitesi

Prof. Dr. Reda ALHAJJ .....
Medipol Üniversitesi

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, alıntı yapılan kaynaklara eksiksiz atıf yapıldığını, referansların tam olarak belirtildiğini ve ayrıca bu tezin TOBB ETÜ Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırlandığını bildiririm.

.

Onur Can Sert

ÖZET

Doktora Tezi

DİNAMİK SOSYAL AĞLARDA AKAN VE ÇOK BOYUTLU VERİ ÜZERİNDEN ANALİZ VE TAHMİN YAPILMASI

Onur Can Sert

TOBB Ekonomi ve Teknoloji Üniveritesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Doç. Dr. Tansel Özyer

Tarih: Nisan 2020

Makine öğrenmesi teknikleri ve bu tekniklerin uygulanabilir olduğu alanlar, veri miktarının artması ve veriye ulaşımın kolaylaşması ile birlikte oldukça ön plana çıkmıştır. Veri kümeleri üzerinde bu yöntemler kullanılarak farklı alanlara yönelik tahmin modellerinin geliştirilmesi mümkündür. Bunun yanında doğal dil işleme yöntemleri, metin verisinin analiz edilmesi ve anlamlandırılması noktasında birçok farklı yöntemi içerisinde bulundurmaktadır.

Yapılan çalışmada, doğal dil işleme yöntemleri kullanılarak, haber ve sosyal medya verisi analiz edilmiştir ve analiz sonuçlarından öznitelik kümeleri oluşturulmuştur. Oluşturulan öznitelik kümeleri ile sayısı fazla olan seyrek öznitelik kümeleri için ölçeklenebilir bir eğitim ve tahmin sistemi ortaya konmuştur. Sistemin geliştirilmesi için, 1 yıllık zaman aralığı içerisinde New York Times web sayfasından 12.560 adet makale ve 4 aylık zaman aralığı içerisinde Twitter isimli sosyal medya platformundan 2.854.333 adet paylaşım toplanmıştır. Toplanan veri üzerinden varlık isimleri tanımlanmış, düşünce analizi yapılmış ve konu modelleri oluşturulmuştur. Geliştirilen sistemin bir başka çıktısı olarak, analizi yapılan metin verileri üzerinden sosyal ağların oluşturulmasını sağlanmıştır ve üretilen sosyal ağların farklı zaman aralıklarındaki değişimleri gözlemlenmiştir. Elde edilen analiz sonuçları ve sosyal ağlar

doğrultusunda öznitelik kümeleri oluşturulmuş ve bu öznitelik kümeleri ile elastik ağ regresyonu temelli bir eğitim yöntemi geliştirilmiştir.
Önerilen bu sistem ile birçok farklı veri kümesinin analiz edilebileceği ve bu analizler doğrultusunda farklı değerleri tahmin etmeye yönelik tahmin modellerinin geliştirilebileceği görülmüştür. Bunun bir örneğini ortaya koymak adına Dow Jones endeksinin yönünün tahmini bir vaka olarak seçilmiştir. Önerilen eğitim yöntemi ile farklı modeller eğitilmiş ve eğitilen bu modeller ile Dow Jones endeksinin hareket yönünün tahmin edilmesine yönelik deneyler yapılmıştır. Bu deneyler sonucunda, önerilen eğitim yönteminin, umut vaat edici sonuçlar veren tahmin modelleri ortaya koyduğu gözlemlenmiştir. Farklı deney gruplarının sonucunda, yüksek oranda tutarlı (70,90% değerine varan) sonuçlar elde edilmiştir. Elde edilen tahmin sonuçlarının aynı zamanda gerçek Dow Jones endeks değerleri ile pozitif bir korelasyon (0,2315 korelasyon katsayına değerine varan) içerisinde olduğu da gözlemlenmiştir. Son kısımda, farklı öznitelik kümeleri ile eğitilen tahmin modellerinin sonuçları birbiri ile karşılaştırılmış ve öne çıkan zaman aralıkları ve öznitelik kümeleri analiz edilmiştir. Deney sonuçları, haber ve sosyal medya verisinin, doğal dil işleme yöntemleri ile analiz edilmesinin ve analiz sonuçlarının tahmin modellerinin eğitimi için kullanılmasının finans alanında tahminler yapmak için değerli olduğunu göstermiştir.

Anahtar Kelimeler: Varlık isimlerinin tanımlanması, Konu modellemesi, Düşünce analizi, Sosyal ağ analizi, Hisse senedi yön tahmini, Makine öğrenmesi.

ABSTRACT

Doctor of Philosophy

ANALYSIS AND PREDICTION IN SPARSE AND HIGH DIMENSIONAL DATA WITH USING DYNAMIC SOCIAL NETWORKS

Onur Can Sert

TOBB University of Economics and Technology
Institute of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Assoc. Prof. Dr. Tansel Özyer

Date: April 2020

Machine learning techniques and applications of these techniques became very popular after the incremental of different data sources and with the ease of accessing the data. Prediction models can be trained with using these datasets which are collected from different sources. In addition, natural language processing techniques are also very useful for data mining and information extraction on text based data.

In this study, with using natural language processing techniques, a large collection of news and social media data is analysed and feature sets are created with results. Then, a scalable prediction system for sparse and high dimensional feature sets to predict stock market movements is built with these feature sets and results. For building that prediction system, 12,560 articles from New York Times covering 1 year time period and 2,854,333 tweets from Twitter covering 4 month time period are collected. The collected data are analysed with named entity recognition, sentiment analysis and topic modelling techniques. As another output of the designed system, social networks are created and analysed according to the various range of timeframes. Feature sets are created and elastic network regression based prediction models are trained with using the natural languages processing results, analysis results and social networks.

With using the proposed approach, different dataset can be analysed and different prediction systems can be created. To show an example of this, predicting direction of the Dow Jones Index, is selected as a case. Different prediction models are trained and used for predicting to stock market movements for Dow Jones Index. As a result of different sets of experiments, the models which are created with the proposed method made promising predictions. In different sets of experiments, highly accurate (up to 70.90% accuracy) predictions are made by the proposed approach. These predicted values also correlated (up to 0.2315 correlation coefficient value) with real Dow Jones Index values. Further, performance tests are made to show scalability of proposed method for various prediction models that are trained with different set of features. Experiment results show that it is possible to make reasonable stock movement prediction by integrating news and related social media data, analysing them using named entity recognition, sentiment analysis and topic modelling techniques together with prediction models which use features that are created from these analysis results.

Keywords: Named entity recognition, Topic modelling, Sentiment analysis, Social network analysis, Stock market movement prediction, Machine learning.

TEŞEKKÜR

Doktora öğrenim hayatım ve çalışmalarım boyunca bilgi birikimini ve deneyimlerini benimle paylaşan, anlayış ve sabırla çalışmalarıma katkıda bulunan danışmanım sayın Doç. Dr. Tansel Özyer'e, tez izleme komitelerime katılarak, verdikleri geri bildirimler ile beni yönlendiren sayın Doç. Dr. Fatma Betül Atalay Satoğlu ve sayın Prof. Dr. Mehmet Kaya'ya, tez savunmamda bulunarak yaptıkları yorumlar ile çalışmama geleceğe yönelik değer katan sayın Prof. Dr. Ali Aydın Selçuk, sayın Prof. Dr. Faruk Polat ve sayın Prof. Dr. Reda Alhajj'a, öğrenim hayatım boyunca tecrübelerinden faydalandığım TOBB Ekonomi ve Teknoloji Üniversitesi Bilgisayar Mühendisliği Bölümü'nün değerli öğretim üyelerine teşekkürlerimi sunarım.

Birlikte girişimcilik dünyasında birçok maceraya atıldığımız ve onlarca başarılı projeye birlikte imza attığımız Barış Okur, Emin Okutan, Cansu Ege Başçıl ve tüm Viveka ailesine ayrıca teşekkür ederim.

Arkadaşlığımızın başlangıcı üniversite hayatlarımızın ilk günlerine dayanan, bana olan desteklerini yıllardır hiç eksik etmeyen canım arkadaşlarım Tunç Akın, Begüm Akın ve Gözde Ünver'e çok teşekkür ederim.

Her zaman yanımda olan ve beraber birçok şeyi paylaştığım biricik dostum Nihan Oya Memlük Çobanoğlu'na, hayatı kendisi için zorlaştırmama rağmen desteğini hiçbir zaman esirgemeyen dostum Baran Çobanoğlu'na, arkadaşlıkları ile hayatı benim için daha keyifli kılan canım arkadaşlarım Pınar Berberoğlu ve Murat Ayhan'a çok teşekkür ederim.

Sadece bu tez çalışması boyunca değil, son yıllarda yaşadığım tüm zorluklara rağmen, bana olan sevgisini ve desteğini hiç eksik etmeyen, elinden gelen her türlü yardımı sunan ve yanımda olan hayat arkadaşım Sinem Laçin'e çok teşekkür ederim.

Son olarak, beni büyüten ve bugünlere gelmemi sağlayan, tüm hatalarıma ve yanlışlarıma rağmen beni karşılıksız seven ve bana olan inançlarını yitirmeyen değerli annem Refika Sert, babam Ahmet Sedat Sert ve kardeşim Anıl Sert'e çok teşekkür ederim.

İÇİNDEKİLER

Sayfa

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

KISALTMALAR

API	**:** Uygulama Programlama Arayüzü (Application Programming Interface)
CRF	**:** Koşullu Rastgele Alanlar (Conditional Random Fields)
HMM	**:** Saklı Markov Modelleri (Hidden Markov Models)
HTTP	**:** Üstmetin Transfer Protokolü (Hypertext Transfer Protocol)
JSON	**:** JavaScript Nesne Gösterimi (JavaScript Object Notation)
KAF	**:** KYOTO Açıklama Biçimi (KYOTO Annotation Format)
LDA	**:** Örtülü Dirichlet Ayrıştırması (Latent Dirichlet Allocation)
LSA	**:** Örtülü Anlam Çözümlemesi (Latent Semantic Analysis)
MPQA	**:** Çok Yönlü Soru Cevaplama (Multi-Perspective Question Answering)
MSAEnet	**:** Çok Adımlı Uyarlanabilir Elastik Ağlar (Multi-Step Adaptive Elastic-Net)
NERC	**:** Varlık İsimlerinin Tanımlanması ve Sınıflandırılması (Named Entity Recognition and Classification)
NLP	**:** Doğal Dil İşleme (Natural Language Processing)
NMF	**:** Matrisleri Negatif Olmayan Çarpanlarına Ayırma (Non-Negative Matrix Factorization)
pLSA	**:** Olasılıksal Örtülü Anlam Çözümlemesi (Probabilistic Latent Semantic Analysis)
PMI	**:** Noktasal Ortak Bilgiler (Pointwise Mutual Information)
POS	**:** Konuşma Bölümü (Part of Speech)
REST	**:** Temsili Durum Transferi (Representational State Transfer)
RMSE	**:** Hataların Ortalama Karekökü (Root Mean Square Error)
SVD	**:** Tekil Değer Ayrışması (Singular Value Decomposition)
SVM	**:** Karar Destek Makineleri (Support Vector Machines)
TF-IDF	**:** Terim Frekansı - Ters Metin Frekansı (Term Frequency - Inverse Document Frequency)
XML	**:** Genişletilebilir İşaretleme Dili (Extensible Markup Language)

1. GİRİŞ

Günümüzde, mobil ve web teknolojilerinde meydana gelen gelişmeler sayesinde, insanlar bilgiye çok daha kolay ve hızlı bir şekilde ulaşabilmektedir. Bu nedenle, dijital gazeteler ve sosyal medya platformları çok büyük oranda önem kazanmıştır. Teknolojik imkanların artması ve internetin tüm dünyada yaygınlaşması ile birlikte geleneksel medyada değişimler meydana gelmiştir. Artık insanlar dijital gazetelerden anlık olarak haber alabilmekte ve gündemi takip edebilmektedir. İnsanların bilgiye hızlı bir şekilde ulaşabilmesinin yanı sıra, teknolojik imkanları olan her bir birey aynı zamanda bir bilgi kaynağına da dönüşmüştür. Kişiler sahip oldukları dizüstü bilgisayar, cep telefonu, tablet, vb... cihazları kullanarak çok hızlı bir şekilde herhangi bir içeriği anında diğer insanlara ulaştırabilmektedir. Bu sayede herkes olaylardan hızlı bir şekilde haberdar olabilmekte ve bu olaylara anlık olarak tepki verebilmektedir.

Dijital gazetelerin artması ve sosyal medyanın popülerleşmesi sonucu her geçen saniye çok büyük miktarda veri üretimi gerçekleşmektedir. Dijital gazetelerde her gün yaklaşık 200 ile 500 arası makale yayınlanmaktadır. Sosyal medyada ise üretilen veri miktarı çok daha fazladır. Sadece 1 dakika içerisinde Facebook isimli sosyal medya ağında 900.000 adet gönderi üretmekte ve Twitter isimli mikroblog internet sitesinde 450.000 adet içerik gönderilmektedir. Veri miktarının bu kadar artmış olduğu bir noktada, bu verilerin üzerinden analizlerin yapılması ve anlamlandırılması çok daha fazla önem kazanmıştır.

Verilerin toplanıp, düzenlenip ve daha sonra da anlamlandırıldığı bu sürece veri madenciliği ismi verilmektedir. Veri madenciliğinin yapılması için bilgisayar bilimlerinde birçok farklı teknik yer almaktadır. Yazılı veriler üzerinde yapılan veri madenciliği işlemleri metin madenciliği konu başlığı altında toplanabilir. Metin madenciliğinin temelini ise doğal dil işleme (natural language processing) teknikleri [1] oluşturmaktadır. Doğal dil işleme, bir diğer adıyla bilişimsel dilbilim, makina öğrenmesi ve yapay zeka yöntemlerinden güç alarak, metinlerin bilgisayarlar tarafından anlamlandırılmasıdır. Doğal dil işleme teknikleri kullanılarak;

- Metin içerisinde yer alan cümlelerin biçimsel analizinin yapılması [2],
- Metin içerisindeki kişilerin, ülkelerin, organizasyonların, vb... yapıların tespit edilmesi [3],
- Metin üzerinden düşünce analizi yapılması [4, 5, 6],
- Metin içerisinde bahsedilen konuların tespit edilmesi [7],
- Metinlerin farklı dillere çevirilerinin yapılması [8] gibi birçok işlem yapılabilmektedir.

Doğal dil işlemlerinin yapılabilmesi ve önerilen yöntemlerin geliştirilebilmesi adına birçok çalışma grubu ve açık kaynak kodlu araç da bulunmaktadır.

Bir diğer taraftan, günümüzde popülerliği oldukça yüksek olan bir diğer konu ise tahmin modellerinin geliştirilmesi ve kullanımıdır. Tahmin modelleri, elde bulunan veriler doğrultusunda öbekleme, sınıflandırma ve regresyon gibi teknikler kullanılarak bir sistemin eğitilmesi sonucu ortaya çıkmaktadır. Daha sonra üretilen bu tahmin modelleri kullanılarak yeni gelen verilerin değerleri tahmin edilmektedir. Bu akıllı sistemler kullanılarak;

- Dijital içerik platformları veya alışveriş sistemleri için öneri sistemlerinin kurgulanması [9, 10, 11],
- Ürün fiyatlarında veya borsa endekslerinde değişim tahminlerin yapılması [12],
- Akıllı sistemler kurgulanarak öngörülemeyen kazaların tahmin edilmesi [13] gibi birçok farklı uygulama geliştirilebilir.

Bu tez kapsamında, doğal dil işleme teknikleri ve tahmin modelleri bir arada kullanılarak bir tahmin sistemi kurgulanmıştır. Bunun için ilk olarak farklı kaynaklardan metin verisi toplanmıştır. Toplanan bu veri üzerinde, varlık isimlerinin tespiti, konu modellemesi ve düşünce analizi olarak adlandırılan doğal dil işleme teknikleri kullanılarak analizler yapılmıştır. Bu analizler doğrultusunda farklı tahmin modellerinin kurgulanması mümkündür. Bu çalışma içerisinde finansal bir tahmin modeli kurgulanmasına yönelik bir vaka ele alınmıştır. Bu vaka doğrultusunda Dow Jones endeksinin yön değişimlerinin tahmin edilmesine karar verişmiştir. Kurgulanan yapıdan elde edilen analiz sonuçlarının farklı kombinasyonları ile seçilen vaka için tahmin modelleri kurgulanmış ve eğitilen tahmin modelleri ile Dow Jones

endeksindeki değişimler tahmin edilmiştir ve bu tahmin sonuçlarının başarımı ölçülmüştür.

Bu tez çalışması şu şekilde düzenlemiştir; Bölüm 2'de benzer çalışmalara ve bu çalışmaların detaylarına yer verilmiştir. Geliştirilen esnek ve ölçeklenebilir sistemin genel mimarisi ve içerisinde yer alan veri toplama ve doğal dil işleme modülleri Bölüm 3'de anlatılmış, bu modüllerin çalışma şekilleri hakkında detaylı bilgiler verilmiştir. Veri toplama ve doğal dil işleme modüllerinin çıktıları Bölüm 4'de farklı açılardan analiz edilmiştir. Bölüm 5'de veri toplama ve doğal dil işleme modüllerinin çıktıları kullanılarak kurgulanmış olan performansı yüksek tahmin modeline yönelik detaylar anlatılmıştır. Geliştirilen tahmin modelleri ile farklı deney düzeneklerinde elde edilen tahmin sonuçları Bölüm 6'da derlenmiş ve değerlendirilmiştir. Son olarak Bölüm 7'de yapılan tez çalışması hakkında genel bir değerlendirme yapılmış ve gelecek çalışmalar ile ilgili bilgiler verilmiştir.

2. İLGİLİ ÇALIŞMALAR

Borsa endeksinin tahminine yönelik çalışmalar incelendiğinde, birçok farklı teknik kullanılarak tahmin modellerinin eğitildiği ve tahmin sonuçlarının oluşturulmasına yönelik çalışmalar yapıldığı görülmüştür. Bu çalışmalar arasında, önerilen yöntemle karşılaştırıldığında benzerlik gösteren çalışmalar iki farklı grupta derlenmiştir. Bu gruplar aşağıdaki gibidir;

- Haber verileri üzerinden analizlerin yapıldığı ve tahmin modellerinin kurgulandığı çalışmalar.
- Sosyal medya verileri üzerinden analizlerin yapıldığı ve tahmin modellerinin kurgulandığı çalışmalar.

Çalışmaların ayrıldığı bu gruplardan ilki olan, tahmin modelinin oluşturulması ve tahmin sonuçlarının alınmasında haber verilerinin kullanıldığı tahmin modelleri, veri kaynağı olarak günlük ve finansal internet gazetelerini kullanmaktadır. Bu gazeteler içerisinde yer alan metinler resmi bir dile ve çok daha yapısal bir anlatım biçimine sahiptir. Bu sebeplerden ötürü, toplanan haber metinleri üzerinde uygulanacak olan doğal dil işleme tekniklerinin de daha iyi çıkarımlar yapması olasıdır.

Haber verilerinin toplandığı ve toplanan bu haber verileri üzerinde doğal dil işleme teknikleri kullanıldıktan sonra elde edilen analiz sonuçları ile tahmin modellerinin oluşturulduğu yöntemlere bakıldığında ilk olarak, Schumaker ve Chen tarafından önerilmiş olan Arizona Finansal Metin Sistemi (Arizona Financial Text System) (AZFinText) [14] isimli çalışma ile karşılaşılmıştır. Bu çalışmada, ilk olarak yayınlamış olan finansal sektöre yönelik olan haberler toplanmış ve bu haberler kelime torbası modeli (bag of words), isim tamlamaları ve varlık isimleri ile işaretlenmiştir. Yapılan işaretlemeler sonucunda her bir finansal makaleye yönelik bir profil oluşturulmuştur. Daha sonra oluşturulan bu finansal makale profilleri ile karar destek makineleri (support vector machines) (SVM) tabanlı bir tahmin modeli eğitilmiş ve S&P 500 isimli Amerika Birleşik Devletleri'nde yer alan hisse senedi endeksine yönelik tahminlerde bulunulmuştur. Tahmin sonuçları incelendiğinde en başarılı modelin 57,10% oranında hisse senetlerinin yönünü başarılı bir şekilde tahmin ettiği

görülmüştür. Buna ek olarak gerçekleştirilen bir simülasyon sonucunda eğer algoritmanın verdiği kararlara göre yatırım yapıldığı takdirde 2,06% oranında kazanç sağlanacağı öngörülmüştür.

Alanda yapılan çalışmalardan bir diğeri ise Wuthrich ve *diğerleri* [15] tarafından yapılmış olan, internet üzerinden yayın yapan finansal gazetelerdeki makalelerin eğitim ve tahmin için kullanıldığı tahmin modelleridir. Yapılan çalışmada, The Wall Street Journal, Financial Times ve benzeri ekonomi merkezli yayın yapan gazeteler anlık olarak takip edilmiş ve bu gazetelerde yayınlanan makalelerde geçen kelimelerden istatistiki modeller oluşturulmuştur. Sonrasında, oluşturulan bu istatistiki modeller ile regresyon temelli tahmin modelleri eğitilmiş ve Amerika, Avrupa ve Asya kıtalarındaki farklı ülkelerin borsalarındaki endeks değişimlerine yönelik tahminler yapılmıştır. Bu çalışmada yapılan deneylerin sonuçları incelendiğinde, 5 farklı endeks için ortalama 43.60% oranında başarılı tahminler yapıldığı görülmüştür.

Bu alanda yapılan bir başka çalışma ise Ding ve *diğerleri* [16] tarafından geliştirilmiş olan, finansal olaylar üzerine derin öğrenme teknikleri kullanılarak eğitilen tahmin modelleridir. Bu tahmin modelleri de eğitim için haber makalelerini kullanmaktadır. Haber makaleleri toplandıktan sonra doğal dil işleme yöntemleri teknikleri arasında yer alan morfolojik analiz teknikleri kullanılarak, haber içerisindeki özne, nesne ve yüklem grupları tespit edilmiştir. Örneğin; (Aktör = Microsoft, Eylem = dava etmek, Nesne = Barnes & Noble) bu gruplara bir örnektir. Toplanan tüm makaleler içerisindeki bu gruplar tespit edildikten sonra, bu veri ile yapay sinir ağları (neural networks) tabanlı tahmin modelleri eğitilmiş ve S&P 500 endeksinin hareketine yönelik tahminlerde bulunulmuştur. Yapılan çalışma içerisinde hem tek tek hisse senetlerinin fiyatlarının hem de S&P 500 endeksinin hareketi tahmin edilmiştir. Bu tahminler sonucunda tahmin edilen S&P 500 endeksinin yönünün 64,21% oranında, seçilen hisse senetlerinin yönünün ise 65,48% oranında başarılı tahmin edildiği gözlemlenmiştir.

Farklı bir yaklaşım ise Kogan ve *diğerleri* [17] tarafından sergilenmiştir. Yaptıkları çalışmada, finansal raporların içerisinde kullanılan dili analiz ederek, hisse senetlerinin değerlerindeki dalgalanmayı tahmin etmeye yönelik bir model kurgulamışlar ve bu hisse senetleri üzerinde risk analizi yapmışlardır. Bu çalışmada, sistemin eğitilmesi ve tahminlerin yapılması için Amerika Birleşik Devletleri'ndeki

farklı borsalarda yer alan farklı şirketlere ait hisse senetleri ve bu şirketlere ait dönemlik finansal raporları kullanmışlardır.

Yapılan bir başka çalışmada ise Kim ve *diğerleri* [18], düşünce analizi temelli bir yöntem önermiştir. Önerilen bu yöntemde ilk olarak, farklı kaynaklardan haber makaleleri toplanmış ve toplanan makaleler üzerinde düşünce analizi yöntemleri kullanılarak pozitif ve negatif anlam katan yapılar tespit edilmiştir. Bir sonraki aşamada ise elde edilen bu bilgiler kullanılarak KOSPI isimli Güney Kore borsa endeksinin değerine yönelik tahminler yapılmıştır. Önerilen yöntem doğrultusunda elde edilen tahmin sonuçları incelendiğinde, yapılan tahminlerin 55,00% oranında gerçek değerler ile örtüştüğü görülmüştür.

Bu çalışmaların yanında, literatürde bulunan farklı çalışmalarda, sadece resmi haber kaynaklarında yayınlanan makalelerin kullanılmadığı çalışmalar da yer almaktadır. Bu çalışmalarda metin madenciliği işlemlerinin yapılabilmesi ve eğitim modelinin oluşturulması adına forumlarda, kişisel internet sitelerinde veya sosyal medya sitelerinde paylaşılan, gazetelerde yer alan metin verisi ile karşılaştırılınca resmi olmaktan uzak olan veriler kullanılmaktadır. Bu doğrultuda toplanan veriler yapısal olarak daha kötü durumda olsa da içerik üreten kişi sayısı çok daha fazla olduğu ve içeriği yayınlama kriterleri çok daha yüzeysel olduğu için veri miktarı çok daha fazladır. Bu da tahmin modellerinin eğitilmesi ve tahmin sonuçlarının alınması noktasında değer yaratan bir durumdur.

Sosyal medya üzerindeki metin verilerinin toplandığı ve toplanan bu metin verileri üzerinde doğal dil işleme teknikleri kullanıldıktan sonra elde edilen analiz sonuçları ile tahmin modellerinin oluşturulduğu yöntemler incelendiğinde birçok farklı yaklaşım olduğu görülmektedir. Bu yaklaşımlardan biri Nguyen ve Shirai'nin çalışmasında [19] yer alan düşünce analizi temelli olan yaklaşımdır. Yapılan çalışmada, Yahoo'nun finans web sitesinin içerisinde yer alan Exxon Mobil, IBM, Dell, eBay ve The Coca-Cola Company'e ait kanalların mesajlaşma panolarında yer alan kullanıcı mesajları toplanmıştır. Daha sonra toplanan bu kullanıcı mesajları üzerinde, konu modellemesi ve düşünce analizi temelli bir algoritma çalıştırılmış ve kullanıcıların düşünceleri ile mesajlar işaretlenmiştir. Algoritmanın çalışması sonucu elde edilen konular ile bu konular hakkındaki kullanıcıların düşünceleri bilgileri bir arada kullanılarak her bir mesaj panolarından kullanıcı verileri toplanan her bir şirket için bir tahmin modeli oluşturulmuş ve oluşturulan bu tahmin modelleri yardımıyla ilgili şirketlerin hisse

senedi değerlerindeki değişimler tahmin edilmeye çalışılmıştır. Bu çalışmada yapılan deneylerin sonuçları incelendiğinde, farklı hisse senetleri için yapılan fiyat tahminlerinde, bu hisse senelerinin gerçek değerine oranla 3,57% ile 14,29% arasında sapma gerçekleştiği gözlemlenmiştir.

Bir başka çalışmada Bollen ve *diğerleri* [20] de düşünce analizi temelli bir yöntem önermiştir. Yapılan çalışmada ilk olarak, Twitter isim mikroblog sitesi üzerinden seçilen kullanıcıların paylaştığı tüm içerikler toplanmıştır. Sonrasında, geliştirilen OpinionFinder isimli düşünce analizi gerçekleştiren araç yardımıyla, paylaştığı içerikler toplanan kullanıcıların paylaşımlar 6 farklı sınıf altında yer alacak şekilde gruplanmıştır. Bu noktada, insanların duygularının doğrudan veya dolaylı olarak ekonomik hareketlerini etkileyeceğinden yola çıkarak bir tahmin modeli eğitmişlerdir. Son olarak eğitilen tahmin modeli ile Dow Jones endeksinin hareketine yönelik tahminlerde bulunulmuştur. Farklı parametreler ile yapılan deneyler sonucunda, çalışma içerisinde önerilen yöntemin 46,70% ile 86,70% arasında değişen yüzdeler ile Dow Jones endeksinin yönünün doğru tahmin edebildiği görülmüştür.

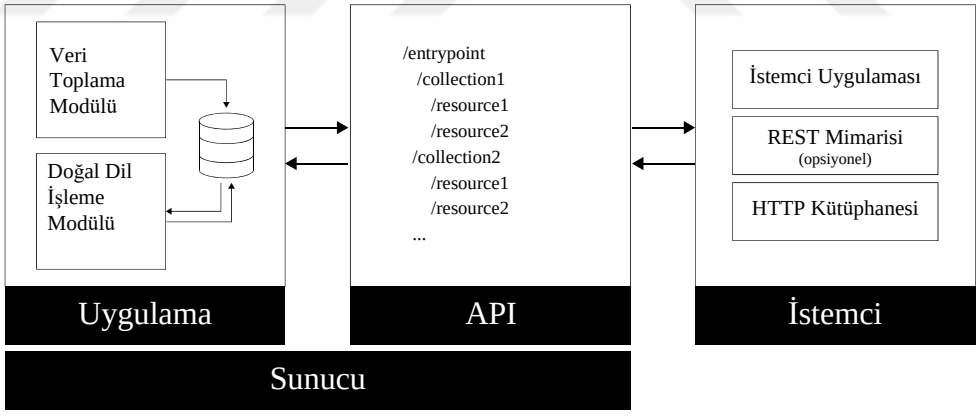
Benzer bir yaklaşım ise Pagolu ve *diğerleri* [21] tarafından sergilenmiştir. Bu çalışmada da Twitter üzerinden veri toplama işi gerçekleştirilmiştir. Ancak bu çalışmada, seçilen belirli kullanıcıların paylaştığı tüm içerikler yerine, belirli tarihler arasında "Microsoft" kelimesini içeren içerikler sosyal medya sitesinden toplanmıştır. Sonrasında, toplanan veri üzerinde düşünce analizi teknikleri uygulanmış ve Microsoft şirketine ait hisse senedinin fiyat değişimleri ile kişilerin sosyal medya üzerinden bu şirket ile ilgili yaptıkları paylaşımların düşünce analizi sonuçlarının arasındaki korelasyon bulunmaya çalışılmıştır. Elde edilen sonuçlar ile bir tahmin modeli eğitilmiş ve sonrasında Microsoft'un hisse senedi değerleri için tahminlerde bulunulmuştur. Microsoft hisse senedi için, önerilen bu yöntem ile farklı öznitelik kümeleri için yapılan tahminlerin 69,01% ile 71,82% arasında başarılı olduğu gözlemlenmiştir.

Farklı bir çalışmada ise Si ve *diğerleri* [22], S&P 100 isimli Amerika Birleşik Devletleri merkezli hisse senedi endeksi içerisinde yer alan şirketlerin hisse senedi isim kısaltmaları ile oluşturdukları belirli bir kelime listesi doğrultusunda Twitter üzerinden veri toplamışlardır. Daha sonra, topladıkları veri üzerinde, konu modellemesi ve düşünce analizi temelli bir öznitelik belirleme işlemi yapmışlardır. Toplanan veriler için, öznitelik kümelerini oluşturduktan sonra ise regresyon temelli

bir tahmin modeli geliştirmiş ve S&P 100 isimli endeksin hareket yönünü geliştirdikleri bu model ile tahmin etmeye çalışmışlardır. Bu çalışmada yapılan deneylerin sonuçları incelendiğinde, S&P 100 endeksinin yönünün en yüksek 68,00% oranında başarılı tahmin edildiği görülmüştür.

3. ÖNERİLEN SİSTEM

Yapılan çalışmada, içerisinde birbiri ile iletişim halinde olan birçok farklı alt modülün yer aldığı, tutarlı, ölçeklenebilir, gelişime açık ve esnek yapıda bir sistem geliştirilmiştir. Geliştirilen bu sistemin temel amacı, ilk olarak farklı kaynaklarda yer alan metin verilerinin, verinin bulunduğu kaynağın yapısına uygun bir yöntemle toplanması ve daha sonra toplanan bu verinin temel doğal dil işleme yöntemleri yardımıyla analiz edilmesidir. Bahsedilen bu işlemleri gerçekleştirebilmek adına, sistem içerisinde yer alacak temel 2 modül ve bu temel modüllerin içerisinde yer alacak farklı alt modüller geliştirilmiştir. Geliştirilmiş olan bu 2 temel modül sırası ile veri toplama modülü ve doğal dil işleme modülüdür. Bu temel modüller, kurgulanan tasarımları sebebi ile yeni alt modüllerin eklenmesine ve var olan alt modüllerin de geliştirilmesine olanak vermektedirler. Geliştirilmiş olan bu sistemin yapısı Şekil 3.1'de görülebilir.



Şekil 3.1 : Geliştirilmiş olan sistemin yapısı.

Şekil 3.1'de görülebileceği üzere geliştirilen bu sistemin sunucu ve kullanıcı katmanı olmak üzere temelde 2 farklı katmanı bulunmaktadır. Sunucu katmanı, temel modüllerin içerisinde bulunduğu, aktif olarak işlemlerin yapıldığı ve elde edilen sonuçların bir web servis aracılığı ile sorgulanabildiği bir yapıdır. Bu sunucu yapısının farklı yapılar ile iletişim kurabilmesi ve uyumlu çalışabilmesi adına, sunucu ile olan iletişim REST (temsili durum transferi) (representational state transfer) [23] tabanlı

bir API (uygulama programlama arayüzü) (application programming interface) ile geliştirilmiştir. Yine aynı sebepten ötürü sunucu katmanında, analizler yapıldıktan sonra bu analizler üzerinden yapılan sorguların JSON (javascript obje notasyonu) (javascript object notation) [24] formatında, tekilleştirilmiş ve standardize edilmiş olarak sonuç vermesi sağlanmıştır.

Geliştirilen sistemdeki bir diğer katman olan kullanıcı katmanında ise sunucu katmanında yer alan veri, bu veriye ait olan analiz sonuçları ve analiz sonuçları üzerinden yapılan sorgular doğrultusunda geliştirilmiş olan uygulamalar veya kütüphaneler yer almaktadır. Bu uygulamalar arasında, örnek vermek gerekirse, metin analiz uygulamaları veya ağ analiz uygulamaları yer almaktadır.

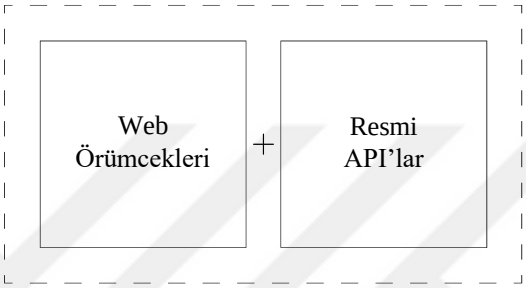
3.1 Veri Toplama Modülü

İnternet üzerinde yer alan farklı servislerde, oldukça hızlı bir şekilde büyüyen bir metin verisi bulunmaktadır. Bu sebeple, internette bulunan metin verisinin toplanabilmesi için de farklı teknikler bulunmaktadır. Bu tekniklerden ilki ve en popüleri ağ gezginleri (web crawler) yardımı ile bu verinin toplanmasıdır. Ağ gezginleri olarak adlandırılan yapılar, internet sitelerini otomatik olarak gezen ve istenilen bilgiyi toplayan programlanmış yapılardır. Bu yapılar oldukça esnektir ve farklı web sayfası yapılarına kolayca adapte olabilirler. Ağ gezginlerinin sahip oldukları olumsuz yönler ise zaman içerisinde bilginin toplandığı web sayfalarında yapılan düzenlemelere göre sürekli olarak güncellenmelerinin gerekliliği ve bazı durumlarda alternatif veri toplama tekniklerine göre daha yavaş çalışabilme ihtimalidir.

Kullanılan bir diğer veri toplama tekniği ise verinin toplanmak istendiği web sayfaları tarafından sunulan resmi web servisleridir. Bu servisler, genel olarak API formatında sunulmakta ve kullanıcıların bu servis üzerinden farklı parametreler ile sorgular yapmasına imkan tanımaktadır. Bu resmi web servisleri üzerinden verinin toplanması diğer veri toplama teknikleri ile kıyaslandığında, özellikle performans anlamında, çok daha yüksek hızda ve kalitede veri toplama imkanı sunmaktadır. Ancak bu servislerin de olumsuz yönleri bulunmaktadır. Bu olumsuz yönlerden ilki, web servisleri, verinin toplanmak istendiği platformun kendisi tarafından geliştirilmeli ve kullanıma açılmasının gerekliliğidir. Günümüzde çeşitli platformların, herkesin kullanımına açık olan web servisleri bulunsa da diğer bir kısmı böyle bir hizmet vermemektedir. Web servisleri ile veri toplamanın yaratabileceği bir diğer olumsuzluk ise toplanabilecek

olan verinin miktarının ve detayının web servisini geliştiren platform sahipleri tarafından belirlenmesidir. Bu sebeple çoğu zaman istenilen miktarda ve detayda veri, web servisleri üzerinden, ekstra bir geliştirme yapılmadan veya farklı yaklaşımlar uygulanmadan toplanamamaktadır.

Geliştirilen sistem içerisinde yer alan veri toplama modülü, farklı ihtiyaçlardan ötürü bahsedilen veri toplama tekniklerinden hem ağ gezginleri yardımı ile veri toplama tekniğini hem de web servisleri üzerinden veri toplama tekniğini kullanmaktadır. Veri toplama modülü ve içerisinde yer alan alt modüller Şekil 3.2'de görülebilir.



Şekil 3.2 : Veri toplama modülü içerisinde yer alan alt modüller.

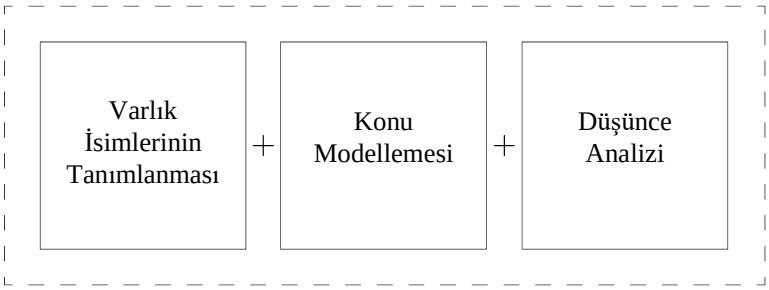
Veri toplama modülü içerisinde yer alan alt modüller kullanılarak 2 farklı kaynaktan veri toplanmaktadır. Bu verilerden ilki New York Times isimli gazetenin web sayfasında bulunan haberlere ve makalelere ait veriler, diğeri ise Twitter [25] isimli mikroblog platformunda yer alan metin verileridir. New York Times web sayfası içerisinde yer alan metin verileri gazetenin yazarları tarafından yazılan, uzun, resmi bir dille yazılmış olan, içerisinde dilbilgisi ve yazım yanlışı neredeyse bulunmayan, yapısal olarak oldukça düzgün olan verilerdir. Ancak, diğer taraftan Twitter içerisinde yer alan metin verisi New York Times verisine kıyasla tam tersi bir noktada bulunmaktadır. Twitter, kişilere 140 karakter uzunluğunda tweet adı verilen girdiler üretme imkanı sunan ve aktif olarak tüm dünyada 320 milyondan fazla kullanıcısı bulunan bir mikroblog platformudur. Bu nedenle Twitter üzerinde bulanan metin verisi, New York Times verisine göre çok daha büyük bir kitle ve düşünce grubunu kapsamanın yanında çok daha kısa, genelde dilbilgisi kurallarının göz ardı edildiği, anlatım bozukluklarının ve yazım yanlışlarının hatta tamamen anlamsız karakter dizilerinin oldukça çok olduğu, sabit bir yapıya sahip olmayan bir veri topluluğudur.

Veri toplama modülü tarafından toplanan bu 2 farklı veri, yapısal olarak birbirlerine taban tabana zıt olarak gözükse de içerikleri açısından değerlendirildiği noktada

birbirlerini tamamlayan verilerdir. Buna bir örnek vermek gerekirse; dünyanın herhangi bir yerinde meydana gelmiş bir toplumsal olay ile ilgili olarak insanlar olayın ana hatları hakkında gazetelerden bilgi edinirken, bu olayın detayları ve olayın yaşandığı yerde bulunan insanların kişisel görüşleri hakkında mikrobloglar ve sosyal medya platformları üzerinden bilgi edinebilirler. Bu noktada, bahsedilen bu 2 farklı veri grubunun birbirleri ile doğrudan ilişki halinde olduğu ve birbirlerini desteklediği söylenebilir.

3.2 Doğal Dil İşleme Modülü

Bir diğer temel modül olan, veri toplama modülü ile toplanan metin verisi veya başka bir kaynaktan elde edilen metin verisi üzerinde analiz yapabilmek adına sistem içerisinde yer alacak doğal dil işleme modülü geliştirilmiştir. Bu modül, varlık isimlerinin tanımlanması, düşünce analizi ve konu modellemesi başta olmak üzere temel doğal dil işleme ve metin madenciliği görevlerini yapmaktan sorumludur. Yapılması hedeflenen bu görevlerin temelinde oldukça karmaşık problemler yer almakta ve bu problemlerin çözümüne dair önerilmiş birçok farklı yaklaşım bulunmaktadır. Bu nedenle geliştirilen bu doğal dil işleme modülü, her biri farklı bir görevden sorumlu olan alt modüllerin oluşturduğu bir yapıya sahiptir. Doğal dil işleme modülü ve içerisinde yer alan alt modüller Şekil 3.3'de görülebilir.



Şekil 3.3 : Doğal dil işleme modülü içerisinde yer alan alt modüller.

3.2.1 Varlık isimlerinin tanımlanması ve sınıflandırılması

Varlık isimlerinin tanımlanması ve sınıflandırılması (named entity recognition and classification) (NERC), metinler içerisinde tanımlanmış olan kişilerin, ülkelerin, yerlerin, kurumların, vb... yapıların doğal dil işleme (natural language processing) (NLP) yöntemleri kullanılarak tespit edilmesidir. Varlık isimlerini tanımlama algoritmaları belirtilen kişi veya kurum gibi alanların dışında daha özelleşmiş olarak

hastalık isimlerinin tespiti veya kimyasal maddelerin isimlerinin tespiti gibi durumlarda da kullanılabilir.

Varlık isimlerini tanımlama yöntemleri, üzerinde çalışacakları metnin yapısına göre ve metnin yazılmış olduğu dile göre de farklılık göstermektedirler. Örneğin bir haber metninin yazı dili, akademik bir çalışmanın yazı dili ve sosyal medyada paylaşılan bir içeriğin yazı dili birbiri ile karşılaştırıldıklarında tamamen farklılık göstermektedir. Bir tanesi çok daha sade, yazım kurallarına uygun ve resmi cümleler içerirken, bir başkası yazım kurallarına uygun olmayan, imla hataları içeren ve bozuk cümlelerden oluşabilir. Bunun yanında metnin yazılmış olduğu dil de aynı zamanda varlık isimlerinin tespit edilmesi noktasında dikkat edilmesi gereken bir başka değişkendir. Cümle yapılarının dillere göre farklılık göstermesi, yazım ve ifade ediliş şekillerinin değişmesi en büyük sebeplerin başında gelmektedir.

Varlık isimlerinin tespiti önemlidir ancak kişi, yer, kurum gibi isimlerin metin içerisinde tespit edilmesinin yanında tespit edilen isimlerin bu kategorilerden hangisi altında yer alması gerektiğinin tespiti de problemin bütününde büyük bir yere sahiptir. Bunun sebebi hem tespit edilen varlık isimlerini kategorize etmek hem de birbiri ile aynı isme sahip olan ancak farklı kategorilerde yer alan durumların birbirinden ayrılmasıdır.

3.2.1.1 Kullanılan öğrenme yöntemleri

Bahsedilen bu problemler göz önüne alındığında varlık isimlerinin tespiti için birçok farklı çözüm önerilmiştir. Bu çözüm önerileri; geçmişte çok daha basit, oluşturulmuş temel isim listelerinin metin içerisindeki isimler ile eşleştirilmesini temel alan yöntemlerken, günümüzde uygulanan çözüm önerileri daha çok makine öğrenme prensiplerini temel alan daha gelişmiş yöntemlerdir. Geçmişten günümüze önerilmiş olan bu çözüm önerilerini gözetimli öğrenme yöntemleri, yarı gözetimli öğrenme yöntemleri ve gözetimsiz öğrenme yöntemleri başlıkları altında toplayabiliriz.

Gözetimli öğrenme yöntemleri

Gözetimli öğrenme yöntemleri, temel olarak, işaretlenmiş varlık isimlerini içeren metinleri, varlık isimlerini belirleyecek kuralları ve durumları içerecek şekilde hazırlanmış olan belirli bir eğitim verisi ile modeller eğitildikten sonra yeni gelen bir metin üzerinde varlık isimlerinin tespit edilmesi şeklinde çalışmaktadır.

Bu yöntemler incelendiğinde, ilk olarak Saklı Markov Modellerini (Hidden Markov Models) (HMM) [26] ve karar ağaçlarını (decision tree) [27, 28] temel alan yöntemler görülmektedir. Bunlara ek olarak karar destek makinelerini temel alan [29, 30], maksimum entropi modellerini (maximum entropy models) kullanan [31] ve koşullu rastgele alanlar (conditional random fields) (CRF) [32] gibi öğrenme temelli, makine öğrenmesinin temellerini içeren yöntemler kullanılarak oluşturulmuş olan sistemlerdir. Bu yöntemler, elde bulunan ve daha önce etiketlenmiş veri doğrultusunda modellerini eğitmekte ve sonraki aşamalarda eğitmiş olduğu bu model üzerinden varlık isimlerini tespit etmektedir. Bu sebeple, gözetimli öğrenme yöntemlerini eğitmek için kullanılacak verinin boyutu ve kalitesi, bu yöntemlerden daha iyi sonuçlar alabilmek adına oldukça büyük önem taşımaktadır.

Yarı gözetimli öğrenme yöntemleri

Bu yöntemler temel olarak gözetimli öğrenme yöntemlerini kapsamaktadır. Ancak ek bir durum olarak, varlık isimlerinin tanımlanacağı metin üzerinde algoritmalar çalıştırılacağı zaman kullanıcılardan varlık isimlerinin bir kısmının işaretlenmesi istenmektedir. Kullanıcılar tarafından yapılan bu öncül işaretlemeler, algoritmayı belirli bir oranda yönlendirmek veya algoritmaların girdi olarak kullanacakları metin yapısı hakkında bilgi sahibi olmasını sağlamak için kullanılmaktadır. Bu durum literatürde önyükleme (bootstrapping) olarak tanımlanmaktadır ve yarı gözetimli öğrenme yöntemlerinin temelini oluşturmaktadır.

Önyükleme olarak adlandırılan bu öncül yardımcı bilgi tanımlama özelliğini kullanan yöntemleri incelediğimizde, bulunmak istenen varlık isminin yapısının bir kurallı ifade (regular expression) olarak tanımlandığı ve bu doğrultuda arama işlemlerinin yapıldığı bir yöntem [33] görülebilir. Buna ek olarak, belirli yapısal kombinasyonların birer kural olarak tanımlandığı (örneğin; “bir yüklem ardından gelecek her bir özne varlık ismi olarak tanımlanmaktadır” şeklinde tanımlanmış olan) ve bu kurallar üzerinden çalıştırılan varlık ismi tanıma yöntemleri [34, 35] de önyükleme içeren yapılara örnek olarak verilebilir. Tüm bu önyükleme içeren yöntemlerin yanı sıra, kişiler tarafından üretilen veya bir kurallar bütünü olarak tanımlanmış olan önyükleme bilgilerini değil, var olan diğer varlık ismi tanıma sistemlerinin çıktılarını bir girdi olarak kabul eden ve ortak önyükleme (mutual bootstrapping) olarak adlandırılan yöntemler [36, 37] de bulunmaktadır.

Gözetimsiz öğrenme yöntemleri

Gözetimsiz öğrenme yöntemleri, varlık isimleri işaretlenmemiş, çok büyük bir metin verisi üzerinde belirli öbekleme algoritmaları çalıştırılarak benzer içeriklerin kümelenmesi mantığı doğrultusunda oluşturulmuş olan sistemlerdir. Bu sistemlerin başarılı çalışabilmesi için çok öznitelik kümesinin oluşturulabilmesi adına büyük miktarda ve yapısal olarak düzgün olan metin verisine ihtiyaç duyulmaktadır. Bu durum hem veri büyüklüğü anlamında problemlere sebep olmakta hem de İngilizce dışındaki dillerin popülerliğinin daha düşük olması ve bu dillerde üretilmiş olan doğal dil işleme içeriklerinin İngilizce ile kıyaslandığında çok daha az olması sebebi ile çoğu durumda verimi düşük olacak şekilde çalışmaktadır.

Bu yöntemler incelendiğinde Princeton Üniversitesi tarafından geliştirilmiş ve çok büyük miktarda içeriğe sahip olan WordNet [38] isimli sözlük veri tabanını kullanarak etiketleme yapan ve daha sonra bu etiketler doğrultusunda varlık ismi tanımlama işlemini yapan bir yöntem [39] görülebilir. Ek olarak, büyük harfle başlama veya cümle içerisinde "bir şehir olarak" veya "bir organizasyon olarak" gibi belirteçler üzerinden tanımlama işlemini yapan yöntemler [40, 41] de gözetimsiz öğrenme yöntemleri arasında sayılabilir. Bu yöntemlerin yanında, dokümanlar içerisinde yer alan varlık isimlerinin birbirleri ile olan ilişkileri bilgisini kullanarak da varlık ismi tespiti yapan yöntemler [42, 43] de bulunmaktadır.

3.2.1.2 Kullanılan özellikler

Bahsedilen yöntemler olan gözetimli öğrenme yöntemleri, yarı gözetimli öğrenme yöntemleri ve gözetimsiz öğrenme yöntemleri içerisinde, varlık isimlerinin tespiti için işaretlenmiş ve gruplanmış olan metin verilerinin yanında belirli kurallar içeren özellik listeleri de kullanmaktadır. Özellik listeleri, varlık isimleri için tanımlanmış olan belirli kurallar doğrultusunda, elde bulunan veriden varlık ismi tespiti işlemi sırasında otomatik olarak üretilmektedir. Bu liste içerisinde yer alan özellikler, temel olarak farklı türlerdeki varlık isimlerinin sahip oldukları yapıları tanımlamaktadırlar.

Bu özellikler mantıksal, sayısal ve yazılı değerler ile ifade edilmektedirler. Varlık isimlerinin tespiti için kullanılan algoritmalarda istenilen büyüklükte bir özellik listesi oluşturulabilir. Kullanılan bu algoritmalar, türetilmiş olan özellik listeleri üzerinden modellerini üretmekte ve sonuç vermektedir. Örneğin bir metin verisi için aşağıdaki gibi tanımlanmış bir kural listesi olabilir;

- Metin içerisindeki her bir kelimenin ilk harfinin büyük olup olmadığı.
- Metin içerisindeki her bir kelimenin karakter sayısı.
- Metin içerisindeki her bir kelimenin tüm harflerini küçük olacak şekilde yazılmış hali.
- Metin içerisindeki her bir numerik veya özel bir karakter içerip içermediğinin bilgisi.

Tanımlanmış bu kural listesi doğrultusunda “The president of Apple eats an apple.” cümlesini ifade etmek istenildiğinde, elde edilecek çıktı şu şekilde olacaktır;

<true,3,”the”>, <false,9,”president”>, <false,2,”of”>, <true,5,”apple”>,
<false,4,”eats”>, <false,2,”an”>, <false,5,”apple”>

Varlık isimlerinin tanımlanması için geliştirilen algoritmaların içerisinde kullanılan özellik listeleri tür olarak bakıldığında, kelime seviyesindeki özellikler (word-level features), listeden okunacak özellikler (list lookup features) ve doküman temelli ve yapısal özellikler (document and corpus features) olarak farklı gruplara ayrılabilir.

Kelime seviyesindeki özellikler

Bu özellikler, kelimeleri temel alan özelliklerdir. Kelimelerin yapıları ve kelimeleri oluşturan karakterlerin üzerinden tanımlanırlar. Örneğin belli başlı tanımlanmış olan özellikler;

- Kelimeyi oluşturan tüm karakterlerin büyük harf olması. (Örnek: TÜBİTAK)
- Kelime içerisindeki karakterlerin büyük ve küçük harfleri karışık olarak kullanması. (Örnek: eBay)
- Kelimeyi oluşturan karakterler arasında noktalama işaretleri olması. (Örnek: I.B.M.)
- Kelime içerisinde kesme işareti, tire veya 've' işareti olması. (Örnek: O'Connor, AT&T)
- Kelimenin bir kalıba uygun olması. (Örnek: yyyy-mm-dd kalıbı, 2018-01-01)
- Kelimeyi oluşturan karakterlerin nümerik ve alfabetik karakterleri karışık olarak içermesi. (Örnek: W3C)
- Kelimelerin belirli bir yapı ile sonlanması. (Örnek: Turkish, English, Irish, Spanish)

Listeden okunacak özellikler

Bu özellikler, belirli kelimelerin birlikte kullanımlarının listelendiği özelliklerdir. Bu özellikler yardımıyla varlık isimlerinin uygun kategoriler altında işaretlenmesi hedeflenmektedir. Örneğin;

- "Paris" ve "şehir" kelimelerinin bir arada kullanılması durumunun bir özellik olarak tanımlandığı bir algoritmada, metin içerisinde bu kelimelerden birinin görülmesi o metin içerisinde bir varlık isminin bulunma ihtimalinin yüksek olduğu gösterecektir.
- "ltd", "şti", "corp" gibi kelimelerin bulunduğu bir liste özellik olarak tanımlanmış ise metin içerisinde bir kurum veya şirketi belirten varlık isminin bulunma ihtimali yüksek olacaktır.

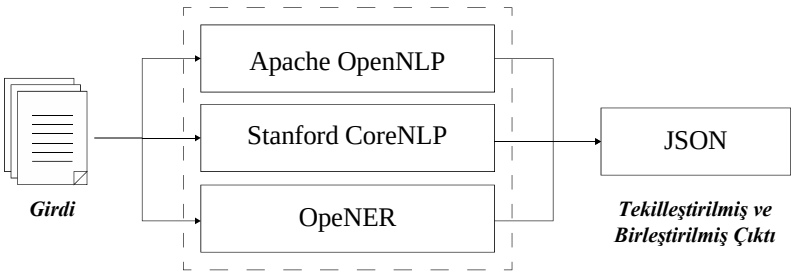
Doküman temelli ve yapısal özellikler

Bu özellikler, bütün bir doküman içerisindeki varlık isimlerinin hangi kategorilerde ve kaç defa kullanıldığı gibi istatistiki bilgileri içermektedir. Bunların yanında ise eğer kullanılan doküman linklerin, kişilerin, vb... yapıların işaretlendiği meta-bilgiler içeriyorsa, bu bilgilerin tespit edilip birer özellik olarak kullanılması da algoritmaların daha başarılı sonuç vermesi için etkili olacaktır.

3.2.1.3 Önerilen sistem

Varlık isimlerinin tanımlanması için kullanılabilecek olan yöntemler, bahsedildiği üzere birbirinden oldukça farklı temellere dayanmaktadır ve bahsedilen yöntemlerin birbirlerine göre farklı avantajları ve dezavantajları bulunmaktadır. Üzerinde çalışılacak olan metin verisinin yapısına veya toplandığı kaynağa göre, var olan algoritmalar birbirlerinden farklı performanslar sergilemektedir. Örneğin, bir cümle içerisindeki varlık isimlerini tanımlarken kullanılan iki yöntem kıyaslandığında, bir yöntem cümle içerisinde geçen markaları varlık ismi olarak tanımlarken bir başka yöntem bu şekilde sonuç vermemektedir. Ancak aynı yöntemler kişileri tanımlarken kıyaslandığında, ilk yöntem ikinci yönteme daha sağlıklı sonuçlar üretebilmektedir. Bu noktada, varlık isimlerinin tanımlanması yapılırken başarı oranını yükseltebilmek adına farklı varlık isimlerini tanıma algoritmalarını bir arada bulunduran ve ölçeklenebilir bir yapı önerilmiştir. Önerilen bu yapı, metin verisini girdi olarak kabul etmekte ve daha sonra içerisinde bulunan farklı varlık isimlerini tanıma araçlarını

birbirlerine paralel olarak çalışacak şekilde kullanarak, farklı yöntemlere göre varlık isimlerini bulmaktadır. Varlık isimlerini farklı yöntemler doğrultusunda bulduktan sonra ise, elde edilen bu sonuçları birleştirerek tek bir sonuç oluşturmakta ve bunu çıktı olarak vermektedir. Oluşturulan bu yapı Şekil 3.4'de görülebilir.



Şekil 3.4 : Önerilen varlık isimlerini tanıma ve sınıflandırma alt modülünün yapısı.

Şekil 3.4'de görüldüğü üzere, önerilen sistem 3 farklı doğal dil işleme aracı ve bu araçlar içerisinde bulunan varlık isimlerini tanımlama algoritmalarını içermektedir. Kurgulanan yapı içerisinde bulunan bu araçlar sırası ile Apache OpenNLP [44], Stanford CoreNLP [45] ve OpeNER [46] isimli araçlardır. Bu doğal dil işleme araçlarının seçilmesi ve kullanılmasının temel sebebi, varlık isimlerinin tanımlanması işlemi için hepsinin birbirinden mantık olarak farklı yaklaşımlar sergilemesinden kaynaklanmaktadır. Bunun yanında, kurgulanan sistem oldukça esnek ve ölçeklenebilir olacak şekilde tasarlandığı için istenildiği takdirde belirtilen bu 3 doğal dil işleme aracının yanına farklı araçlar da eklenerek daha kapsamlı bir yapı kolaylıkla oluşturulabilir.

Tasarlanan sistem içerisinde bulunan ilk doğal dil işleme aracı olan Apache OpenNLP, Apache Vakfı (Apache Foundation) tarafından desteklenerek geliştirilmiş olan, makine öğrenmesi tabanlı algoritmalar kullanarak ana modüllerinin oluşturulduğu, açık kaynaklı ve herkesin kullanımına açık olan bir doğal dil işleme kütüphanesidir. Apache OpenNLP, varlık isimlerinin tanımlanması için daha önceden varlık isimleri içerisinde işaretlenmiş olan metin dosyalarını bir eğitim verisi olarak kullanarak bir sistem oluşturmakta ve bu sistem üzerinden girdi olarak gelen metin dosyaları içerisindeki varlık isimlerini bulmaya çalışmaktadır. Bu doğal dil işleme aracı kendi sahip olduğu eğitim modellerinin yanında, kişilere kendi eğitim modellerini oluşturma ve kullanma olanağı da sunmaktadır. Varlık isimlerinin tanımlanması için kullanılan

bu kütüphane 7 farklı varlık ismi türü için sonuç üretmektedir. Bu türler, sırasıyla; konum, organizasyon, kişi, tarih, saat, parasal değerler ve matematiksel değerlerdir.

Kullanılan bir diğer doğal dil işleme aracı, Stanford CoreNLP kütüphanesidir. Bu doğal dil işleme kütüphanesi, Stanford Üniversitesi içerisinde yer alan doğal dil işleme grubu (Stanford University Natural Language Processing Group) tarafından geliştirilmiş olan bir başka açık kaynaklı araçtır [47]. Stanford CoreNLP içerisinde yer alan ve varlık isimlerinin tanımlanması için kullanılan yöntem, koşullu rastgele alanlar [48] temelli bir tahmin mekanizmasıdır. Koşullu rastgele alanlar, koşullu olasılık kullanarak oluşturmuş olan bir istatistiki modeldir. Stanford CoreNLP kütüphanesi de kendi içerisinde varlık isimlerinin tanımlanmasına yönelik kendi modellerini içermektedir. Bir önceki aşamada kullanılan Apache OpenNLP'ye benzer şekilde konum, organizasyon, kişi, tarih, saat, parasal değerler ve matematiksel değerleri varlık isimleri olarak tanımlayabilmektedir. Stanford CoreNLP, bunlara ek olarak şehir, ülke, eyalet, milliyet, e-posta, din, unvan, internet sitesi adresi, ölüm sebebi, suç iddiası ve ideoloji gibi detaylı varlık isimlerini de tespit edebilmektedir.

Son olarak kullanılan araç ise OpeNER isimli doğal dil işleme aracıdır. OpeNER, Avrupa Birliği Komisyonu tarafından fonlanan ve içerisinde bulunan modüller ile birçok farklı dilde temel doğal dil işleme tekniklerini uygulamaya olanak sağlayan bir araçtır. Bu araç içerisindeki, varlık isimlerini tanımlamaya yönelik kullanılan yöntem ise IXA [49, 50] adı verilen bir altyapıyı temel alan bir yöntemdir. OpeNER tarafından tanımlanan varlık isimleri ise yine konum, organizasyon, kişi, tarih, saat, parasal değerler ve matematiksel değerleri şeklindedir.

Kurgulanan yapı içerisinde varlık isimlerinin tanımlanması için kullanılan bu 3 farklı araç, birbirleri ile tutarlı bir biçimde, metin formatında olmak üzere aynı formatta bir girdi kabul etmektedir. Ancak ürettikleri çıktılar birbirlerinden tamamen farklıdır. Her bir araç kendine ait bir çıktı formatında sonuçlarını üretmektedir. Örneğin, bu 3 farklı doğal dil işleme aracı için aşağıdaki gibi bir metin parçası girdi olarak verilsin;

"... a critical crossing in the suburbs of New York City. Jamey Barbas, the engineer orchestrating the project for the Thruway Authority ..."

Apache OpenNLP doğal dil işleme aracının, girdi olarak verilen bu içerik için ürettiği sonuç basit bir metin dosyası olmaktadır. Bu metin dosyası içerisinde bulunan varlık isimleri, türleri ve metin içerisindeki konumları yer almaktadır. Verilen bu girdi

doğrultusunda Apache OpenNLP doğal dil işleme aracının ürettiği sonuç Şekil 3.5'de görülebilir.

```
[463..466) location New York City
[467..469) person Jamey Barbas
[477..479) organization Thruway Authority
```

Şekil 3.5 : Apache OpenNLP kütüphanesi ile elde edilen örnek çıktı.

Stanford CoreNLP içerisinde yer alan, varlık isimlerinin tanımlanması için kullanılan modül ise verilen aynı girdiye karşılık olarak çok daha karmaşık ve değerlendirilmesi güç bir çıktı üretmektedir. Bu çıktı içerisinde, girdi metninde bulunan tüm kelimeler ve noktalama işaretleri listelenmekte ve varlık isminin bir parçası olarak tanımlanan kelimelerin yanlarına ait oldukları türler bir not olarak yazılmaktadır. Örnekte verilen girdi doğrultusunda Stanford CoreNLP doğal dil işleme aracının ürettiği sonuç Şekil 3.6'da görülebilir.

```
a
critical
crossing
in
the
suburbs
of
New LOCATION
York LOCATION
City LOCATION
.
Jamey PERSON
Barbas PERSON
,
the
engineer
orchestrating
the
project
for
the
Thruway ORGANIZATION
Authority ORGANIZATION
```

Şekil 3.6 : Stanford CoreNLP kütüphanesi ile elde edilen örnek çıktı.

Son olarak OpeNER isimli doğal dil işleme aracı ise diğer iki araçtan tamamen farklı olarak, XML tabanlı KAF (KYOTO Açıklama Biçimi) (KYOTO Annotation Format) [51] isminde çok daha detaylı ve yapılandırılmış bir çıktı üretmektedir. Bu çıktıda ise

varlık isimleri, türleri ve konumları, KAF dosyasının farklı elemanları olarak çıktı dosyasında yer almaktadır. Üretilen çıktının yapılandırılmış olması, elde edilen sonuçların parçalanması ve değerlendirilmesi noktasında kullanım kolaylığı da sunmaktadır. Örnekte verilen girdi doğrultusunda OpeNER doğal dil işleme aracının ürettiği sonuç Şekil 3.7'de görülebilir.

```
<entity eid="e1" type="location">
  <references>
    <!--New York City-->
    <span>
      <target id="t463" />
      <target id="t464" />
      <target id="t465" />
    </span>
  </references>
</entity>
<entity eid="e2" type="person">
  <references>
    <!--Jamey Barbas-->
    <span>
      <target id="t467" />
      <target id="t468" />
    </span>
  </references>
</entity>
<entity eid="e3" type="organization">
  <references>
    <!--Thruway Authority-->
    <span>
      <target id="t477" />
      <target id="t478" />
    </span>
  </references>
</entity>
```

Şekil 3.7 : OpeNER kütüphanesi ile elde edilen örnek çıktı.

Örneklerden de görülebileceği gibi, farklı doğal dil işleme araçları aynı girdi için yapısal olarak birbirinden oldukça farklı sonuçlar üretmektedir. Bu araçların bir arada kullanılabilmesi için 3 farklı araç tarafından üretilen çıktıların parçalandığı, değerli kısımlarının otomatik olarak tespit edildiği ve farklı yapıdaki bu çıktıların aynı ve tutarlı bir yapıya dönüştürüldüğü bir modül geliştirilmiştir. Farklı çıktıları işleyerek tekil bir yapıya dönüştüren bu modül, ürettiği çıktı formatı JSON olacak şekilde kurgulanmıştır. Bu noktada, çıktı olarak JSON formatının seçilmesinin sebebi, günümüzde web servislerinin çok büyük bir kısmı tarafından iletişimin JSON objeleri üzerinden sağlanması, her programlama dili tarafından kolayca okunabilmesi ve alt

parçalarına ayrılabilmesi ve de insan tarafından da kontrol edilebilirliğinin ve okunma kolaylığının yüksek olmasından kaynaklanmaktadır. Geliştirilen bu çıktı birleştirme ve yapılandırma modülünün 3 farklı doğal dil işleme aracının sonuçları doğrultusunda ürettiği sonuç Şekil 3.8'de görülebilir.


```
"namedEntities": [
  ...
  {
    "sentence": "... a critical crossing in the suburbs of New York City",
    "namedEntity": "New York City",
    "category": "LOCATION",
    "startIndex": 463,
    "endIndex": 465
  },
  {
    "sentence": "Jamey Barbas, the engineer orchestrating the project for
the Thruway Authority ...",
    "namedEntity": "Jamey Barbas",
    "category": "PERSON",
    "startIndex": 467,
    "endIndex": 468
  },
  {
    "sentence": "Jamey Barbas, the engineer orchestrating the project for
the Thruway Authority ...",
    "namedEntity": "Thruway Authority",
    "category": "ORGANIZATION",
    "startIndex": 477,
    "endIndex": 478
  },
  ...
]
```


Şekil 3.8 : Kullanılan farklı doğal dil işleme kütüphaneleri doğrultusunda elde edilmiş olan birleştirilmiş ve tekilleştirilmiş örnek çıktı.

3.2.2 Düşünce analizi

Düşünce analizi (sentiment analysis) veya fikir madenciliği (opinion mining), insanların belirli bir konu, haber, olay ya da kişi, kurum, yer gibi varlıklarla ilgili olan düşüncelerinin, değerlendirmelerinin, davranışlarının ve duygularının tespit edilmesidir. Düşünce analizi oldukça zor bir problem olmasının yanında doğru yaklaşımlar uygulandığı takdirde çok kullanışlı bir araç olmaktadır. Düşünce analizi yöntemleri kullanılarak toplanan bir metin verisi üzerinden, daha önce de belirtildiği gibi kişilerin belirli bir konu hakkında nasıl hissettikleri tespit edilebilmekte ve bu sonuçlar sanal ortamda üretilen içeriğin çok hızlı bir şekilde büyüdüğü bir dönemde oldukça değerli bir hale gelmektedir. Bu sayede birçok haber, olay, ürün veya

kampanya ile ilgili, özellikle farklı sosyal medya platformlarının kullanıcıları tarafından üretilen metin verisi kullanılarak, anında geri bildirim alınabilmektedir.

Düşünce analizinin zor bir problem olmasının temel sebebi, çok fazla farklı değişkene bağlı olmasından kaynaklanmaktadır. Bu değişkenlerden biri, bir ürünün veya olayın değerlendirilebilmesi için birçok farklı yöntemin olma ihtimalidir. Basitçe belirli bir aralık arasında puan vermek bu yöntemlerden biri olabilecekken detaylı bir inceleme yazısı yazmak da farklı bir değerlendirme yöntemi olabilir. Verilen bir puan üzerinden düşünce analizi yapmak oldukça kolaydır, bu noktada kişinin düşüncesi net bir şekilde ortadadır. Ancak bir metnin değerlendirilmesi ile düşünce analizi yapılacağı noktada problem çok daha detaylı ve karmaşık bir hale gelmektedir. Bunun bir sebebi insanların düşüncelerini ifade etme şekillerinin oldukça karmaşık olmasından kaynaklanmaktadır. Bir konu hakkında kişiler düşüncelerini ifade ederken konu çok hızlı bir şekilde değişebilir, kişi tarafından bazı metaforlar ile anlatmak istenilen şey anlatılıyor olabilir, düşünceler ironik bir dil ile ifade ediliyor olabilir veya bazı cümlelerde kinaye yapılıyor olabilir. Bu durumda elde bulunan yazılı metin bir insan tarafından okunduğunda oldukça rahat anlaşılabilecek iken algoritmalar tarafından yorumlanması ve sonuç elde edilmesi çok daha zor olmaktadır.

Bir başka sorun ise pozitif veya negatif durumlar için kullanabilecek kelimelerin ve ifadelerin net bir şekilde ayrılıp, gruplandırılamamasından kaynaklanmaktadır. Bazı durumlar için pozitif düşünceleri ifade etmek için kullanılan kelimeler, başka durumlarda ise tamamen negatif düşünceleri ifade etmek için kullanılabilir. Örneğin; "Almış olduğum bilgisayarın açılma süresi çok kısa." cümlesinde "kısa" ile belirtilen durum pozitif bir anlam çıkmasına sebep olurken, "Almış olduğum bilgisayarın pil ömrü çok kısa." cümlesinde ise "kısa" kelimesi negatif bir anlam çıkmasına sebep olmaktadır. Bu gibi durumlar ile oldukça sık bir şekilde karşılaşılabileceği için de sadece kelimeler ve anlamları kullanılarak düşünce analizinin yapılması çok sağlıklı değildir.

Bunların yanında, düşünce analizi yapmak için kullanılacak olan verilerin toplandığı platformlar da oldukça önemlidir. Bunun temel sebebi üzerinde analiz yapılacak olan metni yazan kişilerin, bu metinleri yazdıkları platformlara göre yazım tarzlarının ve metin kalitesinin değişmesinden kaynaklanmaktadır. Örneğin bir gazete için yazılmış olan bir makalenin, çok daha resmi bir dil kullanılarak ve dil bilgisi kuralları göz önünde bulundurularak yazılmış olma ihtimali yüksek iken, sosyal medya

platformlarında yazılmış olan bir yazının çok daha düzensiz ve gelişigüzel yazılmış olma ihtimali oldukça yüksektir. Düşünce analizi veya doğal dil işleme algoritmalarının, düzensiz ve kötü bir dil ile yazılmış olan metinler üzerinden elde edeceği sonuçların doğruluğu da düşük olacaktır. Bu durum da makineler tarafından düşünce analizi yapılacağı noktada büyük bir zorluğa sebep olmaktadır.

3.2.2.1 Düşünce analizi seviyeleri

Düşünce analizi, metin içerisinde tespit edilmek istenen düşüncenin kapsayacağı alana, aktörlere, durumlara göre farklı seviyelerde yapılabilir. Örneğin, kimi durumlar için sadece tek bir ürüne odaklanmakta iken kimi durumlarda ise bir ürün içerisinde yer alan tüm özelliklere odaklanıyor olabilir. Bu durum da düşünce analizinin yapılacağı seviye farklılıklarının ortaya çıkmasının ana sebebidir. Düşünce analizi temelde iki farklı seviyede yapılmaktadır.

Doküman seviyesindeki düşünce analizi

Doküman seviyesinde düşünce analizi (document-level sentiment analysis) yapılacak olan düşünce analizi, elde bir metin verisi olduğu ve elde bulunan bu metin verisinin tek bir ürün, olay, kişi, vb... bir varlık üzerine olduğu varsayımı üzerinden yapılır. Metin verisi içerisinden çıkarılan her olumlu veya olumsuz düşüncenin merkezinde temelde bulunan varlık olduğu kabul edilmektedir. Bu sebeple, doküman seviyesindeki düşünce analizi genelde ürün incelemeleri veya belirli bir servis hakkındaki yorumlar üzerinden yapılmaktadır.

Cümle seviyesindeki düşünce analizi

Cümle seviyesinde düşünce analizi (sentence-level sentiment analysis) yapılacak olan düşünce analizi ise doküman seviyesinde yapılan düşünce analizine oldukça benzer bir şekilde, metin verisinin daha küçük bir parçası üzerinde yapılan düşün analizidir. Bu işlem için kullanılan yöntemler de doküman seviyesinde kullanılan yöntemler ile oldukça yakınlık göstermektedir. Cümle seviyesinde yapılan düşünce analizinin, doküman seviyesinde yapılan düşünce analizine göre sağladığı fayda ise detaylar üzerindeki düşüncelerin tespit edilebilmesine olanak vermesidir. Örneğin doküman seviyesinde yapılan bir düşünce analizi bir ürün hakkında kişinin düşünceleri hakkında bilgi verirken, cümle seviyesinde yapılan düşünce analizi bu ürünün farklı özellikleri hakkında kişinin düşüncelerini öğrenmeye olanak sağlayacaktır.

3.2.2.2 Kullanılan öğrenme yöntemleri

Düşünce analizi probleminin çözümüne yönelik olarak da birçok farklı çözüm önerisi ortaya konulmuştur. Bu çözüm önerileri içerisinde makine öğrenmesi temelli olan gözetimli öğrenme yöntemleri, yarı gözetimli öğrenme yöntemleri ve gözetimsiz öğrenme yöntemleri olmasının yanı sıra tamamen sözlük tabanlı yöntemler de yer almaktadır.

Gözetimli öğrenme yöntemleri

Bu yöntemler, içerisine etiketlenmiş veri kümelerini eğitim verileri olarak kabul eden ve bu eğitim verisi üzerinden oluşturdukları modelleri kullanarak daha sonra girdi olarak gelen verinin sonucunu tahmin etmeye yönelik çalışan sistemlerdir. Düşünce analizi için bu yöntemlerin en başında olasılıksal modelleri temel alan sınıflandırıcılar gelmektedir. Naive Bayes temelli sınıflandırıcılar [52] bu alanda kullanılan öğrenme yöntemlerinin başında gelmektedir. Düşünce analizi yapmak için kullanılan olasılıksal sınıflandırıcılar içerisinde maksimum entropiyi (maximum entropy) gözeten sınıflandırıcılar [53] da bulunmaktadır. Bunlara ek olarak karar destek makinelerini temel alan sınıflandırıcılar [54, 55], yapay sinir ağları [56, 57], karar ağaçları temelli sınıflandırıcılar [58] ve kural tabanlı sınıflandırıcılar (rule-based classifier) [59] da düşüncü analizi yapmaya yönelik yöntemlerin arasında yer almaktadır.

Yarı gözetimli öğrenme yöntemleri

Bu yöntemler, gözetimli öğrenme yöntemleri ile benzeşmektedir. Ancak gözetimli öğrenme yöntemlerinden farklı olarak öğrenme sürecinin sadece belirli noktalarında sisteme dışarıdan bir bilgi akışı ile öğrenme sürecine katkı sağlanmaktadır. Düşünce analizi yapmak için kullanılan bu yöntemler incelendiğinde, Bayesian Ağları'ndan (Bayesian Network) oluşturulmuş olan sınıflandırıcılar [60] bu problemin çözümü için kullanılmaktadır. Buna ek olarak kısmi olarak etiketlenmiş olan verilerin kullanıldığı sınıflandırıcılar [61] ve bazı kelimeler için tanımlanmış olan benzerlik değerleri ve polarite skorlarının kullanılarak analiz yapılan yöntemler [62] de yarı gözetimli öğrenme yöntemleri altında yer almaktadır.

Gözetimsiz öğrenme yöntemleri

Bu grup içerisinde yer alan öğrenme yöntemleri, düşünce analizini yapan sistemlere herhangi bir ön bilginin verilmediği ve verinin sınıflandırılmasının istendiği

yöntemlerdir. Buna yönelik dokümanlar içerisinde bulunan cümleler ve kelimeler doğrultusunda uzaklık tanımlarının yapıldığı ve bu tanımlar doğrultusunda sınıflandırma yapan [63] yöntemler mevcuttur. Bunun yanında dokümanlar içerisinde yer alan konular üzerinden sınıflandırma işlemlerini yapan yöntemler [64] ve dokümanlar arasındaki noktasal ortak bilgilerin (pointwise mutual information) (PMI) kullanılarak düşünce analizinin yapıldığı [65] yöntemler de vardır.

Sözlük tabanlı yöntemler

Makine öğrenmesi yöntemlerinin yanı sıra, doküman içerisinde geçen kelimeler, bu kelimelerin polarite skorları ve kelimelerin istatistiksel dağılımı kullanılarak düşünce analizinin yapıldığı yöntemler de bulunmaktadır. Bu yöntemlerden biri dokümanlar içerisindeki kelimelerin kullanım kalıpları üzerinden düşünce analizi yapan [66] bir yöntemdir. Dokümanlar içerisinde yer alan sıfatların tespiti, bu sıfatların kullanım sıklıkları ve polarite skorları ile analiz yapan [67] yöntemler de mevcuttur. Kullanılan kelimelere ek olarak bu kelimelerin eş anlamlıları, eş anlamlı kelimelerin kullanımı ve polarite skorlarının kullanımı [68] da düşünce analizi yapan yöntemler içerisinde yer almaktadır. Bunlara ek olarak temelini kelime torbası modelinin [69] ve TF-IDF'in (terim sıklığı-doküman sıklığının tersi) (term frequency-inverse document frequency) [70] oluşturduğu çalışmalar da mevcuttur. Daha sade yöntemler olarak ise çok yönlü soru cevaplama (multi-perspective question answering) (MPQA) adı verilen bir yöntemi temel alan [71] ve Princeton Üniversitesi tarafından geliştirilmiş ve çok büyük miktarda içeriğe sahip olan WordNet isimli sözlük veri tabanını kullanarak düşünce analizi yapan farklı bir yöntem [72] de düşünce analizi problemin çözümü için kullanılmaktadır.

3.2.2.3 Kullanılan özellikler

Düşünce analizi yapmak için kullanılan gözetimli öğrenme yöntemleri, yarı gözetimli öğrenme yöntemleri ve gözetimsiz öğrenme yöntemleri içerisinde belirli kurallara dayalı olan özellik listelerini de kullanılmaktadır. Bu özellik listelerinden yararlanılarak kullanılan yöntemler ile, elde edilen sonuçların kalitesinin arttırılması hedeflenmektedir. Düşünce analizi yapmak için geliştirilen algoritmaların içerisinde kullanılan özellik listeleri tür olarak bakıldığında, anlamlı kelimeler ve kullanım sıklıkları, kullanılan kelimelerin türleri, düşünce belirten kelimeler ve cümlelerin tespiti ve olumsuzluk içeren kelimelerin tespiti olarak farklı gruplara ayrılabilir.

Anlamlı kelimeler ve kullanım sıklıkları

Kullanılacak metin verisi içerisinde yer alan kelimelerin, metin içerisindeki kullanım sıklığı verisi düşünce analizi yapmak için değerli bir veridir. Cümlelerin içerisinde yer alan bağlaç, noktalama işaretleri, soru ekleri, vb... kelimeler temizlendikten sonra, geriye kalan anlamlı kısımlar üzerinden anlamlı kelimelerin kullanım sıklığı değerleri hesaplanabilir.

Kullanılan kelimelerin türleri

Analizi yapılacak olan metinlerde yer alan cümlelerin içerisinde geçen isim ve sıfat listeleri, düşünce analizi işlemlerinin yapılabilmesi için oldukça değerlidir. Bu sebeple kelime türlerinin (isim, nesne, yüklem, vb...) tespit edilmesi ve işaretlenmesi büyük önem taşımaktadır. Bu kelimelerin değerli olmasının temel sebebi, doğrudan düşünce analizine yönelik anlam içermelerinden kaynaklanmaktadır. Cümleler içerisinde isim türünde bulunan kelimeler, örneğin bir ürün incelemesi üzerinde düşünce analizi yapılıyorsa, bu ürüne ait olan özelliklerin listesinin tespit edilmesinde yararlı olacaktır. Bir bilgisayar incelemesinde yer alan "hız", "boyut", "pil", vb... isimler, düşünce analizi yapılan incelemenin temel ögesi olan bilgisayarın özellik kümesi hakkında bize bilgi verecektir. Bunun yanında, metin verisi içerisinde sıfat türünde bulunan kelimeler ise metni yazan kişinin olumlu mu yoksa olumsuz mu bir düşünceye sahip olduğu yönünde fikir verecektir. Cümleler içerisinde bulunan isimlerin beraber kullanıldıkları sıfatların incelenmesi sonucunda ise, metni yazan kişinin belirli özellikler için olumlu ya da olumsuz düşüncelere sahip olduğu analizi yapılabilir.

Düşünce belirten kelimeler ve cümlelerin tespiti

Üzerinde düşünce analizi yapılan metinler içerisinde, doğrudan metni yazan kişinin düşüncesinin pozitif mi yoksa negatif mi olduğunun tespiti için kullanılabilecek kelimeler veya cümleler bulunabilir. Bunlar, yoruma açık olmadan doğrudan bahsedilen şey hakkında fikir verecek olan "iyi", "kötü", "muhteşem", "rezalet", vb... kelimeler ya da artık dilde kalıplaşmış olan ve belirli durumlar için kullanılan "hayal kırıklığı yaşamak" gibi cümleler veya deyimler olabilir.

Olumsuzluk içeren kelimelerin tespiti

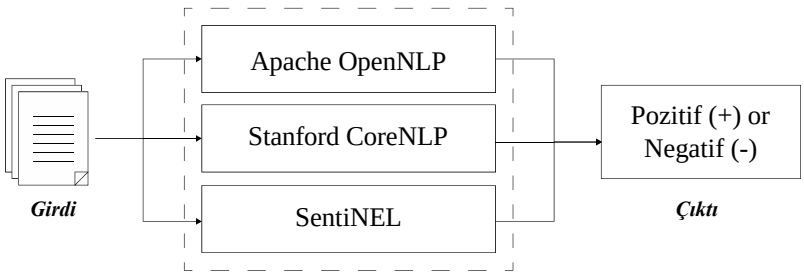
Analizi yapılacak olan metinler içerisindeki pozitif veya negatif anlam veren kelimelerin tespit edilmeleri tek başlarına yeterli olmamaktadır. Bunun sebebi bu

kelimelerin birlikte kullanıldıkları diğer kelimeler doğrultusunda anlamlarının değişme ihtimalinin olmasından kaynaklanmaktadır. Örneğin; "Filmdeki oyunculuklar çok başarılı." ve "Filmdeki oyunculuklar hiç başarılı değil." cümleleri içerisinde yer alan kelimeler tespit edildiğinde temelde aynı kelimelerin bulunduğu görülmektedir. Ancak cümlelerden ilki olumlu bir anlam içermekteyken, diğer cümle içerisinde beraber kullanıldığı kelimelinin anlamını değiştiren "değil" kelimesi bulunduğu için tamamen olumsuz bir anlam içermektedir. Bu sebeple beraber kullanıldığı kelimelerin anlamlarını değiştirebilecek olan olumsuzluk veren kelimelerin tespit edilmesi, düşünce analizi yapılırken doğru sonuçların elde edilebilmesi adına oldukça büyük önem taşımaktadır.

Gözetimsiz öğrenme yöntemleri ise tamamen cümleleri oluşturan yapıları ve cümleler içerisindeki kelimeleri düşünce analizi yapmak için kullanmaktadır. Bu sistemler, metin verisi içerisinde yer alan cümleler içerisindeki sıfatların, sıfat tamlamalarının, ad tamlamalarının, vb... yapıların tespit edilmesi ve bu yapılar içerisindeki kelimelerin sağladıkları olumlu veya olumsuz anlamlar doğrultusunda metni yazan kişinin düşüncesinin tespit edilmesi şeklinde çalışmaktadır. Örneğin, "Cihaz kalitesiz fotoğraf çekiyor." cümlesinde ilk olarak "kalitesiz fotoğraf" sıfat tamlamasının bulunması ve bu sıfat tamlaması içerisindeki "kalitesiz" kelimesi doğrultusunda cümle içerisindeki düşüncesinin anlamsız olduğu sonucunu üretmektedir.

3.2.2.4 Önerilen sistem

Bahsedildiği üzere, metin verisi üzerinde düşünce analizi yapmak için de birçok farklı yaklaşım ve yöntem önerilmiştir. Bu yöntemlerin de kullandıkları eğitim modellerine veya analiz yapmak için üzerinde çalıştırıldıkları metin verisinin yapısına göre birbirlerine sağladıkları farklı avantajları ve dezavantajları bulunmaktadır. Düşünce analizi gibi dilin kullanım esnekliğinin çok etkili olduğu ve sübjektif olan bir konuda, farklı yöntemler doğrultusunda elde edilen farklı sonuçlar daha da ön plana çıkmaktadır. Bu sebeple, düşünce analizi yapmak için de varlık isimlerinin tanımlanması için kurgulanan yapıya benzer bir yapı tasarlanmıştır. Tasarlanan bu yapıda da varlık isimlerinin tanımlanması alt modülünde olduğu gibi birden fazla, farklı yöntemler doğrultusunda düşünce analizi yapan araç kullanılmış ve farklı yöntemlerden elde edilen bu sonuçlar birleştirilip, sabit bir formatta sistemin kullanıcısına sunulmuştur. Oluşturulan bu yapı Şekil 3.9'da görülebilir.



Şekil 3.9 : Önerilen düşünce analizi alt modülünün yapısı.

Şekil 3.9'da görüldüğü üzere, düşünce analizi yapmak için kurgulanan alt modül de temelinde 3 farklı doğal dil işleme aracını bulundurmaktadır. Sistemin içerdiği farklı doğal dil işleme araçları farklı algoritmalar ile düşünce analizi yapmaktadır. Bu alt modülün sahip olduğu yapı da oldukça esnek ve ölçeklenebilirdir. Bunun bir sonucu olarak da istenildiği takdirde sistem içerisinde yer alan farklı düşünce analizi algoritmalarının sayısı kolaylıkla arttırılabilir.

Düşün analizi alt modülünde yer alan ilk doğal dil işleme aracı, varlık isimlerinin tanımlanması için kurgulanmış olan yapıdakine benzer bir şekilde Apache OpenNLP isimli araçtır. Apache OpenNLP, temel olarak bir sınıflandırma algoritmasıyla düşünce analizi yapmakta ve girilen metin verisini pozitif, nötr veya negatif olmak üzere sınıflandırmaktadır. Bu doğal dil işleme aracının kullandığı sınıflandırma algoritması temelinde maksimum entropi ilkesini (maximum entropy principle) [73, 74] barındırmaktadır. Apache OpenNLP ile düşünce analizi yapılabilmesi için sistemin bir eğitim verisi ile eğitilmesine ihtiyaç duyulmaktadır. Bu eğitim verisi içerisinde farklı cümleler ve cümlelerin düşünce analizi skorları yer almalıdır.

Tasarlanan sistem içerisinde bulunan bir diğer doğal dil işleme aracı ise Stanford CoreNLP aracıdır. Stanford CoreNLP aracı, düşünce analizi yapmak için ikili ağaç yapısı formatında, işaretlenmiş olan cümleleri eğitim verisi olarak kullanmakta ve daha sonra bu eğitim verisi üzerinden yinelemeli tensör ağlarını (recursive tensor network) [75] kullanarak analiz sonuçlarını elde etmektedir. Bu doğal dil işleme aracı da verilen metin verisi doğrultusunda çok pozitif, pozitif, nötr, negatif veya çok negatif olmak üzere 5 farklı sonuçtan birini vermektedir.

Kullanılan son araç ise SentiNEL [76] isimli düşünce analizi aracıdır. SentiNEL, kullanılan diğer araçlardan farklı olarak sadece düşünce analizi yapmak için geliştirilmiş olan bir araçtır. Bu araç, bütün bir metin verisi üzerinde düşünce analizi

yapmak yerine metin verisi içerisinde belirtilmiş olan belirli kelimeler veya varlık isimleri üzerine düşünce analizi yapmaktadır. Böylelikle metin verisi içerisinde özellikle belirtilmiş olan bir özne veya obje hakkındaki düşünce analizi sonuçlarını üretebilmektedir. Bu düşünce analizi aracı, IOA ismi verilen karar destek makineleri (support vector machines) tabanlı altyapıyı [77] kullanarak özellikle sosyal medya içerikleri üzerinde düşünce analizi yapması için tasarlanmıştır. SentiNEL isimli düşünce analizi aracı da verilen metin verisi doğrultusunda pozitif, nötr veya negatif olmak üzere 3 farklı sonuçtan birini vermektedir.

Düşünce analizi alt modülünde kullanılan araçlarda, varlık isimleri tanımlama alt modülünde kullanılan araçlardan farklı olarak daha basit bir çıktı formatı yer almaktadır. Bu alt modülde kullanılan araçlar verilen metin verisi doğrultusunda, düşüncenin ait olduğu sınıf bilgisini döndürdüğü için, bu noktada elde edilen sonucun birleştirilmesi ve tekilleştirilmesi de oldukça kolaydır. Bu doğrultuda kurgulanan alt modül girdi olarak verilen metin verisine pozitif veya negatif olmak üzere bir değer vermekte bunu sistemin çıktısı olarak vermektedir.

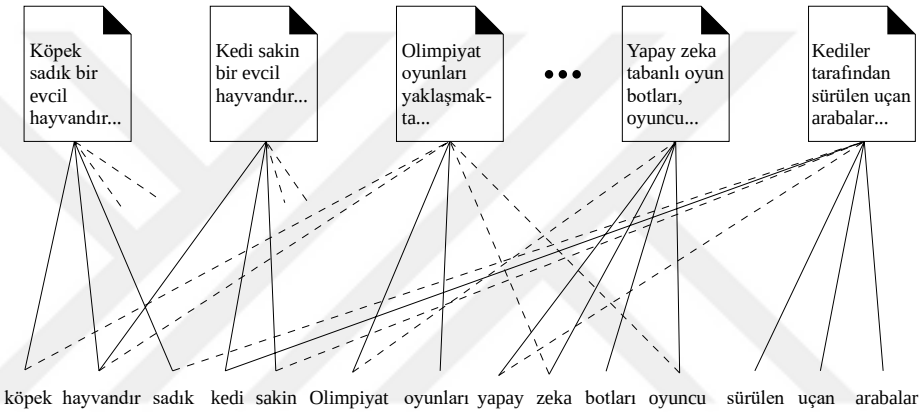
3.2.3 Konu modellemesi

Konu modellemesi (topic modelling), doğal dil işleme ve bilgi çıkarma teknikleri altında bulunan çalışma alanlarından bir diğeridir. Konu modellemesi ile hedeflenen, adından da anlaşılabileceği gibi, metin dosyalarından oluşan bir veri kümesi üzerinden, bu metin dosyaları içerisinde bahsedilen konuların otomatik olarak tespit edilmesi ve sonrasında bu metin dosyalarının, elde edilen bu konular doğrultusunda işaretlenmesidir. Bahsedilen metin dosyaları gazete makaleleri, kitap bölümleri, akademik makaleler ve blog içerikleri başta olmak üzere her türlü yazılı içerik olabilir.

Konu olarak isimlendirilen yapılar, metin verisi içerisinden çıkarılmış, her bir dokümanı temsil etmekte kullanılabilecek ve birbiri ile ilişkili olan kelimelerin bulunduğu öbeklerdir. Konu modellemesi yöntemlerinin uygulanması sonucu elde edilen kelime öbeklerine bir örnek vermek gerekirse; spor ve politika konulu makaleleri içeren bir veri kümesi üzerinde uygulanmış olan bir konu modellemesi algoritmasının sonucu, “futbol”, “hakem”, “şampiyon”, vb... kelimeleri içeren bir kelime öbeği ile “parti”, “seçim”, “aday”, vb... kelimeleri içeren bir başka kelime öbeği olacaktır. Üretilen bu öbekler aynı zamanda yeni dokümanların değerlendirilmesinde ve hangi konuya ait olduğunun belirlenmesinde de kullanılabilir.

Bu noktada, önemli olan bir başka ayrıntı ise bir doküman her zaman tek bir konu içermek zorunda değildir, birden fazla konuyu da içerisinde bulundurabilir. Konu modellemesi algoritmalarının gerçekleştirdiği bir başka iş ise konuları tespit ettikten sonra, dokümanların elde edilen konuların hangisini ne miktarda içerdiğini tespit edip, değerlendirmesini yapmasıdır.

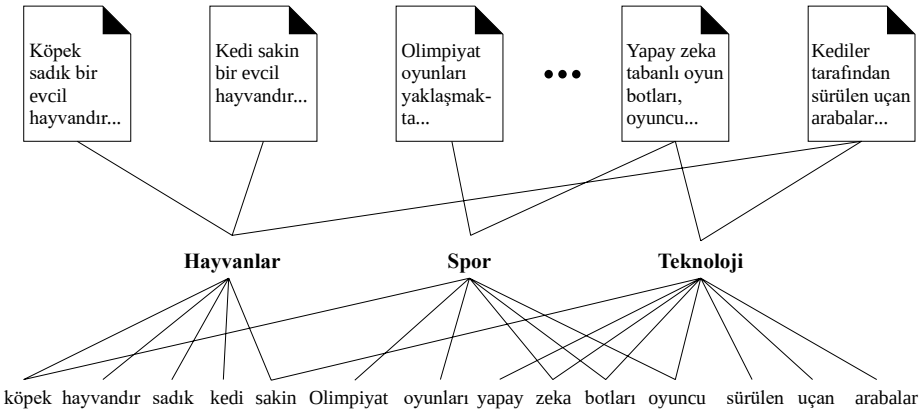
Konu modellemesi ile yapılan, bir başka ifade ile, ilk olarak dokümanlar içerisinde yer alan tüm kelimelerin, bulundukları dokümanlar ile olan ilişkilerini ifade eden bir ağın oluşturulmasıdır. Bu ağ içerisinde yer alan her bir düğüm (node) bir kelime veya bir dokümanı ifade etmekte her bir kenar (edge) ise kelime ve doküman arasındaki ilişkiyi ifade etmektedir. Bu ilişkiyi simgeleyen örnek bir ağ Şekil 3.10'da görülebilir.



Şekil 3.10 : Dokümanlar ve dokümanlar içerisinde yer alan kelimelerin ilişkilerini gösteren örnek ağ.

Şekil 3.10'da görülebileceği üzere her bir doküman içerisinde sayısal anlamda oldukça fazla kelime yer almaktadır. Bu sebeple oluşturulan ağ içerisindeki düğüm ve kenar sayısı da bu doğrultuda sayı olarak yüksek bir değere sahiptir. Buna ek olarak doküman sayısının arttığı noktada bu ağ içerisinde yer alan ilişkilerin sayısı yine aynı şekilde artacak ve daha karmaşık bir hale gelecektir.

Dokümanların ve dokümanlar içerisindeki kelimelerin ilişkilerinin bir ağ şeklinde ifade edildiği bu durumda, ilişki sayılarının miktarı durumu karmaşık bir hale getirmektedir. Bu ilişkileri daha anlaşılabilir bir şekilde ifade etmek adına, birbirine yakın olan kelimeler gruplanarak konular oluşturulmakta ve daha sonra konular ile dokümanlar arasındaki ilişkiler tanımlanmaktadır. Örnek konuların oluşturulduğu ve oluşturulan konuların ağ içerisine eklendiği yeni ağ Şekil 3.11'de görülebilir.



Şekil 3.11 : Dokümanlar, dokümanlar içerisinde yer alan kelimeler ve oluşturulan konuların ilişkilerini gösteren örnek ağ.

Şekil 3.11'de görülebileceği üzere aynı dokümanlar içerisinde bir arada yer alan kelimelerin gruplanması sonucu konular oluşturulmuştur. Sonrasından ise oluşturulan bu konular ara bir katman olarak tanımlanarak, kelime – konu ilişkileri ve konu – doküman ilişkileri ağ içerisine eklenmiştir. Yapılan bu ara katman eklemesi ile elde bulunan ağ üzerinde analiz yapılması da kolaylaştırılmıştır.

Konu modellemesi çalışma alanında bulunan algoritmalar ile elde edilecek olan sonuçlar, elde bulunan metin içerikli dokümanların gruplandırılması, karmaşık bir metin dokümanının alt parçalarına ayrıştırılması, yapısal olmayan veya gürültü içeren metin verilerinin anlamlandırılması gibi problemleri çözmekte kullanılabilir. Bir sonraki aşamada ise elde sunulan bu yapılar, bir öneri sisteminin gerçekleştirilmesi veya profilleme yapan bir sistemin oluşturulması gibi daha karmaşık yapıların temelinde yer alabilirler.

3.2.3.1 Ön işleme yöntemleri

Konu modellemesi yapılmadan önce algoritmalardan maksimum verimin alınabilmesi adına, metin verisi üzerinde çeşitli ön işleme (preprocessing) adımlarının [78, 79] uygulanmasına ihtiyaç vardır. Bu ön işleme adımları, üzerinde doğal dil işleme yöntemleri kullanılacak olan metinler içerisinden gereksiz kısımların atılmasına ve bu metin dokümanlarının daha kolay bir biçimde ifade edilecek şekilde sadeleştirilmesini amaçlamaktadır. Kullanılacak olan ön işleme tekniklerinin çeşitliliği ve değeri, doğal dil işleme problemleri içerisinde, uygulanacakları probleme göre oldukça yüksek olabilir.

Metin bölütleme

Metin bölütleme (text segmentation) veya bir diğer ismiyle simgeleştirme (tokenization), bir metin içerisinde, metni bir araya getiren farklı parçaların tespit edilmesidir. Bu tespit etme işlemi bir doküman içerisindeki paragrafların, paragraflar içerisindeki cümlelerin veya cümlelerin içerisindeki kelimelerin tespit edilmesi şeklinde olabilir. Metin içerisindeki farklı parçaların bulunması için istatistiki modellere başvuran yöntemler [80] veya kelimeler arası mesafelerin tespiti üzerinden bu parçalama işlemini yapan yöntemler [81] olmak üzere birçok farklı yaklaşım bulunmaktadır. Bu parçaların metin verisi içerisinde tespiti, bir sonraki kısımda yapılacak olan ön işleme ve diğer işlemleri doğrudan etkilemektedir. Bu sebeple metin bölütleme işlemi doğal dil işleme yöntemleri uygulanacak veri için oldukça büyük bir öneme sahiptir.

Metnin küçük harfe çevrilmesi

Bu ön işleme yöntemi oldukça basit bir işlemdir. Adından da anlaşılabileceği üzere, bu ön işleme yönteminde gerçekleştirilen elde bulunan metin verisinin tamamen küçük harfe çevrilerek tekilleştirilmesidir. Bunun bir sonucu olarak, ön işleme sonrası aşamalarda kullanılmak istenen doğal dil işleme yöntemleri sırasında karşılaşılma ihtimali olan, büyük/küçük harf duyarlı yapılardan üretilecek yanlış sonuçların önüne geçilmektedir.

Sözcük türü işaretleme

Sözcük türü işaretleme (part-of-speech tagging) (POS tagging), dokümanlar içerisinde yer alan metin verisi üzerinde uygulanan bir işaretleme yöntemidir. Bu işaretleme işlemi, metin verisi içerisindeki cümleler tespit edildikten sonra, bu cümleler üzerinde uygulanmaktadır. Sözcük türü işaretleme ile amaçlanan, dokümanlar içerisinde yer alan her bir kelimenin hangi amaçla kullanıldığını (özne, yüklem, sıfat, zarf, vb...) tespit etmektir. Bu bilgi sonraki aşamalarda sıfat tamlaması gibi farklı yapıların tespit edilmesinde metin verisi içerisinden sadece belirli türdeki kelimelerin seçilip, sonrasındaki işlemlerin bu kelimeler üzerinde çalıştırılması gibi filtreleme işlemleri sırasında kullanılabilir. Sözcük türlerinin işaretlenmesi için karar ağaçları tabanlı [82], maksimum entropi modelleri tabanlı [83] ve kural tabanlı sınıflandırıcıları kullanan [84] yöntemler olmak üzere birçok farklı yöntem bulunmaktadır.

Etkisiz kelimelerin temizlenmesi

Dil içerisinde çok sık kullanılan ve farklı içerikler içerisinde rastlanma ihtimali çok yüksek olan kelimeler etkisiz kelimeler (stop words) olarak adlandırılmaktadır. Bu kelimeler, örneğin İngilizce için incelendiğinde *"are", "most", "the", "this", vb...* kelimelerdir. Bu kelimelerin yoğun ve sık kullanılmalarından ötürü, kimi doğal dil işleme yöntemleri belirli içerikler için uygulanacağı zaman sonuçları olumsuz olarak etkileyebilmektedir [85, 86]. Bu sebeple bir ön işlem olarak etkisiz kelimeler içerikler içerisinden temizlenmektedir. Farklı diller için etkisiz kelimelerin bulunduğu listeler bulunmaktadır ve temizleme işlemleri genelde bu listeler üzerinden yapılmaktadır. Bunun yanında içerikler doğrultusunda etkisiz kelimelerin otomatik olarak belirlendiği çeşitli yöntemler [87, 88] de mevcuttur.

Kurallı ifadelerin filtrelenmesi

Konu modellemesi gibi doğal dil işleme yöntemlerinin uygulanacağı içerikler üzerinden bir başka filtreleme işlemi ise kurallı ifadeler kullanılarak yapılmaktadır. Bu ön işlemde, içerik içerisinde tutulmak istenen veya filtrelenmek istenen yapılar kurallı ifadeler yardımı ile tanımlanarak filtreleme işlemi yapılmaktadır. Bu filtreleme işlemi ile içerik içerisinden özel karakter bulunan kelimeler, içerisinde sayısal karakter bulunduran kelimeler, belirli bir uzunluktan daha kısa olan kelimeler, vb... özel olarak tanımlanan kelimeler temizlenebilir.

Kelime köklerinin bulunması

Kelime köklerinin bulunması (stemming) ön işlemi, temel olarak metin verisi içerisinde bulunan kelimelerin köklerinin tespit edilmesi ve kullanılması işlemidir. Bu işleme, İngilizce'den birkaç örnek vermek gerekirse; *"birds"* kelimesinin kökünün *"bird"* olarak bulunması, *"best"* kelimesinin kökünün *"good"* olarak bulunması veya *"going"* kelimesinin kökünün *"go"* olması kelime kökünü bulmaya yönelik örnekler içerisindedir. İçerikler içerisinde kelimelerin köklerinin bulunması ve kullanılması özellik konu modellemesi gibi istatistik tabanlı yaklaşımlar kullanan yöntemler için oldukça önemlidir. Bunun sebebi kelime köklerinin bulunması ile eklerden, zaman çekimlerinden veya farklı etkenlerden ötürü içerik içerisinde birbirinden farklı şekillerde bulunan aynı köke sahip olan kelimelerin gruplanabilmesidir. Kelime köklerinin bulunması için oldukça popüler Snowball [89] isimli mini bir dil ve Porter

[90] isimli bir algoritma bulunmaktadır. Bu algoritmalara ek olarak, bu probleme odaklanan farklı yaklaşımlar [91, 92] da bulunmaktadır.

3.2.3.2 Konu modellemesi yöntemleri

Konu modellemesi yapmak için önerilen yöntemler, kurallı ifadelerin veya sözlük tabanlı arama yöntemlerinin kullanıldığı kural tabanlı metin madenciliği yöntemlerinin aksine gözetimsiz öğrenme temelli yöntemlerdir. Bir başka deyişle, veri üzerinde konu modellemesi yapılmadan önce herhangi bir etiketleme, eğitim veya benzeri bir işleme ihtiyaç duyulmadan sadece seçilen algoritmaların çalıştırılması ile sonuçlar elde edilebilir. Konu modellemesi yapmak için kullanılan yöntemler arasında, kelime ağırlıkları ve vektörleri üzerinden istatistiki hesaplamalar yapan birçok yöntem bulunmaktadır. Bu yöntemlerden öne çıkanlar arasında matrisleri negatif olmayan çarpanlarına ayırma, örtülü anlam çözümlemesi, olasılıksal örtülü anlam çözümlemesi, örtülü Dirichlet ayrıştırması ve kelime vektörel uzayında örtülü Dirichlet ayrıştırması isimli yöntemler gösterilebilir.

Matrisleri negatif olmayan çarpanlarına ayırma

Matrisleri negatif olmayan çarpanlarına ayırma (non-negative matrix factorization) (NMF) yöntemi elde bulunan bir matris verisi üzerinde çarpanlarına ayırma işlemi yapan ve bu çarpanlar doğrultusunda belirli bir ağırlık ve öznitelik kümesi bulmayı hedefleyen bir yöntemdir [93, 94]. Yapılan bu çarpanlarına ayırma işlemi içerisinde kritik olan elde edilen çarpan sonuçların tamamının pozitif olmasının gerekliliğidir. Bunun sebebi, NMF yönteminin uygulandığı problemlerde elde edilecek çarpanların, bir öznitelik vektörü ve bu özniteliklerin ağırlıklarının vektörü olarak elde edilmek istenmesinden kaynaklanmaktadır.

Bu yöntem, içerisinde ayrıştırma işlemi yapılacak olan veri kümelerini içeren problemler üzerinde uygulanabilir. Bu problemlere örnek olarak, fotoğraflar içerisinde bulunan farklı alanların veya objelerin tespit edilmesi [95] verilebilir. Satın alınan ürünler ve bu ürünler için yazılan değerlendirme yazıları üzerinden kişiler gruplanabilir ve iş birliğine dayalı filtreleme (collaborative filtering) temelli öneri sistemleri [96] kurgulanabilir. Bunların yanında benzer bir yaklaşımla, elde bulunan bir doküman kümesi üzerinde NMF yönteminin uygulanması sonucu bu veri kümesini simgeleyen kelime gurupları ve kelime grupları içerisinde yer alan kelimelerin

ağırlıkları tespit edilebilir [97]. Bu sayede dokümanları simgeleyen konular NMF yöntemi ile tespit edilebilmektedir.

Örtülü anlam çözümlemesi

Konu modellemesi yapmak için kullanılabilecek bir diğer yöntem örtülü anlam çözümlemesi (latent semantic analysis) (LSA) [98] isimli yöntemdir. LSA'nin temel olarak yaptığı, elde bulunan veri kümesinin özniteliklerini filtrelemek ve önemli olanlarını seçmektir. Bunu yapmak için LSA isimli yöntem de, elde bulunan veri üzerinden oluşturulmuş olan öznitelik değer matrisini önce tekil değer ayrışması (singular value decomposition) (SVD) [99] isimli yöntemle çarpanlarına ayırmakta ve daha sonra bu değerleri sadeleştirmektedir.

LSA kullanılarak konu modellemesi yapılabilmesi için ilk olarak veri içerisindeki öznitelikler tanımlanmalı ve veriyi ifade edecek matris oluşturulmalıdır. Metin verisi için oluşturulacak bu matriste her bir satır bir dokümanı her bir kolon ise tüm dokümanlarda geçen her bir kelimeyi temsil etmektedir. Matris içerisinde yer alan değerler ise kelimelerin dokümanlarda geçme sayıları veya TF-IDF değerleri olarak belirlenebilir. Bu noktada TF-IDF [100] değerlerinin kullanılması daha yaygın bir yaklaşımdır. Bunun sebebi, TF-IDF ile hesaplanan değerlerin kelimelerin toplam geçme sayısı ve farklı dokümanlara dağılma değerleri üzerinden hesaplanmasından ötürü sadece kelime sayılarına göre önemli kelimeleri seçme noktasında daha değerli sonuçlar üretmesinden kaynaklanmaktadır.

Dokümanlar, bu dokümanlar içerisinde geçen kelimeler ve bu kelimeler için hesaplanan değerler, LSA için kullanılacak olan girdi matrisini oluşturmaktadır. LSA çıktı olarak ise öznitelikleri sadeleştirilmiş olan bir çıktı vermektedir [101]. Bu matrisde kalan öznitelikler konuları temsil eden kelimeler ve değerler ise bu kelimelerin dokümanlar üzerindeki ağırlıklarını ifade etmektedir. Ancak bu noktada elde edilen bu sonuçlar üzerinden konuları oluşturan kelimelerin gruplanması otomatik olmayan ve kişiler tarafından tamamlanması gereken bir işlemdir.

Olasılıksal örtülü anlam çözümlemesi

Olasılıksal örtülü anlam çözümlemesi (probabilistic latent semantic analysis) (pLSA), LSA tabanlı olasılıksal bir yöntemdir. LSA ile konu modellemesi yapılırken elde bulunan, dokümanları ve kelimeleri içeren matris SVD yöntemi ile çarpanlarına

ayrılırken pLSA yönteminde çarpanlarına ayırma noktasında olasılıksal bir model kullanılmaktadır [102, 103]. Bu model, dokümanlar içerisindeki kelimelerin ve konuların olasılıksal olarak dağılımlarından yola çıkılarak oluşturulmaktadır. Daha sonra LSA'de elde edilen sonuçlara benzer bir şekilde elde edilen sonuçlar üzerinden konulara ait etiketlerin verilmesi ve bunların birbirlerinden ayrılması [104] işlemleri tamamlanmaktadır.

Örtülü Dirichlet ayrıştırması

Örtülü Dirichlet ayrıştırma (latent Dirichlet allocation) (LDA) yöntemi [105, 106, 107], olasılıksal bir model olan pLSA'i temel alan ve modelin üzerine ekstra bir katman olarak da Bayes Teoremi'ni ekleyerek daha gelişmiş bir çözüm ortaya koyan bir yaklaşımdır. LDA'in, pLSA'den bir diğer farklı ise pLSA sadece üzerinde çalıştırıldığı veri kümesi için bir sonuç üretmekte ve bir veri kümesi için üretilmiş olan sonuçlar yeni bir doküman geldiği noktada hiçbir anlam ifade etmemektedir. Ancak LDA ile üretilmiş olan bir model ve bu model içerisinde tespit edilmiş olan konular farklı bir veri kümesinin konu dağılımlarını tespit edebilmek adına tekrar kullanılabilmektedir. Bu noktada model üretilirken kullanılan veri kümesinin büyüklüğü ve veri kalitesi de oldukça büyük önem taşımaktadır. Buna ek olarak aynı zamanda LDA, pLSA'den farklı olarak tespit ettiği her bir konuyu da farklı gruplara bölerek, dokümanları bu konular için puanlayabilmektedir.

Kelime vektörel uzayında örtülü Dirichlet ayrıştırması

Metin verisi içerisindeki kelimeler, vektör uzayında ifade edilebilir ve daha sonra bu vektörel uzaydaki değerlerden yola çıkılarak farklı kelimeler arasındaki uzaklık ilişkileri tanımlanabilir. Bu vektör uzayının oluşturulması için word2vec isimli model [108, 109] kullanılabilir. Kelimeler arasındaki uzaklık, içerik doğrultusunda değişiklik gösterebilir. Bu model kullanılarak, aynı zamanda kelimelerin arasındaki içerik temelli uzaklıklar da hesaplanabilir. Örneğin, sıradan bir metin verisi ile eğitilmiş olan bir word2vec modelinde “kahve” ve “çay” kelimeleri arasındaki mesafe veya “Ankara” ve “İstanbul” kelimeleri arasındaki mesafe yakın olacaktır. Başka bir örnek olarak ise teknolojik içeriğe sahip bir metin verisi ile eğitilmiş olan bir word2vec modelinde “telefon”, “tablet” ve “bilgisayar” gibi kelimeler arasındaki uzaklıklar oldukça düşük olacaktır.

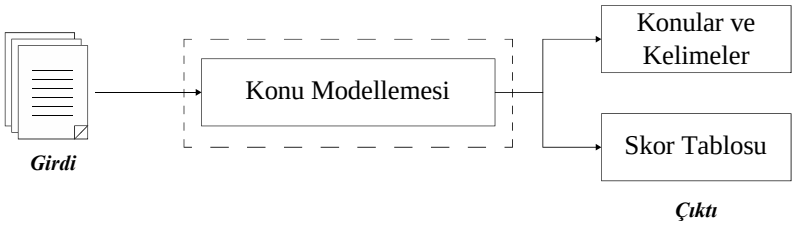
Bu modeller konu modellemesi yapılması noktasında, LDA tarafından üretilmiş olan modellere yardımcı olarak eklenebilir. LDA sonucu üretilen konuların ve içerisinde yer alan kelimelerin, sınırları içerik verisi tarafından belirlenmiş olan uzayda hangi noktaya düştüğü ve yakınında bulunan diğer değerler kullanılarak konu modelleri güçlendirilebilmekte ve genişletilebilmektedir. Bu yaklaşımdan yola çıkılarak geliştirilen ve iki farklı modelin bir arada kullanıldığı ve konu modellemesi yapılan lda2vec [110] isimli bir yöntem de bulunmaktadır.

3.2.3.3 Önerilen sistem

Konu modellemesi yapmak için bahsedilen yöntemler arasından, oldukça popüler olan ve kullanım kolaylığı sağlayan LDA yöntemi, kurgulanan sistem içerisinde konu modellemesi modülünün çekirdeğini oluşturması için seçilmiştir. LDA yönteminin seçilmesinin bir diğer sebebi ise performans ve sonuç kalitesi açısından diğer yöntemlere oranla bir adım ilerde olmasından kaynaklanmaktadır.

Konu modellemesi yapmak için önerilen sistem, varlık isimlerinin sınıflandırılması ve düşünce analizi yapılması için tasarlanan modüllerden farklı olarak, konu modellemesi modülü, farklı yapıları kullanıp sonuçlarını birleştiren bir yapı olarak değil tek bir parçadan oluşacak tekil bir yapı şeklinde tasarlanmıştır. Bunun temel sebebi, farklı algoritmalardan alınacak olan farklı konu modeli sonuçlarının yapısal farklılıklardan ötürü, birleştirilip tek bir model haline getirilmesinin zorluğudur. Bunun yanı sıra farklı algoritmaların ürettiği modeller birleştirilip tek bir model elde edilse bile, elde edilen bu birleştirilmiş model doğrultusunda, dokümanların değerlendirilmesi ve konular için puanlamanın yapılması büyük bir probleme sebep olmaktadır.

Kurgulanan sistem içerisinde konu modellemesi yapması için geliştirilmiş olan LDA tabanlı modül, girdi olarak metin içerikli dokümanlardan oluşan bir veri kümesi kabul etmektedir ve tanımlanan konu sayısı doğrultusunda 2 farklı sonuç kümesi üretmektedir. Bu sonuçlardan ilki, girdi dosyaları üzerinden üretilmiş olan ve sayısı belirtilmiş olan konular ve bu konuları simgeleyen kelime öbekleridir. Her kelime öbeğinin içerisinde, kelimelerin yanı sıra bu kelimelerin içerisinde bulundukları konular için değerlerini simgeleyen ağırlık değerleri de bulunmaktadır. Bu modülden alınan diğer bir sonuç ise üretilen konuların, modüle girdi olarak verilen her bir dokümanın üzerine dağılımını gösteren konu / doküman puanlama tablosudur. Oluşturulan bu yapı Şekil 3.12'de görülebilir.



Şekil 3.12 : Önerilen konu modellemesi alt modülünün yapısı.

LDA algoritmasının popülerliğinden ve başarımından ötürü birçok farklı programlama diline ait kütüphane, konu modellemesi modülü olarak bu algoritmayı kullanıcılarına sunmaktadır. Bu kütüphanelerden biri olan gensim [111, 112] isimli kütüphane Python programlama dili için geliştirilmiş olan kütüphanedir. Gensim, tamamen konu modellemesi üzerine odaklanmış bir kütüphane olmasının yanı sıra LDA algoritmasının performansı olarak değerlendirildiğinde de oldukça yüksek performansla çalışan bir kütüphanedir. Bu sebeplerden ötürü konu modellemesi alt modülünün temelini oluşturacak yapı içerisinde gensim isimli bu kütüphane kullanılmıştır.

4. SOSYAL AĞ ANALİZİ

Sosyal ağ analizi, çizge teorisinden yardım alarak oluşturulan sosyal yapıların analiz edilmesidir [113, 114]. Sosyal ağlar, çizgeleri de oluşturan düğüm ve kenarlardan meydana gelmektedir. Sosyal ağlar içerisindeki düğümler, kişiler, organizasyonalar veya profiller gibi bireysel aktörlerden bir araya gelir. Bu yapılar içerisindeki kenarlar ise düğümler arasındaki ilişkileri ifade etmektedir. Örneğin, Twitter içerisinde yer alan profiller arası ilişkileri gösteren bir sosyal medya ağı [115] bu sosyal yapılar içerisinde yer almaktadır.

Sosyal ağları ifade etmek için oluşturulan çizgelere, sosyogram (sociogram) adı verilmektedir ve bu ağlar üzerinde birçok analiz yapılabilmektedir. Bu analizler yardımı ile farklı aktörler arasında bulunan ancak doğrudan göz önünde olmayan ilişkiler bulunabilir [116]. Örneğin, birbiri ile iş birliğine uygun olan ancak bundan haberdar olmayan yapılar arasındaki ilişkiler sosyal ağ analizi yöntemleri ile tespit edilebilir. Sosyal ağ analizi ile yapılabilecek olan bir başka analiz ise ağ içerisindeki önemi yüksek olan düğümlerin tespit edilmesidir. Bu düğümler sosyal ağlar içerisindeki merkeziyet (centrality) [117] kavramından yararlanılarak tespit edilebilir ve bu analizler yardımı ile bir ağa yön veren düğümler tespit edilebilir. Farklı bir analiz yöntemi olarak ise ağlar içerisinde yer alan alt ağlar tespit edilerek [118], farklı gruplara ulaşılabilir ve gruplar özelinde analizler yapılabilir.

Bu çalışmada sosyal ağ analizleri, 2 farklı kaynaktan toplanmış olan veriler kullanılarak yapılmıştır. İlk olarak, sistem içerisinde geliştirilen veri toplama modülleri kullanılarak New York Times web sitesi üzerinden 01 Ocak 2017 ile 31 Aralık 2017 tarihleri arasında yayınlanmış olan 4 farklı kategoriye (manşet, sürmanşet, ulusal haberler ve uluslararası haberler) ait tüm haberler ve makaleler toplanmıştır. Bu tarihler arasında New York Times web sitesinde yayınlanmış olan 12.560 adet farklı içerik tespit edilmiş ve veri kümesine eklenmiştir. Bir diğer veri kümesi olarak ise 01 Ağustos 2017 ile 30 Kasım 2017 tarihleri arasında Twitter isimli mikroblog platformu üzerinden *"North Korea"* kelimeleri ile yapılmış olan paylaşımlar derlenmiştir. Bu kelime grubunun seçilmesinin sebebi, belirtilen tarihler arasında *"Amerika Birleşik*

Devletleri" ve *"Kuzey Kore"* arasında nükleer silahlanma temelli bir kriz yaşanması ve bu krizin toplumda ve sosyal medyada oldukça fazla etkisinin olmasından kaynaklanmaktadır. Twitter üzerinden belirtilen tarihler arasında konuyla ilgili yapılmış olan 2.854.333 adet paylaşıma ulaşılmıştır ve bu paylaşımlardan oluşan bir başka veri kümesi daha oluşturulmuştur.

Oluşturulan veri kümeleri geliştirilmiş olan sistem yardımı ile hem bir veri tabanı içerisinde hem de JSON dosyaları olarak saklanmıştır. Bu veriler üzerinde daha sonra geliştirilmiş olan varlık ismi tanımlama alt modülü, konu modelleme alt modülü ve düşünce analizi alt modülü çalıştırılmıştır. Bu alt modüllerden elde edilen sonuçlar doğrultusunda var olan veriler güncellenmiş ve sonraki aşamalarda yer alan analizlerde bu güncellenmiş veriler kullanılmıştır.

Elde edilen veri üzerinde, sosyal ağ analizleri varlık isimleri temelli, konu modellemesi temelli ve düşünce analizi temelli analizler olmak üzere 3 farklı kategoride yapılmıştır ve analizlerden elde edilen çıktılar belirli zaman aralıkları içerisinde değerlendirilmiştir.

4.1 Varlık İsimleri Temelli Analizler

Varlık isimleri temelli analizler, New York Times web sitesi üzerinden toplanan haber ve makale verisi üzerinde gerçekleştirilmiştir. Bu analizlerin gerçekleştirilebilmesi için ilk olarak geliştirilmiş olan sistemdeki varlık isimlerinin tespiti alt modülü yardımı ile haberler içerisinde yer alan tüm varlık isimleri tespit edilmiştir. Varlık ismi tespiti alt modülünün çalıştırılması sonucu 41.184 adet varlık ismi tespit edilmiş ve daha sonra tespit edilen bu varlık isimlerinin birbirleri arasındaki ilişkilerin zaman içerisindeki değişimleri gözlemlenmiştir.

4.1.1 Varlık ismi ağlarının oluşturulması

New York Times web sitesi üzerinden toplanan veri içerisinde bulunan varlık isimlerinin tespit edilmesinin ardından, bu varlık isimlerinin hangi günlerde, hangi sıklıkta geçtiğine ve birbirleri arasındaki ilişkilerine yönelik analizler yapılmıştır. Bu ilişki ağlarını oluşturabilmek için ilk olarak varlık isimlerinin hangi haberler içerisinde ne sıklıkla geçtiği temel bilgisini gösteren tablolar oluşturulmuştur. Oluşturulan bu tablolar temel olarak haber id'si, haber tarihi, tespit edilen varlık ismi ve bu varlık

isminin ilgili haberde kaç defa geçtiği bilgisini her bir haber–varlık ismi ikilisi için listelemektedir. Bu listenin temel yapısı Şekil 4.1'de görülebilir.

$Haber_1$	$tarih(Haber_1)$	Varlık İsmi$_1$	$adet(Haber_1, Varlık\ İsmi_1)$
$Haber_1$	$tarih(Haber_1)$	Varlık İsmi$_2$	$adet(Haber_1, Varlık\ İsmi_2)$
⋮	⋮	⋮	⋮
$Haber_1$	$tarih(Haber_1)$	Varlık İsmi$_n$	$adet(Haber_1, Varlık\ İsmi_n)$
$Haber_2$	$tarih(Haber_2)$	Varlık İsmi$_1$	$adet(Haber_2, Varlık\ İsmi_1)$
$Haber_2$	$tarih(Haber_2)$	Varlık İsmi$_2$	$adet(Haber_2, Varlık\ İsmi_2)$
⋮	⋮	⋮	⋮
$Haber_2$	$tarih(Haber_2)$	Varlık İsmi$_n$	$adet(Haber_2, Varlık\ İsmi_n)$
⋮	⋮	⋮	⋮
$Haber_i$	$tarih(Haber_i)$	Varlık İsmi$_1$	$adet(Haber_i, Varlık\ İsmi_1)$
$Haber_i$	$tarih(Haber_i)$	Varlık İsmi$_2$	$adet(Haber_i, Varlık\ İsmi_2)$
⋮	⋮	⋮	⋮
$Haber_i$	$tarih(Haber_i)$	Varlık İsmi$_n$	$adet(Haber_i, Varlık\ İsmi_n)$

Şekil 4.1 : Haberler ve haberler içerisinde tespit edilen varlık isimleri ilişkilerinin gösterildiği örnek yapı.

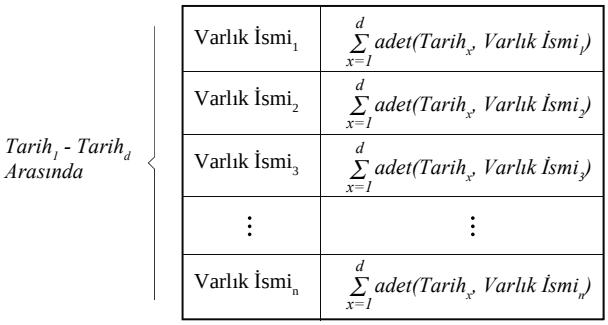
Varlık ismi ağları oluşturulmadan önce, haber–varlık ismi tablosunda dikkat edilmesi gereken bir diğer nokta ise tabloda yer alan bazı haberlerin aynı tarihte geçiyor ve benzer varlık isimlerini içeriyor olmasıdır. Bu durum, tespit edilmiş olan bazı varlık isimlerinin diğerlerine kıyasla daha popüler olduğunu ve haberler içerisinde daha sık kullanıldığını göstermektedir. Bu durumda, bu varlık isimlerini temsil eden düğümlerin varlık ismi ağları içerisindeki ağırlık ve önem değerleri de farklılık göstermektedir. Bu ağırlık değerlerinin doğru olarak belirlenebilmesi adına aynı gün içerisinde yer alan haberlerdeki aynı varlık isimleri birleştirilerek tek bir yapı haline getirilmiştir. Varlık isimleri üzerinde uygulanmış olan bu birleştirme işlemi sonucunda bir varlık ismi listesi oluşturulmuş ve varlık isimlerinin haberlerin toplandığı tüm tarihler içerisindeki geçme sayıları güncellenmiştir. Oluşturulan bu yeni varlık ismi listesinin temel yapısı Şekil 4.2'de görülebilir.

$Tarih_1$	$Varlık\ İsmi_1$	$adet(Tarih_1, Varlık\ İsmi_1)$
$Tarih_1$	$Varlık\ İsmi_2$	$adet(Tarih_1, Varlık\ İsmi_2)$
⋮	⋮	⋮
$Tarih_1$	$Varlık\ İsmi_n$	$adet(Tarih_1, Varlık\ İsmi_n)$
$Tarih_2$	$Varlık\ İsmi_1$	$adet(Tarih_2, Varlık\ İsmi_1)$
$Tarih_2$	$Varlık\ İsmi_2$	$adet(Tarih_2, Varlık\ İsmi_2)$
⋮	⋮	⋮
$Tarih_2$	$Varlık\ İsmi_n$	$adet(Tarih_2, Varlık\ İsmi_n)$
⋮	⋮	⋮
$Tarih_i$	$Varlık\ İsmi_1$	$adet(Tarih_i, Varlık\ İsmi_1)$
$Tarih_i$	$Varlık\ İsmi_2$	$adet(Tarih_i, Varlık\ İsmi_2)$
⋮	⋮	⋮
$Tarih_i$	$Varlık\ İsmi_n$	$adet(Tarih_i, Varlık\ İsmi_n)$

Şekil 4.2 : Aynı tarihlerde yayınlanmış olan haberler ve bu haberler içerisinde tespit edilen varlık isimleri ilişkilerinin gösterildiği örnek yapı.

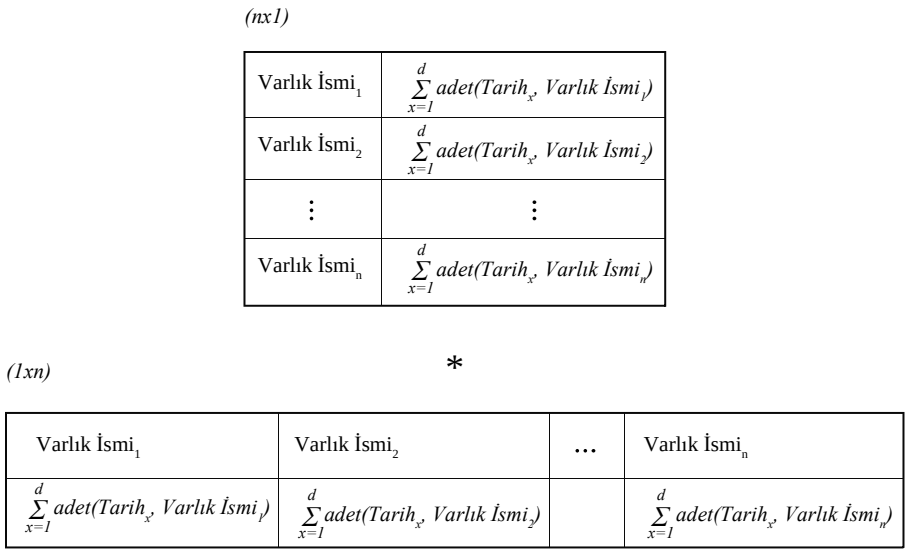
Şekil 4.2'deki yapı doğrultusunda, varlık ismi sayılarına ait tablo oluşturulduktan sonra bu tablodaki değerler kullanılarak aynı zaman dilimi içerisinde geçen varlık isimleri tespit edilmekte ve bu doğrultuda varlık isimlerini içeren ağlar oluşturulmaktadır. Bu ağlar oluşturulurken kullanılacak olan zaman dilimleri günlük, haftalık veya aylık olacak şekilde oldukça esnek bir şekilde belirlenebilir. Bu sayede yapılmak istenen analizler farklı zaman dilimleri ve aralıkları ile kolayca yapılabilir. Farklı zaman dilimleri doğrultusunda analiz yapılmak istendiğinde, belirlenen zaman diliminin genişliği doğrultusunda varlık isimleri ve bu varlık isimlerinin zaman dilimi içerisinde geçme sayıları yeniden gözden geçirilmelidir. Bu durumda, aynı zaman dilimi içerisine düşen farklı günlerdeki aynı varlık isimlerinin de bir önceki adımda olduğu gibi birleştirilmesi gerekmektedir. Bu duruma bir örnek vermek gerekirse, örneğin haftalık zaman dilimleri doğrultusunda analiz yapılmak istendiğinde, aynı hafta içerisinde farklı günlerde geçen *"United States"* varlık ismi, tek bir varlık ismi olacak şekilde birleştirilmeli ve her hafta içerisindeki toplam bulunma sayısı hesaplanmalıdır. Her bir zaman dilimi içerisindeki varlık ismi için ilgili hesaplamalar yapıldığında, bu

zaman dilimine ait olan varlık isimlerinin listelendiği bir tablo elde edilecektir. Oluşturulan bu yeni varlık ismi listesinin temel yapısı Şekil 4.3'de görülebilir.



Şekil 4.3 : Varlık isimlerinin belirli bir zaman aralığında geçme sıklığının listelendiği örnek yapı.

Yapısı Şekil 4.3'de ifade edilen bu yeni varlık ismi tablosu, varlık ismi ağlarının oluşturulmasında temel parça olarak yer almaktadır. Kullanılacak bu tablo, içerisinde varlık isimlerini ve farklı zaman dilimlerini tutmaktadır. Bu noktada varlık ismi ağlarının oluşturulabilmesi adına ihtiyaç duyulan bir diğer değer ise farklı varlık isimleri arasındaki ilişkilerin ağırlıklarıdır. Bu ağırlık değerleri, elimizde bulunan farklı zaman dilimleri doğrultusunda oluşturulmuş olan varlık ismi sayıları tablosunun kendi transpozu ile matris çarpımı sonucu elde edilmektedir. Yapılacak olan matris çarpımının yapısı Şekil 4.4'de görülebilir.



Şekil 4.4 : Varlık isimleri ve tarihlere göre geçme sayılarını bulunduran matris ile bu matrisin transpozu ile çarpımının örnek gösterimi.

Şekil 4.4'de görüldüğü üzere elimizde bulunan içerisinde belirlenmiş bir zaman dilimi için n adet farklı varlık isminin kaç kere geçtiğini gösteren $(n\ x\ 1)$ büyüklüğünde bir matris bulunmaktadır. Bu matrisin transpozu alınarak $(1\ x\ n)$ büyüklüğünde bir başka matris elde edilmiş ve daha sonra bu iki matris arasında bir matris çarpımı işlemi gerçekleştirilmiştir. Bu noktada, $(n\ x\ 1)$ ve $(1\ x\ n)$ büyüklüğündeki bu iki matrisin çarpım işleminin yapılması sonucu $(n\ x\ n)$ büyüklüğünde bir matris elde edilmiştir. Varlık ismi değer matrisinin kendi transpozu ile çarpımı sonucu elde edilen $(n\ x\ n)$ büyüklüğünde olan bu matris her bir varlık ismi ikilisinin arasındaki mesafeyi simgelemektedir. Bu matris içerisinde aynı zaman dilimi içerisinde sıkça bir arada geçen varlık isimlerinin değerleri oldukça yüksek olurken aksine aynı zaman dilimi içerisinde pek fazla bir arada bulunmayan varlık isimleri arasındaki değerler oldukça küçüktür. Elde edilen bu matris, varlık isimleri arasındaki uzaklık matrisi olarak kullanılmaktadır. Matris çarpımı sonucu elde edilen uzaklık matrisinin temel yapısı Şekil 4.5'de görülebilir.

(nxn)

	Varlık İsmi$_1$	Varlık İsmi$_2$	⋯	Varlık İsmi$_n$
Varlık İsmi$_1$	*ağırlık(Varlık İs$_1$,Varlık İs$_1$)*	*ağırlık(Varlık İs$_1$,Varlık İs$_2$)*		*ağırlık(Varlık İs,Varlık İs$_n$)*
Varlık İsmi$_2$	*ağırlık(Varlık İs$_2$,Varlık İs$_1$)*	*ağırlık(Varlık İs$_2$,Varlık İs$_2$)*		*ağırlık(Varlık İs$_2$,Varlık İs$_n$)*
Varlık İsmi$_3$	*ağırlık(Varlık İs$_3$,Varlık İs$_1$)*	*ağırlık(Varlık İs$_3$,Varlık İs$_2$)*		*ağırlık(Varlık İs$_3$,Varlık İs$_n$)*
⋮	⋮	⋮		⋮
Varlık İsmi$_n$	*ağırlık(Varlık İs$_n$,Varlık İs$_1$)*	*ağırlık(Varlık İs$_n$,Varlık İs$_2$)*		*ağırlık(Varlık İs$_n$,Varlık İs$_n$)*

Şekil 4.5 : Varlık isimleri için matris çarpımı sonucu oluşturulmuş olan komşuluk matrisinin yapısı.

4.1.2 Varlık ismi ağ analizleri

Varlık ismi ağları oluşturulduktan sonra, bu ağlar üzerinde oluşturulmuş oldukları zaman aralıkları doğrultusunda analizler yapılmıştır. Analizlerin yapılması için zaman aralıkları aylık olacak şekilde seçilmiştir. Bu durumda toplanmış olan 1 yıllık veri içerisinde tespit edilmiş olan 41.184 varlık ismi doğrultusunda 12 farklı zaman diliminde varlık ismi ağları oluşturulmuştur. Bu varlık ismi ağları incelendiğinde elde bulunan varlık isimlerinin tümünün ağlar oluşturulurken birer düğüm olarak kullanılması ve bu varlık isimleri kümesinde çok fazla uç değer bulunmasından ötürü

oldukça karmaşık yapılar oluştuğu gözlemlenmiştir. Bu sebeple daha anlamlı analizler yapabilmek adına elde edilen ağlar *"united nations"* (birleşmiş milletler) varlık ismi çevresinde filtrelenmiş ve bu varlık ismi ile ilişkisi bulunmayan varlık isimleri oluşturulan ağlardan silinmiştir. Bu filtreleme yapılırken *"birleşmiş milletler"* varlık isminin seçilmesinin sebebi, yapılmak istenen analizin ülkeler, liderleri ve ülkelerdeki toplumsal olaylar çerçevesinde yapılmak istemesinden kaynaklanmaktadır.

Filtreleme işlemi yapıldıktan sonra varlık ismi ağları içerisinde 6.314 adet varlık ismi kalmıştır. Her bir aya ait olan ve filtrelenmiş 6.314 adet varlık ismini içeren ağlar üzerinden ilk olarak, farklı zaman dilimlerine geçişler esnasında ortaya çıkan ve kaybolan varlık isimlerinin analizi yapılmıştır. Ağ analizlerinin yapılabilmesi için Pajek [119, 120] isimli, büyük ağların analizlerinin yapılabilmesi ve görselleştirilebilmesine olanak sağlayan bir uygulama kullanılmıştır. İlk yapılan analiz ile zaman içerisinde ülkeler arası ilişkilerde önem kazanan ve kaybeden aktörler gözlemlenmiştir. 2017 yılının ilk 6 aylık zaman dilimi içerisinde geçişler yapılırken ortaya çıkan ve kaybolan en önemli 20 varlık ismi Çizelge 4.1'de görülebilir.

Çizelge 4.1 : 2017 yılının ilk 6 ayı içerisinde ortaya çıkan ve kaybolan en önemli 20 varlık ismi.

Aylık Zaman Aralığı – 2017 Yılı İçin	**Ortaya Çıkan**	**Kaybolan**
Ocak – Şubat Ayları Arasında	Mosul (193), Kushner (96), Qaddafi (84), Khalil (78), Rashidiya (72), Churkin (47), Sharif (42), Anastasiades (40), Quabba (36), Libyan (35), Malaysia (30), Tripoli (30), Flynn (26), General McMaster (26), Al Bab (25), Unama (25), Congo (24), Khmer Rouge (24), Muslim Majority (24), Nicosia (24), Mohamed Sharif (24), Elliniko (24), Huong (24)	Jammeh (227), Gambia (152), Barrow (91), Senegal (49), Juba (48), Central African Republic (36), Kiir (36), Rafsanjani (35), Machar (33), Paris (30), Puebla (30), Donald J. Trump (28), Wadi Barada (28), Hacking Team (27), Zhou (24), Vincenzetti (21), Dakar (21), Gambian (20), West African (20), Yahya Jammeh (20)

Çizelge 4.1 : 2017 yılının ilk 6 ayı içerisinde ortaya çıkan ve kaybolan en önemli 20 varlık ismi. (devam)

Şubat – Mart Ayları Arasında	Haiti (256), Khalaf (111), South Sudanese (70), Schumer (59), South Africa (58), Darfur (54), Kislyak (54), India (46), Yambio (42), Dubai (41), Liberia (38), Heritage Foundation (36), Mali (36), Groves (36), Moon (34), Levinson (34), Delattre (34), S.G. (32), Nepalese (30), Gayflor (28)	Kushner (96), Qaddafi (84), Khalil (78), Rashidiya (72), Myanmar (62), Churkin (47), Sharif (42), Anastasiades (40), Quabba (36), Libyan (35), Greek (33), Gaza (32), Tripoli (30), Baghdad (27), General McMaster (26), Al Bab (25), Unama (25), Muslim Majority (24), Nicosia (24), Mohamed Sharif (24), Elliniko (24)
Mart – Nisan Ayları Arasında	Kosovo (157), Roma (98), Kony (56), Unmik (56), Aung San Suu Kyi (51), Resistance Army (49), Ugandan (35), Page (34), Zarrab (33), Khan Sheikhoun (32), Xi (29), Office of Legal Affairs (28), Lombardi (27), Chechnya (23), Italy (22), Giuliani (20), Podobnyy (20), Skype (19)	Khalaf (111), Apartheid (106), South Sudanese (70), Schumer (59), Netanyahu (57), Kislyak (54), Mosul (46), Flynn (44), Yambio (42), Dubai (41), Senate (39), Mali (36), Groves (36), Levinson (34), S.G. (32), Nigeria (31), Wang (31), Nepalese (30), Canada (29), Iraqi (29)
Nisan – Mayıs Ayları Arasında	Catalan (306), Kanku (136), Tedros (49), Duterte (47), California (42), Flynn (35), Sweden (35), Elizabeth (34), Monusco (34), Abe (30), German (29), Buddhist (29), Daba (28), Philippines (26), Ma Ba Tha (26), Nabarro (25), O'Brien (24), Callamard (24), Priesner (24), Green Party (23)	Unmik (56), Pinheiro (54), Cambodia (35), Council (34), Page (34), Zarrab (33), Xi (29), Idlib (28), Office of Legal Affairs (28), Lombardi (27), Cambodian (23), Italy (22), Khmer Rouge (22), Giuliani (20), Ban Ki-moon (20), Podobnyy (20), Skype (19), Erdogan (19), Spicer (19)

Çizelge 4.1 : 2017 yılının ilk 6 ayı içerisinde ortaya çıkan ve kaybolan en önemli 20 varlık ismi. (devam)

Mayıs – Haziran Arasında	Manning (354), Bloomberg (56), Srebrenica (50), Mosul (41), Italy (40), FARC (33), Human Rights Council (30), Bosnian Serb (30), Erdogan (28), Cappelaere (28), Mattis (26), U.S.D.B. (26), General dela Rosa (26), Colombia (24), Qatar (24), Tice (24), Mars (22), Macron (20), Shamdasani (20), Turkish (19), Figueres (19), Corallo (19)	Catalan (306), Ugandan (178), Kanku (136), Kony (69), Sharp (68), Central African Republic (61), Myanmar (57), Putin (49), Tedros (49), Uganda (40), Ethiopia (39), Sweden (35), Elizabeth (34), Monusco (34), Haiti (32), Abe (30), Buddhist (29), Dandong (28), Daba (28), Philippines (26), Ma Ba Tha (26)
Haziran – Temmuz Arasında	Othman (98), Hebron (91), Liu (90), Hammarskjold (70), Unesco (63), Myanmar (57), Jewish (56), Ndola (56), Bachelet (48), Abe (41), Kamwina Nsapu (40), Burundi (38), Cohen (35), Nganza (35), Inada (34), Nkurunziza (32), Kasai (30), Ismail (30), Shiite (28), Deutsche Bank (26)	Manning (354), Bloomberg (56), Flynn (52), Srebrenica (50), FARC (33), Al-Hussein (30), Human Rights Council (30), Bosnian Serb (30), Erdogan (28), Cappelaere (28), Mattis (26), U.S.D.B. (26), General dela Rosa (26), Tice (24), Mars (22), Shamdasani (20), Figueres (19), Corallo (19), Raby (18), Cardinal Pell (18)

Oluşturulan Çizelge 4.1, analizlerin yapıldığı tarih aralığı içerisinde gerçekleşen değişimlerle ilgili olarak fikir vermektedir. Örneğin, değerler incelendiğinde 2017 yılının Şubat ayı içerisinde 1 yıl süren bir krizin ardından Haiti'nin devlet başkanının seçilmesi sonucu *"Haiti"* varlık isminin önemi oldukça yükselmiş ve ilerleyen aylar içerisinde bu varlık isminden bahsedilmesi azalarak Haziran ayında tamamen kaybolmuştur. Bir başka örnek olarak ise Amerika Birleşik Devlet'lerinde, ülke sırlarını sızdırmak suçundan ötürü tutuklanmış olan, *"Chelsea Manning"* isimli

askerin affedilmesi ve hapishaneden çıkmasından ötürü Mayıs ayı içerisinde *"Manning"* varlık ismi oldukça sık bahsedilir olmuş ve bir sonraki ay tamamen etkisini yitirip kaybolmuştur.

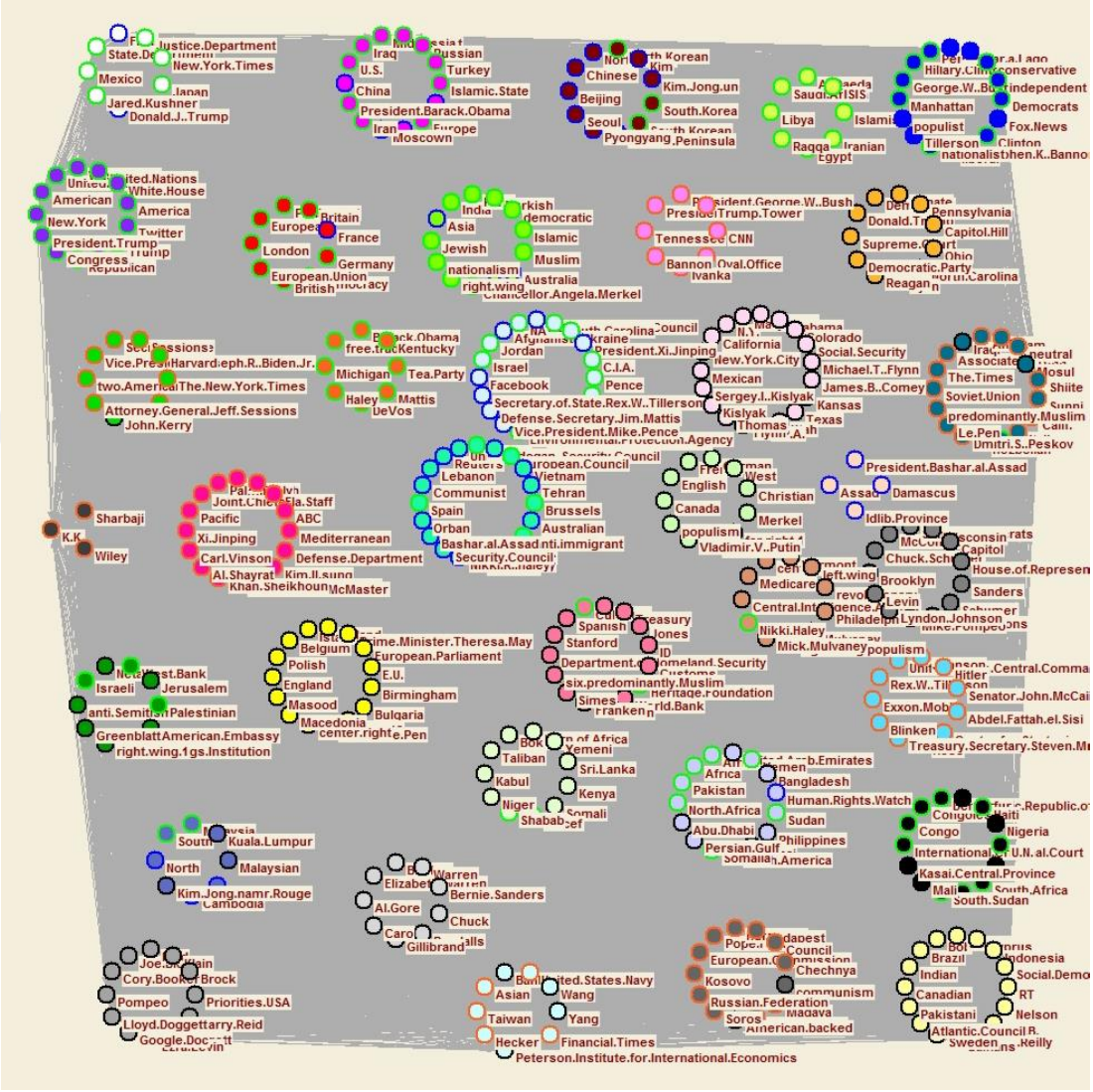
Bir başka analiz olarak, elimizde bulunan varlık isimleri ve bu varlık isimlerinin zaman içerisinde değişimlerinden yola çıkarak birbiri ile ilişkili varlık isimleri ve bu varlık isimlerine ait alt ağları tespit edilmiştir. Bunun yapılabilmesi için, analizlerde kullanılan varlık isimleri arasından en popüler olan 500 varlık ismi seçilmiştir ve bu varlık isimleri üzerinden işlemler yapılmıştır.

En popüler 500 varlık isminin tespitinden sonra, bu varlık isimlerinin aylık periyotlar içerisindeki değişimleri birer öznitelik olarak kullanılarak, hiyerarşik öbekleme yöntemleri [121] bu varlık isimlerinin bulunduğu veri kümesi üzerinde uygulanmış alt varlık ismi ağları oluşturulmuştur.

Hiyerarşik öbekleme yöntemleri, aşağıdan yukarıya (agglomerative) [122, 123] ve yukarıdan aşağıya (divisive) [124] olmak üzere 2 farklı grupta yer almaktadır. Aşağıdan yukarıya olan yöntemler, ilk olarak elde bulunan her bir elemanı tekil bir öbek olarak kabul etmekte ve birbirleri ile yakın olan öbekleri teker teker birleştirerek en uygun öbek sayısını bulmayı hedeflemektedir. Yukarıdan aşağıya olan hiyerarşik öbekleme yöntemler ise aşağıdan yukarıya olan yöntemlerin aksine ilk olarak tüm elemanların bulunduğu bir öbeği ele alarak sürece başlamakta ve her bir aşamada birbirinden maksimum uzaklığa sahip olan öbekleri birbirinden ayırarak en uygun öbek sayısını bulmayı hedeflemektedir.

En popüler 500 varlık isimleri için aylık olarak öbekleme işlemleri tamamlandıktan sonra elde edilen öbekler bir düzlemde, aynı öbek içerisinde bulunan varlık isimleri dairesel bir şekil oluşturacak şekilde görselleştirilmiştir. Alt ağların gösterildiği bu görsellerde yer alan düğümlerin sahip olduğu her bir renk farklı bir öbeği temsil etmektedir. Bu düğümlerin çeperlerinin sahip olduğu renkler ise bu düğümün durumunu ifade etmektedir. Aynı zaman yapılan görselleştirmede bir düğümün sahip olabileceği durumlar olan, varlık isminin ortaya çıkması turuncu çeper ile, ortadan kaybolması siyah çeper ile, bir önceki aya oranla öneminin artması yeşil çeper ile ve bir önceki aya oranla öneminin azalması mavi çeper ile ifade edilmiştir. Aynı alt ağda yer alan varlık isimleri bu renk değerlerinin takip edilmesi ile ayrıştırılabilir. Bu görselleştirilmenin yapılması için yine Pajek içerisinde bulunan özellikler kullanılmıştır.

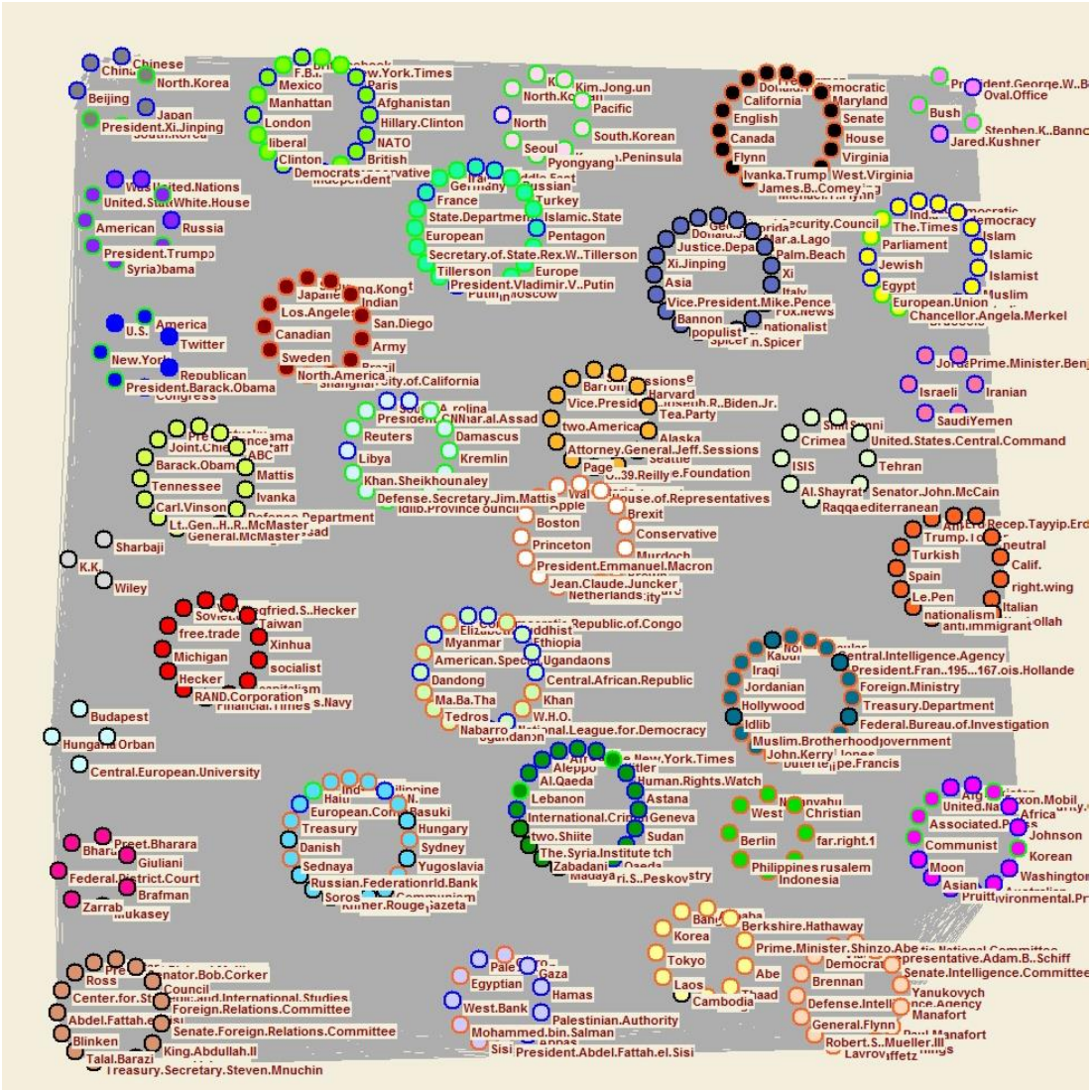
Bu alt ağlara bir örnek olarak, 2017 yılının Mart ayından Nisan ayına kadar geçen süre içerisinde, en popüler 500 varlık ismine ait olan ve alt ağların hiyerarşik öbekleme yöntemleri doğrultusunda tespit edilerek görselleştirildiği ağ Şekil 4.6'da görülebilir.



Şekil 4.6 : 2017 yılının Mart ayı ve Nisan ayı arasında, en popüler 500 varlık isminin oluşturdukları alt ağlar.

Şekil 4.6'da görülebileceği üzere birbirleri ile ilişkisi olan varlık isimleri aynı alt ağlarda yer alacak şekilde konumlanmıştır. Örneğin, *"Avrupa Birliği", "İngiltere", "Polonya", "Belçika", "Macaristan"* gibi varlık isimleri bir öbek altında yer alırken bir başka öbek içerisinde *"Amerika", "Beyaz Saray", "Başkan Trump", "Kongre", "Cumhuriyetçiler"* varlık isimlerinin olduğu görülebilir. Bu alt ağlar, zaman içerisinde ortaya çıkan veya kaybolan ilişkilerin görüntülenmesinde ve yorumlanmasında oldukça yardımcı olmaktadır.

Farklı bir örnek olarak, 2017 yılının Nisan ayından Mayıs ayına kadar geçen süre içerisinde, en popüler 500 varlık ismine ait olan ve alt ağların hiyerarşik öbekleme yöntemleri doğrultusunda tespit edilerek görselleştirildiği ağ Şekil 4.7'de görülebilir.



Şekil 4.7 : 2017 yılının Mart ayı ve Nisan ayı arasında, en popüler 500 varlık isminin oluşturdukları alt ağlar.

Bir sonraki aşamada yapılan analizlere ek olarak, zaman içerisinde en fazla önem kazanan ve kaybeden varlık isimlerinin yanı sıra, belirlenen zaman aralığı içerisinde ortaya çıkan ve kaybolan tüm varlık isimlerinin istatistiki bilgileri de incelenmiştir. Bu doğrultuda, toplanan ve üzerinde analiz yapılan veri içerisinde tespit edilmiş olan varlık isimlerinin 12 aylık farklı zaman dilimi içerisinde ortaya çıkan ve kaybolan tüm değerlerin ait analizler yapılmıştır. Bu analizler sonucunda elde edilen sonuçlar Çizelge 4.2'de 1'de görülebilir.

Çizelge 4.2 : 2017 yılı içerinde, aylık periyotta ortaya çıkan ve kaybolan varlık ismi sayıları.

Aylık Zaman Aralığı – 2017 Yılı İçin	**Ortaya Çıkan**	**Kaybolan**
Ocak – Şubat Ayları Arasında	793	580
Şubat – Mart Ayları Arasında	704	758
Mart – Nisan Ayları Arasında	522	721
Nisan – Mayıs Ayları Arasında	525	559
Mayıs – Haziran Arasında	548	532
Haziran – Temmuz Arasında	742	534
Temmuz – Ağustos Arasında	643	718
Ağustos – Eylül Arasında	849	594
Eylül – Ekim Arasında	613	859
Ekim – Kasım Arasında	604	640
Kasım – Aralık Arasında	712	619

Çizelge 4.2 incelendiğinde, haber verisinin toplandığı 2017 yılı içerisinde Mart ayı ile Haziran ayı arasında diğer aylara oranla çok daha az varlık isminin etki kazandığı ve Nisan ayı ile Temmuz ayı arasında ise diğer aylara oranla çok daha az varlık isminin etkisini yitirdiği gözlemlenmektedir. Bunun yanı sıra Ağustos ayı ile Eylül ayı arasındaki dönemin yeni olayların ve bu olaylara bağlı olarak yeni varlık isimlerinin ortaya çıkması ve Eylül ayı ile Ekim ayı arasındaki dönemin ise var olan olayların etkisini yitirmesi ve kaybolması noktasında, diğer aylardan farklı olarak, çok daha fazla öne çıktığı görülmektedir.

Yapılan analizlere ek, bir başka analiz olarak, tanımlanan zaman aralığı doğrultusunda ortaya çıkan veya kaybolan varlık ismi ikilileri önem sırasına göre tespit edilebilir. Varlık isimlerinin önem değerleri, birbiri ile beraber bulunma sayıları üzerinden hesaplanmaktadır. İki varlık isminin beraber bulundukları haber sayısı, belirlenen periyotlar doğrultusunda düzenli bir artış ve azalış izliyorsa, bu iki varlık ismi arasında bir ilişki tanımlanabilmektedir. Bu ilişkiler gözlemlenerek, varlık isimlerinin birbirleri ile etkileşimleri veya farklı varlık ismi etkileşimleri üzerindeki rolleri gözlemlenebilir. Bu doğrultuda, toplanan haber verisi için aylık zaman dilimleri doğrultusunda ortaya çıkan ve kaybolan en önemli varlık ismi ikilileri Çizelge 4.3'de görülebilir.

Çizelge 4.3 : 2017 yılı içerinde, aylık periyotta ortaya çıkan ve kaybolan en önemli varlık ismi ikilileri.

Aylık Zaman Aralığı – 2017 Yılı İçin	**Ortaya Çıkan**	**Kaybolan**
Ocak – Şubat Ayları Arasında	Trump – Flynn	Mexico – Nafta
Şubat – Mart Ayları Arasında	Trump – Ryan	Trump – Bannon
Mart – Nisan Ayları Arasında	Trump – CNN	Trump – Ryan
Nisan – Mayıs Ayları Arasında	Trump – Flynn	Fox News – O'reilly
Mayıs – Haziran Arasında	Army – Manning	Aleppo – Assad
Haziran – Temmuz Arasında	Trump – Putin	Army – Manning
Temmuz – Ağustos Arasında	Trump – Bannon	China – Liu
Ağustos – Eylül Arasında	Trump – Democrats	Iran – Mullah Mansour
Eylül – Ekim Arasında	Spain – Catalonia	Trump – Bannon
Ekim – Kasım Arasında	Lebanon – Hariri	Spain – Catalonia
Kasım – Aralık Arasında	Trump – Jerusalem	Senate – House

Çizelge 4.3'de listelenen sonuçlar, farklı zaman aralıkları için değerli olan varlık isimleri ilişkilerine ve olaylara dair yönlendirici bilgiler vermektedir. Örneğin, Çizelge 4.3 incelendiğinde 2017 yılının Eylül ayı ve Ekim ayı arasındaki süreçte İspanya'da Katalanlar'ın yaşadıkları bölgede yaşanan olaylardan ötürü *"Spain"* ve *"Catalonia"* varlık isimleri arasındaki ilişki oldukça güçlenmiş olduğu ve daha sonrasında kaybolduğu görülmektedir veya Haziran ayı ile Temmuz ayı arasında Amerika Birleşik Devletleri ile Rusya arasındaki sıkça gerçekleşen görüşmelerden ve gelişen ilişkilerden ötürü *"Trump"* ve *"Putin"* varlık isimleri arasındaki ilişkinin güçlendiği görülmektedir. Bir başka örnek olarak ise Temmuz ve Ağustos ayı arasında, Amerika Birleşik Devletleri Başkanı Donald Trump'ın en yakınında bulunan insanlardan biri olan ve baş stratejisti olarak görev yapan Steve Bannon'ı kovmasından ötürü, *"Trump"* ve *"Bannon"* varlık isimleri arasındaki bağ oldukça güçlenmiştir. Ancak ilerleyen dönemlerde bu olayın etkilerinin azalması ile birlikte Eylül ve Ekim ayları arasındaki dönemde bu iki varlık ismi arasında bağ ortadan kaybolmuştur.

Bir sonraki aşamada elde edilen bu varlık ismi ikililerinin, üzerinde analiz yapılan tüm zaman aralığındaki değişimleri de gözlemlenmiştir. 2017 yılına ait en önemli varlık ismi ikililerinin, tüm aylar doğrultusundaki değişimlerinin gösterildiği sonuçlar Çizelge 4.4'de görülebilir.

Çizelge 4.4 : 2017 yılına ait en önemli varlık ismi ikililerinin, tüm aylar doğrultusundaki değişimleri.

	Ocak	Şubat	Mart	Nisan	Mayıs	Haziran	Temmuz	Ağustos	Eylül	Ekim	Kasım	Aralık
Trump – Flynn		+3471	-2479	-992	+2185	-947	-1239					
Mexico – Nafta	+1262	-1262					+68	-68				
Trump – Ryan			+5375	-5375								
Trump – Bannon		+1600	-1600	+2791	-2791			+6127	-4744	-1383		
Trump – CNN				+10583	-10221	+246	-608	+337	+572	-706	+248	+250
Fox News – O'Reilly				+3792	-3792							
Army – Manning						+2832	-2832					
Aleppo – Assad	+33	+55	-50	+14	+1772	-1824		+14	-14		+16	+1
Trump – Putin	+1256	-419	-284	+479	+198	-1230	+4005	-3440	-451	+207	+1217	-1005
China – Liu						+2390	-2390					
Trump – Democrats	+4103	-339	+3336	+4022	+918	-1664	-159	-2173	+2803	-623	-208	+576
Iran – Mullah Mansour								+1368	-1368			
Spain – Catalonia										+2783	-2783	+620
Lebanon – Hariri						+13	-13				+1755	-1040
Trump – Jerusalem	+428	-243	-62	-123	+730	-730	+7	-7	+34	-8	-26	+3608
Senate – House					+978	+247	-1225			+495	+2979	-3474

Varlık ismi ağları üzerinde yapılan son analiz olarak, istenilen zaman aralıkları içerisinde oluşturulan geçici ağlar üzerinde yer alan en merkezi ve önemli varlık isimleri tespit edilmiş ve bu varlık isimleri gözlemlenmiştir. Burada en merkezi varlık isimlerinin tespit edilebilmesi adına, varlık isimlerinin içerisinde bulundukları ağların özvektör merkeziliği (eigenvector centrality) [125] değerli hesaplanmış ve bu değere göre varlık isimleri seçilmiştir. Analiz yapılan veri içerisinde, 12 aylık farklı zaman dilimindeki her ay içerisinde bulunan en önemli 25 adet varlık ismi Çizelge 4.5'de görülebilir.

Çizelge 4.5 : 2017 yılı içerisinde yer alan ve aylık olarak en merkezi konumda bulunan 25 varlık ismi.

Aylık Zaman Aralığı – 2017 Yılı İçin	**Varlık İsimleri**
Ocak – Şubat Ayları Arasında	Donald Trump, Flynn, Bannon, White House, Pence, Abe, Mar-a-Lago, Michael T. Flynn, Kushner Stephen K. Bannon, Russia, American, Japanese, National Security Council, Miller, Islam, One-China, McCain, General McMaster, President George W. Bush, Sweden, Ninth Circuit, Defense Secretary Jim Mattis, Taiwan, Palm Beach
Şubat – Mart Ayları Arasında	Ryan, Schumer, Democrats, California, Reid, Justice Department, Donald Trump, Democratic Party, Senate, liberal, Democratic, Warren, Capitol Hill, Levin, McConnell, Kislyak, DeVos, Republican , North Carolina, Klain, Nationalist, Greenblatt, Pennsylvania, Tea Party, EPA
Mart – Nisan Ayları Arasında	Senate Schumer, Ryan, Democrats, Supreme Court, Reid, California, Donald Trump, Republican, Democratic Party, Warren, Capitol Hill, McConnell, Levin, Flynn, Kislyak, Texas, North Carolina, Netanyahu, Klain, Merkel, Pennsylvania, Congress, Obama

Çizelge 4.5 : 2017 yılı içerisinde yer alan ve aylık olarak en merkezi konumda bulunan 25 varlık ismi. (devam)

Nisan – Mayıs Ayları Arasında	Senate House, Flynn, California, James B. Comey, Republican, Democrats, Netanyahu, Michael T. Flynn, Capitol Hill, Virginia, Jerusalem, Canada, German, Congress, Donald Trump, French, Ivanka Trump, Maryland, West Wing, Lavrov, Supreme Court, White House, Russia, Duterte, Senate Intelligence Committee
Mayıs – Haziran Arasında	Manning, Army, U.S.D.B., Iraq, Raby, Chelsea, Quantico, Chelsea Manning, American, WikiLeaks, Qatar, Kuwait, Julian Assange, Casey Brian, Afghanistan, Fort Leavenworth, Strangio, Fort Huachuca, Iraqi, Afghan, Starbucks, Iceland, Watkins, Susan
Haziran – Temmuz Arasında	Putin, Trump, Russia, Kremlin, Russian, Ukraine, President Vladimir V. Putin, Moscow, Hamburg, Cohen, Deutsche Bank, Trump Tower, Crimea, McCain, Spicer, Syria, Liu, China, Vrablic, Soviet Union, Donald J. Trump, Vladivostok, United States, Tillerson, Germany
Temmuz – Ağustos Arasında	Bannon, Charlottesville, Kelly, Senate, Cohn, White House, Mattis, Bedminster, Donald Trump, Phoenix, Pence, Ivanka Trump, South Carolina, Kentucky, Afghanistan, Republican Party, Oval Office, Clinton, North Korea, Gary D. Cohn, Nationalist, General McMaster, CNN
Ağustos – Eylül Arasında	Donald Trump, Democrats, Justice Department, Moore, Sessions, Republican, Texas, Mueller, White House, Senate, Hillary Clinton, Manafort, Stephen K. Bannon, Congress, President Trump, Price, Conservative, Tennessee, Washington, Miller, John F. Kelly, CIA, Los Angeles

Çizelge 4.5 : 2017 yılı içerisinde yer alan ve aylık olarak en merkezi konumda bulunan 25 varlık ismi. (devam)

Eylül – Ekim Arasında	Catalonia, Catalan, Spain, Puigdemont, Rajoy, Madrid, Spanish, Barcelona, European Union, Parliament, Democratic, Catalans, Belgium, Independent, Belgian, Europe, European, Nationalism, Right Wing, Britain, Nationalist, France, British, European Council
Ekim – Kasım Arasında	Hariri, Lebanon, Saudi Arabia, Hezbollah, Riyadh, Saudi, Iran, Lebanese, Prince Mohammed, Macron, France, Yemen, Mohammed Bin Salman, Kushner, Israel, Syria, United States, Trump, Nasrallah, Middle East, Jordan, King Salman, American, Paris, Israeli
Kasım – Aralık Arasında	Jerusalem, Moore, Trump, Flynn, Alabama, FBI, Jones, Israel, Palestinian, Republican, Senate, United States, Comey, Democrats, Conservative, American, Oval Office, Russia, President Trump, Jewish, Tel Aviv, East Jerusalem, Abbas, Kislyak, McFarland

Çizelge 4.5'de yer alan değerler incelendiğinde ve burada yer alan değerler Çizelge 4.1'de bulunan değerler ile karşılaştırıldığında, bir varlık isminin çok sayıda geçiyor olması ve merkezi olması arasındaki fark daha net bir biçimde anlaşılmaktadır. Örneğin, 2017 yılının Ocak ayı içerisinde *"Mosul"* veya *"Rashidiya"* gibi Orta Doğu ile ilgili varlık isimleri en sık geçen varlık isimleri olarak görülürken, bu ay içerisinde merkezi olarak yer alan varlık isimlerine bakıldığında *"Trump", "Flynn"* veya *"Bannon"* gibi Amerika Birleşik Devletleri içerisinde siyasi olarak öne çıkan isimlerin bu listede yer aldığı görülmektedir. Bir başka örnek olarak, 2017 yılı Mart ayı içerisinde Sırbistan ve Kosova arasında tekrar bir gerilim yaşanmasından ötürü en sık geçen varlık isimleri arasında *"Kosova"* veya *"UNMIK (United Nations Mission in Kosovo)"* olduğu görülmektedir. Ancak 2017 yılının Mart ayı içerisinde en merkezi olarak yer alan varlık isimlerine bakıldığında *"Senate Schumer"* veya *"Supreme Court"* gibi yine Amerika Birleşik Devletleri içerisinde yer alan kişilerin öne çıktığı

görülmektedir. Bunun temel sebebi, toplanan haberlerin Amerika Birleşik Devletleri kökenli bir gazeteden toplanmasından kaynaklanmaktadır. Bunun bir sonucu olarak, bazı varlık isimlerinin belirli dönemlerde gerçekleşen olaylardan ötürü popüler hale gelmekte ve sayıca çok fazla kez haberlerde yer almaktadır. Ancak Amerika Birleşik Devletleri içerisindeki figürler daha düzenli bir şekilde farklı zaman dilimlerindeki haberlerde yer almakta ve farklı varlık isimleri ile daha fazla ilişki içerisinde olmaktadır. Bu nedenle de daha düzenli olarak ve daha fazla bağlantısı olan bu varlık isimleri varlık ismi ağları içerisinde daha merkezi olarak konumlanmaktadır.

4.2 Konu Modelleri Temelli Analizler

Bir sonraki aşamada, New York Times verisi üzerinde konu modellemesi yapılmıştır. Konu modellemesi yapabilmek için de ilgili algoritmaları içeren konu modellemesi alt modülü veri üzerinde çalıştırılmış ve sonuçlar alınmıştır. Bu sonuçlar doğrultusunda, 2017 yılı içerisinde yapılmış olan haberlerin, konu başlıkları üzerinden dağılımları gözlemlenmiş ve analizler yapılmıştır.

4.2.1 Konu modellerinin oluşturulması

Konu modellemesi alt modülün çalışması sonucu, sistem tarafından 2 ayrı çıktı üretilmektedir. Bu çıktılardan ilki, tespit edilen konulara ait kelimelerin ve bu kelimelerin konu içerisinde sahip oldukları ağırlıkların bir listesidir. Örnek olarak, 3 konuya ait olan ve en yüksek ağırlık değerine sahip ilk 10 kelime Şekil 4.8'de görülebilir.

Örnek Konu$_1$	*Örnek Konu*$_2$	*Örnek Konu*$_3$
('0.009*"puerto" +	('0.116*"lebanon" +	('0.004*"iraqi" +
0.008*"catalan" +	0.090*"hezbollah" +	0.004*"myanmar" +
0.008*"rico" +	0.085*"hariri" +	0.004*"rohingya" +
0.006*"spain" +	0.070*"lebanes" +	0.003*"sergeant" +
0.006*"catalonia" +	0.030*"sinai" +	0.003*"kurdish" +
0.004*"rajoy" +	0.023*"egyptian" +	0.002*"jammeh" +
0.004*"kurdish" +	0.021*"tomb" +	0.002*"serbia" +
0.004*"separatist" +	0.018*"honduran" +	0.002*"kurd" +
0.003*"cuba" +	0.017*"sufi" +	0.002*"falcon" +
0.003*"madrid" + ...)	0.013*"sisi" + ...)	0.002*"gambia" + ...)

Şekil 4.8 : Konu modellemesine ait üç örnek çıktı.

Konu modellemesi sonucu elde edilen bu kelime grupları içerisindeki kelimeler 10 adet kelime ile sınırlı değildir. Konu modellemesinin yapıldığı veri kümesi içerisinde yer alan tüm kelimeler, tespit edilen tüm konular içerisinde belirli bir ağırlığa sahip olmaktadır. Ancak çoğu kelimenin konular için sahip oldukları ağırlıklar oldukça düşük olduğu için, konular içerisindeki yüksek ağırlıklı belli sayıdaki kelimeler dışında kalan kelimeler önemlerini yitirmektedir.

Sistem tarafından, konu modellemesi alt modülünün çalışması sonucu elde edilen bir diğer çıktı ise tespit edilen konuların farklı haberler üzerine dağılımının gösterildiği bir matrisdir. Bu matris içerisinde yer alan her bir değer, yüzdesel anlamda haberin hangi konuya ait olduğunu simgelemektedir. Oluşturulan bu matris, (*haber sayısı x tespit edilen konu*) büyüklüğündedir ve matris içerisinde yer alan her bir hücredeki değer 0 ile 1 arasında değişmektedir. Bu değerlerin 0 ile 1 arasında değişiyor olmasının temel sebebi, üretilen modelin olasılıksal bir model olmasından ve konuların haberler içerisindeki ulaşılma olasılıklarını simgelemesidir. Bu durumda, elde edilen çıktıdaki her bir haberi temsil eden satırlardaki hücre değerlerinin toplamı ise 1'e eşittir. Sistemden çıktı olarak alınan bu matrisin temel yapısı Şekil 4.9'da görülebilir.

	$Konu_1$	$Konu_2$	⋯	$Konu_t$
$Haber_1$	$skor(Haber_1, Konu_1)$	$skor(Haber_1, Konu_2)$		$skor(Haber_1, Konu_t)$
$Haber_2$	$skor(Haber_2, Konu_1)$	$skor(Haber_2, Konu_2)$		$skor(Haber_2, Konu_t)$
$Haber_3$	$skor(Haber_3, Konu_1)$	$skor(Haber_3, Konu_2)$		$skor(Haber_3, Konu_t)$
⋮	⋮	⋮		⋮
$Haber_n$	$skor(Haber_n, Konu_1)$	$skor(Haber_n, Konu_2)$		$skor(Haber_n, Konu_t)$

Şekil 4.9 : Haber – Konu skor matrisinin gösterildiği örnek yapı.

Konu modellemesi ile ilgili yapılan tüm analizler Şekil 4.8'de yer alan kelime listesi yapısı ve Şekil 4.9'da yer alan haberler ve tespit edilen konular arasındaki ilişkilerin gösterildiği skor matrisi üzerinden yapılmaktadır.

4.2.2 Konu modellemesi analizleri

Geliştirilen sistemde konu modellemesi için kullanılan LDA yöntemi, istatistiksel bir model üretmektedir ve bu istatistiksel modelin üretilebilmesi için kaç farklı konunun tespit edilmek istendiğinin belirtilmesi gerekmektedir. Bu sebeple konu modellemesi

alt modülünün aynı veri üzerinde çalışması için 25, 50 ve 100 konu olmak üzere 3 farklı konu sayısı değeri tanımlanmıştır ve konu modellemesi alt modülünden sonuçlar bu değerler doğrultusunda alınmıştır.

İlk olarak, tüm haber verisi üzerinde 25 konu için çalıştırılan konu modellemesine ait sonuçlar arasında en sık geçen 10 konu ve bu konulara ait en yüksek ağırlığa sahip 10 kelimenin listelendiği sonuçlar Çizelge 4.6'da görülebilir.

Çizelge 4.6 : 2017 yılında yayınlanan tüm haberler üzerinden oluşturulmuş 25 konu arasında en popüler olan 10 konu ve bu konu ile ilgili olan 10 kelime.

Önem Sırası	**İlgili Kelimeler**
1	trump, house, white, campaign, obama, administration, former, meet, officials, news
2	republican, republicans, bill, senate, house, democrats, trump, health, care, senator
3	percent, company, plan, market, bank, economic, money, financial, increase, business
4	court, case, judge, rule, federal, justice, legal, department, administration, order
5	group, right, women, political, white, black, country, protest, think, support
6	police, officer, kill, shoot, fire, attack, city, arrest, death, report
7	military, force, islamic, american, syria, attack, group, fight, government, iraq
8	family, live, children, father, mother, life, tell, home, want, write
9	north, korea, trump, nuclear, south, american, china, korean, missile, trade
10	home, water, storm, hurricane, residents, food, puerto, island, house, live

Çizelge 4.6'da görüldüğü üzere, 2017 yılı içerisindeki tüm haberler üzerinde 25 konu içerecek şekilde yapılan konu modellemesi doğrultusunda en popüler konuların Amerikan Birleşik Devletleri'nin iç siyaseti, Kuzey Kore ile olan nükleer silahlanma krizi ve Orta Doğu'da gerçekleşen saldırılar gibi oldukça geniş bir kapsama alanı olan ve yaygın konular olduğu görülmektedir.

Bir sonraki aşamada, tüm haber verisi üzerinde 50 konu için konu modellemesi yapılmıştır. Bu doğrultuda elde edilen konu modellerine ait sonuçlar arasında en sık

geçen 10 konu ve bu konulara ait en yüksek ağırlığa sahip 10 kelimenin listelendiği sonuçlar Çizelge 4.7'de görülebilir.

Çizelge 4.7 : 2017 yılında yayınlanan tüm haberler üzerinden oluşturulmuş 50 konu arasında en popüler olan 10 konu ve bu konu ile ilgili olan 10 kelime.

Önem Sırası	**İlgili Kelimeler**
1	trump, house, white, obama, administration, campaign, former, tell, meet, clinton
2	bill, health, care, plan, house, republicans, insurance, republican, senate, budget
3	republican, democrats, republicans, party, senate, senator, democratic, vote, moore, trump
4	company, percent, workers, market, job, economic, industry, economy, technology, growth
5	build, city, fire, home, residents, house, live, area, local, land
6	party, government, political, country, minister, power, opposition, prime, leader, former
7	food, around, cook, restaurant, find, small, use, shop, place, look
8	russian, russia, intelligence, investigation, information, officials, campaign, report, security, email
9	court, judge, justice, case, rule, legal, supreme, right, federal, decision
10	team, game, play, players, season, coach, league, sport, football, player

Çizelge 4.7'de görülebileceği üzere konu sayısı 50 konuya arttırıldığında yine benzer konular sonuçlar içerisinde bulunmaktadır. Ancak bir önceki sonuçlardan farklı olarak bu konuların daha detaylı olduğu görülmektedir. Örneğin, 25 konu ile oluşturulan konu modellemesi sonuçları içerisinde Amerika Birleşik Devletleri'nin iç siyaseti ile ilgili olan konuların yeni konu modellemesi içerisinde daha detaylı olarak yer aldığı görülmektedir. Bunun yanından en fazla geçen konuların farklı modeller için birbirlerine benzerlik gösterdikleri gözlemlenmektedir. Bunun temel sebebi, konu modellemesinin olasılıksal bir sonuç vermesinden ötürü yoğun geçen kelime gruplarının benzer şekilde dağılmasından kaynaklanmaktadır. Ancak bu noktada konu sayısının arttırılmaya devam edilmesi ile birlikte elde edilen konu modelleri içerisinde yer alan konuların daha özel konuları içermesi beklenmektedir. Örneğin, 50 konu ile

oluşturulan modelin en popüler sonucuna bakıldığında bir önceki konu modelinden farklı olarak spor ve restoranlar ile ilgili konuların yeni model içerisinde belirdiği görülmektedir.

Son olarak, tüm haber verisi üzerinde 100 konu için konu modellemesi yapılmıştır. Bu doğrultuda elde edilen konu modellerine ait sonuçlar arasında en sık geçen 10 konu ve bu konulara ait en yüksek ağırlığa sahip 10 kelimenin listelendiği sonuçlar Çizelge 4.8'de görülebilir.

Çizelge 4.8 : 2017 yılında yayınlanan tüm haberler üzerinden oluşturulmuş 100 konu arasında en popüler olan 10 konu ve bu konu ile ilgili olan 10 kelime.

Önem Sırası	**İlgili Kelimeler**
1	trump, obama, clinton, campaign, political, think, twitter, white, former, presidential
2	police, officer, shoot, kill, department, charge, case, report, crime, murder
3	party, political, germany, election, vote, national, government, country, elections, minister
4	military, force, american, afghanistan, officials, afghan, troop, fight, taliban, kill
5	trump, investigation, intelligence, campaign, comey, russian, russia, committee, officials, flynn
6	trump, american, meet, foreign, tillerson, leaders, countries, administration, world, visit
7	study, percent, find, university, research, professor, think, data, less, might
8	trump, white, house, administration, bannon, kushner, staff, adviser, secretary, officials
9	north, korea, south, korean, nuclear, missile, test, military, american, weapons
10	prison, charge, case, arrest, sentence, court, right, jail, release, trial

100 konu için oluşturulan konu modelleri içerisinde yer alan en popüler 10 konu Çizelge 4.8'de incelendiğinde, önceki konu modellerine göre çok daha detaylı konuların ortaya çıktığı görülmektedir. Belirli bir konu gurubunu tanımlayan genel kelimelerin yanında, olaylar doğrultusunda daha detaylı kelimeler de bu konular

içerisinde bulunmaktadır. Örneğin, Amerika Birleşik Devletleri'nin Afganistan'da yaptığı operasyonlar ile ilgili bir konunun tespit edildiği veya Almanya'dan gerçekleşecek seçimlere yönelik bir konunun tespit edildiği sonuçlar arasında görülmektedir.

Konu modelleri için yapılan bir başka analizde ise konuların farklı zaman dilimleri doğrultusunda nasıl dağıldığı ve değiştiği gözlemlenmiştir. Bunun için de aylık zaman dilimleri içerisinde, her bir ay için konu modelleri oluşturulmuştur ve daha sonra bu konu modelleri içerisinde bulunan konular arasındaki değişimler gözlemlenmiştir. Örneğin ilk olarak, 2017 yılının Temmuz ayı haberleri üzerinde 25 konu için konu modellemesi yapılmıştır. Bu doğrultuda elde edilen konu modellerine ait sonuçlar arasında en sık geçen 10 konu ve bu konulara ait en yüksek ağırlığa sahip 10 kelimenin listelendiği sonuçlar Çizelge 4.9'da görülebilir.

Çizelge 4.9 : 2017 yılının Temmuz ayında yayınlanan tüm haberler üzerinden oluşturulmuş 25 konu arasında en popüler olan 10 konu ve bu konu ile ilgili olan 10 kelime.

Önem Sırası	**İlgili Kelimeler**
1	police, officer, shoot, city, kill, tell, family, live, find, black
2	bill, republican, care, senate, republicans, health, house, vote, senator, repeal
3	federal, department, rule, trump, sessions, office, administration, case, court, legal
4	house, water, build, live, home, local, beach, food, place, town
5	company, china, chinese, market, business, billion, build, energy, deal, american
6	party, government, political, country, minister, force, opposition, court, protest, power
7	trump, meet, campaign, russian, email, information, investigation, lawyer, donald, russia
8	health, drug, program, percent, company, study, agency, plan, service, care
9	trump, white, house, scaramucci, chief, news, priebus, staff, administration, former
10	islamic, city, group, iraq, syria, force, fight, militants, fighters, organization

Çizelge 4.9'da bulunan konular incelendiğinde, Amerikan polisinin bir vatandaşı vurmasına yönelik bir konunun bulunduğu, Amerika Birleşik Devletleri ile Çin arasında gerçekleşecek olan antlaşmalara yönelik bir konunun olduğu veya Orta Doğu'da Irak ile Suriye arasında olan bir gerginliğe yönelik bir konunun olduğu görülmektedir.

Konuların değişimlerini gözlemleyebilmek adına 2017 yılının Ağustos ayı haberleri üzerinde 25 konu için konu modellemesi yapılmıştır ve konu modellemesi doğrultusunda elde edilen konular Temmuz ayı konuları ile karşılaştırılmıştır. Bu doğrultuda elde edilen konu modellerine ait sonuçlar arasında en sık geçen 10 konu ve bu konulara ait en yüksek ağırlığa sahip 10 kelimenin listelendiği sonuçlar Çizelge 4.10'da görülebilir.

Çizelge 4.10 : 2017 yılının Ağustos ayında yayınlanan tüm haberler üzerinden oluşturulmuş 25 konu arasında en popüler olan 10 konu ve bu konu ile ilgili olan 10 kelime.

Önem Sırası	**İlgili Kelimeler**
1	trump, white, house, bannon, charlottesville, news, political, violence, campaign, statement
2	court, case, federal, department, judge, justice, rule, legal, administration, right
3	police, officer, kill, city, arrest, protest, attack, violence, rally, shoot
4	trump, republican, party, republicans, senate, democrats, senator, bill, house, congress
5	north, korea, south, nuclear, trump, missile, korean, military, american, japan
6	china, chinese, trade, government, foreign, american, country, minister, party, countries
7	houston, storm, texas, flood, water, hurricane, harvey, city, home, emergency
8	attack, insurance, government, migrants, barcelona, authorities, asylum, report, terrorist, spain
9	city, vote, election, york, cuomo, million, system, plan, mayor, well
10	force, islamic, fighters, swift, syrian, syria, fight, nations, government, area

Çizelge 4.10'da yer alan, 2017 yılının Ağustos ayına ait konu modellemesi sonuçları incelendiğinde ve Temmuz ayı içerisinde yer alan popüler konular ile karşılaştırıldığında Amerika Birleşik Devletleri'ndeki polis şiddetine dair haberlerin veya Amerika Birleşik Devletleri'nin Çin ile gerçekleşecek olan ticari antlaşmalara yönelik konuların hala güncelliğini koruduğu ve popüler kaldıkları görülmektedir. Bunun yanında, Ağustos ayı içerisinde, Amerika Birleşik Devletleri ile Kuzey Kore arasında nükleer silahlanma temelli bir kriz yaşanmasından dolayı veya Charlottesville isimli şehirde yaşanan Afrikalı-Amerikalı vatandaşlara yönelik ırkçı bir saldırı yaşanmasından dolayı yeni konuların ortaya çıktığı da görülmektedir.

4.3 Düşünce Analizi Temelli Analizler

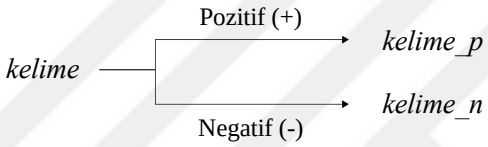
Düşünce analizi temelli analizler, toplanmış olan sosyal medya verisi üzerinde yapılmıştır. Bu analizlerin yapılabilmesi adına, ilk olarak geliştirilmiş olan sistemdeki veri toplama modülü aracılığıyla Twitter üzerinden veri toplama işi yapılmıştır. Toplanan veri, *"North Korea"* kelimeleri ile 2017 yılının Ağustos ve Kasım ayı arasında yapılmış olan paylaşımları içerecek şekilde toplanmıştır. *"North Korea"* kelime grubunun ve belirtilen tarih aralığının seçilmesinin temel sebebi, bu tarih aralığında Amerika Birleşik Devletleri ile Kuzey Kore arasında gerçekleşmiş olan nükleer silah krizinin analiz edilmek istenmesinden kaynaklanmaktadır.

Veri toplama işlemi sonucunda toplamda 2.854.333 adet tweet elde edilmiştir. İlk olarak bu veriler üzerinde, olaylarda yer alan baş aktörlerin tespit edilebilmesi adına varlık isimlerini tespit etmeye yönelik geliştirilmiş olan alt modül çalıştırılmıştır ve toplamda 19.120 adet varlık ismi tespit edilmiştir. Daha sonra bu varlık isimlerine yönelik Twitter üzerinden paylaşım yapan insanların, ilgili varlık isimleri ile ilgili düşünce analizleri yapılmıştır. Düşünce analizinin yapılabilmesi adına toplanmış olan tweet verisi, tespit edilmiş olan varlık isimleri ile birlikte geliştirilmiş olan düşünce analizi alt modülüne bir girdi olarak verilmiş ve bu alt modül çalıştırılmıştır. Bu alt modül ürettiği sonuçlarda, girdi olarak verilen ve tweet'ler içerisinde geçen varlık isimlerini bir diğer girdi olan tweet verisi doğrultusunda pozitif veya negatif olmak üzere işaretlemiştir.

4.3.1 Düşünce analizi ağlarının oluşturulması

Twitter üzerinden "North Korea" kelime grubu doğrultusunda toplanan verinin içerisinde bulunan varlık isimleri ve bu varlık isimleri ile birlikte yapılan düşünce

analizinin ardından, bu varlık isimlerinin hangileri pozitif iken hangilerinin negatif olduğu, varlık isimlerinin hangi günlerde, hangi sıklıkta geçtiğine ve birbirleri arasındaki diğer ilişkilere yönelik analizler yapılmıştır. Bu ağların oluşturulabilmesi için ilk olarak varlık isimleri, düşünce analizi sonuçları doğrultusunda gruplanmış ve etiketlenmiştir. Bu noktada yapılan varlık isimlerini etiketleme işlemi, düşünce analizi alt modülünden alınan sonuçlar doğrultusunda ilgili varlık isimlerinin sonuna bir ek eklenmesi şeklindedir. Örneğin, *"Trump"* olarak tespit edilmiş olan bir varlık ismi, düşünce analizi işlemi gerçekleştirildikten sonra pozitif durumlar için *"Trump_p"* ve negatif durumlar için *"Trump_n"* olacak şekilde sistem tarafından güncellenmektedir. Bu noktada düşünce analizi değerleri kullanılarak yapılan etiketleme işleminin temel çalışma yapısı Şekil 4.10'da görülebilir.



Şekil 4.10 : Düşünce analizi değerleri doğrultusunda varlık isimleri üzerinde yapılan etiketleme işlemi.

Tespit edilmiş olan varlık isimleri, düşünce analizi alt modülü tarafından güncellendikten sonra elde edilen bu güncellenmiş değerler ile yeni ağlar oluşturulmuştur. Bu ağlar da varlık ismi analizlerinde olduğu gibi belirli zaman aralıkları doğrultusunda varlık isimlerinin bir arada geçme sayıları ele alınarak oluşturulan ağlar doğrultusunda yapılmıştır.

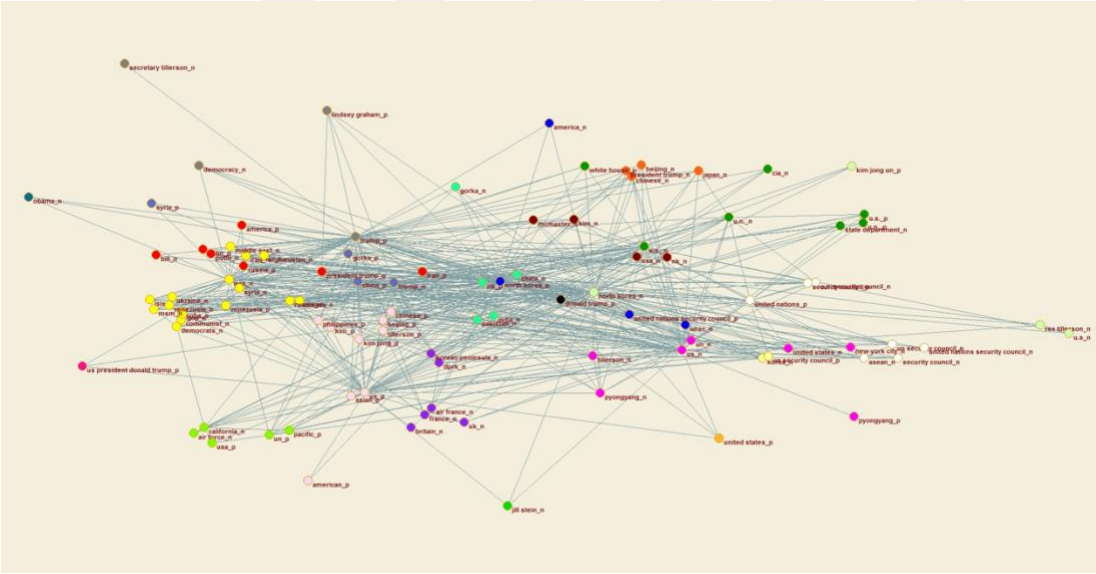
4.3.2 Düşünce analizi ağ analizleri

Belirli bir zaman aralığı içerisinde ve seçilmiş bir konu çerçevesinde Twitter üzerinden toplanmış olan veri içerisinde tespit edilen varlık ismi sayısı, haber verisinin kapsadığı zaman aralığı, konu çeşitliliği ve bu doğrultuda haber verileri içerisinde tespit edilmiş olan varlık ismi sayısı ile kıyaslandığında aradaki farkın oldukça az olduğu görülmektedir. Bunun sebebi sosyal medya üzerinden görüşlerini ifade eden insanların zaman zaman konu kapsamından uzaklaşarak farklı konulara yönelmesinden ve fikir ifade eden kişi sayısının oldukça fazla olmasından kaynaklanmaktadır. Bu noktada daha sağlıklı düşünce analizi ağları oluşturabilmek adına en fazla geçen 500 varlık ismi ile ağlar oluşturulmuştur.

En fazla geçen 500 varlık isminin yanı sıra, ağlar oluşturulurken kullanılan zaman aralığı değeri olarak da haftalık periyotlar belirlenmiştir. Zaman aralığının haftalık olarak belirlenmesindeki temel sebep, analizi yapılan ağların belirli bir konu çerçevesinde oluşturulmuş olması ve oluşturulan bu ağların arasındaki geçişlerin daha net bir şekilde görüntülenme istediğinden kaynaklanmıştır. Bu doğrultuda nükleer silah krizinin en yoğun olduğu 5 haftalık sürece ait ağlar ele alınmıştır. Değerlendirilen bu 5 hafta aşağıda belirtildiği gibidir;

- Hafta 1: 01 Ağustos – 06 Ağustos 2017
- Hafta 2: 07 Ağustos – 13 Ağustos 2017
- Hafta 3: 14 Ağustos – 20 Ağustos 2017
- Hafta 4: 21 Ağustos – 27 Ağustos 2017
- Hafta 5: 28 Ağustos – 03 Eylül 2017

Oluşturulmuş olan bu 5 farklı ağ içerisinde ek olarak Louvain topluluk tespiti algoritması [126] yardımı ile alt topluluklar bulunmuştur. Örneğin, Hafta 1'e ait olan ağ ve bu ağda yer alan alt topluluklar Şekil 4.11'de görülebilir.



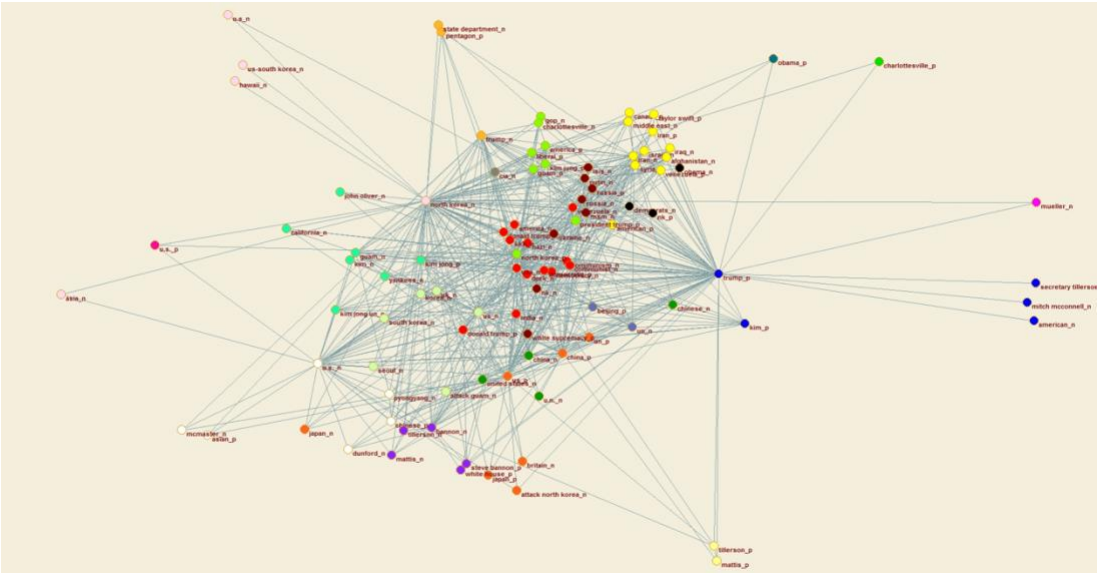
Şekil 4.11 : Hafta 1'e ait varlık isimlerini içeren düşünce analizi ağı ve bu ağda yer alan alt topluluklar.

Oluşturulan ağlar içerisinde yer alan her bir alt topluluk, görsellerde farklı bir renk ile ifade edilmiştir. Şekil 4.10'de yer alan ağ incelendiğinde, ağın oluşturulduğu zaman aralığı içerisinde yer alan varlık isimleri, bu varlık isimlerinin oluşturdukları alt ağlar ve bu varlık isimlerinin konuşulma şekilleri görülmektedir. Örneğin, *"white house",*

"state department", "u.s." ve *"cia"* gibi birbiri ile yakın ilişki içerisinde olması beklenen varlık isimleri aynı alt topluluk altında yer almaktadır. Bir başka alt topluluk incelendiğinde, *"asian", "beijing", "philippines"* ve *"kim jong"* gibi Asya kıtası içerisindeki varlık isimlerinin bir arada olduğu görülmektedir.

Bunun yanında, varlık isimleri arasındaki ilişkiler düşünce analizi sonuçları ile birlikte gözlemlendiğinde, varlık isimlerinin birbirleri ile olumlu veya olumsuz bir ilişki içinde olduğu yönünde bir analiz yapılmasına olanak sağlamaktadır. Örneğin, Hafta 1'e ait verinin toplandığı dönemde Amerika Birleşik Devletleri ile Kuzey Kore arasındaki nükleer silahlanma krizi yeni başladığı için, bu ülkeleri simgeleyen varlık isimleri arasındaki düşünce analizi sonuçlarının genelde negatif değerler üzerinden birbirlerine bağlandıkları görülmektedir. Bu ilişkilere örnek olarak, Amerika Birleşik Devletleri için verinin toplandığı dönemde güvenlikten sorumlu devlet bakanı olan Rex W. Tillerson'u simgeleyen *"tillerson"* varlık ismi ile Kuzey Kore'nin başkentini simgeleyen *"pyongyang"* varlık isminin arasındaki negatif ilişki gösterilebilir. Bir başka örnek olarak ise o dönem içerisinde Amerika Birleşik Devletleri ile Rusya arasında iyi bir ilişki olması sebebiyle *"trump", "putin", "russia"* ve *"america"* varlık isimleri arasında pozitif bir ilişki olması gösterilebilir.

Bir başka örnek ağ olarak ise Hafta 1'e ait olan ağ ve bu ağda yer alan alt topluluklar Şekil 4.12'de görülebilir.



Şekil 4.12 : Hafta 3'e ait varlık isimlerini içeren düşünce analizi ağı ve bu ağda yer alan alt topluluklar.

Şekil 4.12'de görülen ağ ve bu ağ içerisindeki alt topluluklara bakıldığında benzer varlık isimlerinin yine aynı alt topluluklar altında yer aldığı görülebilir. Örneğin, Amerika Birleşik Devletleri ile Kuzey Kore arasındaki nükleer silahlanma krizinin büyümesi ve bu sebeple Stephen K. Bannon, Rex W. Tillerson ve James N. Mattis gibi Amerika Birleşik Devletleri adına yetkili farklı aktörlerin devreye girmesi ile birlikte yeni alt topluluklar oluşmuştur. Bu topluluğun *"bannon", "tillerson", "mattis"* ve *"white house"* gibi varlık isimlerini içerdiği görülmektedir. Buna ek olarak İran, Suriye, Irak ve Afganistan gibi Orta Doğu ülkelerinin de bu konuşmalarda birer aktör olarak yer almaya başladığı ve bu ülkeleri simgeleyen *"iran", "iraq", "syria", "afghanistan"* ve *"middle east"* gibi varlık isimlerinin bir alt topluluk oluşturduğu görülebilir. Düşünce analizi odaklı bakıldığında ise örnek olarak Güney Kore ile Kim Jong-Un'u temsil eden varlık isimlerinin negatif bir ilişki içerisinde olduğu veya Trump ve Kuzey Kore arasındaki ilişkinin negatif bir ilişki olduğu görülmektedir.

Yapılan analizlere ek olarak, oluşturulan ağların yardımıyla haftalar arası geçişler sırasında etkisini yitiren varlık isimleri ve ortaya çıkan varlık isimleri ağların farkı alınarak tespit edilmiştir. Elde edilen bu değerler Çizelge 4.11'de görülebilir.

Çizelge 4.11 : Twitter verisinin toplandığı süre içerisinde haftalık olarak ortaya çıkan ve kaybolan varlık isimleri.

Haftalık Zaman Aralığı	**Ortaya Çıkan**	**Kaybolan**
1. Hafta – 2. Hafta Arasında	air_force_n, american_p, asean_n, beijing_n, bill_n, britain_n, cuba_n, france_n, iran_p, isis_n, peninsula_n, pakistan_n, philippines_p, putin_p, pyongyang_p, rex_tillerson_n, syria_p, un_security_council_p, un_security_council_n, donald_trump_p, jill stein_n, gorka_p, tillerson_n, lindsey_graham_p	american_n, bannon_n, north_korea_n, xi_jinping_p, congress_n, donald_trump_p, donald_trump_n, fbi_n, guam_p, guam_n, kim_jong-un_p, kim_jong-un_n, mattis_n, mccain_n, mueller_n, obama_p, pentagon_p, washington_p, mitch_mcconnell_n, moon_p, haley_p, mattis_n, jonathan_soble_p

Çizelge 4.11 : Twitter verisinin toplandığı süre içerisinde haftalık olarak ortaya çıkan ve kaybolan varlık isimleri. (devam)

2. Hafta – 3. Hafta Arasında	north korea_n, xi_jinping_p, congress_n, fbi_n, kim_jong-un_p, kim_jong-un_n, mccain_n, donald_trump_p, donald_trump_n, united_nations_p, usa_p, washington_p, moon_p, gorka_n, haley_p, mattis_n	american_p, asia_n, north_korea_n, britain_n, dunford_n, hawaii_n, iran_p, isis_n, israel_n, japan_p, mattis_p, seoul_n, south_korea_n, ukraine_n, yankees_n, tillerson_n, bannon_p, guam_n
3. Hafta – 4. Hafta Arasında	american_p, american_n, asia_n, asia_p, north_korea_n, bannon_n, california_n, dunford_n, guam_n, hawaii_n, iran_p, kim_jong-un_p, mattis_p, mattis_n, mueller_n, obama_p, obama_n, pentagon_p, state_department_n, ukraine_n, white_house_p, yankees_n, mcmaster_n, mcconnell_n, tillerson_n, bannon_p, guam_n	afghanistan_p, australia_n, egypt_n, harvey_n, kim_jong-un_p, london_n, los_angeles_n, mexico_n, navy_n, nigeria_n, north_korea_n, pakistan_p, pakistan_n, tillerson_n, russian_p, russian_n, japan_p, japan_n, seoul_p, syria_p, united_nations_p, us_navy_n, usa_p, gorka_p, washington_p, yemen_p, south_korea_p
4. Hafta – 5. Hafta Arasında	afghanistan_p, australia_n, britain_n, cia_n, egypt_n, guam_p, isis_n, kim_jong-un_p, london_n, los_angeles_n, mexico_n, middle_east_n, navy_n, nigeria_n, north_korea_n, pakistan_p, tillerson_n, russian_p, russian_n, japan_p, japan_n, seoul_p, seoul_n, syria_p, tillerson_p, venezuela_p, washington_p, yemen_p, gorka_n	california_n, cuba_n, france_n, germany_p, guam_n, harvey_p, iran_p, kim_jong-un_p, korea_p, mattis_p, mattis_n, mnuchin_n, obama_p, obama_n, pentagon_p, donald_trump_p, pyongyang_p, putin_p, south_korea_n, texas_p, un_security_council_p, un_security_council_n, united states_p, washington_n, white_house_p, moon_p, ted_cruz_p, mattis_p

Çizelge 4.11 incelendiğinde, haftalık periyotlar içerisinde yeni varlık isimlerinin ortaya çıktığı ve kaybolduğu görülmektedir. Bu varlık isimlerinin çoğu, toplanan verinin nükleer silahlanma krizi ile ilgi olmasından ötürü bu krizin tarafları olan Amerika Birleşik Devletleri, Asya ülkeleri ve Birleşmiş Milletler ile ilgili olan varlık isimlerinden oluşmaktadır. Buna ek olarak, bu çizelgede ortaya çıkan ve kaybolan varlık isimlerinin aynı zamanda düşünce analizi değerleri doğrultusunda pozitif ya da negatif olmak üzere hangi değerler doğrultusunda ortaya çıktığı ve kaybolduğu da belirtilmektedir.

5. TAHMİN MODELİ

Çalışmanın önceki aşamalarında, farklı alt modüller içeren, esnek ve ölçeklenebilir bir sistem oluşturulmuştur. Daha sonra bu sistem kullanılarak biri haber sitesi bir diğer sosyal medya platformu olmak üzere 2 farklı kaynaktan veri toplanmış ve farklı kaynaklardan toplanan bu veriler üzerinden de yine geliştirilen sistem içerisindeki doğal dil işleme alt modülleri yardımı ile varlık isimleri, konu modellemesi ve düşünce analizi temelli analizler yapılmıştır.

Bir sonraki aşamada, elde edilen analiz sonuçları derlenerek bir tahmin sistemi kurgulanmıştır. Kurgulanan bu tahmin sisteminin temel dayanak noktası birbiri ile benzer örüntülere sahip farklı olayların, farklı zamanlarda birbirine yakın sonuçlar doğurabileceği hipotezidir. Bunu gerçekleştirebilmek adına, farklı kaynaklardan toplanan veriler doğrultusunda dünyada yaşanan gelişmeleri ve insanların bu gelişmeler doğrultusunda sosyal medya platformları üzerinden verdikleri tepkileri analiz edip, elde edilen analiz sonuçları içerisindeki örüntüler bir öznitelik kümesi olacak şekilde kullanılan ve gelecekte yaşanacak benzer durumlarda da anlık olarak tahmin yapabilecek olan bir sistem tasarlanmıştır. Bir önceki aşamada yapılan analizler için sistem tarafından toplanan veri Amerika Birleşik Devletleri merkezli bir gazete olan New York Times üzerinden toplanmış olması sebebiyle, yayınlanan haberlerin çok büyük bir kısmı da Amerika Birleşik Devletleri'ni temel almakta ve ilgilendirmektedir. Bundan dolayı bir sonraki aşamada tahmin edilecek değer olarak New York Menkul Kıymetler Borsası'nda işlem gören dünyanın en büyük 30 şirketinin hisse senetlerinden oluşan Dow Jones endeksinin yönü seçilmiştir ve tahmin modeli bu değeri tahmin edecek şekilde kurgulanmıştır.

5.1 Tahmin Modeli İçin Oluşturulan Öznitelikler

Geliştirilecek bu tahmin modelini oluşturmak için, doğal dil işleme alt modüllerinden elde edilmiş sonuçları olan, varlık ismi dağılımları ve ağları, konu modelleri ve düşünce analizi ağları kullanılmıştır. Eğitilecek olan modele girdi olarak verilecek olan veri yapısı belirtilen bu alt modüllerden elde edilecek olan öznitelik kümelerini içeren

yapılar olacak şekilde kurgulanmıştır. Bu öznitelik kümelerinden her biri tahmin modeli içerisinde kanal olarak adlandırılmıştır. Örneğin, *n* adet varlık isminin kullanılacağı bir öznitelik için oluşturulan girdi yapısı *d* farklı tarihi ve bu tarihler içerisinde seçilen varlık isimlerinin dağılımını ifade edecek bir matris olacak şekilde tasarlanmıştır. *n* farklı varlık ismi ve *d* farklı tarih için oluşturulmuş olan örnek bir öznitelik yapısı Şekil 5.1'de görülebilir.

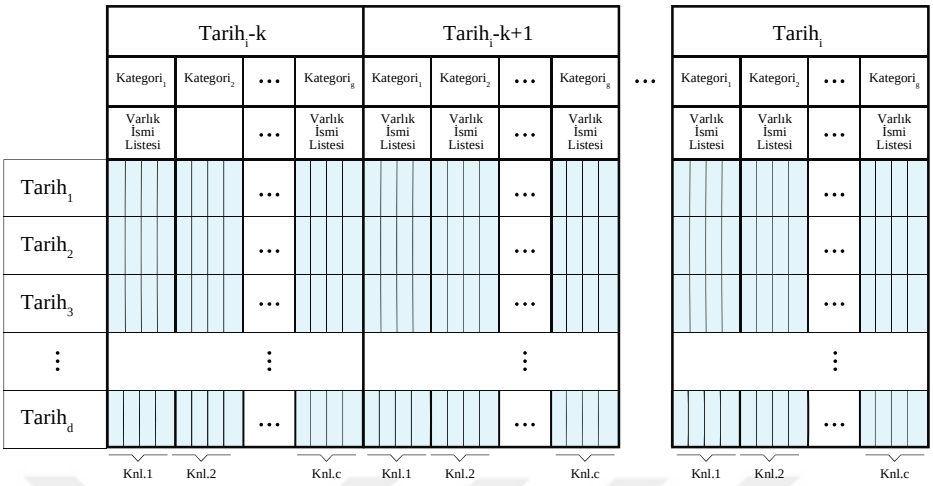
	Varlık İsmi$_1$	Varlık İsmi$_2$	⋯	Varlık İsmi$_n$
Tarih$_1$	*adet(Tarih$_1$, Varlık İsmi$_1$)*	*adet(Tarih$_1$, Varlık İsmi$_2$)*		*adet(Tarih$_1$, Varlık İsmi$_n$)*
Tarih$_2$	*adet(Tarih$_2$, Varlık İsmi$_1$)*	*adet(Tarih$_2$, Varlık İsmi$_2$)*		*adet(Tarih$_2$, Varlık İsmi$_n$)*
Tarih$_3$	*adet(Tarih$_3$, Varlık İsmi$_1$)*	*adet(Tarih$_3$, Varlık İsmi$_2$)*		*adet(Tarih$_3$, Varlık İsmi$_n$)*
⋮	⋮			⋮
Tarih$_d$	*adet(Tarih$_d$, Varlık İsmi$_1$)*	*adet(Tarih$_d$, Varlık İsmi$_2$)*		*adet(Tarih$_d$, Varlık İsmi$_n$)*

Şekil 5.1 : Tarih – Varlık İsmi değerlerinin listelendiği matrisi gösteren örnek yapı.

Varlık isimlerini ve bu varlık isimlerinin tarihlere dağılımını içeren bu özniteliklere ek olarak farklı öznitelikler de tanımlanmıştır. Konu modelleri doğrultusunda elde edilen konuların farklı tarihlere dağılımını gösteren matrisler ve benzer bir şekilde farklı tarihler içerisinde düşünce analizi yapılmış olan varlık isimlerinin bulunma sayılarını içeren matrisler de tahmin modeli içerisinde birer öznitelik olarak oluşturulmuş ve kullanılmıştır.

Tahmin modeli içerisinde kullanılacak olan temel öznitelikler belirlendikten sonra, bu özniteliklerin sınırlarına yönelik geliştirmeler de yapılmıştır. Örneğin tahmin modeli içerisinde kullanılacak olan varlık isimlerinin tarihler üzerindeki dağılımlarının gösterildiği öznitelik için, belirli bir tarihte geçen varlık isimlerinin bulundukları güne ek olarak bulunduğu günden sonra gelen farklı günlere yönelik etkilerinin de olacağı varsayılmıştır. Farklı bir örnek olarak ise, toplanan haberlerin her biri gazete içerisinde manşet, sürmanşet, ulusal haberler ve uluslararası haberler olmak üzere 4 farklı kategori içerisinde yer almaktadır. Farklı kategorilerde yer alan haberlerin içerisindeki varlık isimlerinin olaylara yönelik etkilerinin farklı olabileceği varsayılmıştır. Bu doğrultuda geçmişe yönelik *k* farklı tarih için 4 farklı haber kategorisinde yer alan *n*

farklı varlık ismi ve *d* farklı tarih için oluşturulmuş olan örnek bir öznitelik yapısı Şekil 5.2'de görülebilir.

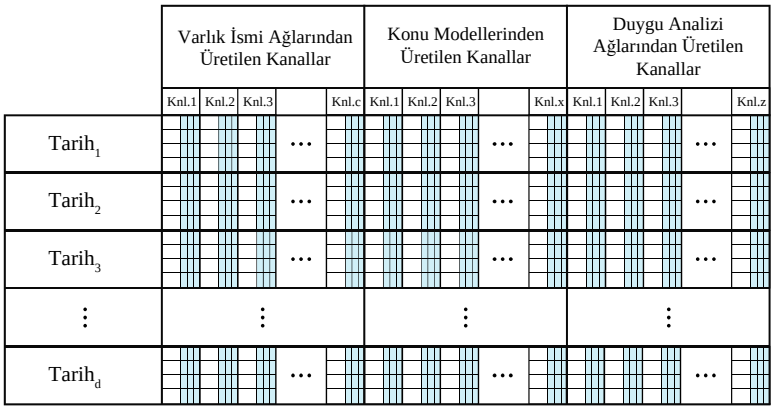


Şekil 5.2 : Tarih – Varlık İsmi – Haber Kategorisi değerlerinin listelendiği genişletilmiş matrisi gösteren örnek yapı.

Varlık isimleri ağları, konu modellemesi sonuçları ve düşünce analizi ağlarına yönelik sınırlarının genişletilmesi için yapılan çalışmalar sonucu kanal olarak adlandırılan birçok öznitelik varyasyonu elde edilmiştir. Elde edilen bu öznitelik varyasyonları aşağıdaki gibidir;

- Varlık ismi ağları için;
    - Farklı alt ağlar doğrultusunda varlık isimlerinin dağılımları
    - Varlık isimlerinin son *k* gün içerisindeki dağılımları
    - Farklı kategoriler altında varlık isimlerinin dağılımları
- Konu modelleri için;
    - Farklı konu sayısı ile oluşturulan konuların model içerisindeki dağılımları
    - Konuların son *k* gün içerisindeki konuların model içerisindeki dağılımları
- Düşünce analizi ağları için;
    - Farklı sayılardaki varlık isimlerinin üzerinde yapılmış olan düşünce analizi ve sonuçlarının dağılımları
    - Düşünce analizi sonrası varlık isimlerinin son *k* gün içerisindeki dağılımları

Tüm bu öznitelik varyasyonlarının oluşturulması sonucu, tahmin modeli için girdi olarak verilen matris içerisinde birçok kanal yer alabilmektedir. Tüm bu kanallar için, *d* farklı tarihe ait olan girdi matrisinin yapısı Şekil 5.3'de görülebilir.



Şekil 5.3 : Girdi olarak tasarlanan veri yapısı ve bu veri yapısının sahip olduğu özniteliklerin tümünün listelendiği matrisi gösteren örnek yapı.

5.2 Öznitelik Uzayının Küçültülmesi

Öznitelikler için yapılan tüm bu varyasyon eklemeleri ile beraber, tahmin modeli için oluşturulacak olan girdi matrisi oldukça detaylı bir yapı haline gelmiştir. Aynı zamanda farklı kanalların girdi matrisi içerisinde kullanım seçeneklerinden ötürü oluşturulacak olan girdinin oldukça esnek ve detaylı olmasının da önü açılmıştır. Ancak bu iyileştirmeler ile beraber tahmin modeli içerisinde kullanılacak öznitelik sayısı da oldukça büyük bir oranda artmıştır. Öznitelik sayısındaki bu artış, tahmin modelinin çalışması esnasında, algoritmanın performansını olumsuz olarak etkileyecektir. Bu sebepten ötürü, tahmin modelinin performansını arttırmaya yönelik bir yapı kurgulanmıştır.

Kurgulanan tahmin modeli yapısı, girdi matrisi içerisinde yer alan kanalların ve bu kanallar içerisinde yer alan öznitelik değerlerinin bir arada kullanılması ve bir tahminde bulunulması yerine, her bir kanalın kendi içerisinde değerlendirilmesinin yapılması ve daha sonra buradan elde edilen sonuçların tekrar birbirleri ile birlikte değerlendirilerek nihai sonucun elde edilmesi şeklinde kurgulanmıştır. Önerilen bu yöntem kademeli bir tahmin modeli olarak da adlandırılabilir.

Bu yapı, her bir kanal içerisinde öznitelikler için sayısal değerler tutmaktadır. Önerilen yöntemde bahsedildiği üzere, kanalların kendi içlerinde değerlendirilebilmesi için bu kanallar içerisinde yer alan özniteliklerin sahip oldukları sayısal değerler kullanılarak

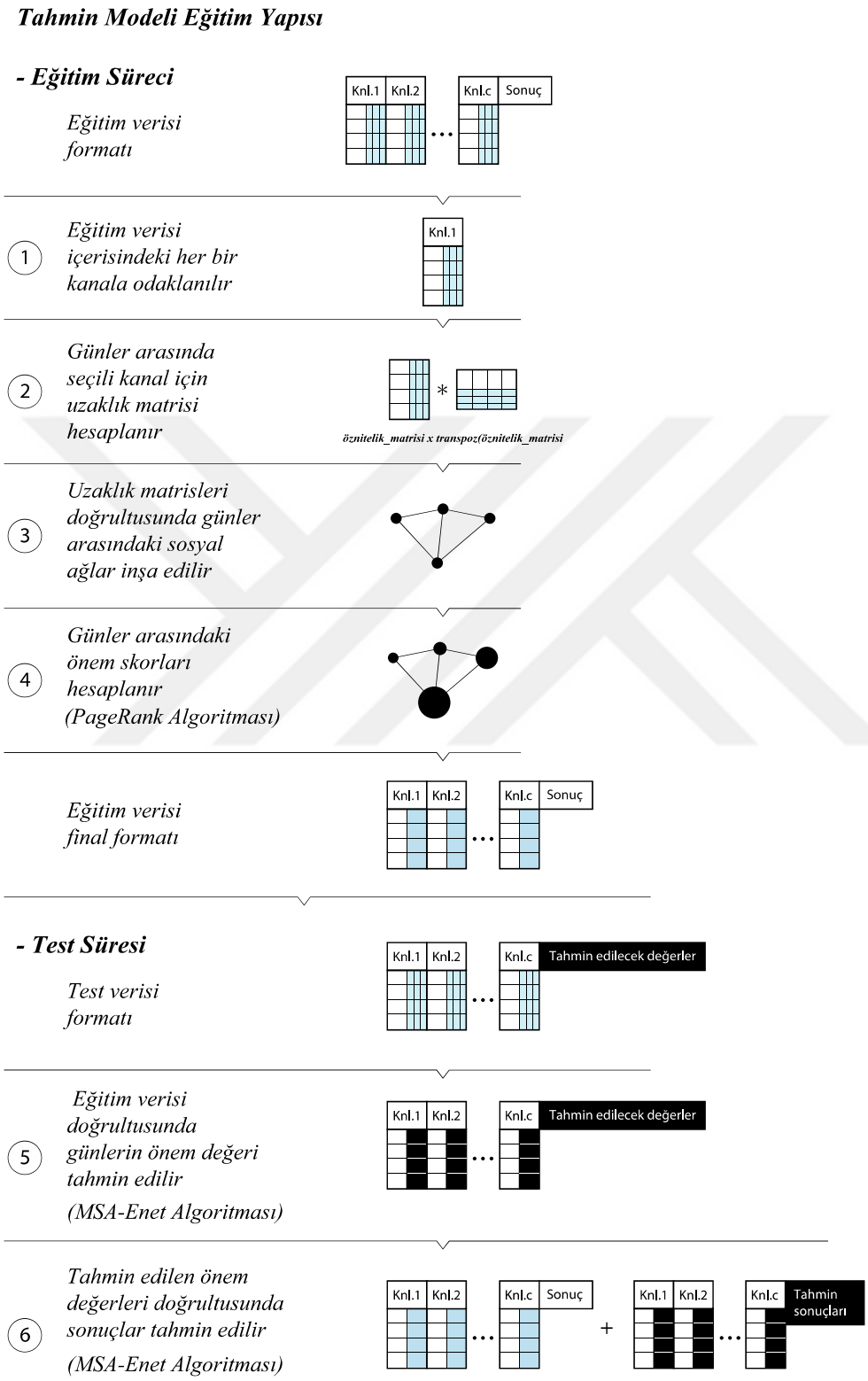
kanalların değerlendirilmesi ve bu doğrultuda bir maliyet fonksiyonu tanımlanması gerekmektedir. Bu fonksiyonu tanımlayabilmek adına, öneri sisteminin çalışabilmesi için girdi olarak verilen matriste yer alan her bir günün farklı kanallar için önem değerlerinin tespit edilmesi önerilmiştir. Günlerin önem değerlerini bulmak için, ilk olarak elde bulunan girdi matrisi kanallar doğrultusunda parçalara ayrılmıştır. Daha sonra her bir kanal içerisinde yer alan günler ve günlere ait olan öznitelik değerlerinden oluşan matris kendi transpozu ile çarpılarak bir matris oluşturulmuştur. Oluşan bu matris, düğümleri girdi matrisi içerisindeki günler olan ve günler arasında kenarların ise öznitelik değerleri doğrultusunda elde edilmiş olan ağırlıklar olduğu bir komşuluk matrisidir.

Her bir kanal için oluşturulan, günler arasındaki ilişkileri gösteren bu komşuluk matrisleri kullanılarak yönlü çizgeler oluşturulmuş ve oluşturulan bu çizgeler üzerinde PageRank algoritması [127] uygulanmıştır. PageRank algoritması, üzerinde çalıştığı çizge doğrultusunda ziyaret edilme olasılığı daha yüksek olan düğümleri tespit etmekte ve çizgenin içerisinde yer alan tüm düğümlere bu doğrultuda puan vermektedir. Günler için üretilmiş olan çizgeler üzerinde çalıştırıldığında ise önemi yüksek olan günlerin tespit edilmesini sağlamakta ve günlere puan atamaktadır. Farklı kanallar için oluşturulmuş olan gün çizgelerinin üzerinde çalıştırılmış olan PageRank algoritması ile günlerin, kanallar içerisindeki önem puanları belirlenmiştir. Aynı zamanda kanal içerisindeki her bir günün önem puanı, o güne ait farklı özniteliklerin değerleri doğrultusunda hesaplandığı için büyük bir öznitelik kümesinin tek bir puan değerine sıkıştırılmış halini simgelemektedir. Elde bulunan her bir kanal için bu sıkıştırma işlemi yapıldığında, oldukça büyük olan öznitelik uzayının çok daha küçük ve kanal sayısı kadar olan bir öznitelik kümesine dönüşmesi sağlanmıştır.

5.3 Sonuçların Üretilmesi

Tahmin modelinde kullanılacak olan girdi matrisi üzerinden öznitelik uzayının küçültülmesinin ardından, tahminler bu yapı üzerinden yapılacak şekilde kurgulanmıştır. Bunun için MSAEnet (çok adımlı uyarlanabilir elastik ağlar) (multi-step adaptive elastic-net) [128] isimli algoritma kullanılmıştır. Elastik ağlar regresyon temelli tahmin modelleridir. Aşırı öğrenmenin önüne geçmek adına geliştirilmiş olan, kement regresyonu (lasso regression) ve sırt regresyonu (ridge regression) birleşiminden ortaya çıkmış olan bir yöntemdir. MSAEnet yöntemi aynı zamanda

özniteliklerin kendi içlerinde önem sırasının belirlenmesine ve ağırlıklandırılmasına da olanak sağlamaktadır.

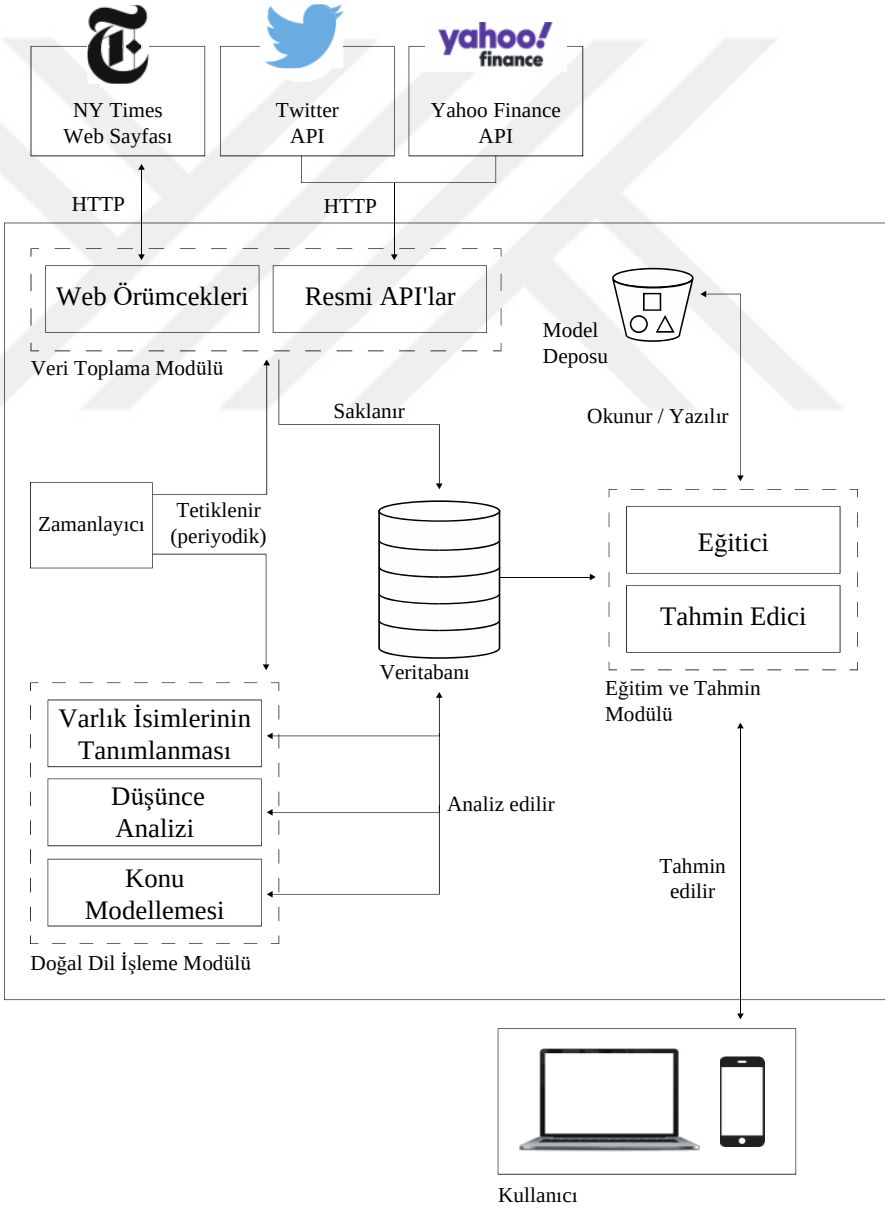


Şekil 5.4 : Tahmin modelinin yapısı ve çalışma mantığının tanımlandığı şema.

Kurgulanan model üzerinden tahmin yapılacağı noktada, ilk olarak test verisi olarak yeni gelen verinin sahip olduğu kanalların değerleri tahmin modeli üzerinden eğitim verisi doğrultusunda tahmin edilmekte ve sıkıştırılmış düzleme uyacak şekilde güncellenmektedir. Bunun üzerine, elde edilen öznitelikleri güncellenmiş olan test verisi üzerinden tekrar tahmin modeli çalıştırılarak nihai sonuçların tahmini yapılmaktadır. Kurgulanan tahmin modelinin yapısı ve tüm akışı Şekil 5.4'de görülebilir.

6. GENEL SİSTEM MİMARİSİ

Çalışma içerisinde geliştirilen tüm parçalar, birbiri ile uyum içerisinde çalışacak şekilde bir sistem haline getirilmiştir. Bu sistem içerisinde verilerin toplandığı gezicileri, verilerin saklandığı alanları, analiz ve tahmin işlemlerinin yapıldığı modülleri olmak üzere farklı parçaları içermektedir. Kurgulanan sistem ve mimarisi ana hatları ile Şekil 6.1'de görülebilir.



Şekil 6.1 : Kurgulanan sistemin ana hatlarıyla mimarisi.

Sistem içerisindeki veri toplama modülü, zamanlayıcı tarafından belirli aralıklar ile otomatik olarak veya sistemi kullanan kullanıcı tarafından tetiklenmektedir. Bu modül her tetiklendiğinde, belirtilen kaynaklardan istenen verileri toplamakta ve belirtilen alanlara kaydetmektedir. Bunun için geliştirilen ağ gezginleri veya veri toplanacak olan web sitelerinin resmi uygulama programlama arayüzleri kullanılmaktadır. Kurgulanan sistem içerisinde veri toplanan web sayfaları ve servisler, New York Times, Twitter ve Yahoo Finance olarak belirlenmiştir. Veri toplama modülünün hangi periyotlar ile tetikleneceği, hangi bilgilerin çekileceği ve toplanan verinin nereye kaydedileceği kullanıcı tarafından tanımlanabilmektedir.

Bir sonraki aşamada, doğal dil işleme modülü yine benzer bir şekilde zamanlayıcı tarafından belirli aralıklar ile otomatik olarak veya sistemi kullanan kullanıcı tarafından tetiklenebilmektedir. Bu modül daha önce tanımlanmış olan, varlık isimlerinin tanımlanması ve sınıflandırılması, düşünce analizi ve konu modellemesi alt modüllerini içermektedir. Bu alt modüller, veri toplama modülü tarafından toplanmış ve belirtilen alana kaydedilmiş olan verileri okumakla ve ilgili doğal dil işleme yöntemlerini bu metin verileri üzerinde çalıştırmakla sorumludur. Doğal dil işleme modülü içerisinde tanımlanan bu alt modüller çalıştırıldıktan sonra, çıktı olarak elde edilen sonuçlar ile verilerin saklandığı alan güncellenmektedir.

Veriler toplandıktan ve doğal dil işleme modülü ile belirlenen analiz işlemleri yapıldıktan sonra, eğitim ve tahmin modülü devreye girmektedir. Eğitim ve tahmin modülü, sistemin kullanıcısı tarafından istenildiğinde aktif hale getirilmektedir. Bu modül içerisinde iki farklı alt modül bulundurmaktadır. Bu modüllerden ilki elde bulunan veri doğrultusunda tahmin modellerinin eğitilmesinden sorumlu olan eğitici alt modülü, diğeri ise var olan eğitilmiş modelleri kullanarak yeni bir sonuç tahmini yapmakla yükümlü olan tahmin edici alt modülüdür. Eğitici alt modülü, belirtilen eğitim verisi ile üretilen tahmin modelinin eğitimini tamamladıktan sonra bu modeli daha sonra kullanılmak üzere sistem içerisinde bir alan kayıt etmektedir. Tahmin edici alt modülü ise eğitici tarafından kayıt edilen bu modelleri kullanarak tahminlerini gerçekleştirmektedir. Ancak önerilen yöntem içerisinde, sistem içerisindeki veri anlık olarak artış gösterdiği için günlük olarak yeni modellerin eğitimine

ihtiyaç duyulmuştur. Bu nedenle eğitilen tahmin modelleri ile birlikte bu modellerin temel künye bilgilerini tutacak olan bir dosya da modellerin yanında kayıt altına alınmıştır. Modellerin künye bilgilerini tutan bu dosyalar sayesinde tahmin yapmaktan sorumlu tahmin edici alt modülü de hangi modeli kullanması gerektiği bilgisine ulaşabilmektedir.

Kurgulanan sistemin bir diğer uç noktasında ise tekil kullanıcılar bulunmaktadır. Tekil kullanıcılar farklı cihazlar ve arayüzler aracılığı ile tahmin edici alt modülüne erişebilmekte ve istedikleri tahmin bilgilerini sistem üzerinden üretebilmektedir. Sistem tarafından üretilen bu tahmin sonuçları, son kullanıcının kendisine yine aynı cihazlar üzerinden iletilmektedir.

7. YAPILAN DENEYLER

Yapılan deneyler, New York Menkul Kıymetler Borsası'nda işlem gören dünyanın en büyük 30 şirketinin hisse senetlerinden oluşan Dow Jones endeksinin 2017 senesi içerisindeki hareket yönünün tahmin edilmesi doğrultusunda yapılmıştır. Tahminlerin yapılabilmesi için tahmin modeline girdi olarak verilecek olan matrisler, New York Times ve Twitter olmak üzere 2 farklı kaynaktan toplanan veriler ve bu verilerin geliştirilen doğal dil işleme araçları ile analizlerinden elde edilen sonuçları kullanılarak oluşturulmuştur.

Girdi matrislerinin oluşturulması için kullanılan veriler, 01 Ocak 2017 ile 31 Aralık 2017 arasında New York Times'da yayınlanmış olan 12.560 makale ve 01 Ağustos 2017 ile 30 Kasım 2017 arasında *"North Korea"* kelime grubu ile yapılmış olan sorgular sonucunda Twitter üzerinden toplanmış olan 2.854.333 adet tweet'dir. Bu veriler toplandıktan sonra geliştirilen sistem içerisindeki doğal dil işleme alt modülleri kullanılarak varlık ismi ağları, konu modellemesi sonuçları ve düşünce analizi ağları oluşturulmuş ve girdi matrisi içerisinde yer alan kanallar bu analiz sonuçları ile oluşturulmuştur.

Yapılan deneyler içerisinde 2 farklı ana deney ve bu deneylerin içerisinde farklı alt deneyler yer almaktadır. Bu ana deneylerden ilki; 01 Ocak 2017 ile 30 Mart 2017 arasında yayınlamış olan New York Times makalelerini ve bu makalelerden elde edilmiş varlık ismi ağları ve konu modellemesi sonuçları ile oluşturulmuş farklı konfigürasyonlara sahip kanalları eğitim verisi olarak kullanmıştır. Elde edilen farklı konfigürasyonlar ve farklı öznitelik varyasyonları ile eğitilen bu tahmin sisteminin, bir sonraki günkü Dow Jones endeksinde gerçekleşecek hareketin yönünü tahmin etmesi istenmiştir.

Bir sonraki ana deneyde ise, Twitter üzerinden toplanan verinin ve bu veri üzerinden üretilmiş olan düşünce analizi ağlarının da farklı konfigürasyonlar ile girdi matrisine birer kanal olarak eklenerek, girdi matrisinin genişletilmesi ve genişletilmiş olan bu girdi matrisi doğrultusunda modeller eğitilmiş ve Dow Jones endeksinin yönüne ilişkin tahminlerin yapılmıştır. Bu deney ise Twitter verisi daha dar bir zaman aralığını

kapsadığı için 01 Ağustos 2017 ile 31 Ağustos 2017 arasındaki 1 aylık süre eğitim verisi olarak kullanılacak, sonrasındaki 3 aylık veri ise test verisi olarak kullanılacak şekilde kurgulanmıştır.

Yapılan tüm deneylerde, tahmin edilmeye çalışılan değerlerin bulunduğu Dow Jones endeksi hafta sonları ve resmi tatillerde aktif olmadığı için yapılan tahminler sadece endeksin aktif olduğu günleri kapsamaktadır. Deneyler sırasında, endeks yönünün tahmin edileceği her günden önce, sistem yeni gelen veriler ile tekrar eğitilip, modelin kendi içerisinde sürekli gelişmeye devam etmesi ve güncel kalmasına da olanak sağlanmıştır.

Yapılan ana deneyler içerisinde, aynı zamanda tahminlerin yapılması için kullanılan girdi matrisinde yer alan kanalların farklı konfigürasyonlar ve farklı öznitelik varyasyonları ile kullanılması ile alt deney kümeleri de oluşturulmuştur. Alt deney kümeleri içerisinde farklı yapılardan elde edilen bu sonuçlar da performans ve isabet açısından birbirleri ile kıyaslanmıştır.

7.1 Değerlendirme Kriterleri

Yapılan tüm ana ve alt deneylerde, eğitilen tahmin modelleri tarafından üretilen tahmin sonuçlarının birbirleri ile kıyaslanabilmesi için 3 farklı değerlendirme metriği önerilmiştir. Bu metriklerden ilki yapılan tahminler doğrultusunda elde edilen sonuçların, gerçek değerlere oranla yüzdesel olarak doğruluk oranıdır. Kullanılan bir diğer değerlendirme metriği, tahmin sonuçlarını ve gerçek sonuçları kullanılarak elde edilen hataların ortalama karekökü (root mean square error) (RMSE) [129] değeridir. Bu metrik ile tahmin sonucu elde edilen değerin arasındaki fark bulunarak, tahmin edilen değerin gerçek değerden ne kadar saptığı gözlemlenmektedir. Son olarak kullanılan değerlendirme metriği ise gerçek sonuçlar ile tahmin edilen sonuçlar arasındaki korelasyon değeridir. Korelasyon değeri -1 ve 1 arasında değişen sayısal bir değerdir. Korelasyon değeri olarak elde edilen negatif değerler, karşılaştırılan iki değer listesi arasında negatif bir ilişkinin olduğunu gösterirken, pozitif değerler bu durumun tam tersi olarak karşılaştırılan iki değer listesi arasında pozitif bir ilişkinin olduğunu göstermektedir. Korelasyon sonucunda negatif bir ilişkinin olması, aralarında korelasyon bulunan bir liste içerisinde yer alan bir değer artış gösterirken diğerinin azalış göstermesi, pozitif bir ilişkinin olması ise bu liste içerisinde yer alan değerlerin birlikte artış veya azalış göstermesi şeklinde tanımlanabilir.

Dow Jones endeksindeki değişim tahminleri sonrası başarılı kabul edilebilecek sonuçlar için tanımlanan bu 3 değerlendirme metriğinin aşağıdaki belirtildiği gibi olması beklenmektedir.

- Başarım oranı metriği için yüzdesel olarak yüksek bir değere sahip olması.
- Hataların ortalama karekökü metriği için mümkün oldukça düşük bir değere sahip olması.
- Tahmin sonuçları ile endeksteki gerçek değişim değerleri arasındaki korelasyonun pozitif yönlü olması.

7.2 Deney Sonuçları

7.2.1 Deney I: 01 Nisan 2017 – 31 Aralık 2017 tarihleri arasında Dow Jones endeksi değişimlerinin tahmin edilmesi

İlk ana deney grubu sadece haber verisi üzerinden, doğal dil işleme aracının alt modülleri ile yapılan analizler sonucu, Dow Jones endeksindeki değişimi tahmin etmeyi hedeflemektedir. Bu doğrultuda yapılan analizler ve oluşturulan öznitelikler doğrultusunda farklı alt deneyler tanımlanmış, bu doğrultuda tahmin modelleri eğitilmiş ve endeksin yönüne ilişkin tahminler yapılmıştır.

7.2.1.1 Varlık ismi ağlarından üretilen kanallar ile yapılan tahminler

İlk olarak yapılan deneylerde, tahmin modelleri, varlık ismi ağlarından oluşturulan kanallar ile eğitilmiş ve tahmin sonuçları üretilmiştir. Bu noktada tahmin modeli, varlık ismi ağlarından oluşturulmuş 3 farklı eğitim verisi varyasyonu ile eğitilmiştir. Girdi matrisinin oluşturulması için kullanın kanal varyasyonlarının tamamı New York Times makalelerinden üretilmiş olan varlık ismi ağlarındaki değerler kullanılarak oluşturulmuştur ve bu kanallar sırası ile;

- Varlık ismi ağlarında tespit edilmiş olan tüm 41.184 adet varlık isminin kullanıldığı bir kanal,
- Varlık ismi ağlarındaki en sık geçen 700 adet varlık isminin kullanıldığı bir kanal,
- Varlık ismi ağlarındaki *"birleşmiş milletler"* düğümü çevresinde yer alan varlık isimlerinden oluşturulmuş olan ve 6.314 adet varlık ismini içeren bir alt ağın kullanıldığı bir kanaldır.

Bu 3 farklı tahmin modelinin eğitilmesi ve sonrasında bu eğitilmiş olan bu modeller üzerinden yapılan tahminler doğrultusunda elde edilmiş sonuçlar Çizelge 7.1'de görülebilir.

Çizelge 7.1 : Farklı büyüklükteki varlık ismi ağlarının, bulundukları gün verisini kullanarak eğitilmiş olan tahmin modelleri ile yapılmış olan tahmin sonuçları.

	Doğru Tahmin Sayısı	**Yanlış Tahmin Sayısı**	**Başarı Oranı**	**RMSE**	**Korelasyon Değeri**
Tüm Varlık İsimlerini İçeren Tahmin Modeli	108	81	57,14%	113,57	0,0710
En Sık Geçen 700 Varlık İsmini İçeren Tahmin Modeli	116	73	61,38%	138,81	0,0487
"Birleşmiş Milletler" Alt Ağının Varlık İsimlerini İçeren Tahmin Modeli	121	68	64,02%	70,72	0,1026

Çizelge 7.1'de yer alan sonuçlar incelendiğinde, başarımı en düşük olan sonuçların tüm varlık isimleri kullanılarak oluşturulan tahmin modeli tarafından üretildiği görülmektedir. Bunun temel sebebi, tüm varlık isimleri öznitelik kümesi içinde kullanıldığında bazı varlık isimlerinin oldukça uç değerler olmasından ve bu sebeple oluşturulan girdi verisinin oldukça seyrek bir hale gelmesindendir. Girdi verisi içerisinde kullanılan özniteliklerin seyreklik değerleri belirli bir eşiğin üzerine çıktığı takdirde tahmin modelinin ürettiği sonuçları olumsuz yönde etkilemekte ve başarımını düşürmektedir.

Uç değer olarak adlandırılabilecek varlık isimleri ve bu varlık isimlerinin etkilerini ortadan kaldırabilmek adına, en sık geçen varlık isimleri ile farklı bir tahmin modeli oluşturulmuştur. Varlık ismi ağları içerisinde en sık geçen 700 varlık ismi kullanılarak oluşturulan bu kanal ile eğitilen tahmin modelinin sonuçlarına bakıldığında, başarı oranının, tüm varlık isimlerinin kullanıldığı tahmin modeline oranla arttığı gözlenmektedir. Ancak diğer 2 değerlendirme metriği olan RMSE ve korelasyon

değerleri incelendiğinde, bu değerlerin başarı oranı değerinin aksine olumsuz yönde etkilendiği görülmektedir. Bunun sebebi, en sık geçen varlık isimlerinin seçilmesi ile daha değerli bir öznitelik kümesi oluşturulsa da bu sık geçen varlık isimlerinin günler içerisindeki dağılımları, bu varlık isimlerinin popülerliği ve haberlerde oldukça sık geçmesi sebebiyle birbirine oldukça benzemektedir. Dolayısıyla, sadece sık geçen varlık isimleri seçildiğinde günleri içerisindeki önemli olabilecek diğer detaylar tamamen yok olmaktadır.

Hem uç değerlerin etkilerini ortadan kaldırmak hem de az sayıda bulunan ama sonuçlara etkisi olan ve de uç değer olarak nitelendirilemeyecek varlık isimlerinin filtrelenmesinin önüne geçmek adına mevcut olan varlık ismi ağının içerisinde yer alan bir alt ağ, kullanılan varlık ismi ağına bir alternatif olarak oluşturulmuştur. Oluşturulan bu alt varlık ismi ağı, ekonomiye büyük etkisi olan varlık isimlerini içerme ihtimalinin yüksek olması adına *"birleşmiş milletler"* düğümünü merkez alacak ve bu düğümün etrafında şekillenecek bir şekilde oluşturulmuştur. Böylelikle bu alt ağda yer alan varlık isimleri büyük oranda ülkeler ve bu ülkeler içerisinde yer alan önemli insanlar ve kurumlardan oluşmuştur.

Seçilmiş *"birleşmiş milletler"* düğümünü merkez alan bu alt varlık ismi ağı kullanarak oluşturulan yeni kanal ile eğitilmiş olan tahmin modelinin sonuçlarını incelediğimizde, diğer sonuçları incelenen diğer 2 modele oranla tüm değerlendirme metrikleri doğrultusunda çok daha iyi tahminler yapıldığı görülmektedir. Bu modelden üretilen sonuçların diğer 2 modele göre başarı oranı daha yüksek, hataların ortalama karekökü değeri daha düşük ve tahmin edilen değerlerin gerçek değerler ile arasındaki korelasyon değeri pozitiftir.

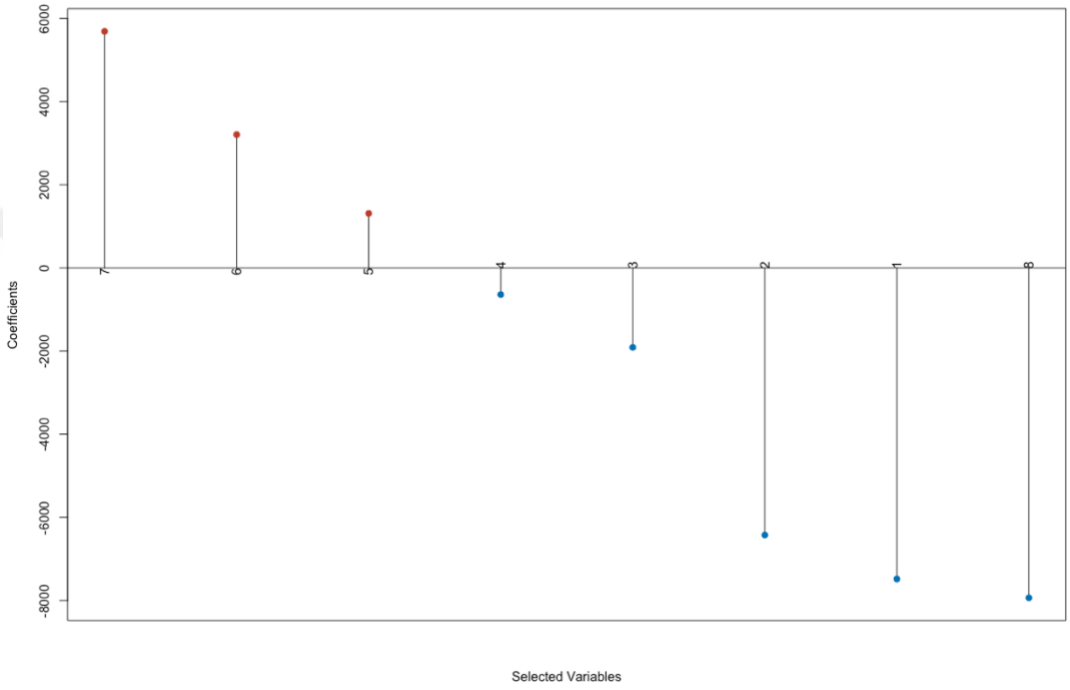
Deneyin bir sonraki aşamasında, tahmin modeli genişletilmiş zaman aralıkları içerisinde yer alan kanallar ile eğitilmiş ve tahmin sonuçları bu model kullanılarak elde edilmiştir. Bu aşamada zaman aralığının genişliği olarak bir haftalık bir süreç kullanılmıştır. Genişletilmiş zaman aralıkları doğrultusunda eğitilen ilk grupta bulunan tahmin modelleri, varlık ismi ağları için oluşturulmuş 3 farklı kanalın 7 gün öncesine kadar olan değerlerini içerecek şekilde eğitilmiştir. Bu sebeple tahmin modelinin üretilmesi esnasında kullanılan girdi verisinde, bir önceki modeller için kullanılan girdi verisine kıyasla ek 7 günü temsil edecek 7 ek kanal daha bulunmaktadır. Bu doğrultuda eğitilen 3 farklı tahmin modelinin sonuçları Çizelge 7.2'de görülebilir.

Çizelge 7.2 : Farklı büyüklükteki varlık ismi ağlarının, son 7 gün verisini kullanarak eğitilmiş olan tahmin modelleri ile yapılmış olan tahmin sonuçları.

	Doğru Tahmin Sayısı	**Yanlış Tahmin Sayısı**	**Başarı Oranı**	**RMSE**	**Korelasyon Değeri**
Tüm Varlık İsimlerinin Son 7 Gün İçerisindeki Değişimlerini İçeren Tahmin Modeli	112	77	59,25%	87,68	0,0538
En Sık Geçen 700 Varlık İsminin Son 7 Gün İçerisindeki Değişimlerini İçeren Tahmin Modeli	114	75	60,32%	352,60	-0,0734
"Birleşmiş Milletler" Alt Ağının Varlık İsimlerinin Son 7 Gün İçerisindeki Değişimlerini İçeren Tahmin Modeli	129	60	68,25%	70,59	0,1481

Çizelge 7.2'deki sonuçlardan görülebileceği üzere, 7 gün öncesinin bilgileri ile oluşturulan modellerin başarı oranı, 1 günlük zaman aralığındaki bilgiler ile oluşturulan modellerin sonuçları ile kıyaslandığında pozitif yönde değişim göstermiştir. Özellikle 1 günlük zaman aralığını içeren modeller arasında da doğruluk oranı en yüksek olan, "Birleşmiş Milletler" alt ağının varlık isimleri ile oluşturulan modelden elde edilen RMSE ve korelasyon değerlerinde de iyileşmeler olmuştur. Ancak en sık geçen 700 varlık isminin son 7 günlük zaman aralığı içerisindeki dağılımlarını içeren kanallar ile oluşturulan modelde, RMSE ve korelasyon değerlerinde negatif yönde değişimler gerçekleşmiştir. Bunun temel sebebi, en sık geçen 700 varlık ismi ile oluşturulan ve 1 günlük zaman aralığındaki dağılımları içeren model içerisindeki değerler bile oldukça birbirine benzer şekilde yer alırken, bu aralık son 7 güne çıktığında bu ayırt ediciliğin daha da fazla ortadan kalkmasıdır.

Bu sonuçlar doğrultusunda, borsa endeksinde meydana gelen hareketlerin sadece günlük olaylardan etkilenmediği, daha önce meydana gelen olayların da bu hareketler üzerinde etkisinin olduğu görülmektedir. Son 7 günlük zaman aralığı içerisinde varlık isimleri kullanılarak oluşturulan tahmin modelleri arasından başarı oranı en yüksek olan modelin içerisinde yer alan ve her bir günü temsil eden kanalların hesaplanan önem katsayısı değerleri Şekil 7.1'de görülebilir.



Şekil 7.1 : Son 7 günün varlık isimleri kullanılarak eğitilmiş olan tahmin modeli içeresinde yer alan kanallar ve bu kanalların önem katsayıları.

Bu değerler incelendiğinde, belirli bir zaman dilimi içerisinde gerçekleşmiş olan olayların etkisini zaman içerisinde yitirmekte olduğu ve tahmin modeli için değerli bilgiler içerdiği görülmektedir. Bu değerler arasında yer alan en düşük katsayı değerinin, tahminin yapıldığı güne ait olan katsayı değeri olmasının sebebi ise piyasalar kapandıktan sonra yeni gelişmelerin yaşanmaya ve gelişmeler doğrultusunda haberlerin yayınlanmaya devam etmesi ancak bu haberlerin bulunduğu gün içerisindeki endeks değerini etkileyememesidir.

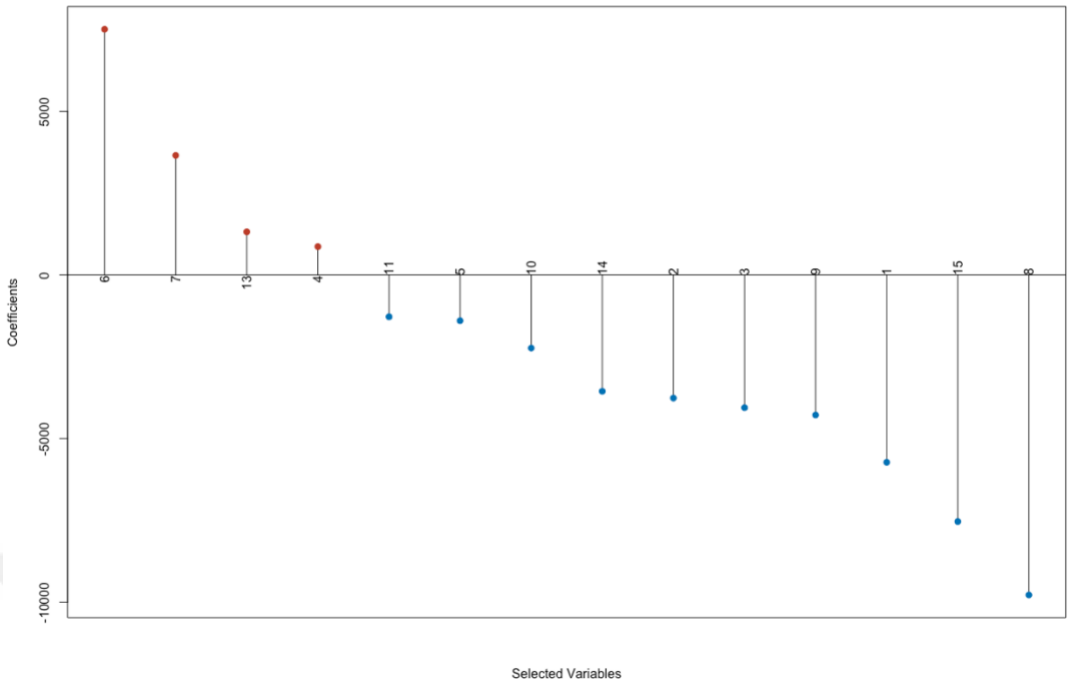
Zaman aralığının, bulunulan günden önceki 7 günü kapsayacak şekilde genişletilmesi ve tahmin modellerinin bu doğrultuda oluşturulması, sonuçları olumlu şekilde etkilemiştir. Bu sebepten ötürü zaman aralığı bir seviye daha genişletilmiştir. Bunun

sonucunda varlık ismi ağlarını içeren yeni kanallar, bulunulan günden önceki son 14 günü kapsayacak şekilde oluşturulmuştur. Bu yeni kanallar doğrultusunda, 3 farklı büyüklükteki varlık ismi ağı ile oluşturulan tahmin modellerinin sonuçları Çizelge 7.3'de görülebilir.

Çizelge 7.3 : Farklı büyüklükteki varlık ismi ağlarının, son 14 gün verisini kullanarak eğitilmiş olan tahmin modelleri ile yapılmış olan tahmin sonuçları.

	Doğru Tahmin Sayısı	**Yanlış Tahmin Sayısı**	**Başarı Oranı**	**RMSE**	**Korelasyon Değeri**
Tüm Varlık İsimlerinin Son 14 Gün İçerisindeki Değişimlerini İçeren Tahmin Modeli	108	81	57,14%	86,09	0,0742
En Sık Geçen 700 Varlık İsminin Son 14 Gün İçerisindeki Değişimlerini İçeren Tahmin Modeli	107	82	56,61%	419,81	-0,0027
“Birleşmiş Milletler” Alt Ağının Varlık İsimlerinin Son 14 Gün İçerisindeki Değişimlerini İçeren Tahmin Modeli	123	66	65,08%	73,72	0,0089

Varlık ismi ağları kullanılarak, bulunulan günden önceki son 14 günü kapsayacak şekilde üretilen kanallar ile oluşturulan tahmin modellerinin Çizelge 7.3’de ye alan sonuçları incelendiğinde, bir noktadan sonra zaman aralığını genişletmenin, başarı oranını arttırmadığı görülmektedir. Son 7 günü kapsayacak şekilde üretilen kanalları içeren tahmin modelleri, son 14 günü kapsayan modellere oranla daha başarılı tahminlerde bulunmuştur. Tahmin modellerini oluşturmak için kullanılan kanallar arasındaki önem sırası ve bu tahmin modelleri arasından en başarılı sonuçları üreten modelin içerisinde yer alan kanalların hesaplanan önem katsayısı değerleri Şekil 7.2'de görülebilir.



Şekil 7.2 : Son 14 günün varlık isimleri kullanılarak eğitilmiş olan tahmin modeli içeresinde yer alan kanallar ve bu kanalların önem katsayıları.

Son 14 günün varlık isimleri kullanılarak eğitilmiş olan tahmin modelinde kullanılan kanalların önem katsayısı değerleri incelendiğinde, yine benzer bir şekilde zaman aralığı içerisinde geriye gittikçe kanalların model içerisindeki önem katsayılarının düştüğü gözlemlenmektedir. Ancak belirli bir noktadan sonra kanallar arasındaki önem farkının ortadan kalktığı da görülmektedir.

Hem kanalların katsayı değerleri hem de oluşturulan tahmin modellerinin başarı oranları göz önüne alındığında, bulunulan günden 7 gün öncesine kadar olan zaman aralığı içerisindeki varlık ismi ağları kullanılarak bir tahmin modeli oluşturmak en uygun yöntem olarak gözükmektedir.

Son olarak, kanalları oluştururken detay seviyesini arttırmak ve daha başarılı tahmin sonuçları elde etmek adına kategoriler doğrultusunda alt kanallar oluşturulmuştur. Bunun için, her bir kanal içerisindeki değerlerin oluşturulduğu varlık ismi ağları, bulundukları haberlerin kategorilerine göre güncellenmiş ve bu güncellenen değerler ile alt kanallar üretilmiştir. Haber kategorileri olarak New York Times makalelerinde yer alan manşet, sürmanşet, ulusal haberler ve uluslararası haberler kategorileri kullanılmış ve her kanaldan 4 ayrı alt kanal oluşturulmuştur. Kategori alt kanalları

kullanılarak, daha önce yapılan deneyler sonucunda en başarılı sonuçların gözlemlendiği "birleşmiş milletler" düğümünün merkez olarak alındığı varlık ismi ağlarına ait kanallar oluşturulmuş ve 3 farklı tahmin modeli eğitilmiştir. Bu modeller, belirtilen varlık ismi ağı üzerinden 3 farklı zaman aralığı (bulunulan gün, bulunulan günden 7 önceki güne kadar ve bulunulan günden 14 önceki güne kadar) içerisinde bulunacak şekilde eğitilmiştir. Oluşturulan bu tahmin modellerine ait sonuçlar Çizelge 7.4'de görülebilir.

Çizelge 7.4 : "Birleşmiş Milletler" merkezli varlık ismi alt ağının, farklı haber kategorileri ve farklı zaman dilimlerinde bulunan değerlerini kullanarak eğitilmiş olan tahmin modelleri ile yapılmış olan tahmin sonuçları.

	Doğru Tahmin Sayısı	**Yanlış Tahmin Sayısı**	**Başarı Oranı**	**RMSE**	**Korelasyon Değeri**
"Birleşmiş Milletler" Alt Ağının Varlık İsimlerini 4 Alt Kategori Altında İçeren Tahmin Modeli	108	81	57,14%	86,09	0,0742
"Birleşmiş Milletler" Alt Ağının Varlık İsimlerinin Son 7 Gün İçerisindeki Değişimlerini 4 Alt Kategori Altında İçeren Tahmin Modeli	107	82	56,61%	419,81	-0,0027
"Birleşmiş Milletler" Alt Ağının Varlık İsimlerinin Son 14 Gün İçerisindeki Değişimlerini 4 Alt Kategori Altında İçeren Tahmin Modeli	123	66	65,08%	73,72	0,0089

Sonuçları Çizelge 7.4'de görülebilecek olan, haber kategorileri doğrultusunda oluşturulan alt kanallar ile eğitilen tahmin modellerinin sonuçları incelendiğinde,

sonuçların daha kötü olduğu görülmektedir. Buna ek olarak, öznitelik kümesi içerisindeki detaylılığın artışının, bir noktadan sonra tahmin modelleri üzerinde negatif etkisinin olduğu sonucu da bu tahmin değerlerinden yola çıkarak söylenebilir. Tüm modellerin başarı oranları, hataların ortalama karekökü değeri ve tahminler ile gerçek değerler arasındaki korelasyon değerleri önceki modeller ile kıyaslandığında, olumsuz bir doğrultuda değişiklik göstermiştir. Bu tahmin modellerinden elde edilen sonuçların olumsuz yönde etkilenmesinin sebeplerinden bir diğeri, bazı zaman aralıkları için toplanmış olan verinin tüm kategorilerde yeterli seviyede içerik bulundurmaması olarak yorumlanabilir. Bu durum tahmin modellerinin yanlış yönde tahminler yapmış olmasına sebep olmuştur. Sonuçların olumsuz yönde etkilenmesinin bir başka sebebi ise, kategoriler doğrultusunda alt kanalların oluşturulması sonucu kanallar içerisindeki varlık ismi değerleri dağılımlarının birbirlerini oldukça benzer bir hale gelmesi ve bu durumunda modelleri olumsuz yönde etkilemesidir.

7.2.1.2 Konu modellerinden üretilen kanallar ile yapılan tahminler

Bu alt deney serisinde kullanılan tahmin modelleri, konu modelleri doğrultusunda oluşturulan kanallar ile eğitilmiş ve tahmin sonuçları eğitilen bu modeller kullanılarak elde edilmiştir. Konu modelleri ile eğitilmiş olan bu tahmin modelleri 3 farklı eğitim verisi varyasyonu ile eğitilmiştir.

Tahmin modellerinin eğitilmesi için kullanılacak olan girdi matrisinin oluşturulması için kullanılan kanal varyasyonlarının tamamı New York Times verisi kullanılarak üretilmiş olan konu modelleri ve bu konu modelleri içerisindeki değerler doğrultusunda oluşturulmuştur. Tahmin modelinin eğitilmesi için kullanılan konu modellemesi temelli oluşturulan bu kanallar sırası ile;

- 25 farklı konu ile konu modellemesinin yapıldığı ve konu dağılımının bulunduğu bir kanal,
- 50 farklı konu ile konu modellemesinin yapıldığı ve konu dağılımının bulunduğu bir kanal,
- 100 farklı konu ile konu modellemesinin yapıldığı ve konu dağılımının bulunduğu bir kanaldır.

Konu modelleri temelli kanallar kullanılarak eğitilmiş olan bu 3 farklı tahmin modelinden elde edilmiş olan tahmin sonuçları Çizelge 7.5'de görülebilir.

Çizelge 7.5 : Farklı büyüklükteki konu modellerinin, bulundukları gün verisini kullanarak eğitilmiş olan tahmin modelleri ile yapılmış olan tahmin sonuçları.

	Doğru Tahmin Sayısı	**Yanlış Tahmin Sayısı**	**Başarı Oranı**	**RMSE**	**Korelasyon Değeri**
25 Konu İçeren Konu Modeli Değerleri ile Eğitilen Tahmin Modeli	106	83	56,08%	71,34	-0,0600
50 Konu İçeren Konu Modeli Değerleri ile Eğitilen Tahmin Modeli	116	73	61,38%	71,38	0,0556
100 Konu İçeren Konu Modeli Değerleri ile Eğitilen Tahmin Modeli	104	85	55,02%	71,81	-0,0612

Elde edilen sonuçlar incelendiğinde, 50 farklı konu ile oluşturulan konu modellemesinin kullanıldığı kanallar ile eğitilen tahmin modelinin diğer 2 tahmin modeline göre daha başarılı tahmin sonuçları ürettiği görülmektedir. Bunun sebebi, küçük olan konu modeli için, 25 konunun tüm bir yıl içerisindeki haberleri temsil etmek için çok düşük kalması ve oluşturulan konu kümelerinin çok geniş bir şekilde haberleri kapsaması olarak yorumlanmıştır. Diğer taraftan 100 konu ile oluşturulan konu modelinin, 50 konu içeren modele göre olumsuz sonuçlar vermesi ise 100 konu grubunun fazla detaylı olacak şekilde haberleri gruplaması ve bu sebeple artan detayların tahmin modelini olumsuz olarak etkilemesinden kaynaklanması olarak görülmektedir.

Bu deney serisinin bir sonraki aşaması olarak tahmin modelleri, genişletilmiş zaman aralığı içerisinde yer alan kanallar ile eğitilmiş ve eğitilen bu tahmin modelleri kullanılarak tahmin sonuçları elde edilmiştir. Bu tahmin modellerinin eğitilmesi için kullanılacak kanalların oluşturulabilmesi adına ilk zaman aralığı 7 gün olarak belirlenmiştir. Genişletilmiş zaman aralıkları doğrultusunda eğitilen ilk grupta bulunan tahmin modelleri, farklı konu sayıları için oluşturulmuş olan 3 farklı konu modelinin bulunduğu gün ve 7 gün öncesine kadar olan değerlerini içerecek şekilde

eğitilmiştir. Bu doğrultuda eğitilen 3 farklı tahmin modelinin sonuçları Çizelge 7.6'da görülebilir.

Çizelge 7.6 : Farklı büyüklükteki konu modellerinin, son 7 gün verisini kullanarak eğitilmiş olan tahmin modelleri ile yapılmış olan tahmin sonuçları.

	Doğru Tahmin Sayısı	**Yanlış Tahmin Sayısı**	**Başarı Oranı**	**RMSE**	**Korelasyon Değeri**
25 Konu İçeren Konu Modelinin Son 7 Gün İçerisindeki Değerleri ile Eğitilen Tahmin Modeli	105	84	55,56%	72,03	0,0513
50 Konu İçeren Konu Modelinin Son 7 Gün İçerisindeki Değerleri ile Eğitilen Tahmin Modeli	111	78	58,73%	74,03	-0,0311
100 Konu İçeren Konu Modelinin Son 7 Gün İçerisindeki Değerleri ile Eğitilen Tahmin Modeli	96	93	50,79%	73,56	-0,0030

Çizelge 7.6'daki sonuçlardan görülebileceği üzere, 7 gün öncesinin bilgileri ile oluşturulan modellerin başarı oranı, 1 günlük zaman aralığındaki bilgiler ile oluşturulan modellerin sonuçları ile kıyaslandığında negatif yönde değişim göstermiştir. Tüm 7 günlük modellerden elde edilen tahmin sonuçlarının, 1 günlük modellerden elde edilen sonuçlara göre başarımı daha düşüktür. Bunun sebebi, farklı günler içerisinde yayınlanan haberlerdeki konu dağılımlarının sürekli değişim göstermesi olarak yorumlanmıştır. Günler içerisinde yer alan konu dağılımlarının değişiklik göstermesinden ötürü, 7 günlük konu modellerini içerecek şekilde oluşturulan öznitelik kanallarının ve bu kanallar ile eğitilen modellerin sonuçları olumsuz yönde etkilediği görülmüştür.

Bu alt deney gurubunda son olarak ise konu modelleri üzerinden oluşturulan tahmin modelleri için zaman aralığını arttırmanın etkilerini gözlemlemek adına, daha geniş bir zaman aralığı içerisindeki konu modellemesi değerleri kullanılarak da tahmin modelleri eğitilmiştir. Eğitilen bu modeller için zaman aralığı 14 gün olarak seçilmiştir. 14 gün olarak seçilen bu zaman aralığı doğrultusunda eğitilen tahmin modelleri, farklı konu sayıları için oluşturulmuş olan 3 farklı konu modelinin bulunduğu gün ve 14 gün öncesine kadar olan değerlerini içerecek şekilde eğitilmiştir. Bu doğrultuda eğitilen 3 farklı tahmin modeli ile elde edilen tahmin sonuçları Çizelge 7.7'de görülebilir.

Çizelge 7.7 : Farklı büyüklükteki konu modellerinin, son 14 gün verisini kullanarak eğitilmiş olan tahmin modelleri ile yapılmış olan tahmin sonuçları.

	Doğru Tahmin Sayısı	**Yanlış Tahmin Sayısı**	**Başarı Oranı**	**RMSE**	**Korelasyon Değeri**
25 Konu İçeren Konu Modelinin Son 14 Gün İçerisindeki Değerleri ile Eğitilen Tahmin Modeli	99	90	52,38%	76,17	-0,0704
50 Konu İçeren Konu Modelinin Son 14 Gün İçerisindeki Değerleri ile Eğitilen Tahmin Modeli	105	84	55,56%	77,57	-0,1428
100 Konu İçeren Konu Modelinin Son 14 Gün İçerisindeki Değerleri ile Eğitilen Tahmin Modeli	98	91	51,85%	77,86	-0,0668

Çizelge 7.7'de yer alan sonuçlar incelendiğinde, farklı konu modellerinin ve bu konu modellerinin 14 gün öncesine kadar olan değerlerinin eğitim için kullanıldığı tahmin modellerinden elde edilen sonuçların, beklenildiği gibi daha da başarısız olduğu görülmektedir. Eğitilen bu tahmin modellerinden üretilen tahmin sonuçlarında, başarı

oranının düştüğü görülmektedir. Bu tahmin modellerinde, başarı oranının düşmesinin yanı sıra bu modeller kullanılarak tahmin edilen değerler ile gerçek değerler arasındaki korelasyon da negatif yönlüdür. Bunun temel sebebi olarak da 7 günlük zaman aralığındaki konu modelleri ile eğitilen modellerde olduğu gibi, zaman aralığı arttıkça konu modelleri doğrultusunda yapılan gruplamaların değerinin kaybolması olarak gösterilebilir.

7.2.1.3 Varlık ismi ağlarından ve konu modellerinden üretilen kanallar ile yapılan tahminler

Önceki deney serileri incelendiğinde, varlık ismi ağları kullanılarak oluşturulan tahmin modellerinin iyi olarak değerlendirilebilecek tahminlerde bulunduğu gözlemlenmiştir. Bunun yanında konu modelleri ile oluşturulan kanalların ise başarılı tahmin modellerinin eğitilmesi için yeterli olmadığı görülmüştür. Ancak konu modelleri ile üretilen kanallar içerisinde yer alan değerler, tahmin modelleri için tamamen değersiz değildir. Konu modelleri iler üretilen bu kanallar farklı metotlar ile üretilmiş olan kanallar ile birlikte kullanılarak, oluşturulan tahmin modelinin başarı oranını arttırmakta yardımcı olabilirler. Bu yaklaşımdan yola çıkarak, bir sonraki adımda, varlık ismi ağlarından ve konu modellerinden oluşturulan kanalların bir arada kullanılarak hibrit tahmin modelleri eğitilmiştir.

Hibrit tahmin modelleri oluşturulurken, ilk olarak, daha önce yapılan alt deneylerde varlık ismi ağları ile oluşturulan modeller arasından en iyi sonuçların gözlemlendiği kanallar seçilmiştir. Seçilen bu kanallar oluşturulacak bu alt deney grubunda eğitilecek olan tahmin modellerinin temelini oluşturmaktadır. Bu doğrultuda temel olarak kullanılan kanallar, *"birleşmiş milletler"* düğümünün merkez olarak alındığı varlık ismi ağlarının 7 günlük zaman aralığı değerlerini içeren kanallardır. Eğitilecek tahmin modelinin temeli olarak alınan bu kanallara yardımcı olarak kullanılacak kanallar ise konu modelleri ile oluşturulan tahmin modelleri arasından en iyi sonuçları veren modeli oluşturan kanallar olarak seçilmiştir. Bu durumda konu modellemesi temelli analizler doğrultusunda seçilen kanallar; 25, 50 ve 100 konu olmak üzere 3 farklı büyüklükte olan ve zaman aralığı olarak sadece bulunulan günü içeren konu modeli kanallarıdır. Farklı analizler doğrultusunda üretilmiş olan farklı kanalların kombinasyonları ile eğitilen bu tahmin modellerinin sonuçları Çizelge 7.8'de görülebilir.

Çizelge 7.8 : “Birleşmiş Milletler” merkezli varlık ismi alt ağının, farklı büyüklükteki konu modelleri ve farklı zaman dilimlerinde bulunan değerlerini kullanarak eğitilmiş olan tahmin modelleri ile yapılmış olan tahmin sonuçları.

	Doğru Tahmin Sayısı	**Yanlış Tahmin Sayısı**	**Başarı Oranı**	**RMSE**	**Korelasyon Değeri**
“Birleşmiş Milletler” Alt Ağının Varlık İsimlerinin Son 7 Gün İçerisindeki Değerleri ve 25 Konu İçeren Konu Modeli Değerleri ile Eğitilen Tahmin Modeli	123	66	65,08%	70,28	0,0422
“Birleşmiş Milletler” Alt Ağının Varlık İsimlerinin Son 7 Gün İçerisindeki Değerleri ve 50 Konu İçeren Konu Modeli Değerleri ile Eğitilen Tahmin Modeli	134	55	70,90%	69,98	0,2315
“Birleşmiş Milletler” Alt Ağının Varlık İsimlerinin Son 7 Gün İçerisindeki Değerleri ve 100 Konu İçeren Konu Modeli Değerleri ile Eğitilen Tahmin Modeli	119	70	62,96%	69,77	0,1972

Çizelge 7.8'de yer alan sonuçlardan görülebileceği üzere, hibrit tahmin modellerinin başarı oranları diğer modellerin başarı oranlarını ile karşılaştırıldığında daha iyi gözükmektedir. Özellikle korelasyon değerlerinde büyük oranda iyileşme görülmektedir. Kendi deney serilerinde en iyi sonuçların alındığı, “birleşmiş milletler” düğümünün merkez olarak alındığı varlık ismi ağlarının 7 günlük zaman aralığı

değerlerini içeren kanallar ile 50 konu içeren konu modeli kanalının bir arada kullanılarak eğitildiği tahmin modeli, tüm deney serileri içerisindeki en yüksek başarı oranına ve en yüksek korelasyon değerine sahip olacak şekilde tahmin sonuçlarını üretmiştir.

Eğitilen bu tahmin modellerinden yola çıkarak, varlık ismi ağları kullanılarak oluşturulan kanalların, tahmin modellerinin eğitilmesi için tek başına değerli oldukları ancak bu kanalların konu modellerinden üretilmiş kanallar ile birlikte kullanılarak daha güçlü tahmin modelleri oluşturulabileceği söylenebilir.

7.2.2 Deney II: 01 Eylül 2017 – 30 Kasım 2017 tarihleri arasında Dow Jones endeksi değişimlerinin tahmin edilmesi

İkinci ana deney grubu, sadece haber verisini değil buna ek olarak sosyal medya verisini de kullanmaktadır. Toplanan bu 2 farklı veri kümesini kullanarak, geliştirilen doğal dil işleme aracının alt modülleri ile yapılan analizler sonucu, Dow Jones endeksindeki değişimi tahmin etmeyi hedeflemektedir. Bu doğrultuda yapılan analizler ve oluşturulan öznitelikler doğrultusunda farklı alt deneyler tanımlanmış, bu doğrultuda tahmin modelleri eğitilmiş ve endeksin yönüne ilişkin tahminler yapılmıştır.

7.2.2.1 Düşünce analizi ağlarından üretilen kanallar ile yapılan tahminler

Yapılan bu deney serisinde, tahmin modellerinin eğitilmesi için düşünce analizi ile oluşturulan ağlar ve bu ağlar içerisinde yer alan varlık ismi değerlerinden oluşturulan kanallar kullanılmıştır. Girdi matrisinin oluşturulması için kullanın kanal varyasyonlarının tamamı Twitter üzerinden Amerika Birleşik Devletleri ile Kuzey Kore arasında gerçekleşmiş olan nükleer silah krizini ile ilgili toplanmış olan veri ile üretilmiştir. Bu verinin sosyal medya kullanıcıları tarafından üretilmiş olması sonucu, konu ile ilgisiz olarak birçok varlık ismi içerikler içerisinde yer edinmiştir. Bunun bir sonucu olarak farklı büyüklüklerdeki ağların, düşünce analizi ağı içerisinden bir alt ağ olarak çıkarılması ve kullanılması mümkündür. Bu sebeple en sık geçen 500 varlık ismi ve en sık geçen 5.000 varlık ismi olmak üzere, 2 alt ağ seçilmiş ve deneyler bu alt ağlar kullanılarak yapılmıştır.

Eğitilen tahmin modelleri ilk olarak, düşünce analizi ağlarında en sık geçen 500 varlık isminin farklı zaman aralıklarındaki dağılımları ile oluşturulmuş 3 farklı eğitim verisi varyasyonu ile eğitilmiştir. Bu modellerin eğitilmesi için kullanılan kanallar sırası ile;

- Düşünce analizi ağlarında en sık geçen 500 varlık ismi ile sadece bulunulan günü kapsayacak şekilde üretilen kanallar,
- Düşünce analizi ağlarında en sık geçen 500 varlık ismi ile bulunulan günden önceki son 7 günü kapsayacak şekilde üretilen kanallar,
- Düşünce analizi ağlarında en sık geçen 500 varlık ismi ile bulunulan günden önceki son 14 günü kapsayacak şekilde üretilen kanallardır.

Bu 3 farklı tahmin modelinin eğitilmesi ve kullanılması ile elde edilmiş sonuçlar Çizelge 7.9'da görülebilir.

Çizelge 7.9 : En popüler 500 varlık ismini içeren düşünce analizi ağlarının, farklı zaman dilimlerinde bulunan değerlerini kullanarak eğitilmiş olan tahmin modelleri ile yapılmış olan tahmin sonuçları.

	Doğru Tahmin Sayısı	**Yanlış Tahmin Sayısı**	**Başarı Oranı**	**RMSE**	**Korelasyon Değeri**
Düşünce Analizi Ağları İçerisinde En Sık Geçen 500 Varlık İsminin Değerlerini İçeren Tahmin Modeli	38	25	60,32%	75,09	0,0943
Düşünce Analizi Ağları İçerisinde En Sık Geçen 500 Varlık İsminin Son 7 Gün İçerisindeki Değerlerini İçeren Tahmin Modeli	39	24	61,90%	77,53	0,0954
Düşünce Analizi Ağları İçerisinde En Sık Geçen 500 Varlık İsminin Son 14 Gün İçerisindeki Değerlerini İçeren Tahmin Modeli	39	24	61,90%	78,82	0,1992

Çizelge 7.9'da yer alan sonuçlar incelendiğinde, bu deney serisi için kullanılan eğitim verisinin büyüklüğü, bir önceki deney serisi için kullanılan eğitim verisine oranla çok

daha küçük olmasına rağmen sonuçların kabul edilebilir seviyelerde olduğu gözlemlenmektedir. Başta başarı oranı olmak üzere, diğer değerlendirme kritel, bu değerlerin kabul edilebilir seviyelerde olduğu gözlemlenmektedir. Bu sonuçlar doğrultusunda, ekonomi üzerinde doğrudan etkisi olabilecek olan konuların tespiti, bu konular için ilgili veri kümesinin farklı kaynaklardan toplanan veriler ile oluşturulması ve oluşturulan bu veri kümeleri üzerinde yapılan düşünce analizi işlemleri doğrultusunda bir eğitim kümesinin oluşturulabileceği söylenebilir. Bu sürecin bir sonraki aşaması olarak ise oluşturulan bu eğitim verisi kullanılarak farklı tahmin modellerinin eğitilebileceği ve eğitilen bu tahmin modelleri kullanılarak borsa endeksinde meydana gelebilecek hareketlerin tahmin edilebileceği sonucu çıkarılabilir.

Bunun yanında, tahmin modellerinin başarı oranlarının değişen zaman aralıklarına göre kayda değer bir farklılık göstermediği de Çizelge 7.9'da yer alan sonuçlar incelendiğinde görülmektedir. Bu durumun başlıca sebebi, tahmin modelinin eğitilmesi sırasında girdi olarak verilen kanalların oluşturulması için kullanılan düşünce analizi ağlarının belirli bir konu üzerine odaklanması ve bu konu içerisindeki aktörleri simgeleyen varlık isimlerinin sürekli sabit kalmasından ötürü, tanımlanan zaman aralığı değişse bile elde edilen değerlerin kendini tekrarlaması olarak yorumlanmaktadır.

Bu deney serisi içerisindeki bir sonraki alt deney grubunda eğitim modelleri, düşünce analizi ağlarında en sık geçen 5.000 varlık isminin farklı zaman aralıklarındaki dağılımları doğrultusunda oluşturulmuş 3 farklı öznitelik kümesi kullanılarak oluşturulmuş olan eğitim verisi varyasyonu ile eğitilmiştir. Bu modellerin eğitilmesi için kullanılan kanallar sırası ile;

- Düşünce analizi ağlarında en sık geçen 5.000 varlık ismi ile sadece bulunulan günü kapsayacak şekilde üretilen kanallar,
- Düşünce analizi ağlarında en sık geçen 5.000 varlık ismi ile bulunulan günden önceki son 7 günü kapsayacak şekilde üretilen kanallar,
- Düşünce analizi ağlarında en sık geçen 5.000 varlık ismi ile bulunulan günden önceki son 14 günü kapsayacak şekilde üretilen kanallardır.

Bu 3 farklı tahmin modelinin eğitilmesi ve kullanılması ile elde edilmiş sonuçlar Çizelge 7.10'da görülebilir.

Çizelge 7.10 : En popüler 5.000 varlık ismini içeren düşünce analizi ağlarının, farklı zaman dilimlerinde bulunan değerlerini kullanarak eğitilmiş olan tahmin modelleri ile yapılmış olan tahmin sonuçları.

	Doğru Tahmin Sayısı	**Yanlış Tahmin Sayısı**	**Başarı Oranı**	**RMSE**	**Korelasyon Değeri**
Düşünce Analizi Ağları İçerisinde En Sık Geçen 5.000 Varlık İsminin Değerlerini İçeren Tahmin Modeli	39	24	61,90%	74,88	0,0966
Düşünce Analizi Ağları İçerisinde En Sık Geçen 5.000 Varlık İsminin Son 7 Gün İçerisindeki Değerlerini İçeren Tahmin Modeli	40	23	63,49%	77,15	0,0999
Düşünce Analizi Ağları İçerisinde En Sık Geçen 5.000 Varlık İsminin Son 14 Gün İçerisindeki Değerlerini İçeren Tahmin Modeli	39	24	61,90%	79,22	0,1915

Çizelge 7.10'da yer alan sonuçlar bir önceki deney serisinde yer alan sonuçlar ile karşılaştırıldığında çok büyük bir fark söz konusu değildir. Tüm değerlendirme kriterleri, düşünce analizi ağları içerisinde en sık geçen 5.000 varlık ismi ile eğitilen modeller için de oldukça benzer şekilde sonuç vermiştir. Bu doğrultuda düşünce analizi ağları ile eğitilen tahmin modellerinde, değerleri olan ve fark yaratan varlık isimlerinin ağın merkezinde yer alan varlık isimleri olduğu sonucuna ulaşılabilir.

7.2.2.2 Varlık ismi ağlarından, konu modellerinden ve düşünce analizi ağlarından üretilen kanallar ile yapılan tahminler

Düşünce analizi ağlarından oluşturulan kanallar yardımıyla eğitilen tahmin modellerinin başarı oranları kabul edilebilir seviyededir. Bu kanallara ek olarak, varlık

ismi ağları ve konu modelleri ile oluşturulan kanallar da eklenerek oldukça kapsamlı bir tahmin modeli oluşturulabilir. Bu yaklaşımdan yola çıkarak, varlık ismi ağları, konu modelleri ve düşünce analizi olmak üzere farklı 3 alandaki analiz sonuçlarından oluşturulmuş kanalların yer aldığı bir hibrit bir tahmin modeli eğitilmiştir. Bu modelin eğitilmesi için kendi deney serilerinde ortalamanın üstünde bir başarı oranı yakalayan kanallar seçilmiştir. Bu kanallar sırası ile;

- Varlık ismi ağlarındaki *"birleşmiş milletler"* düğümü çevresinde yer alan varlık isimlerinden oluşturulmuş olan ve bulunulan günden önceki son 7 günü kapsayacak şekilde üretilen kanallar,
- 50 farklı konu ile konu modellemesinin yapıldığı ve konu dağılımının bulunduğu bir kanal,
- Düşünce analizi ağlarında en sık geçen 500 varlık ismi ile bulunulan günden önceki son 7 günü kapsayacak şekilde üretilen kanallardır.

Oluşturulan bu tahmin modelinin eğitilmesi ve kullanılması ile elde edilmiş sonuçlar Çizelge 7.11'de görülebilir.

Çizelge 7.11 : En başarılı kanallar ile oluşturulmuş olan hibrit tahmin modeli ile yapılmış olan tahmin sonuçları.

	Doğru Tahmin Sayısı	**Yanlış Tahmin Sayısı**	**Başarı Oranı**	**RMSE**	**Korelasyon Değeri**
"Birleşmiş Milletler" Alt Ağının Varlık İsimlerinin Son 7 Gün İçerisindeki Değerleri, 50 Konu İçeren Konu Modeli Değerleri ve Düşünce Analizi Ağları İçerisinde En Sık Geçen 500 Varlık İsminin Son 7 Gün İçerisindeki Değerleri ile Eğitilen Tahmin Modeli	42	21	66,66%	68,67	0,2040

Son olarak eğitilen bu tahmin modelinin Çizelge 7.11'de yer alan sonuçları incelendiğinde, başarı oranı ve diğer metriklerin olumlu yönde değiştiği gözlemlenmektedir. Özellikle ortalama hata değeri büyük oranda düşüş göstermiştir. Bunun sebebi, doğru kanalların bir arada kullanılması sonucu, tahmin modelinin daha az sapan tahminlerde bulunmasıdır. Buna ek olarak korelasyon değeri de pozitif yönde değişim göstermiştir. Bu hibrit model, doğru kanalların bir arada kullanılmasının tahmin modelinin performansına büyük ölçüde etki edeceğini göstermektedir.

7.3 Performans

Tahmin modellerinin eğitilmesi ve tahmin sonuçların üretilebilmesi adına, ana ve alt deney grupları içerisinde yapılan tüm işlemler (örneğin; matris çarpımlarının yapılması, veri kümesi için öznitelik kümelerinin oluşturulması, öznitelikler doğrultusunda çizgelerin oluşturulması ve bu çizgelerin dolaşılması) hesaplama yoğunluğu yüksek olan işlemlerdir. Ancak bu işlemlerin yüksek performans ve hızda yapılabilmesi adına R ve Python gibi dinamik programlama dilleri için optimize edilmiş ve yüksek hızda bu işlemleri yapmaya yardımcı olan kütüphaneler yer almaktadır. Buna ek olarak, deneylerden sonuç alınabilmesi adına yapılması gereken işlemler yapıları gereği paralelleştirilmeye de olanak vermektedir. Bu noktada kaynaklar el verdiği takdirde, işlemler birden fazla işlemci gücü kullanılarak çok daha hızlı bir şekilde tamamlanabilmesi mümkündür.

Deneylerin yapılabilmesi adına 6 adet 2.2Ghz Intel i7 işlemciye sahip olan ve 16GB belleği bulunan bir bilgisayar kullanılmıştır. Kullanılan bu bilgisayar üzerinde, önerilen eğitim ve tahmin sistemin performansının değerlendirilebilmesi ölçümler de yapılmıştır. Sistemin performansını gözlemlemek adına yapılan bu ölçümler, 3 farklı büyüklükteki eğitim verisi üzerinde gerçekleştirilmiştir. Performans ölçümlerin üzerinde yapıldığı verileri oluşturmak için kullanılan kanallar aşağıdaki gibidir;

- **Küçük Ölçekli Eğitim Verisi:** Varlık ismi ağlarında tespit edilmiş olan varlık isimleri ile oluşturulmuş veri kümesi. İçerisinde 1 adet kanal bulundurmaktadır.
- **Orta Ölçekli Eğitim Verisi:** Varlık ismi ağlarındaki tespit edilmiş olan varlık isimleri ile bulunulan günden önceki son 7 günü kapsayacak şekilde üretilen kanallar kullanılarak oluşturulmuş veri kümesi. İçerisinde toplam 8 adet kanal bulundurmaktadır.

- **Büyük Ölçekli Eğitim Verisi:** Varlık ismi ağlarındaki tespit edilmiş olan varlık isimleri ile bulunulan günden önceki son 7 günü kapsayacak şekilde üretilen kanallar, 50 farklı konu ile konu modellemesinin yapıldığı ve konu dağılımının bulunduğu bir kanal ve düşünce analizi ağlarında en sık geçen 500 varlık ismi ile bulunulan günden önceki son 7 günü kapsayacak şekilde üretilen kanallar kullanılarak oluşturulmuş veri kümesi. İçerisinde toplam 24 adet kanal bulundurmaktadır.

Belirlenen farklı ölçeklerdeki veri kümeleri için deneyler yapılırken, özellikleri tanımlamış olan bilgisayarın sahip olduğu bütün hesaplama gücü bu işlem için ayrılmıştır ve sonuçlar bu doğrultuda alınmıştır. Performans değerleri, veri kümeleri kullanılarak yeni bir modelin eğitilmesi için geçen sürelerin saniye değerleri olarak belirlenmiştir. Bu değerler hesaplanırken farklı ölçekteki eğitim verilerinin her biri için, tutarsızlıklarının önüne geçebilmek adına, 10 farklı değer hesaplanmış ve bu değerlerin ortalaması alınmıştır. Performans testleri sonucunda elde edilmiş sonuçlar Çizelge 7.12'de görülebilir.

Çizelge 7.12 : Farklı ölçekteki eğitim verileri için yapılan performans testi sonuçları.

	Ortalama Süre (saniye)
Küçük Ölçekli Eğitim Verisi (Toplam 1 Kanal)	15,09
Orta Ölçekli Eğitim Verisi (Toplam 8 Kanal)	70,84
Büyük Ölçekli Eğitim Verisi (Toplam 24 Kanal)	176,03

Test sonuçlarından da görülebileceği üzere ortalama bir kişisel bilgisayarda bile modellerin eğitim sürelerinin oldukça makul bir aralıkta bulunduğu görülmektedir. Bu süreler veri kümesi çok daha büyüdüğü ve zaman dilimi olarak çok daha geniş bir aralığı kapsadığı takdirde bile çok büyük bir probleme yol açmayacaktır. Bunun yanında, çok daha yüksek performansa ve hızlı şekilde sonuç alınmaya ihtiyaç duyulduğu senaryolar ile karşılaşıldığında ise var olan sistem bulut servis sağlayıcılarından biri üzerinde yer alan daha güçlü bir bilgisayar üzerine kolayca kurulabilir ve çalıştırılabilir.

8. SONUÇ VE ÖNERİLER

Geçmişten günümüze bakıldığında, her geçen gün teknolojik gelişimin bir kat daha arttığı görülmektedir. Bu gelişim özellikle bilgisayar bilimleri üzerinde çok daha etkin olmaktadır. Bunun bir sonucu olarak da hem insanların veri üretebileceği araçlar hem de bu veriye erişim yolları artmıştır. Haber web siteleri, ürün değerlendirme ve yorumlama web siteleri, günlük olarak da değerlendirilebilecek bloglar, sosyal medya ağları, medya platformları ve benzeri farklı oluşumlar bu içerik kaynaklarına birer örnek teşkil etmektedir. Bu kadar çok sayıda farklı kaynaktan, farklı içeriklere ulaşılabilmesi bilgisayar bilimleri alanında çalışan birçok araştırmacının da ilgisini çekmiştir. Bunun sebebi, internet ortamında bulunan bu farklı yapıdaki içerikler toplanabildiği, sonrasında ise toplanan bu veri etkili ve verimler teknikler aracılığıyla işlendiği takdirde birçok anlamlı sonuca ulaşılabilmesindir. Elde edilen bu anlamlı sonuçlar doğrultusunda hem geçmişe yönelik analiz çalışmaları yapılabilmekte hem de geleceğe yönelik karar mekanizmaları kurgulanabilmektedir.

Yapılan bu çalışmada, doğal dil işleme ve makine öğrenmesini temel alan yöntemler bir arada kullanılarak metin verisini işleyen ve bu doğrultuda bir tahmin modeli ortaya koyan bir sistem geliştirilmiştir. Geliştirilen bu sistemin ilk aşamasında New York Times ve Twitter olmak üzere iki farklı kaynaktan metin verileri toplanmıştır ve doğal dil işleme yöntemleri kullanılarak toplanan bu metin verisi işlenmiştir. Daha sonra işlenen verilerden, makine öğrenmesi alanı içerisinde yer alan regresyon temelli yöntemler ile eğitilen modeller yardımıyla, finans alanında tahminler yapılmıştır. Yapılan bu tahminler, New York Menkul Kıymetler Borsası'nda işlem gören dünyanın en büyük 30 şirketinin hisse senetlerinden oluşan Dow Jones endeksindeki değişimler içermektedir. Aynı zamanda yapılan bu tahminler, farklı öznitelik kümelerine ve bu öznitelik kümelerinin seçilme şekline göre oluşturulmuş olan değişik senaryolar için tekrarlanmış ve elde edilen sonuçlar birbiri ile kıyaslanmıştır.

İlk olarak verilerin farklı kaynaklardan toplanabilmesi adına ağ gezgini adı verilen uygulamalar geliştirilmiştir. Bu uygulamaların amacı belirtilen kaynaklardan belirli

periyotlar ile istenilen verilen otomatik olarak toplanması ve istenilen yere kaydedilmesidir. Bunun yapılabilmesi adına, ilk aşamada verinin toplanmak istenildiği kaynakların sağlamış oldukları resmi uygulama programlama arayüzleri kullanılmıştır. Bu arayüzler, tanımlanan değişkenler doğrultusunda isteği gönderen kişiye talep edilen veriyi hazırlamak ve döndürmek ile yükümlüdür. Ancak bu arayüzlerin olumsuz tarafı içerisinde istek üst limit değerleri bulundurması ve bu sebeple sınırlı miktarda veriyi sonuç olarak döndürmesidir. Bu problemi çözmek adına, kaynaklar tarafından sağlanan resmi uygulama programlama arayüzlerine alternatif olacak gezginler tasarlanmıştır. Bu gezginler yardımıyla New York Times web sitesi üzerinden 01 Ocak 2017 ile 31 Aralık 2017 tarihleri arasında yayınlanmış olan 4 farklı kategoriye ait 12560 adet haber ve makale toplanmıştır. Buna ek olarak ise Twitter'da kullanıcılar tarafından, 01 Ağustos 2017 ile 30 Kasım 2017 tarihleri arasında *"Amerika Birleşik Devletleri"* ve *"Kuzey Kore"* ile ilgili paylaşılmış olan 2.854.333 adet tweet toplanmıştır.

Bir sonraki aşamada, farklı kaynaklardan toplanmış olan metin verilerini analiz edebilmek adına, farklı doğal dil işleme yöntemlerinin içerisinde toplandığı bir uygulama geliştirilmiştir. Geliştirilen bu uygulama girdi olarak metin verilerini kabul etmekte ve bu metin verileri üzerinde metin bölütleme, sözcük türü işaretleme, etkisiz kelimelerin temizlenmesi, kurallı ifadelerin filtrelenmesi gibi ön işleme yöntemlerini uygulayabilmektedir. Ek olarak geliştirilmiş olan bu uygulama ön işleme yöntemlerinin yanı sıra varlık isimlerinin tanımlanması ve sınıflandırılması, düşünce analizi ve konu modellemesi gibi çok daha karmaşıklığı yüksek olan işlemlerin yapılmasına da olanak vermektedir. Bu işlemlerin yapılması noktasında, literatürde birçok farklı yöntem ve yaklaşım bulunmaktadır. Farklı yöntemler ile yapılan analizler sonucunda elde edilen sonuçların da birbirine göre farklılık gösterebildiği gözlemlenmiştir. Bu nedenle geliştirilen uygulama içerisinde, farklı yöntemleri temel alan Apache OpenNLP, Stanford CoreNLP, OpeNER ve SentiNEL gibi farklı doğal dil işleme kütüphaneleri birbirleri ile uyum içerisinde çalışacak şekilde entegre edilmiştir. Bu sayede sonuçlar üretilirken birçok farklı uygulama çıktısı farklı katmanlar olarak birbiri üzerine eklenerek daha tutarlı bir çıktı elde edilmesine olanak sağlanmıştır.

Varlık isimlerinin tanımlanması, düşünce analizi ve konu modellemesi işlemleri toplanan metin verisi üzerinde uygulandıktan sonra elde edilen sonuçlar farklı

analizler yapılmasına olanak sağlamaktadır. Bu sonuçlardan yararlanarak, tespit edilen farklı varlık isimleri arasındaki ilişkiler tespit edilmiştir. Varlık isimleri arasındaki ilişkiler doğrultusunda sosyal ağlar ortaya konmuş ve bu sosyal ağlar içerisinde yer alan ilişkilerin zaman içerisindeki değişimleri gözlemlenmiştir. Dönemsel olarak ortaya çıkan ve kaybolan varlık isimleri tespit edilmiş ve farklı varlık isimleri arasındaki güçlenen ve zayıflayan ilişkiler gözlemlenmiştir. Bu analizlere ek olarak, düşünce analizinden elde edilen sonuçlar da bir katman olarak eklenmiştir. Düşünce analizi sonucunda elde edilen metin verisinin pozitif veya negatif olduğuna yönelik sonuçlar doğrultusunda farklı varlık isimlerinin birbirleri arasındaki ilişkiler de bu yönüyle analiz edilmiştir. Birbiri ile pozitif ilişkisi ve negatif ilişkisi bulunan varlık isimleri bu analiz sonucunda ortaya konmuştur. Metin verisinin analizinde kullanılan bir diğer yöntem olan konu modellemesiyle ise ana konu başlıkları ve bu konu başlıklarını oluşturan kelime grupları ortaya konmuştur. Dönemsel olarak öne çıkan konu başlıkları, bu konu başlıklarının zaman içerisindeki değişimi ve yine bu başlıkların birbirleri ile olan ilişkilerine yönelik analizler ortaya konmuştur.

Farklı doğal dil işleme yöntemleri ile elde edilen sonuçlar, yapılan analizlere yön vermelerinin yanı sıra farklı tahmin modellerinin eğitilmesi noktasında da kullanılmıştır. Bu sonuçlar ile farklı öznitelik kümeleri ortaya konmuştur. Öznitelik kümeleri oluştururken varlık isimlerinin birbirleri ile olan ilişkileri, düşünce analizi doğrultusunda varlık isimleri arasındaki ilişkilerin değişimi ve konu modellemesi sonucu ortaya çıkan doküman skorları temel alınmıştır. Oluşturulan bu öznitelik kümeleri farklı varyasyonlar ile kullanılarak modellerin eğitimi esnasında farklı yaklaşımlar ortaya konmuştur. Modellerin eğitimi esnasında kullanılan öznitelik sayısının oldukça fazla olmasından ve bu durumun kurgulanan sistemin performansını olumsuz yöntem etkiliyor olmasından ötürü, özniteliklerin sayısının azaltılması ve sistemin verimini arttırmaya yönelik çalışmalar yapılmıştır. Bu çalışmalar sonucunda, her bir tarih için öznitelik kümelerinden bir ağ oluşturulup, günlerin önem değerlerinin PageRank algoritması yardımıyla hesaplanmasına ve bu sonuçlar ile sadeleştirilmiş bir öznitelik kümesi oluşturulmasına karar verilmiştir. Tahmin modelleri ise sadeleştirilmiş olan bu öznitelik kümeleri üzerinden eğitilecek şekilde tekrar kurgulanmıştır.

Tahmin modelleri üretmek için kullanılacak veri yapısı ve eğitim yöntemi tanımlandıktan sonra yapılacak deneyler belirlenmiştir. Deneyler, farklı metin

verilerini ve öznitelikleri kullanacak şekilde 2 ana grup altında olacak şekilde gruplanmıştır. Bu ana deney grupları altında farklı alt deney grupları da tanımlanmıştır. Bu deneyler içerisinde Dow Jones endeksinde gerçekleşecek olan değişimler tahmin edilmiştir ve bu elde edilen sonuçlar 3 farklı değerlendirme metriği doğrultusunda değerlendirilmiştir. Bu değerlendirme metrikleri, sonuçların yüzdesel olarak doğruluk oranı, elde edilen hataların ortalama karekökü ve tahmin ile gerçek sonuçlar arasındaki korelasyon değeri olarak belirlenmiştir.

Tanımlanan ilk ana deney grubu içerisinde, New York Times web sitesinden çekilen veriler doğrultusunda 01 Nisan 2017 ile 31 Aralık 2017 tarihleri arasında Dow Jones endeksi değişimleri tahmin edilmiştir. Bu doğrultuda, doğal dil işleme yöntemleri sonucunda elde edilen varlık isimlerinden oluşturulan ağların ve alt ağlarının bulunulan günde, son 7 gündeki ve son 14 gündeki değişimleri, düşünce analizi sonuçları ve konu modellemesi sonuçlarının farklı konu sayısına göre alınan sonuçları farklı öznitelik kümeleri olacak şekilde eğitim verisi içerisinde tanımlanmıştır. Elde edilen tahmin modellerinin başarımları incelendiğinde, bu oranın 70,90% değerine kadar çıktığı ve tahmin sonuçları ile gerçek sonuçlar arasında pozitif bir korelasyon değeri olduğu gözlemlenmiştir. Başarımı en yüksek olan tahmin modeli, "Birleşmiş Milletler" varlık ismini temel alan ve bu varlık ismi doğrultusunda ulaşılan alt Varlık ismi ağlarının son 7 gün içerisindeki değişimlerini ve 50 konu içeren konu modeli değerlerini içeren tahmin modelidir. Bu deneyin bir parçası olarak, yaşanan gündelik olayların endeks üzerindeki etki süreleri de gözlemlenmiş ve 7 günlük bir süre içerisinde yaşanan olayların etkilerinin devam ettiği de tespit edilmiştir. Bunlara ek olarak, bu deney grubu için başarımı en düşün olan tahmin modeli, 50,79%'lik başarım oranı ile sadece konu modellemesi sonuçlarını kullanarak eğitilen bir tahmin modelidir. Bunun bir sonucu olarak, konu modellerinin endekste gerçekleşen değişimlere olan etkisinin, varlık ismi ağları ve düşünce analizi sonuçlarına göre daha düşük olduğu, bu özniteliklerin destekleyici rolde kullanılabileceği sonucuna ulaşılmıştır.

İkinci ana deney gurubu içerisinde, mevcut metin verilerinin üzerine Twitter üzerinden çekilen veri kümesi de eklenerek 01 Eylül 2017 ile 30 Kasım 2017 tarihleri arasında Dow Jones endeksi değişimleri tahmin edilmiştir. Bu deney gurubu içerisinde kullanılan ek Twitter verisi, ilk deney grubu içerisinde tek başına kullanılan New York Times verisine oranla çok daha farklı yapıdadır. Verinin bir sosyal ağ platformu

üzerinden toplanmış olmasının bir sonucu olarak, yapısal anlamda yazım kurallarının çok daha az önemsendiği, çok daha fazla yazım yanlışının bulunduğu ve analiz etmesi oldukça zor olan bir veri ortaya çıkmıştır. Ancak öte yandan çok daha geniş bir grubu temsil ediyor olmasından ötürü bu veri oldukça büyük bir öneme sahiptir. Nükleer silahlanmaya yönelik ortaya çıkan krizin etkilerinin gözlemlenmesi adına Twitter üzerinden "Amerika Birleşik Devletleri" ve "Kuzey Kore" ile ilgili olan tweet'ler toplanmıştır. Toplanan bu sosyal medya metin verisi geliştirilen araçlar ile analiz edilmesinin ardından eğitim verisi içerisine ek kanallar olarak eklenmiştir. Ek kanallar ile genişletilen eğitim verileri ile bu deney grubuna ait olan alt deneyler yapıldığında, başarım oranının 66,66% değerine kadar çıktığı görülmüştür. Ulaşılan en yüksek başarım oranı bir önceki deney grubunda ulaşılan en yüksek orana göre daha düşük kalsa da tüm tahmin modelleri belirli bir ortalamanın üzerinde başarım sağlamıştır. Bu alt deney grubunda elde edilmiş olan en düşük başarım oranı 60,32% değeri olmuştur. Bunun yanında, ikinci deney grubu içerisinde eğitilmiş olan tüm tahmin modellerinden elde edilen tahmin sonuçları ile gerçek sonuçlar arasında bulunan korelasyon değerlerinin de pozitif olduğu görülmektedir. Bu deney grubundan elde edilen sonuçlar, sosyal medya üzerinden toplanan güncel verilerin ve bu veriler kullanılarak elde edilen özniteliklerin tahmin modelleri için değerli olduğunu göstermiştir.

Önerilen yöntem doğrultusunda gerçekleştirilen farklı deneyler sırasında, eğitilen tahmin modelleri için performans testleri de yapılmıştır. Bu testler doğrultusunda, önerilen yöntemin çalışma zamanı ve verimlilik açısından bir sorun yaratmayacağı test sonuçları ile desteklenmiştir. Yapılan performans testleri sonucunda, öznitelik kümesi oldukça büyük olan, 24 kanala sahip bir eğitim verisi kullanılarak yapılan eğitim işleminin, ortalama bir kişisel bilgisayar yardımıyla ortalama 176 saniye içerisinde tamamlanabildiği görülmüştür. Bu sürenin farklı kaynaklar yardımıyla çok daha fazla düşürülebileceği de görülmektedir.

Yapılan çalışmada, günlük gazete verisi ve insanların sosyal medya üzerinden gündeme verdikleri tepkilerin yer aldığı metin verisi kullanılarak finans alanında bir tahmin modeli ortaya konabileceği gösterilmiştir. Bu çalışmanın sonraki aşamalarında, ilk olarak kullanılan veri kümesi çok daha geniş bir zaman aralığını kapsayacak şekilde genişletip daha fazla özniteliğe sahip ve daha detaylı tahmin modelleri üretilebilir. Verinin toplandığı kaynaklara ek olarak farklı gazetelerden, sosyal medya platformlarından ve diğer kaynaklardan veri toplanarak, eğitim esnasından kullanılan

veri kümesi zenginleştirilebilir. Yeni kaynaklardan toplanan veriler doğrultusunda farklı analiz sonuçları ve öznitelikler de ortaya konabilir. Bunların yanı sıra, ortaya konan yöntem finans alanı dışında yer alan farklı senaryolar için de tahmin modellerinin eğitilmesi için kullanılabilir. Bu senaryolar, seyrek ve yüksek sayıda özniteliğe sahip olan durumları içermektedir.

KAYNAKLAR

[1] **Hirschberg, J., Manning, C. D.,** (2015). "Advances in natural language processing". *Science*, 349(6245), 261-266.

[2] **Natalya, F. N.,** (2004). "Semantic integration: a survey of ontology-based approaches". *ACM SIGMOD Record*, 33(4), 65-70.

[3] **Nadeau, D.**, **Sekine, S.,** (2007). "A Survey of Named Entity Recognition and Classification". *Lingvisticae Investigationes*, 30(1), 3-26.

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

[5] **Liu, B., Zhang, L.,** (2012). "A Survey of Opinion Mining and Sentiment Analysis". *Mining Text Data, Mining Text Data*, ISBN: 978-1-4614-3222-7, 415-463.

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

[7] **Jacobi, C., van Atteveldt, W. H., Welbers, K.,** (2016). "Quantitative analysis of large amounts of journalistic texts using topic modelling". *Digital Journalism*, 4(1), 89-106.

[8] **Ney, H., Niessen, S., Och, F. J., Sawaf, H., Tillmann, C., Vogel, S.,** (2000). "Algorithms for statistical translation of spoken language". *IEEE Transactions on Speech and Audio Processing*, 8(1), 24-36.

[9] **Zukerman, I., Albrecht, D. W.,** (2001). "Predictive Statistical Models for User Modeling". *User Modeling and User-Adapted Interaction*, 11(1-2), 5-18.

[10] **Golbeck J.,** (2006). "Generating Predictive Movie Recommendations from Trust in Social Networks". *In Proceedings of the 4th International Conference on Trust Management (iTrust'06)*, 93-104.

[11] **Bennett, J., Lanning, S.,** (2007). "The Netflix Prize". *KDD Cup and Workshop in conjunction with KDD*.

[12] **Kimoto, T., Asakawa, K., Yoda, M., Takeoka, M.,** (1990). "Stock market prediction system with modular neural networks". *1990 IJCNN International Joint Conference on Neural Networks*.

[13] **Chong, M., Abraham, A., Paprzycki, M.,** (2005). "Traffic accident analysis using machine-learning paradigms". *Informatica 2005*, 29, 89–98.

[14] **Schumaker, R. P., Chen, H.,** (2009). "Textual analysis of stock market prediction using breaking financial news: The AZFin text system". *ACM Transactions on Information Systems*, 27(2), Article No. 12.

[15] **Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., Zhang, J.,** (1998). "Daily stock market forecast from textual web data". *1998 IEEE International Conference on Systems, Man, and Cybernetics (SMC'98 Conference Proceedings)*, 3, 2720-2725.

[16] **Ding, X., Zhang, Y., Liu, T., Duan, J.,** (2015). "Deep learning for event-driven stock prediction". *In Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*, 2327-2333.

[17] **Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., Smith, N. A.,** (2009). "Predicting Risk from Financial Reports with Regression". *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 272-280.

[18] **Kim, Y., Jeong, S. R., Ghani, I.,** (2014). "Text Opinion Mining to Analyze News for Stock Market Prediction". *International Journal of Advances in Soft Computing and its Applications*, 6(1), 1–13.

[19] **Nguyen, T. H., Shirai, K.,** (2015). "Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction". *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1354–1364.

[20] **Bollen, J., Mao H., Zeng X.,** (2011). "Twitter mood predicts the stock market". *Journal of Computational Science*, 2(1), 1–8.

[21] **Pagolu, V. S., Reddy, K. N., Panda G., Majhi B.,** (2016). "Sentiment analysis of Twitter data for predicting stock market movements". *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*.

[22] **Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., Deng, X.,** (2013). "Exploiting topic based twitter sentiment for stock prediction". *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 24–29.

[23] **Fielding, R.**, (2000). "Representational state transfer". Architectural Styles and the Design of Network-based Software Architecture, *Doktora Tezi*, University of California, ISBN: 0-599-87118-0, 76-85.

[24] **European Association for Standardizing Information and Communication Systems (ECMA)**, (1999). "ECMA-262: ECMAScript Language Specification". 3rd edition.

[25] "Twitter API". https://developer.twitter.com/en/docs/api-reference-index, *alındığı tarih: 18 Eylül 2019.*

[26] **Bikel, D. M., Miller, S., Schwartz, R., Weischedel R.**, (1997). "Nymble: a High-Performance Learning Name-finder". *5th Conference on Applied Natural Language Processing*, 194-201.

[27] **Sekine, S.**, (1998) "Description of the Japanese NE System Used for MET-2". *Message Understanding Conference 7* (MUC-7).

[28] **Szarvas, G., Farkas, R., Kocsor, A.,** (2006). "A multilingual named entity recognition system using boosting and c4.5 decision tree learning

algorithms". *Proceedings of the 9th International Conference on Discovery Science (DS'06)*, 267-278.

[29] **Asahara, M., Matsumoto, Y.,** (2003). "Japanese Named Entity extraction with redundant morphological analysis". *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (NAACL '03)*, 1, 8-15.

[30] **Tsochantaridis, T., Hofmann, T., Joachims, T., Altun, Y.,** (2004). "Support vector machine learning for interdependent and structured output spaces". *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*.

[31] **Borthwick, A. E.,** (1999). "A Maximum Entropy Approach to Named Entity Recognition", *Doktora Tezi*, New York University, ISBN: 0-599-47232-4.

[32] **Lafferty, J. D., McCallum, A., Fernando, C. N. P.,** (2001). "Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data". *Proceedings of the 18th International Conference on Machine Learning (ICML '01)*, 282-289.

[33] **Brin, S.,** (1998). "Extracting Patterns and Relations from the World Wide Web". *International Workshop on The World Wide Web and Databases (WebDB '98)*, 172-183.

[34] **Cucerzan, S., Yarowsky, D.,** (1999). "Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence". *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

[35] **Yangarber, R., Lin, W., Grishman, R.,** (2002). "Unsupervised learning of generalized names". *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1 (COLING '02)*, 1-7.

[36] **Riloff, E., Jones, R.,** (1999). "Learning Dictionaries for Information Extraction using Multi-level Bootstrapping". *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI'99)*, 474-479.

[37] **Pasca, M., Lin, D., Bigham, J., Lifchits, A., Jain, A.,** (2006). "Organizing and Searching the World Wide Web of Facts—Step One: The One-Million Fact Extraction Challenge". *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI'06)*, 1400-1405.

[38] **Miller, G. A.,** (1995). "WordNet: A Lexical Database for English". *Communications of the ACM*, 38(11), 39-41.

[39] **Alfonseca, E., Manandhar, S.,** (2002). "An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery". *Proceedings of the 1st International Conference on General WordNet*, 41-42.

[40] **Evans, R.,** (2004). "A framework for named entity recognition in the open domain". *Recent Advances in Natural Language Processing III: Selected Papers from RANLP*, 260, 267–274.

[41] **Hearst, M. A.,** (1992). "Automatic acquisition of hyponyms from large text corpora". *Proceedings of the 14th Conference on Computational Linguistics - Volume 2 (COLING'92)*, 2, 539-545.

[42] **Cimiano, P., Völker, J.,** (2005). "Towards Large-Scale, Open-Domain and Ontology-Based Named Entity Classification". *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, 43-47.

[43] **Shinyama, Y., Sekine, S.,** (2004). "Named Entity Discovery Using Comparable News Articles". *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, 848–853.

[44] "Apache OpenNLP Library". https://opennlp.apache.org, *alındığı tarih: 18 Eylül 2019.*

[45] "Stanford CoreNLP – Natural Language Library", https://stanfordnlp.github.io/CoreNLP, *alındığı tarih: 18 Eylül 2019.*

[46] "The OpeNER Project". https://www.opener-project.eu, *alındığı tarih: 18 Eylül 2019.*

[47] **Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.,** (2014). "The Stanford CoreNLP natural language processing toolkit". *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55 - 60.

[48] **Finkel, J. R., Grenager, T., Manning, C.,** (2005). "Incorporating non-local information into information extraction systems by Gibbs sampling". *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*.

[49] **Agerri, R., Cuadros, M., Gaines, S., Rigau, G.,** (2013). "OpeNER: Open Polarity Enhanced Named Entity Recognition". *Sociedad Española para el Procesamiento del Lenguaje Natural*, 51, 215-218.

[50] **Agerri, R., Bermudez, J., Rigau, G.,** (2014). "IXA pipeline: Efficient and ready to use multilingual NLP tools". *Proceedings of the 9th Language Resources and Evaluation Conference*, 3823–3828.

[51] **Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., Aliprandi, C.,** (2009). "KAF: a generic semantic annotation format". *Proceedings of the GL2009 Workshop on Semantic Annotation*.

[52] **Kang, H., Yoo, S. J., Han D.,** (2012). "Senti-lexicon and improved Naive Bayes algorithms for sentiment analysis of restaurant reviews". *Expert Systems with Applications*, 39, 6000-6010.

[53] **Kaufmann, J. M.,** (2012). "JMaxAlign: A Maximum Entropy Parallel Sentence Alignment Tool". *Proceedings of the 24th International Conference on Computational Linguistics (COLING'12): Demonstration Papers*, 277-288.

[54] **Chen, C. C., Tseng, Y.,** (2011). "Quality evaluation of product reviews using an information quality framework". *Decision Support Systems*, 50, 755–768.

[55] **Li, Y., Li, T.,** (2013). "Deriving market intelligence from microblogs". *Decision Support Systems*, 55, 206-217.

[56] **Ruiz, M. E., Srinivasan P.,** (1999). "Hierarchical neural networks for text categorization". *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, 281-282.

[57] **Ng, H. T., Goh, W. B., Low, K. L.,** (1997). "Feature selection, perceptron learning, and a usability case study for text categorization". *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97)*, 67-73.

[58] **Yi, H., Li, W.,** (2011). "Document sentiment classification by exploring description model of topical terms". *Computer Speech and Language*, 25(2), 386-403.

[59] **Medhat, W., Yousef, A. H., Mohamed, H. K.,** (2008). "Combined Algorithm for Data Mining Using Association Rules". *Ain Shams Journal of Electrical Engineering.*

[60] **Ortigosa-Hernández, J., Rodríguez, J. D., Alzate, L., Lucania, M., Inza, I., Lozano, J. A.,** (2012). "Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers". *Neurocomputing*, 92, 98-115.

[61] **He, Y., Zhou, D.,** (2011). "Self-training from labeled features for sentiment analysis". *Information Processing and Management: An International Journal*, 47(4), 606-616.

[62] **Read, J., Carroll, J.,** (2009). "Weakly supervised techniques for domain-independent sentiment classification". *In Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion (TSA '09)*, 45-52.

[63] **Youngjoong, K., Jungyun, S.,** (2000). "Automatic text categorization by unsupervised learning". *Proceedings of the 18th Conference on Computational Linguistics - Volume 1 (COLING'00)*, 1, 453-459.

[64] **Xianghua, F., Guo, L., Yanyan, G., Zhiqiang, W.,** (2013). "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon". *Knowledge-Based Systems*, 37, 186-195.

[65] **Turney, P. D.,** (2002). "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews". *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, 417-424.

[66] **Jian, J., Yanquan, Z.,** (2011). "Sentiment Polarity Analysis based multi-dictionary". *2011 International Conference on Physics Science and Technology (ICPST'11)*.

[67] **Fahrni, A., Klenner, M.,** (2008). "Old Wine or Warm Beer: Target-Specific Sentiment Analysis of Adjectives". *Proceedings of the Symposium on Affective Language in Human and Machine (AISB)*, 60–63.

[68] **Kim, S., Hovy, E.,** (2004). "Determining the Sentiment of Opinions". *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, 1367–1373.

[69] **Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., Potts, C.,** (2011). "Learning word vectors for sentiment analysis". *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*, 1, 142-150.

[70] **Martineau, J., Finin, T.,** (2009). "Delta TFIDF: An improved feature space for sentiment analysis". *Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media.*

[71] **Wilson, T., Wiebe, J., Hoffmann, P.,** (2005). "Recognizing contextual polarity in phrase-level sentiment analysis". *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*, 347-354.

[72] **Baccianella, S., Esuli, A., Sebastiani, F.,** (2010). "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining". *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10).*

[73] **Manning, C. D., Schütze, H.,** (1999). "Foundations of Statistical Natural Language Processing". ISBN:0-262-13360-1.

[74] **Berger, A. L., Della Pietra, V. J., Della Pietra, S. A.,** (1996). "A maximum entropy approach to natural language processing". *Computational Linguistics*, 22(1), 39-71.

[75] **Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., Potts, C.,** (2013). "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, 1631–1642.

[76] "SentiNEL: Sentiment Analysis from Tweets". https://github.com/D2KLab/sentinel, *alındığı tarih: 18 Eylül 2019.*

[77] **Li, P., Xu, W., Ma, C., Sun, J., Yan, Y.,** (2015). "IOA: Improving SVM Based Sentiment Classification Through Post Processing". *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 545-550.

[78] **Uysal, A. K., Gunal, S.,** (2014). "The impact of preprocessing on text classification". *The impact of preprocessing on text classification*, 50(1), 104-112.

[79] **Vijayarani, S., Ilamathi, J., Nithya.,** (2015). "Preprocessing Techniques for Text Mining - An Overview". *International Journal of Computer Science & Communication Networks*, Vol 5(1), 7-16.

[80] **Beeferman, D., Berger, A., Lafferty, J.,** (1999). "Statistical models for text segmentation". *Machine Learning - Special Issue on Natural Language Learning*, 34(1-3), 177-210.

[81] **Kozima, H.,** (1993). "Text segmentation based on similarity between words". *Proceedings of the 31st annual meeting on Association for Computational Linguistics (ACL '93)*, 286-288.

[82] **Marquez, L., Rodriguez, H.,** (1998). "Part-of-speech tagging using decision trees". *Proceedings of the 10th European Conference on Machine Learning (ECML'98)*, 25-36.

[83] **Ratnaparkhi, A.,** (1996). "A Maximum Entropy Model for Part-Of-Speech Tagging". *Conference on Empirical Methods in Natural Language Processing.*

[84] **Brill, E.,** (1992). "A simple rule-based part of speech tagger". *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLC '92)*, 152 - 155.

[85] **Silva, C., Ribeiro, B.,** (2003). "The importance of stop word removal on recall values in text categorization". *Proceedings of the International Joint Conference on Neural Networks*, 3, 1661-1666.

[86] **Dolamic, L., Savoy, J.,** (2010). "When stopword lists make the difference". *Journal of the American Society for Information Science and Technology*, 61(1), 200-203.

[87] **Lo, R. T., He, B., Ounis, I.,** (2005). "Automatically Building a Stopword List for an Information Retrieval System". *Journal of Digital Information Manage (JDIM)*, 3, 3-8.

[88] **Zou, F., Wang, F. L., Deng, X., Han, S., Wang, L. S.,** (2006). "Automatic construction of Chinese stop word list". *Proceedings of the 5th WSEAS International Conference on Applied Computer Science (ACOS'06)*, 1009-1014.

[89] **Porter, M.** "Snowball: A language for stemming algorithms". (2001), http://snowball.tartarus.org/texts/introduction.html, *alındığı tarih: 18 Eylül 2019.*

[90] **Willett, P.,** (2006). "The Porter Stemming Algorithm: Then and Now". *Electronic Library and Information Systems*, 40, 219-223.

[91] **Lovins, J. B.,** (1968). "Development of a stemming algorithm". *Mechanical Translation and Computational Linguistics*, 11, 22-31.

[92] **Hull, D. A.,** (1996). "Stemming algorithms: A case study for detailed evaluation". *Journal of the American Society for Information Science*, 47(1), 70-84.

[93] **Lee, D. D., Seung H. S.,** (2000). "Algorithms for Non-negative Matrix Factorization". *Proceedings of the 13th International Conference on Neural Information Processing Systems (NIPS'00)*, 535-541.

[94] **Gillis, N.,** (2014). "The Why and How of Nonnegative Matrix Factorization". *Regularization, Optimization, Kernels and Support Vector Machines*, ISBN: 9781482241402, 257-291.

[95] **Lee, D. D., Seung H. S.,** (1999). "Learning the parts of objects by non-negative matrix factorization". *Nature*, 401, 788-791.

[96] **Luo, X., Zhou, M., Xia, Y., Zhu, Q.,** (2014). "An Efficient Non-Negative Matrix-Factorization-Based Approach to Collaborative Filtering for

Recommender Systems". *IEEE Transactions on Industrial Informatics*, 10(2), 1273 - 1284.

[97] **Xu, W., Liu, X., Gong, Y.,** (2003). "Document Clustering Based on Non-negative Matrix Factorization". *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03)*, 267-273.

[98] **Dumais, S. T.,** (2005). "Latent semantic analysis". *Annual Review of Information Science and Technology*, 38(1), 188–230.

[99] **Golub, G. H., Reinsch, C.,** (1971). "Singular Value Decomposition and Least Squares Solutions". *Linear Algebra*, ISBN: 978-3-662-38854-9, 134-151.

[100] **Ramos, J. E.,** (2003). "Using TF-IDF to Determine Word Relevance in Document Queries". *1st International Conference on Machine Learning.*

[101] **Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman R.,** (1990). "Indexing by latent semantic analysis". *Journal of the American Society for Information Science*, 41(6), 391-407.

[102] **Hofmann, T.,** (1999). "Probabilistic latent semantic analysis". *Proceedings of the 15th Conference on Uncertainty in AI*, 289–296.

[103] **Hofmann, T.,** (2001). "Unsupervised Learning by Probabilistic Latent Semantic Analysis". *Machine Learning*, 42(1-2), 177-196.

[104] **Brants, T., Chen, F., Tsochantaridis, I.,** (2002). "Topic-based Document Segmentation with Probabilistic Latent Semantic Analysis". *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM'02)*, 211-218.

[105] **Blei, D. M., Ng, A. Y., Jordan, M. I.,** (2003). "Latent Dirichlet allocation". *Journal of Machine Learning Research*, 3, 993-1022.

[106] **Griffiths, T. L., Steyvers, M.,** (2004). "Finding scientific topics". *Proceedings of the National Academy of Science*, 101, 5228–5235.

[107] **Blei, D. M., McAuliffe, J. D.,** (2007). "Supervised Topic Models". *Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'07)*, 121-128.

[108] **Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.,** (2013). "Distributed Representations of Words and Phrases and Their Compositionality". *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13)*, 2, 3111-3119.

[109] **Mikolov, T., Chen, K., Corrado, G., Dean, J.,** (2013). "Efficient Estimation of Word Representations in Vector Space". *Proceedings of Workshop at International Conference on Learning Representations*.

[110] **Moody, C. E.,** (2016). "Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec".

[111] "Gensim Library". https://radimrehurek.com/gensim, *alındığı tarih: 18 Eylül 2019.*

[112] **Rehurek, R., Sojka, P.,** (2010). "Software Framework for Topic Modelling with Large Corpora". *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45-50.

[113] **Wasserman, S., Faust, K.,** (1994). "Social Network Analysis: Methods and Applications". ISBN: 978-0-521-38707-8.

[114] **Otte, E., Rousseau, R.,** (2002). "Social Network Analysis: A powerful strategy, also for the information sciences". *Journal of Information Science*, 28(6), 441-453.

[115] **Kubica, J., Moore, A., Cohn, D., Schneider, J.,** (2003). "Finding underlying connections: A fast method for link analysis and collaboration queries". *International Conference on Machine Learning (ICML 2003).*

[116] **Staddon, J.,** (2009). "Finding hidden connections on LinkedIn an argument for more pragmatic social network privacy". *Proceedings of the 2nd ACM Workshop on Security and Artificial Intelligence (AISec '09)*, 11-14.

[117] **Freeman, L. C.,** (1978). "Centrality in social networks conceptual clarification". *Social Networks*, 1(3), 215-239.

[118] **Cross, R., Borgatti, S. P., Parker, A.,** (2002). "Making Invisible Work Visible: Using Social Network Analysis to Support Strategic Collaboration". *California Management Review*, 44(2), 25-46.

[119] "Pajek: Program for Large Network Analysis". http://mrvar.fdv.uni-lj.si/pajek/, *alındığı tarih: 18 Eylül 2019.*

[120] **Batagelj, V., Mrvar, A.,** (2013). "Pajek - Analysis and Visualization of Large Networks". *Graph Drawing Software*, ISBN: 978-3-642-18638-7, 77-103.

[121] **Johnson, S. C.,** (1967). "Hierarchical clustering schemes". *Psychometrika*, 32(3), 241–254.

[122] **Savaresi, S. M., Boley, D. L., Bittanti, S., Gazzaniga, G.,** (2002). "Cluster Selection in Divisive Clustering Algorithms". *2nd SIAM International Conference on Data Mining*, 299-314.

[123] **Gowda, K. C., Krishna, G.,** (1978). "Agglomerative clustering using the concept of mutual nearest neighbourhood". *Pattern Recognition*, 10(2), 105-112.

[124] **Murtagh, F., Legendre, P.,** (2014). "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?". *Journal of Classification*, 31(3), 274–295.

[125] **Bonacich, P.,** (1987). "Power and Centrality: A Family of Measures". *American Journal of Sociology*, 92(5), 1170-1182.

[126] **Meo, P. D., Ferrara, E., Fiumara, G., Provetti, A.,** (2011). "Generalized Louvain Method for Community Detection in Large Networks". *11th International Conference on Intelligent Systems Design and Applications.*

[127] **Page, L., Brin, S., Motwani, R., Winograd, T.,** (1998). "The PageRank citation ranking: Bringing order to the web". *Technical Report, Stanford Digital Library Technologies Project.*

[128] **Xiao, N., Xu, Q.,** (2015). "Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection". *Journal of Statistical Computation and Simulation*, 85(18), 3755-3765.

[129] **Levinson, N.,** (1946). "The Wiener (Root Mean Square) Error Criterion in Filter Design and Prediction". *Journal of Mathematics and Physics*, 25(1-4), 261-278.

ÖZGEÇMİŞ

Ad – Soyad **:** Onur Can SERT

Uyruğu **:** T.C.

Doğum Tarihi ve Yeri **:** 29.05.1988 Altındağ / Ankara

E-posta **:** onurcansert@gmail.com

ÖĞRENİM DURUMU:

- **Lisans** **:** 2010, TOBB Ekonomi ve Teknoloji Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği
- **Yüksek Lisans** **:** 2012, TOBB Ekonomi ve Teknoloji Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği
- **Doktora** **:** 2020, TOBB Ekonomi ve Teknoloji Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği

MESLEKİ DENEYİM VE ÖDÜLLER:

Yıl	Yer	Görev
2018 – …	Accenture – The Dock	Yazılım Mühendisi
2018 – 2014	Viveka	Yazılım Mühendisi
2014 – 2010	TOBB Ekonomi ve Teknoloji Üniversitesi	Tam Burslu Lisansüstü Öğrencisi

YABANCI DİL: İngilizce

TEZDEN TÜRETİLEN YAYINLAR, SUNUMLAR VE PATENTLER:

- Ceyhan, K., Kurtulmaz, E., **Sert, O. C.** ve Özyer, T., (2018). "Bitcoin movement prediction with text mining". *26th Signal Processing and Communications Applications Conference (SIU 2018),* 2-5 Mayıs 2018, İzmir, Türkiye.
- **Sert, O. C.**, Şahin, S. D., Özyer, T. ve Alhajj, R., (2020). "Analysis and prediction in sparse and high dimensional text data: The case of Dow Jones stock market". *Physica A: Statistical Mechanics and its Applications*, 545.

DİĞER YAYINLAR, SUNUMLAR VE PATENTLER:

- **Sert, O. C.,** Dursun, K. ve Özyer, T., (2011). "Ensemble of Multi-objective Clustering Unified with H-Confidence Metric as Validity Metric". *The 2011 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM 2011)*, 25-27 Temmuz 2011, Kaohsiung, Tayvan.
- **Sert, O. C.,** Dursun, K., Özyer, T., Jida, J. ve Alhajj, R., (2012). "The Unification and Assessment of Multi-Objective Clustering Results of Categorical Datasets with H-Confidence Metric". *J.UCS: Journal of Universal Computer Science*, 18(4), 507-531.
- Erdoğan, A. E., Yılmaz, T., **Sert, O. C.**, Akyüz M., Özyer, T. ve Alhajj, R., (2017). "From Social Media Analysis to Ubiquitous Event Monitoring: The case of Turkish Tweets". *The 2017 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM 2017)*, 31 Temmuz – 3 Ağustos 2017, Sydney, Avustralya.
- Dastjerd, N. K., **Sert, O. C.**, Özyer, T. ve Alhajj, R., (2019). "Fuzzy Classification Methods Based Diagnosis of Parkinson's disease from Speech Test Cases". *Current Aging Science*, 12(2), 100-120.