
GAZİ ÜNİVERSİTESİ
Gazi
1926

TÜRKİYE BÜYÜK MİLLET MECLİSİ GENEL KURUL TUTANAKLARININ YAPAY ZEKA TABANLI METİN ANALİZİ

Mesut KÖRPE

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

HAZİRAN 2023

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Mesut KÖRPE
19/06/2023

TÜRKİYE BÜYÜK MİLLET MECLİSİ GENEL KURUL TUTANAKLARININ YAPAY ZEKA TABANLI METİN ANALİZİ

(Doktora Tezi)

Mesut KÖRPE

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Haziran 2023

ÖZET

Türkiye Büyük Millet Meclisi (TBMM) Genel Kurulunda görüşmelerin yazılı tutanak haline getirilmesi ile yıllar boyunca binlerce sayfa metin verisi oluşmuştur. Bu tezde, TBMM Genel Kurul tutanaklarının yapay zeka tabanlı metin analizi üç farklı konuda gerçekleştirilmiştir. Kutuplaşma ölçütü olarak metin sınıflandırması doğruluk değerinin kullanıldığı birinci çalışmada siyasi partilerin kutuplaşması incelenmiştir. Parlamento tutanakları ile siyasi parti kutuplaşması çalışmaları zaman aralığı olarak bir yılı kullanır. Bu yaklaşım ile yıl içinde oluşan kutuplaşmayı değerlendirmek ve zaman serisi analizi yapmak mümkün değildir. Çalışmada önerilen 12-aylık hareketli kutuplaşma ölçütü ile parlamento tutanaklarından aylık temelde kutuplaşma ölçümü olanaklı olmuştur. Aylık ölçümler 2011-2023 yılları arasında kutuplaşmalarda en çok öne çıkan unsurların ideolojiler ve seçim ittifakları olduğunu göstermiştir. İkinci çalışmada konuşmalar milletvekillerinin demografik ve siyasi aidiyetlerine göre analiz edilmiştir. Çalışma, klasik makine öğrenimi ve derin öğrenme tekniklerinin performansını karşılaştırmaktadır. Demografik özelliklerin konuşma içeriğine yansıması hata analizi ve en iyi özellik analizi ile incelenmiştir. Sınıflandırma sonuçlarına göre milletvekilleri için en ayırt edici özellikler parti durumu aidiyeti (%92), parti aidiyeti (%84), cinsiyet (%82) ve meslektir (%67). Üçüncü çalışmada word2vec, GloVe ve fastText kelime yerleştirme algoritmaları karşılaştırılmıştır. Kelime vektörleri benzerlik değerleri kullanılarak illerinin, ülkelerin ve kabine üyelerinin kendi aralarındaki benzerlikleri incelenmiştir. Çalışma sonucuna göre iller ve ülkelerin kelime benzerliği coğrafi komşuluklarına, ekonomik, sosyal ve tarihi yakınlıklarına göre artmaktadır. Kabine üyelerinin kelime benzerliğini ise aynı konuda veya yakın tarihli kabinelerde bakanlık yapmaları artırmaktadır. Üç çalışma birlikte değerlendirildiğinde parlamento tutanaklarının makine öğrenmesi, derin öğrenme, doğal dil işleme gibi yapay zeka tabanlı yöntemler ile analizi toplumun yararına değerli bilgiler sağlar.

Bilim Kodu	:	92432
Anahtar Kelimeler	:	Parlamento tutanakları, TBMM, makine öğrenmesi, kutuplaşma, yazar analizi, kelime yerleştirme
Sayfa Adedi	:	155
Danışman	:	Doç. Dr. Hüseyin POLAT

ARTIFICIAL INTELLIGENCE BASED TEXT ANALYSIS OF GRAND NATIONAL ASSEMBLY OF TÜRKİYE PLENARY SESSION MINUTES

(Ph. D. Thesis)

Mesut KÖRPE

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

June 2023

ABSTRACT

The transcription of the debates in the plenary session of the Grand National Assembly of Türkiye (GNAT) has generated thousands of pages of textual data over the years. In this thesis, artificial intelligence-based text analysis of the minutes of GNAT was carried out on three different topics. In the first study, in which the accuracy of text classification was used as a measure of polarization, the polarization of political parties was examined. Political party polarization studies with parliamentary minutes use one year as a time interval. In this approach, it is impossible to perform a time series analysis and evaluate polarization over a year. Using the 12-month moving polarization measure presented in the research, it is possible to evaluate polarization from parliamentary minutes on a monthly basis. Monthly measures showed that the most prominent aspects on polarizations are ideologies and electoral alliances between 2011 and 2023. In the second study, the speeches were analyzed according to the demographics and political affiliations of the deputies. The study compares the performance of classical machine learning and deep learning techniques. The reflection of demographic features on speech content is analyzed by error analysis and best feature analysis According to the classification accuracies, the most distinguished characteristics of MPs are party status affiliation (92%), party affiliation (84%), gender (82%), and occupation (67%). In the third study, word embedding algorithms Word2vec, GloVe, and fastText were compared. Using word vector similarity values, the similarities between provinces, countries, and cabinet members were examined. According to the results of the study, the word similarity of provinces and countries increases according to their geographical neighborhood, economic, social and historical proximity. The word similarity of the cabinet members is increased by the fact that they were ministers in the same subject or recent cabinets. When the three studies are evaluated together, valuable information for the benefit of society can be obtained by analyzing parliamentary minutes with artificial intelligence based methods such as machine learning, deep learning, and natural language processing.

Science Code : 92432

Key Words : Parliamentary debates, GNAT, machine learning, polarization, authorship analysis, word embedding

Page Number : 155

Supervisor : Assoc. Prof. Dr. Huseyin POLAT

TEŞEKKÜR

Araştırma ve tez çalışmamda beni cesaretlendiren, yönlendiren ve destekleyen danışmanım Doç. Dr. Hüseyin POLAT hocama minnettarım. Tezin her aşamasında sabırla ve içtenlikle yardım ettiği için çok teşekür ederim. Kendisiyle çalışmaktan onur duydum.

Tez izleme komitemde yer alarak tez çalışmama sundukları değerli katkılar ve destekleri için Sayın Prof. Dr. Necaattin Barışcı ve Sayın Doç. Dr. Baha ŞEN hocalarıma saygılarımı ve sonsuz şükranlarımı sunarım. Tez çalışmam boyunca hocalarımın yapıcı yönlendirmelerinden ve engin bilgi birikimlerinden faydalandım.

Tezin gözden geçirilip tez yazım biçimine uygun hale getirilmesinde büyük emekleri olan çalışma arkadaşım Gülcan YILDIZ'a, çalışmanın siyasi sonuçlarının değerlendirmesinde katkı sunan Dr. Betül AYDOĞAN ÜNAL'a yardımları için teşekkürlerimi sunarım.

Tüm hayatım boyunca bana destek olan tüm aileme; Babam İsmet KÖRPE'ye, annem Zeynep KÖRPE'ye ve dostlarıma sonsuz teşekür ederim.

Yoğun doktora çalışmalarım esnasında büyük fedakârlıklarda bulunarak her zaman yanımda olan sevgili eşim Ayşegül KÖRPE'ye teşekkür ederim.

Canım oğullarım İsmet Eren ve İlyas Erdem'e…

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil **Sayfa**

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklamalar
ADF	Augmented Dickey-Fuller
AK Parti	Adalet ve Kalkınma Partisi
ANN	Artificial Neural Network
AP	Author Profiling
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag of Words
CBOW	Continous Bag of Words
CHP	Cumhuriyet Halk Partisi
CHS	Cumhurbaşkanlığı Hükümet Sistemi
CNN	Convolution Neural Network
CMP	Comparative Manifesto Project
DBOW	Distributed Bag Of Words
DDİ	Doğal Dil İşleme
DÖ	Derin Öğrenme
DL	Deep Learning
DTW	Dynamic Time Warping
DVM	Destek Vektör Makinaları
DZB	Dinamik Zaman Bükmesi
ESA	Evrişimli Sinir Ağı
FFNN	Feed Forward Neural Networks
GNAT	Grand National Assembly of Türkiye
HDP	Halkların Demokratik Partisi
İBSA	İleri Beslemeli Sinir Ağları
LDSE	Low-Dimensionality Statistical Embedding
LR	Lojistik Regresyon
LSA	Latent Semantic Analysis

Kısaltmalar	Açıklamalar
LSTM	Long Short-Term Memory
MÖ	Makine Öğrenimi
ML	Machine Learning
MHP	Milliyetçi Hareket Partisi
NLP	Natural Language Processing
OOV	Out-Of-Vocabulary
PCA	Principal Component Analysis
PV	Paragraf Vektörleri
SVD	Singular Value Decomposition
SVM	Support Vector Machines
SDK	Sözlük Dışı Kelime
TBA	Temel Bileşen Analizi
TF-IDF	Term-Frequency Inverse Document Frequency
TBMM	Türkiye Büyük Millet Meclisi
TSA	Tekrarlayan Sinir Ağları
UKSB	Uzun Kısa Süreli Bellek
YPO	Yazar Profili Oluşturma
YSA	Yapay Sinir Ağları

1. GİRİŞ

İnternetin yaygınlaşması ve verilerin dijitalleşmesi ile birlikte veri kaynaklarının miktarında ve çeşitliliğinde büyük oranda artış meydana gelmiştir. Bu verilerin ortaya çıkması, Grafik İşlemci Üniteleri (Graphics Processing Unit, GPU) gibi hesaplama kaynaklarının ucuzlaması, hesaplama ve büyük veri kaynaklarına erişimin kolaylaşması ile devasa veri miktarına ihtiyaç duyan klasik makine öğrenmesi, derin öğrenme gibi yöntemlerinin verimli kullanılmasının önü açılmıştır. Kelime vektörleri, büyük dil modelleri gibi Doğal Dil İşleme (DDİ) alanları bu büyük veri ve derin öğrenme devriminin bir sonucu olarak ortaya çıkmıştır.

Parlamentoların görevi yasama, denetim ve temsil işlevleridir. Parlamentolar yasama işlevini toplumsal ihtiyaçları karşılayan, sosyal ve kamusal ilişkileri düzenleyen yasaları çıkararak, değiştirerek ve yürürlükten kaldırarak yerine getirir. Denetim görevini ise yürütme organının yasalara uygun, toplum yararına kararlar ve eylemler almasını kontrol ederek yapar.

Vatandaşların milletvekilleri aracılığı ile temsil edildiği parlamentolarda seçilmiş olan milletvekilleri seçmenlerine karşı sorumludur. Seçmenlerin çıkarlarının korunduğu, görüşlerinin temsil edildiği yer parlamentolardır. Parlamentoların kanun yapma, hükümeti denetleme, vatandaşı temsil etme görevlerinin görüşüldüğü, tartışıldığı, karar bağlandığı nihai yer ise genel kurullardır. Birçok parlamentoda genel kurul konuşmaları transkript edilip metin verisi olarak parlamentoların İnternet sayfasında vatandaşların erişimine sunulur. Parlamento görüşmelerinin otomatik analizi için oluşturulmuş veri kümesi örnekleri mevcuttur (Erjavec ve diğerleri, 2023; Odell, 2020/2022; *parlparse,* 2013/2023). Yıllar içinde artarak büyük veri tanımına uyacak boyutlara gelen bu metin verisinin analiz edilip yorumlanması ve anlamlandırılması ile toplumun yararına değerli bilgiler ortaya çıkabilir. Büyük metin verisinin miktarı insan emeğini aşan yöntemler ile analiz edilmeyi gerektirmektedir. Büyük metin verisinin istatiksel yöntemler, makine öğrenmesi, derin öğrenme, doğal dil işleme gibi hesaplamalı yaklaşımlarla analiz edilmesi ve anlamlı bilgilerin ortaya çıkarılıp yorumlanması Hesaplamalı Metin Analizinin ana konusudur. Bu çalışmalar alan yazında veri olarak metin veya metinden veriye (Grimmer ve Stewart, 2013) olarak da adlandırılmaktadır. Yapay zekaya dayalı yöntemler büyük metin verisinin anlamlı bilgilere dönüştürülmesi için çok elverişlidir.

Konuşmalardan milletvekillerinin duygu analizi, belirli bir konudaki tutumlarının tahmin edilmesi, genellikle parti etiketinin ideolojiyle eşleştirilmesi sonucu ortaya çıkan ideoloji tespiti, siyasi partilerin veya milletvekillerinin kutuplaşması veya pozisyon alması, parlamento tartışmaları kullanılarak en sık yapılan çalışmalardır (Abercrombie ve Batista-Navarro, 2019). Dzieciątko (2019) metin sınıflandırması ile Polonya Parlamentosu tutanaklarından milletvekillerinin mutluluk, öfke, üzgünlük gibi duygu durumlarını çıkarmışlardır. Eskişar ve Çöltekin (2022) TBMM Genel Kurul tutanaklarında partilere ait milletvekillerinin korku, üzgünlük, mutluluk, tiksinti gibi duygularını analiz etmişlerdir. Rudkowsky ve diğerleri (2018)parlamento tutanaklarında kelime vektörlerini kullanarak metinlerin duygu analizini pozitif negatif olarak sınıflandırmışlardır.

Parlamento tutanaklarının hesaplamalı metin analizi ile belirgin konuların ele alındığı çalışmalar vardır. Müller-Hansen ve diğerleri (2021) 70 yıllık parlamento konuşmalarını termal enerji ve madencilik konusu bağlamında konu modeli ile analiz etmiştir. Fraccaroli ve Giovannini (2020) parlamento konuşmalarında merkez bankaları (Bank of England, the European Central Bank ve Federal Reserve) ile ilgili konuşmalara odaklanmış ve ülkelerde belirsizlik ve enflasyon artıkça duygu durumunun negatif olduğu sonucuna varmışlardır. Frid-Nielsen (2018) milletvekili konuşmalarından Avrupa parlamentosunda mülteci konusundaki tutumlarını analiz etmişlerdir.

Çalışmada TBMM Genel Kurul tutanaklarının yapay zeka tabanlı metin analizi üç ayrı çalışma ile gerçekleştirilmiştir.

Birinci çalışmada metin sınıflandırması ile siyasi partilerin kutuplaşmasının aylık olarak ölçülmesi için yeni bir yöntem önerilmiştir. Çalışmada kutuplaşma ifadesi ve ölçüsü parti çiftleri arasındaki ilişkiyi belirtir.

Parlamento tartışmaları üzerine veri olarak metin çalışmalarında, parti kutuplaşması ve ideoloji tespiti kilit görevlerden biridir. Parti kutuplaşması, siyasi konularda elitlerin veya kitlelerin birbirlerine karşı aldıkları pozisyonların çıkarılması olarak ifade edilebilir. Hesaplamalı parti kutuplaşmasında kullanılan yöntemler çeşitlilik göstermektedir. Denetimsiz metin ölçekleme yöntemlerinin yanı sıra denetimli makine öğrenmesi, metin sınıflandırma ve derin öğrenme yöntemleri parti kutuplaşması veya ideoloji tespiti için yaygın olarak kullanılmaktadır.

Zaman serisi verileri, birbirini izleyen zaman aralıklarındaki veri noktaları dizisidir. Zaman serilerinin analizi değişikliklerin izlenmesini, örüntülerin tanınmasını ve veri noktalarının sonraki adımlarının tahmin edilmesini sağlar. Kutuplaşma çalışmalarında, ölçüm ayrık takvim yılı verilerine dayanır. Bu yaklaşım, yıllar veya on yıllar boyunca kutuplaşmayı incelemekte fayda sağlar ancak yıl içindeki kutuplaşmayı incelemek için yetersizdir. Ayrıca yıllık kutuplaşma ölçümü sonucu elde edilen veri noktalarının sayısının azlığı, kutuplaşmanın zaman serilerini analiz etmek için de bir zorluk oluşturur.

Büyük miktarda veri ve Uygulama Programlama Arabirimleri (UPA) sayesinde, Twitter gibi sosyal medya verileriyle aylık, haftalık ve hatta günlük veri kümeleri oluşturulabilir. Bu veriler bir yıldan daha kısa aralıklarla kutuplaşmayı ölçmek için kullanılabilir. Buna karşın, parlamento tartışmaları veri kümelerinde yeterli miktarda ayrık aylık veri bulunmamakta ve veri kümelerinin boyutunda değişkenlik görülmektedir. Parlamento görüşmeleri veri kümelerinin bu yapısı, kutuplaşmanın kesikli takvim ayları veri kümeleri gibi ölçülmesini zorlaştırmaktadır.

McCleary, Hay, Meidinger ve McDowall (1980: 331)'a göre zaman serisi analizi için minimum gereklilik 40-50 zaman gecikmesidir. Hanke ve Wichern (2013: 80) zaman serileri için gerekli veri noktası sayısını mevsimsel periyodun iki ila altı katı olarak formüle etmiştir. Aylık ölçümlerde mevsimsel periyot 12 olduğuna göre 40-50 veri noktasına 3 ya da 4 yıllık ölçümle ulaşılabilir ve zaman serisi analizi mümkün olur. Kutuplaşmayı parlamento tartışmalarından ölçerken, 12 aylık hareketli aralık kullanılabilir. Böylece kutuplaşmayı aylık zaman serilerinde gözlemlemek ve analiz etmek mümkün olur. Bu yaklaşımın bir sonucu olarak, kutuplaşmanın bir yıl içindeki seyrini ve aylık zaman serileri analizini olanaklı hale getirir. Bu çalışmada, hareketli bir gösterge olarak noktadan noktaya 12 aylık kutuplaşma ölçütünü kullanılmış ve bu yaklaşım parlamento alanına uyarlanmıştır.

Çalışmada, siyasi partilerin pozisyonları, lojistik regresyon makine öğrenmesi algoritmaları ile parti üyeliği sınıflandırması kullanılarak belirlenmiştir. Yüksek sınıflandırma doğruluğu partilerin ayırt edilebilirliğini gösterdiği için yüksek kutuplaşma anlamına gelir Her sınıflandırma görevi, Türkiye Büyük Millet Meclisi'ndeki bir yıllık genel kurul konuşmalarından iki partinin ikili sınıflandırmasıdır. Bir parti çiftinin zaman serisini elde etmek için her ay için bir yıllık görüşmelerden oluşan bir alt veri kümesi oluşturulmuştur.

Kutuplaşma farklılığına konuşma içeriğinin yanı sıra veri kümesinin yapısı da neden olabilir. Veri kümesi yapısının aylık kutuplaşma üzerindeki etkisini regresyon analizi kullanarak ortadan kaldırılmıştır. Regresyon analizinde negatif binom regresyonu kullanılmış, bağımlı değişken ardışık aylardaki kutuplaşma farkının mutlak değeri, bağımsız değişken ardışık iki veri seti miktarının Jaccard mesafesidir.

Çalışmada regresyon analizi sonucu elde edilen nihai zaman serilerini kullanarak bir parti çiftinin diğeri üzerindeki etkisini araştırılmıştır. Zaman serileri arasındaki şekil benzerliği bize nedensellik konusunda bir ipucu vermektedir. İki zaman serisi arasındaki benzerlik Dinamik Zaman Bükmesi (DZM) (Dynamic Time Warping, DTW) algoritması ile ölçülmüştür (Berndt ve Clifford, 1994).

İki değişken arasında korelasyonun olması nedenselliğin olduğu anlamına gelmez. Bir değişken diğer bir değişkeni tahmin etmede yardımcı oluyorsa bu değişkenin Granger-nedeni (Granger, 1969) olur. İki parti çifti kutuplaşması arasındaki nedenselliğin olup olmadığı, nedensellik varsa bu nedenselliğin yönü Granger'ın nedensellik algoritması ile belirlenmiştir.

İkinci çalışmada metin sınıflandırması ve en iyi özellik analizi ile milletvekillerinin demografik özellikleri analiz edilmiştir.

Demografik özelliklerin tahmini, metin sahiplerinin yaş, cinsiyet, eğitim, meslek, görüş, siyasi görüş gibi demografik özelliklerine odaklanır. Yazarın demografik veya kişisel özelliklerin metinden tahmin edilmesine Yazar Profili Oluşturma (YPO) (Author Profiling, AP) adı verilir. Doğal dil işleme (DDİ) (Natural Language Processing, NLP), klasik makine öğrenimi (MÖ (Machine Learning, ML) ve derin öğrenme (DÖ) (Deep Learning, DL) yaklaşımlarının birlikte kullanımı, metinden demografik özelliklerin tahmin edilmesi için kullanılabilir. Bu alandaki çalışmaların çoğu probleme, metin sahiplerinin özelliklerinin önceden tanımlanmış sınıflar olduğu denetimli metin sınıflandırması olarak yaklaşmaktadır.

Demografik özelliklerin tahmini adli tıp, pazarlama ve diğer alanlar için çok önemli olabilir (Rangel, Rosso, Potthast, Stein ve Daelemans, 2014) . Adli tıpta, metin sahiplerinin cinsiyet ve yaş gibi özellikleri bir soruşturma için kanıt olabilir. Pazarlama alanında ise müşterilerin demografik özelliklerinin bilinmesi şirketlerin pazarlama politikalarını belirlemede yardımcı

olur. Politikacılar için, kamuoyunun belirli konulardaki görüşlerini tahmin etmek hükümet politikalarına veya seçim kampanyalarına rehberlik eder.

YPO'da çalışılan veriler ve dil çok yönlü değildir. Çoğu çalışma sosyal medya verilerine dayanmakta ve İngilizce derlemler kullanmaktadır. Bu çalışmada, TBMM Genel Kurul görüşmelerinden milletvekillerinin demografik özellikleri tahmin edilmiş ve demografik özelliklerin konuşma içeriklerine yansıması en iyi özellik analizi ile incelenmiştir. İlk olarak, 2012-2020 yılları arasında TBMM Genel Kurul oturumlarında milletvekillerinin kürsü konuşmalarının transkripsiyonunu içeren bir derlem oluşturulmuştur. Milletvekillerinin demografik özellikleri olarak cinsiyet, yaş, eğitim, meslek, parti üyeliği, parti statüsü üyeliği ve seçim bölgesi belirlenmiştir. Doküman temsili için Kelime Torbası (Bag of Word, BoW), Terim Frekansı-Ters Doküman Frekansı (Term Frequency - Inverse Document Frequency, TF-IDF), Düşük Boyutlu İstatistiksel Gömme (Low-Dimensionality Statistical Embedding, LDSE), Paragraf Vektörleri (PV) ve önceden eğitilmiş kelime vektörleri gibi farklı doğal dil işleme yöntemleri kullanılmıştır. Ardından, her bir demografik özellik için derlemden elde edilen alt veri kümeleri oluşturmuştur. Bu alt veri kümeleri Lojistik Regresyon (LR), Destek Vektör Makinesi (DVM), İleri Beslemeli Sinir Ağı (İBSA) (Feed Forward Neural Networks, FFNN) ve BERT (Bidirectional Encoder Representations from Transformers) modeli gibi makine öğrenmesi ve derin öğrenme algoritmaları ile eğitilmiştir.

Parlamento görüşmeleri yedi demografik özellik ile analiz edilmiştir. İlk görevde milletvekillerinin cinsiyeti tahmin edilmiştir. Yaş sınıflandırmasında milletvekilleri *40 yaş altı, 40-50, 50-60 ve 60 yaş üstü* olarak kategorize edilmiştir. Eğitim durumu tahmini *profesör veya doçent, doktora, yüksek lisans* veya *lisans* kategorileri ile yapılmıştır. Meslek sınıflandırmasında *hukuk, ekonomi ve finans, tıp* ve *mühendislik* görev kategorileridir. Seçim bölgesinin demografik özelliklerini elde etmek için milletvekillerinin seksen bir seçim bölgesi yedi coğrafi bölgeye dönüştürülmüştür. Parti üyeliği sınıflandırması için, milletvekilleri 2012-2020 yılları arasında parlamentoda var olan dört siyasi partiye sınıflandırılmıştır. Parti durumu tahmini, görüşmelerin *hükümete* mi yoksa *muhalefet* partisine ait milletvekiline mi olduğunu belirlemek için dokümanların ikili sınıflandırılmasıdır. Tüm görevlerde, sınıflandırma doğruluğu modellerin bir değerlendirmesidir. Yanlış sınıflandırılmış bir örneğin sınıflandırılmış kategoriye yakın olduğu varsayılarak sınıflar arasındaki ilişkileri incelenmiştir.

Özellik seçimi ve TF-IDF değerlerinin sınıflar üzerindeki dağılımını kullanarak her bir özelliğin kategorilerinin en önemli terimleri bulunmuştur. Böylece milletvekili konuşmalarının demografik özelliklerine göre içerik analizi yapılmıştır.

Kutuplaşma ve YPO çalışmaları denetimli yöntemler kullanıldığı için konuşmaların parti aidiyeti ve milletvekillerinin demografik özelliklerine göre etiketlendiği veriye ihtiyaç duyar. Veri için TBMM İnternet sayfasının milletvekilleri profil sayfasındaki özgeçmiş ve konuşma metinleri kullanılmıştır(https://www.tbmm.gov.tr/milletvekili/liste). Bu veri 2011 yılından itibaren mevcuttur ve oluşturulan derlem 2011-2023 yılını kapsar.

Üçüncü çalışmada kelime yerleştirme algoritmaları kullanılarak elde edilen kelime vektörleri kullanılmıştır. Çalışma etiketli veriye ihtiyaç duymaz ve 1994-2023 yılları arasındaki tutanak metinleri kullanılmıştır.

Kelime yerleştirme; bir kelimenin dağılım hipotezine göre temsil edilmesidir. Dağılım hipotezi yakın anlamlı kelimelerin benzer bağlamda konumlanacağını anlatır (Harris, 1954). Böylece bir kelime bağlamıyla, anlamıyla birlikte sayısallaştırılır. Bu dağılım, analojiler ve kelime benzerlik gibi kelimeler arasındaki ilişkileri anlamak için kelime vektörleri üzerinde cebirsel işlemlerin kullanılmasını mümkün kılar. Kelime yerleştirmede, kelime vektörleri yüksek boyutlu olabilir. Her bir boyut, kelimenin belirli bir özelliğine veya karakteristiğine karşılık gelir.

Dâhili (intrinsic) görevler, kelime yerleştirmede kullanılan kelime vektörlerinin vektör uzayındaki konumundan doğrudan yararlanan kelime benzerliği, ilişkili kelimeler (relatedness), analoji gibi tekil sözcük çiftlerinin ya da sözcük gruplarının arasındaki ilişkileri konu alan görevlerdir. Kelime vektörlerinin kalitesi bu görevlerden elde edilen başarım ile değerlendirilir.

Kelime vektörleri düşünüldüğünde hârici (extrinsic) görevler, bilgi çıkarımı, doküman sınıflandırma, soru-cevap sistemleri, varlık ismi tanıma sistemleri gibi DDİ uygulamalarında bir cümleyi, paragrafı ya da dokümanı oluşturan kelimeleri temsilen kelime vektörleri kullanılır. Kelime vektörlerinin semantiği ve bağlamı temsil yeteneği bu görevlerin performansını önemli ölçüde artırır.

Anlamsal kelime vektörleri (Bojanowski, Grave, Joulin ve Mikolov, 2017; Landauer ve Dumais, 1997; Mikolov, Chen, Corrado ve Dean, 2013; Pennington, Socher ve Manning, 2014) bir sözlüğü oluşturan kelimelerin eğitildiği bütün bir derlemde taşıdığı anlama göre vektör uzayında temsil edilmesini sağlar. Tekil bir kelimenin bir tane vektörle temsil edildiği yaklaşımda yazılışları aynı, anlamları farklı olan eş sesli kelimeler aynı vektörle temsil edilir. Bu durum eş sesli kelimelerin temsilinde anlam kaybına yol açabilir.

Bağlamsal kelime vektörlerinde (Devlin, Chang, Lee ve Toutanova, 2018; Howard ve Ruder, 2018) kelimeler içinde bulunduğu bağlama göre zenginleştirilerek, bağlamdaki anlamı yakalaması sağlanır. Aynı kelime farklı metin dizilerinde farklı vektör değerlerine sahiptir. Kelimenin bağlamı ile birlikte bulunduğu duygu analizi, soru-cevap, makine çeviri gibi hârici görevlerde bağlamsal kelime vektörleri başarımı artırır ancak analoji, kelime benzerliği gibi dâhili görevlerde kelimeler tekil olarak ele alındığı için bağlamı ile birlikte bulunmaz. Bu yüzden çalışmada dâhili görevler için anlamsal kelime vektörleri tercih edilmiştir.

Anlamsal kelime vektörleri ile bir kelimeye yakın anlamlı ya da ilişkili kelimelerin incelendiği çalışmada derlem olarak TBMM Genel Kurul tutanakları kullanılmıştır. Derlem Temmuz 1994-Nisan 2023 yılları arasında TBMM Genel Kurulunda yapılan konuşmalın deşifre edilmiş metinlerden oluşur. Konuşma metinlerinde geçen kelimelerin en uzun madde başları (lema) kullanılarak kelimeler normalize edilmiştir. Bu sayede kelime çeşitliliği azaltılmış ve aynı anlamdaki fakat ek almış kelimeler tek bir temsil ile gösterilmiştir. Ön işlem aşamasından sonra word2vec (Mikolov, Chen, Corrado ve Dean, 2013), GloVe (Pennington, Socher ve Manning, 2014) ve fastText (Bojanowski, Grave, Joulin ve Mikolov, 2017) modelleri ile eğitilen derlemden elde edilen kelime vektörleri eş anlamlı kelimeler, ek alarak türemiş kelimeler ile kelime benzerliği ve ilişkili kelimeleri bulmadaki başarımları ile test edilmiştir. Ek alarak türemiş kelimeler eş anlamlı kelimelere göre morfolojik olarak yakın olduğu için bu iki veri kümesinin sonuçları arasındaki fark, kelime yerleştirme algoritmalarından elde edilen sonuçların dilin morfolojik yapısına göre karşılaştırılmasını sağlamıştır.

Kelime yerleştirme algoritmalarının kelime analojilerini çıkarmadaki başarısı için eş anlamlı ve türemiş kelimelere ilaveten siyasi parti genel başkanları ve siyasi partiler, siyasi parti grup

başkanvekilleri ve siyasi partiler, ülkeler ve başkentlerden oluşan veri kümeleri ile test edilmiştir.

Kelime benzerliği, ilişkili kelimeler ve kelime analojilerinden elde edilen sonuçlarındaki başarım değerlerine göre seçilen word2vec algoritmasından elde edilen kelime vektörleri kullanılarak Türkiye'deki 81 ilin ikili kombinasyonlarının kelime benzerliği incelenmiş, birbirine benzer olan iller arasındaki benzerliği etkileyen etmenler araştırılmıştır. Aynı yöntemle 148 ülkenin ikili kombinasyonları için benzerlik değerleri ile ülke benzerliği incelenmiştir.

Kabine üyelerinin kelime benzerliği için 52-66. hükümetlerde görev alan kabine üyeleri incelenmiştir. Aynı görevi alan ve aynı görevi almayan kabine üyelerinin benzerlik değerleri T testi ile karşılaştırılarak aynı konuda kabinede görev almanın kabine üyeleri arasındaki benzerlik değerlerini artırdığı gözlenmiştir.

İller, ülkeler ve kabine üyelerini aynı düzlemde gösterilmesi için 300 boyutlu kelime vektörleri t-sne(Van der Maaten ve Hinton, 2008) ile 2 boyuta indirgenerek saçılım grafiği ile gösterilmiş ve kelime yakınlığına göre oluşan kavram kümeleri incelenmiştir.

2. İLGİLİ ÇALIŞMALAR

Siyasi kutuplaşma çalışmaları veri kümelerinin içeriklerine göre iki kategoriye ayrılabilir: kitlesel ve elit. En yaygın kitlesel kutuplaşma çalışmaları seçmenlerin sosyal medya verilerine dayanmaktadır. Sosyal medya kullanıcılarından yeterli veri toplayarak, kitlesel siyasi kutuplaşma bir ay (Garimella ve Weber, 2017) veya iki ay aralıklarla (Morales, Borondo, Losada ve Benito, 2015) ölçülebilir. Anket soruları da kitlesel kutuplaşmanın bir başka kaynağıdır (Ertan, Çarkoğlu ve Aytaç, 2022; Torcal, Santana, Carty ve Comellas, 2020).

Elit kutuplaşmasını ölçen birçok çalışma Karşılaştırmalı Manifesto Projesini (Comparative Manifesto Project, CMP) kullanmaktadır. CMP (https://manifesto-project.wzb.eu) 50'den fazla ülkede 1000'den fazla parti için parti seçim manifestosu veri kümeleri sağlamaktadır. CMP 1945'ten 2022'ye kadar geniş bir yıl aralığını kapsar ve on yılların kutuplaşması ele alındığında zaman serisi analizi sunar ancak aylık kutuplaşma ölçütü için uygun değildir ve yalnızca seçim dönemlerini içerir (dört veya beş yıl).

Parlamento tartışmaları, elitlerin ve politikacıların kutuplaşmasını incelemek için yaygın olarak kullanılan bir başka veridir. Otomatik parti üyeliğinin bulunması; ideoloji ve kutuplaşma tespiti için yaygın olarak çalışılmıştır.

Goet (2019) on yıllar boyunca Birleşik Krallık parlamentosundaki milletvekillerinin konuşmalarından parti kutuplaşmasını ölçmüştür. Goets, kutuplaşmayı ölçmek için wordfish (Slapin ve Proksch, 2008) ve wordshoal (Lauderdale ve Herzog, 2016) gibi denetimsiz metin ölçekleme yöntemlerini ve denetimli makine öğrenimi yöntemlerini karşılaştırmıştır. Çalışmada kullanılan parlamento görüşmeleri 1811'den 2019'a kadar 6,2 milyondan fazla konuşmayı kapsamaktadır. Kutuplaşma sonuçlarıyla ilgili olarak, metin ölçekleme yaklaşımları yerine denetimli makine sınıflandırmasını önerilmiştir. Nanni, Glavas, Ponzetto ve Stuckenschmidt (2019) politik metin ölçekleme yöntemleri ile kelimelerin semantiğini sağlayan kelime yerleştirmeleri (Mikolov, Sutskever, Chen,Corrado ve Dean, 2013) kullanmış ve bu sayede kelimelerin sıklığına ek olarak kelimelerin semantiğini de elde etmişlerdir. Rheault, Beelen, Cochrane ve Hirst (2016) kelime yerleştirme (Pennington, Socher ve Manning , 2014) ve alana özgü sözlük kullanarak on yıllar boyunca duygusal kutuplaşmayı ölçmüştür. (Sakamoto ve Takikawa, 2017) çalışmalarında siyasi partilerin

belirli konulardaki pozisyonlarını belirlemiştir. ABD ve Japon parlamentolarındaki konuşmaları analiz etmişler ve yıllar boyunca parlamentolardaki kutuplaşmaları karşılaştırmışlardır.

Hix ve Høyland (2013) parti kutuplaşmasını iki faktörle ilişkilendirmiştir: partilerin ideolojik eğilimi ve partilerin koalisyon oluşumundaki rolü. Hirst, Riabinin, Graham, Boizot-Roche ve Morris (2014) Kanada Parlamentosu tartışmalarını analiz etmiş ve milletvekilinin ideolojisini liberal ya da muhafazakâr olarak tahmin etmiştir. Hirst ve diğerleri ayrıca milletvekillerinin parti statülerini (muhalefet veya hükümet) analiz etmiş ve sınıflandırmanın ideolojiden ziyade parti statüsüne, "muhalefet-hükümet" daha duyarlı olduğu sonucuna varmıştır.

Yu, Kaufmann ve Diermeier (2008) parlamento görüşmelerini Destek Vektör Makinaları (DVM) (Support Vector Machines, SVM) ile parti üyeliklerine göre sınıflandırmıştır. Parti üyelikleri milletvekillerinin ideolojilerini temsil edecek şekilde ölçülmüş, Cumhuriyetçi parti milletvekilleri muhafazakâr ideolojiye, Demokrat parti milletvekilleri liberal ideolojiye sahip olarak kabul edilmiş ve yıllar içinde bu ideolojiler arasındaki kutuplaşma ölçülmüştür. İki meclisli parlamentolarda parti aidiyeti parlamento statüsüne bağlı olarak değişebilmektedir. İki meclisli sisteme sahip parlamentoları karşılaştırmışlar ve partizanlığın veya kutuplaşmanın Temsilciler Meclisi'nde Kongre'ye göre daha yüksek olduğunu göstermişlerdir.

Lapponi, Soyland, Velldal ve Oepen (2018) Norveç Parlamentosu görüşmeleri veri kümesi Talk of Norway'i tanıtmış ve konuşmaları altı kabine dönemi boyunca DVM ile parti üyeliğine göre sınıflandırmıştır. Peterson ve Spirling (2018) 1935-2013 yılları arasındaki İngiliz Parlamentosu görüşmelerinde kutuplaşmayı ölçmek için denetimli makine sınıflandırmaları kullanmıştır. Peterson ve Spirling kutuplaşma ölçütü için dört farklı makine sınıflandırma algoritması kullanmışlar ve neredeyse aynı sonuçları elde etmişlerdir. Gentzkow, Shapiro, Taddy ve diğerleri (2016) ABD Kongre konuşmalarındaki kutuplaşmayı on yıllar boyunca sınıflandırma olasılıkları kullanarak ölçmüştür. Çalışmada kutuplaşma, sınıflandırma doğruluğu yerine her bir tahminin olasılıkları ile ölçülmüştür.

Kutuplaşmanın zaman içindeki seyrini parlamento konuşmalarından ölçen çalışmalar, zaman aralığı olarak yıllık veya on yıllık periyotlar kullanmışlardır. Bu çalışma zaman

aralığı olarak bir yıldan daha kısa aylık zaman aralığını önermektedir. Çalışmada Parlamento tutanaklarından aylık kutuplaşmayı ölçmek için 12 aylık hareketli aralıklar kullanılmıştır.

İkinci çalışmada TBMM milletvekilleri konuşmalarında Yazar Profili Çıkarma (YPO) çalışılmıştır. Bir metnin sahibinin demografik özelliklerinin otomatik olarak belirlenmesi birçok araştırmacının ilgisini çeken ortak bir problemdir. Analiz edilen derlemler, doküman temsiline (özellik çıkarımı) yönelik yaklaşımlar ve klasik makine öğrenimi ve derin öğrenme modelleri gibi sınıflandırma teknikleri bu alandaki çalışmaları çeşitlendirmektedir.

Demografik özelliklerin tahmininde metnin içeriği ve yazarın yazım tarzı öne çıkan noktalardır. Herring ve Paolillo (2006) bloglardaki cinsiyet ve tür farklılıklarını araştırmış ve blog sahiplerinin özelliklerinin üslup ve metnin içeriği olarak iki ana boyuta bağlı olduğu sonucuna varmıştır.

Yazar Analizi (YA) üzerine yapılan önceki çalışmalar yazarın yazı stili özelliklerini sınıflandırmıştır (Elmanarelbouanani ve Kassou, 2014; Holmes, 1994; T.R. Reddy, Vardhan ve P.V. Reddy, 2016; Stamatatos, 2009; Stamatatos, Fakotakis ve Kokkinakis, 2001; Zheng, Li, Chen ve Huang, 2006). Bu özellikler üslup belirteçleri (style markers) olarak adlandırılır. Stil belirteçleri sözcüksel (lexical) özellikler, karakter tabanlı özellikler ve sözdizimsel (syntatic) özellikler olarak kategorize edilebilir. Sözcüksel özelliklerde metin, bölütlerden oluşan bir dizi olarak ele alınır ve bölütler genellikle kelimeleri gösterir. Toplam kelime sayısı, ortalama kelime uzunluğu, cümle uzunluğu sayısı, kelime uzunluğu sayısı, derlemdeki en sık kullanılan kelimeler, kelime n-gramları, kelime zenginliği ve Hapax legomenon sözcüksel özellik türleridir. Karakter tabanlı özelliklerde metin bir karakter dizisi olarak görülür. Karakter n-gramları, rakamlar, büyük harfler, beyaz boşluklar, sekme boşlukları ve diğer özel karakterler karakter tabanlı özellikleri oluşturur. Cümlelerin sözdizimsel kalıpları, metindeki bir yazarın tarzını belirleyebilir. Sözdizimsel özellikler altında Part-of-Speech (POS) n-gram, noktalama işareti sıklığı ve işlev kelimesi sıklığı kategorize edilebilir. Alan yazındaki Yazar Profil Oluşturma çalışmaları, doküman temsilinde kullanılan optimum özellik setlerinin farklı alanlara, yazarlara veya derlemlere göre değişebileceğini göstermektedir (Iqbal ve diğerleri, 2010; Reddy ve diğerleri, 2016).

Kullanılan kelimeler metnin konusunu veya içeriğini belirler. Bu nedenle, kelime n-gramları içerik tabanlı özellikleri oluşturur. Yazarın profili spor, ekonomi, düğün ve yaşam tarzı gibi

konulara bağlı olabilir. Örneğin metinin konusu yazarın cinsiyet tahmini etkileyebilir. Janssen ve Murachver (2004) cinsiyet ve konular arasında yakın bir ilişki olduğunu göstermiştir.

PAN (https://pan.webis.de, Son Erişim Tarihi: 27 Nisan 2023), Yazar Analizi alanında bir dizi bilimsel etkinlik ve yarışma dizisidir. PAN, 2007 yılından bu yana 21 farklı etkinlikte 22 ortak göreve ev sahipliği yapmıştır. (Bevendorff ve diğerleri, 2020). Lim, Goh ve Thing (2013) PAN 2013'te verilen derlemden yazarların cinsiyetini ve yaşını tahmin etmişlerdir. İçerik tabanlı özelliklerin stil tabanlı özelliklerden daha ayırt edici olduğunu gözlemlemişlerdir. Rangel, Rosso, Potthast, Stein ve Daelemans (2014) çalışmaların çoğunda içerik tabanlı özelliklerin (kelimeler) stil tabanlı özelliklerden (cümle uzunluğu, POS n-gram, karakter n-gram, vb.) daha iyi performans gösterdiği sonucuna varmışlardır. Konu sınıflandırma ve bilgi erişimi gibi klasik metin sınıflandırma çalışmalarında, durak kelimelerin kaldırılması mantıklıdır çünkü bu kelimelerin bir metnin konusunu tahmin etmede hiçbir etkisi yoktur. Ancak her kelime yazarın yazım tarzının bir unsuru olabilir. Örneğin, Agun, Volkan ve Yilmazel (2017) Türkçe Yazarlık Atıf (Authorship Attribution) çalışmasında işlev sözcüklerinin doküman temsilinde güçlü özellikler olduklarını göstermişlerdir.

Standart değerlendirme kaynaklarının eksikliği, YPO için kıyaslama yöntemlerinde bir zorluktur. PAN YPO yarışmaları bu alandaki çalışmaların önünü açmaktadır. Yazar Profilleme alanındaki çalışmaların çoğu İngilizce dilinde ve sosyal medya verileri üzerinde gerçekleştirilmiştir. Koppel, Argamon ve Shimoni (2002) İngiliz Ulusal Derleminden (Consortium, 2007) cinsiyet tahmini yapmak için sözcüksel ve sözdizimsel özellikler kullanmıştır. Mukherjee ve Liu (2010) klasik makine öğrenmesi algoritmalarıyla yazarın cinsiyetini sınıflandırmak için blogları kullanmış ve %89 doğruluk elde etmiştir. Bartle ve Zheng (2015) aynı veri kümesinde pencereli Tekrarlayan Evrişimli Sinir Ağı (Recurrent Convolution Neural Network, WRCNN) derin öğrenme tekniğini kullanmış ve %86 doğruluk elde etmiştir. Cinsiyet ve yaş tahminleri alanyazında çeşitli veri setleri ile gerçekleştirilmiştir. Lin (2007) çevrimiçi sohbet günlüklerini kullanarak, Estival, Gaustad, Pham, Radford ve Hutchinson (2007) ise İngilizce e-postaları kullanarak cinsiyet ve yaş tahmini yapmıştır. Fatima ve diğerleri (Fatima vd., 2018) SMS mesajlarını yaş, cinsiyet, eğitim, ana dil ve meslek tahmini için analiz etmiştir. Fatima, Hasan, Anwar ve Nawab (2017) çok dilli YPO için Facebook verilerini kullanmıştır. Boulis ve Ostendorf (2005)

12000 transkripsiyonlu telefon konuşması için %93 doğrulukla cinsiyet sınıflandırması üzerinde çalışmıştır. Nguyen, Smith ve Rosé (2011) doğrusal regresyon kullanarak meme kanseri forumlarından, telefon konuşmalarından ve kullanıcı bloglarından yazarın yaşını tahmin etmiştir. Dwi ve Hauff (2015) Twitter mesajlarını klasik makine öğrenmesi teknikleriyle seçim tahmini yapmak için kullanmıştır Kaati, Lundeqvist, Shrestha ve Svensson (2017. Google bloglarından ve LinkedIn'den cinsiyet ve yaş tahmini yapmıştır. Cinsiyet için DVM ile %83 ve Evrişimli Sinir Ağları (ESA) (Convolution Neural Network, CNN) ile %77, yaş için ise ESA ile %34 ve DVM ile %44 doğruluk elde etmişlerdir. Kapočiute-Dzikienė, Utka ve Šarkutė (2015) cinsiyet ve yaş tahmini için Litvanya edebi metinleri ve parlamento metinleri üzerinde çalışmışlardır. Edebi metinlerde parlamento metinlerine göre daha yüksek doğruluk oranlarına ulaşmışlardır. Conover, Gonçalves, Ratkiewicz, Flammini ve Menczer (2011) DVM kullanarak Twitter kullanıcılarının siyasi parti hizalamasını tahmin etmiştir. Durdurma sözcüklerine ek olarak, tweetlerden hashtag'leri, bahsedenleri (mention) ve URL'leri de çıkarmışlar ve %79,2 doğruluğa ulaşmışlardır. Flores, Pavan ve Paraboni (2022) Brezilya hükümeti tarafından sağlanan vatandaş bilgilendirme hizmetinin talep metinlerini kullanmışlardır. Uzun Kısa Süreli Bellek (UKSB) (Long Short-Term Memory, LSTM) ile yaş için %67, eğitim için %69, cinsiyet için %87 ve meslek için 0.63 doğruluk elde etmişlerdir. García-Díaz, Colomo-Palacios ve Valencia-García (2022) cinsiyet, yaş ve ideoloji tahmini için 2020 yılında İspanyol politikacıların tweetlerini kullanmıştır. En iyi makro F1 skorları, cinsiyette Çift Yönlü Geçitli Tekrarlayan Birimler (Bidirectional Gated Recurrent Units, BiGRU) kullanılarak önceden eğitilmiş kelime vektörleri ile %72,022 olmuştur. Yaşta, en iyi makro F1 skoru aynı özellik seti ve cinsiyet sınıflandırıcısı ile %46,687 olmuştur. İkili siyasi spektrumda (sol-sağ), çok katmanlı algılayıcılar ile önceden eğitilmiş kelime vektörleri %98,036 makro F1 skoru ile en iyi sonucu elde etmiştir.

Sınırlı sayıda araştırmacı Türkçe metinler üzerinde YPO çalışması yapmıştır. Amasyalı ve Diri (2006) Türkçe günlük gazete makalelerini yazar, tür ve cinsiyet açısından kategorize etmiştir. Kadın yazarlar için 140 ve erkek yazarlar için 490 dokümana karakter düzeyinde n-gram modelleri uygulamışlar ve DVM ile %96 doğruluk elde etmişlerdir. Deniz ve Kiziloz (2017) Türkçe dokümanların yazar, tür ve cinsiyet açısından metin sınıflandırması için n-gram modelini uygulamıştır. Çalışmalarında 30 yazar tarafından yazılmış 300 Türkçe günlük gazete makalesi kullanmışlardır. Cinsiyet sınıflandırmasında beş erkek ve beş kadın yazar kullanmışlardır. Sınıflandırma görevlerinde DVM, Naive Bayes ve Rastgele Ormanlardan

daha iyi performans göstermiştir. Kucukyilmaz, Cambazoglu, Aykanat ve Can (2006) 200 erkek ve 200 kadın kullanıcıya ait çevrimiçi eşler arası Türkçe sohbet mesajlarından cinsiyet tahmini yapmışlardır. NB sınıflandırıcısı ile %81,5 doğruluk elde etmişlerdir. Türkmen ve diğerleri Türkmen, Diri, Biricik ve Doğan (2011) psikoloji bölümü öğrencileri tarafından yapılan halka açık görüşmelerden yaş grubu, cinsiyet, medeni durum ve eğitim seviyesini tahmin etmiştir. Sınıflandırma, k-En Yakın Komşular (K-Nearest Neighbors, k-NN), Naive Bayes (NB) ve DVM makine öğrenmesi algoritmaları ile gerçekleştirilen kelime köklerine ve 2-gramlı kelimelere dayanmaktadır. Cinsiyet için kNN ile %78,1, medeni durum için DVM ile %86,1, eğitim seviyesi için NB ile %95,2 ve yaş grubu için DVM ile %87,3 doğruluk oranlarına ulaşmışlardır. Ciot, Sonderegger ve Ruths (2013), İngilizce olmayan bağlamlarda Twitter metin verilerinden DVM ile cinsiyet tahmininde bulunmuştur. Analiz için Fransızca, Endonezce, Türkçe ve Japonca dillerini kullanmışlar ve İngilizce ile karşılaştırılabilir sonuçlar bulmuşlardır. Türkçe'nin doğruluğu diğer dört dil arasında en yüksektir (%87). "Tüm diller arasında, Türkçe dilinin Twitter genel popülasyonundan alınan bir veri kümesi üzerinde tüm Twitter cinsiyet çıkarımı literatüründe elde edilen en yüksek doğruluk olduğunu" belirtmişlerdir. Yilmaz ve Abul (2018), 2017 Türkiye anayasa referandumunda Twitter kullanıcılarının siyasi eğilimlerini DVM, Karar Ağacı ve Rastgele Ormanlar ile tahmin etmiştir. En iyi sınıflandırma doğruluğunu doğrusal DVM ile %89,9 olarak bildirmişlerdir.

Parlamento verileri, özellikle de parlamento görüşmeleri üzerinden milletvekillerinin özelliklerini tahmin etmek için çeşitli çalışmalar yapılmıştır. Dahllöf (2012) İsveç parlamentosunda 2003-2010 yılları arasında yedi yıllık oturumlarda yapılan konuşmalara dayanarak milletvekillerinin yaş, cinsiyet ve siyasi eğilimlerini tahmin etmiştir. Dahllöf, DVM ile farklı demografik grupları analiz etmiş ve cinsiyet için %81,2, siyasi eğilim için %89,4 ve yaş için %78,9 doğruluk oranına ulaşmıştır. Przybyla ve Teisseyre (2014) konuşmacının cinsiyet, eğitim, parti üyeliği ve doğum yılı gibi özelliklerini tahmin etmek için Polonya parlamentosundaki konuşmaları analiz etmiştir. En iyi doğruluk (%97) cinsiyet sınıflandırmasında Rastgele Orman tarafından elde edilmiştir. Doğum yılını analiz ederken regresyon analizi kullanmışlar ve En Yakın Komşu regresyonu ile 6,48 karesel hata elde etmişlerdir. Dunn, Argamon, Rasooli ve Kumar (2016) 1995-2013 yılları arasında ABD Meclisi ve Senatosu'nda yapılan konuşmaları kullanarak demografik özelliklerin profil tabanlı bir tahminini geliştirmiştir. Milletvekillerinin cinsiyetini, yaşını, askerlik hizmetini, coğrafi konumunu, milliyetini ve dinini analiz etmişlerdir. Ayrıca kongre konuşmalarından

parti üyeliği gibi parlamentoya özgü özelliklerin tahminini de gerçekleştirmişlerdir. Høyland ve diğerleri Høyland, Godbout, Lapponi ve Velldal (2014) DVM kullanarak Avrupa Parlamentosu görüşmelerinden beş parti için parti üyeliklerini tahmin etmiş ve %55 genel doğruluk elde etmiştir. Yu, Kaufmann ve Diermeier (2008) ABD Temsilciler Meclisi ve Senato konuşmalarında parti aidiyeti ve ideolojilerini sınıflandırmak için DVM kullanmış ve Temsilciler Meclisi'nde %80, Senato'da %86 doğruluk elde etmiştir. Hirst, Riabinin, Graham, Boizot-Roche ve Morris (2014) Kanada parlamento görüşme transkriptlerini kullanarak klasik makine öğrenimi algoritmalarıyla liberalleri muhafazakârlardan ayırarak ideolojiyi belirlemiştir. Milletvekillerinin parti statüsü bağlılığını araştırarak, parlamento konuşmasının ideolojiden ziyade parti statüsüne daha duyarlı olduğu sonucuna varmışlardır. Lapponi, Soyland, Velldal ve Oepen (2018) ToN (Talk of Norway) parlamento görüşmeleri veri kümesini tanıtmış ve DVM kullanarak parti üyeliği sınıflandırması geliştirmiştir. Peterson ve Spirling (2018) İngiliz Parlamentosu görüşmelerini kullanarak partiler arasındaki kutuplaşmayı ölçmüştür. Düşük doğruluk oranı, parti üyeliği sınıflandırma görevinde düşük kutuplaşma anlamına gelmektedir. Yu (2014) çalışmasında iki grup ortalaması arasındaki farkı ölçen Cohen's d (Cohen, 2013) istatistiksel etki büyüklüğünü kullanarak milletvekillerinin cinsiyetine göre dilin nasıl farklılaştığını belirlemek için 1989 ve 2008 yılları arasındaki kongre konuşmalarını incelemiştir. Cinsiyet profili çıkarma çalışmalarında (Koppel ve diğerleri, 2002; Newman, Groom, Handelman ve Pennebaker, 2008; Yu, 2014), beklendiği gibi kadın milletvekillerinin daha fazla duygusal sözcük ve daha az artikel kullandığı, erkek milletvekillerinin ise daha fazla isim ve uzun sözcük ile daha az şahıs zamiri kullandığı görülmüştür.

Üçüncü çalışmada TBMM Genel Kurul görüşmelerinde kelime benzerliği çalışılmıştır. Kelime vektörleri elde edilirken sayma tabanlı ve tahmine dayalı yöntemler kullanılmaktadır. Sayma tabanlı yöntemlerden bir kelimenin derlemdeki dokümanlarda geçme sıklığına dayalı yöntemlerde doküman-kelime matrisleri (Salton, Wong ve Yang, 1975) oluşturulur. Bu matrisin her bir satırı derlemdeki dokümanlarla ilişkisini gösterir kelime vektörünü verir. Sütunlar ise bir dokümanın kelimelere bağlı ilişkisini gösterir bir vektördür. Bu matriste kelimenin doküman sıklığının (tf) yanında diğer dokümanlarda geçme sıklığının tersi (idf) gibi metrikler de kullanılarak doküman temsili için başarılı sonuçlar elde edilebilir. Ancak bu yöntemden elde edilen kelime vektörleri kelimenin bir uzaydaki anlamsal temsilini vermez.

Gizli anlam analizi (Latent Semantic Analysis-LSA), kelime anlamı temsillerinde en çok kullanılan sayma tabanlı yöntemlerden birisidir (Hu, Cai, Wiemer-Hastings, Graesser ve McNamara, 2007; Landauer ve Dumais, 1997). LSA, dokümanlardan oluşan derlemi girdi olarak alır ve kelime-doküman sıklık matrisini oluşturur. Bu matris çok yüksek boyutlu ve seyrek olabilir, bu da dokümanları doğrudan analiz etmeyi veya karşılaştırmayı zorlaştırır. Bu sorunu gidermek için bir matris faktorizasyon metodu olan Tekil Değer Ayrıştırma (TDA) (Singular Value Decomposition, SVD) uygulanır. TDA sonucu sol tekil matris, tekil değerlerin köşegen matrisi ve sağ tekil matris ortaya çıkar. Sol tekil U matrisi, derlemdeki terimler ve kavramlar arasındaki ilişkiyi temsil eder. U matrisindeki her sütun bir kavrama, her satır ise bir terime karşılık gelir. Matristeki değerler, bir terim ile bir kavram arasındaki ilişkinin gücünü gösterir. Köşegen matrisi S, faktorizasyona uğrayan kelime-doküman matrisinin tekil değerlerini içerir. S matrisinde tekil değerler, derlemdeki her kavramın önemini temsil eder. V matrisinin transpozu dokümanlar ile derlemdeki kavramlar arasındaki ilişkiyi temsil eder. Matriste her satır bir dokümana, her sütun da bir kavrama karşılık gelir. Matristeki değerler, bir doküman ile bir kavram arasındaki ilişkinin gücünü gösterir.

TDA kullanılarak, terim-doküman matrisin altında yatan gizli anlamsal değerler her bir kelime için belirlenebilir. Bu durumda bir kelime vektörünün boyutu seçilen gizli (latent) konu veya kavram sayısı kadardır. Böylece her bir kelimenin bulunan gizli kavramlarla veya konularla ilişkisini gösteren kelime vektörleri elde edilir.

Pencereye dayalı sayma tabanlı yöntemlerden, büyük bir derlemde bir sözcük ile sözlükteki diğer sözcüklerin belli bir aralıkta, birlikte yer alma sıklığını kullanarak birliktelik matrisi oluşturur. Birliktelik matrisinin her bir satırı bir kelime vektörünü temsil eder. Bu yöntemde bir kelime vektörünün boyutu sözlükteki kelime sayısı kadar olacaktır. Sözlük sayısın çok olduğu derlemlerde bu durum birçok değerin sıfıra eşit olduğu çok uzun kelime vektörlerinin oluşmasına, vektör uzayında yüksek boyutluluğa (high-dimension) ve vektörlerin derinliğinin azalarak seyrekliğe (sparsity) yol açar. Yüksek boyutlu uzaylarda, güvenilir istatistiksel tahminler elde etmek için gereken veri miktarı, boyut sayısı ile katlanarak geometrik olarak artar. Bu durum artan hesaplama karmaşıklığı ve anlamsal temsilin kaybolmasına yol açar. Bir diğer dezavantajı ise derleme yeni giren kelimelerin analize dâhil edilmesi ise güçtür.

Yapay sinir ağları düşük boyutlu kelime vektörleri elde etmek için kullanılabilir(Collobert ve Weston, 2008; Mikolov ve diğerleri, 2013). Kelime yerleştirme için yapay sinir ağları ilk olarak (Bengio, Ducharme, ve Vincent, 2000) tarafından önerilmiştir. İleri Beslemeli Sinir Ağı Dil Modeli (Feed Forward Neaural Network Language Model-FFNNLM) doğrusal bir iz düşüm ve doğrusal olmayan bir saklı katmandan oluşmaktadır. Model, verilen kelimelerden, bağlamdan hedef kelimeyi tahmin etmeye çalışır. Mikolov ve diğerleri (2013) bu modeldeki hesaplama karmaşıklığına çözüm bulduğu word2vec modelini tanıtmıştır. Yapay sinir ağları kelime vektörlerinin çıkarılmasında gizli anlam analizi gibi klasik yöntemlerin yerini almaya başlamıştır. Özellikle word2vec gibi tahmine dayalı yöntemler bu sayede popüler olmuştur. Li, Fu, Masud ve Huang (2016) çalışmalarında word2vec modeli ile elde ettikleri kelime vektörlerini yerel bilgi olarak değerlendirmişler ve kelimenin içinde geçtiği dokümandan elde ettikleri vektörle (global bilgi) birlikte değerlendirerek "çoklu-bağlamsal karışık yerleştirme" olarak adlandırdıkları kelime vektörleri önermişlerdir.

Boyut indirgeme için oto kodlayıcıların kullanıldığı çalışmalar da vardır. Kaynar, Aydın ve Görmez (2017) boyut düşürme tekniklerini karşılaştırdıkları çalışmalarında boyut düşürmek için derin öğrenme tabanlı oto kodlayıcıları kullanmışlardır.

GLoVe modeli ise hem sayma tabanlı sistemlerin hem de tahmine dayalı yöntemlerin avantajlarını bir araya getirdiğini iddia etmektedir. Altszyler, Sigman, Ribeiro ve Slezak (2016) çalışmalarında küçük doküman derleminde LSA ve word2vec modelini karşılaştırmışlar ve LSA'nin küçük veri kümeleri için daha iyi sonuçlar verdiğini göstermişlerdir. Çalışmada derlemdeki kelime sayısı 10^6'dan fazla olduğunda, tahmin tabanlı word2vec modelinin, sayma tabanlı LSA modeline göre benzerlik bulma konusunda daha başarılı olduğu gösterilmiştir.

Levy, Goldberg ve Dagan (2015) word2vec modelinin kelime benzerliği konusunda GloVe'dan daha başarılı olduğunu göstermişlerdir. Naili, Chaibi ve Ghezala (2017) word2vec CBOW modelinin sık geçen kelimeler için, word2vec Skip-gram modelinin seyrek geçen kelimeler için daha başarılı olduğunu göstermişlerdir. Semantik sözlükler bir kelimenin eş anlamı, üst anlamı, alt anlamı gibi bilgileri içerir. Faruqui ve diğerleri (2014) kelime vektörlerinin kalitesini artırmak için vektör uzayı kelime temsillerinden çıkan sonuçları WordNet (Miller, 1995), FrameNet (Baker, Fillmore ve Lowe, 1998) ve

ParaPharase (Ganitkevitch, Van Durme ve Callison-Burch, 2013) gibi semantik sözlüklerle birlikte değerlendirerek “retrofitting” adını verdikleri kelime vektörlerini elde etmişlerdir.

BERT (Bidirectional Encoder Representations from Transformers) (Devlin ve diğerleri, 2018) ve ULMFiT (Universal Language Model Fine-tuning) (Howard ve Ruder, 2018) kelime yerleştirme için kullanılan derin öğrenmeye dayalı modellerdir. BERT Dönüştürücü (Transformer) tabanlı bir modeldir Model bir cümlede eksik kelimeleri tahmin etmek için maskelenmiş dil modelini ve “sonraki cümleyi tahmin etme” görevlerini kullanarak geniş bir derlemde eğitilir. Bu modeller ile bağlamsal kelime vektörleri elde edilir, kelime vektörünün değeri kullanıldığı bağlama göre değiştir. Bu yönüyle word2vec ve GloVe gibi statik, bağlama göre değişmeyen modellere göre avantajlıdır. Metin sınıflandırma, soru-cevap, makine çeviri gibi kelimenin bağlamı ile birlikte bulunduğu görevlerde çok başarılıdır. Kelime benzerliği ve kelime analojilerinin çıkarılmasında kelimeler bir cümle içinde ya da bağlamı ile birlikte bulunmaz. Oysa BERT gibi bağlamsal kelime vektörleri kullanan modellerin asıl kazancı kelimeyi bağlamı ile birlikte değerlendirmesidir. Bu yüzden kelime analojileri ve kelime benzerliği için GloVe ve word2vec modelleri BERT gibi bağlamsal kelime vektörü elde eden modellere göre daha çok tercih edilebilir.

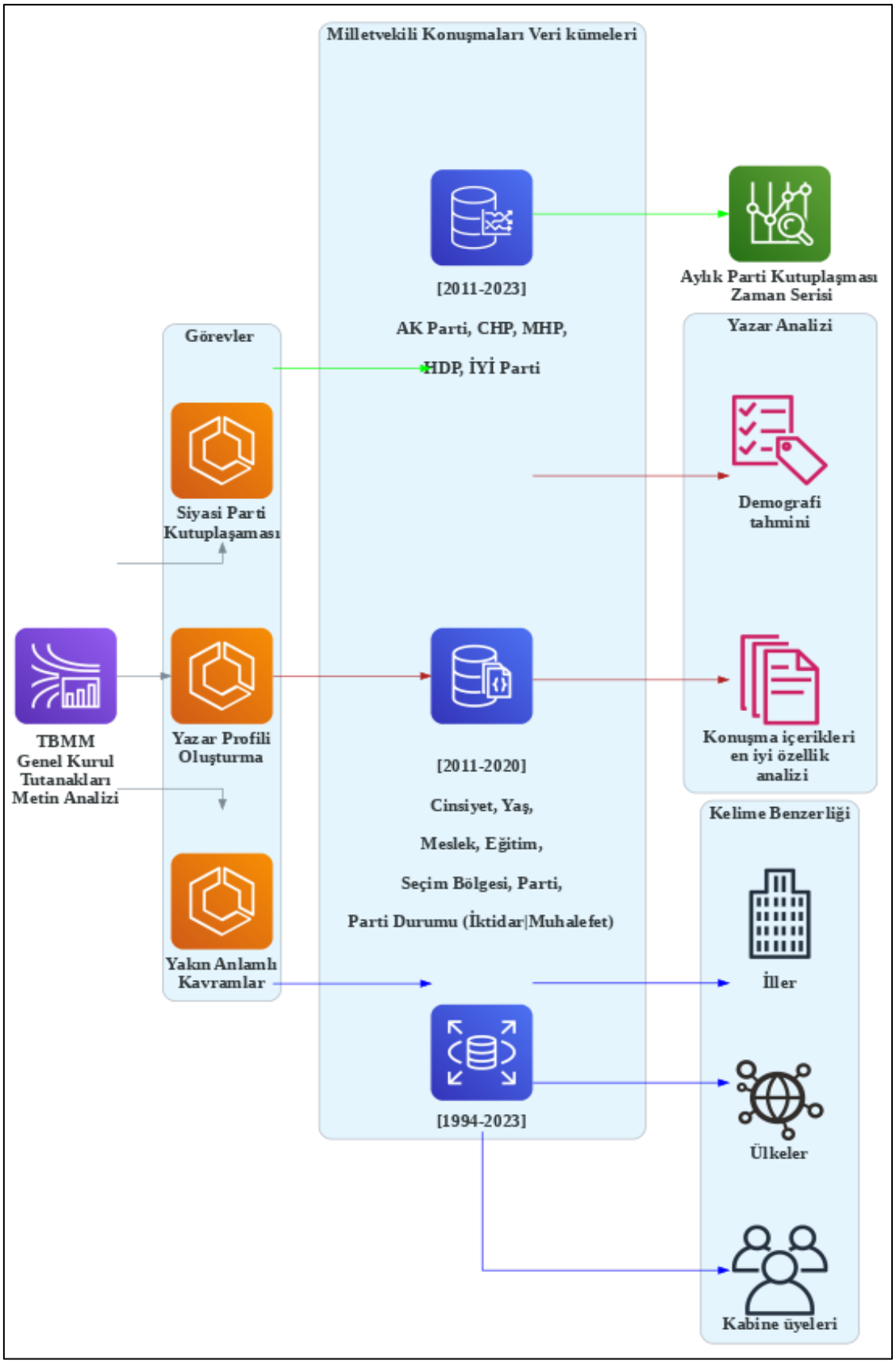
Bu çalışmada, TBMM Genel Kurul tutanaklarındaki birbirine yakın anlamlı kavramlar çıkarılırken yapay sinir ağları temelinde modellenen word2vec modeli, global birlikte geçme matrisi temelinde modellenen GloVe modeli kullanılmıştır. Çıkan sonuçlar Türk Dil Kurumu güncel sözlüğü ile karşılaştırıldığında birbiri ile yakın anlamlı kavramların ve analojilerin word2vec ve GloVe modeli ile başarılı bir şekilde elde edildiği görülmüştür.

3. VERİ VE YÖNTEM

TBMM'de görüşmenin yapıldığı güne ait genel kurul tutanağı https://www.tbmm.gov.tr/Tutanaklar/TutanakMetinleri adresi ile vatandaşların erişimine açıktır. Bu veri ham metin verisi halindeki bir günlük tutanaktır. TBMM'de ayrıca her milletvekiline özgü özgeçmiş ve genel kurul tutanaklarının bulunduğu profil sayfası da mevcuttur (https://www.tbmm.gov.tr/milletvekili/liste, Son Erişim Tarihi: 18 Temmuz 2023). TBMM'de dönem iki genel seçim arasını belirtir ve her dönemde Genel Kurulda konuşan milletvekilleri kısmen yenilenir. Milletvekillerinin konuşmaları 2011 yılı ve 24. Dönem itibari ile mevcuttur. Çalışmada milletvekillerine ait profil sayfasındaki konuşmalar ve üst bilgiler konuşmaların etiketlenmesi, anatosyonu için kullanılarak 2011 ve 2023 yıllarını kapsayan TBMM Genel Kurul Tutanakları Derlemi oluşturulmuştur. Siyasi parti kutuplaşması ve Yazar Profili oluşturma çalışması bu derlemden oluşturulan veri kümesi ile gerçekleştirilmiştir.

Bir güne ait birleştirilmiş tutanak metinlerinden 1994-2023 yılları arasındaki bir günlük konuşmalardan (birleşim) oluşan ikinci bir derlem oluşturulmuştur. Kelime yerleştirme algoritmaları kullanılarak kelime vektörleri elde etmek için 1994-2023 yılını kapsayan etiketsiz milletvekili konuşmalarından oluşan bu derlem kullanılmıştır. Derlem (https://www.tbmm.gov.tr/Tutanaklar/TutanakMetinleri , Son Erişim Tarihi: 18 Temmuz 2023) sitesindeki birleşimlere ait metin verisinden elde edilmiştir.

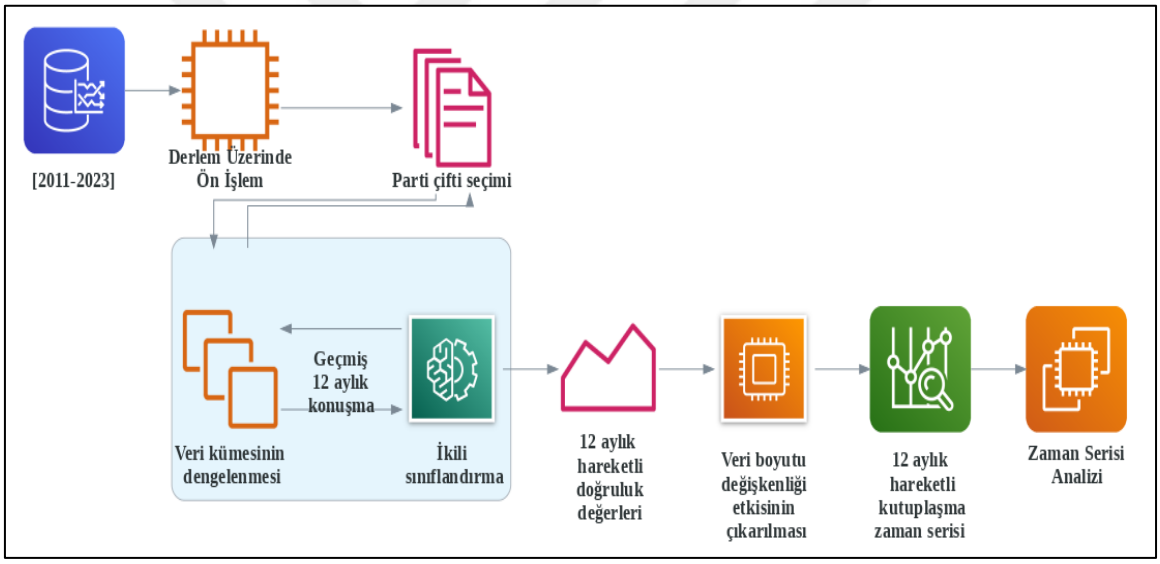
Şekil 3.1'de TBMM Genel Kurul tutanaklarının yapay zeka tabanlı analizinde yapılan 3 ayrı çalışmanın genel şeması verilmektedir. Çalışmalarda kullanılan veri kümelerinin kapsadığı tarihler ve görevlerin ana hatları şekilde gösterilmiştir.



Şekil 3.1. TBMM Genel Kurul tutanakları metin analizi çalışmasının genel gösterimi

3.1. TBMM Genel Kurul Görüşmelerinde 12 Aylık Hareketli Siyasi Parti Kutuplaşması

Dünyadaki siyasal yapılarda ABD veya İngiltere gibi iki ana siyasi partiden oluşan ikili bir yapı olduğu gibi Almanya gibi ikiden çok siyasi partinin siyasette ağırlığının olduğu örnekler de vardır. Türkiye'deki siyasi hayat ikili değil çok partili bir siyasal sistemdir. Bu durum kutuplaşma çalışmalarını daha karmaşık hale getirir. Çalışmada, 2011 ve 2023 yılları arasında TBMM'de yapılan genel kurul konuşmaları kullanılmıştır. Derlem, 2011-2018 yılları arasında dört partinin (AK Parti, CHP, MHP, HDP) konuşmalarından oluşmaktadır. 27. Dönemde (2018-2023) ise, meclise yeni giren İYİ Parti de dâhil olmak üzere beş partinin konuşmalarını kapsamaktadır. Derlemde milletvekillerinin 80700 genel kurul konuşması bulunmaktadır.



Şekil 3.2. Parlamento tutanaklarından 12 aylık hareketli parti kutuplaşması adımları

Şekil 3.2 çalışmanın adımlarını göstermektedir. Parlamento tutanaklarından aylık polarizasyonun ölçüldüğü yöntem iki ana fazdan oluşmaktadır. Birinci adımda bir ay için geriye dönük 1 yıllık veri kümesi oluşturulmuş ve metin sınıflandırması ile polarizasyon ölçülmüştür. Bu işlem her bir parti çifti için gerçekleştirilmiştir. Bu adım çalışmada Şekil 3.3'de gösterilmiştir.

İkinci adımda ise oluşturulan alt veri kümelerinin yapısının sınıflandırmaya etkisi regresyon analizi ile bulunarak polarizasyon sonuçlarından çıkarılmıştır. Bu aşama çalışmada Şekil 3.5'de gösterilmiştir.

Zaman serileri Şekil 3.3 ve Şekil 3.5'deki adımlar sonucunda elde edilen nihai polarizasyon sonuçlarından elde edilmiştir. Elde edilen zaman serileri kullanılarak Türkiye'deki siyasi partilerin birbirleri ile ilişkileri analiz edilmiştir.

3.1.1. Metin sınıflandırması ile kutuplaşma ölçümü

Kutuplaşma çalışmasının ilk adımı konuşma metinlerinin ön işlemden geçirilmesidir. Ön işlem aşamasında, sınıflandırıcının yalnızca konuşmaların içeriğine odaklanmasını sağlamak için konuşmacı milletvekilinin adı, soyadı, seçim bölgesi, parti liderinin adı ve soyadı ile parti adları dokümanlardan çıkarılmıştır. Sınıflandırmada metin normalizasyonunu sağlamak amacıyla kelimeler küçük harfe dönüştürülmüş ve sözcük başları (lemma) kullanılmıştır. Şekil 3.3'de, derlemdeki tüm konuşmaları ön işleme tabi tutmak için *"onIslem"* fonksiyonu kullanılmıştır.

Parlamento genel kurul görüşmeleri söz konusu olduğunda, takip eden bazı haftalarda, hatta aylarda parlamentonun çalışmadığı ve herhangi bir konuşma yapılmadığı dönemler olmaktadır. Çalışmada her ay için geçmiş bir yılı kapsayan veri kümeleri ile kutuplaşma ölçülmektedir. Oysa kimi aylarda yeterli veri bulunmamaktadır. Yıllık kutuplaşma doküman düzeyinde ölçülürse, belirli dönemler aylar veya partiler için yeterli veri olmaması sorunu oluşur. Bu nedenle, doküman düzeyinde oluşturulmuş veri kümelerinde veri artırımına ihtiyaç vardır. Bir milletvekilinin konuşmasını paragraflara dönüştürmek, doküman sınıflandırması için doğal bir veri artırma yolu olabilir. Cambridge sözlüğüne (https://dictionary.cambridge.org/dictionary/english/paragraph , Son Erişim Tarihi: 18 Temmuz 2023) göre paragraf; "bir metnin en az bir cümleden oluşan ve yeni bir satırda başlayan kısa bir bölümü ve genellikle tek bir olay, açıklama, fikir ve benzerlerini kapsar."

Tek bir olayın, tanımın, fikrin vb. temsilcisi olarak paragraflar, bir dokümandaki atomik anlamsal varlıklar olarak tanımlanabilir ve siyasi partilerin görüşlerinin temsili için tek bir metinsel birim olarak düşünülebilir. Öte yandan, birkaç kelimeden oluşan kısa paragraflar çoğunlukla bir konuyu temsil etmez. Bu nedenle, çalışmada tek bir konuyu daha kararlı bir şekilde temsil etmek için en az yirmi kelime uzunluğundaki paragraflar seçilmiştir. Paragraf seviyesi kullanılarak, veriler doküman seviyesine göre neredeyse altı kat artırılmış ve 80 700'den 487 862'ye çıkarılmıştır. Derlemde ortalama paragraf uzunluğu altmış sekiz

kelimedir. Derlemin önceden işlenmiş konuşmalar, Şekil 3.3'de *"paragraflaraBol"* fonksiyonu ile paragraflara dönüştürülmüştür.

Beş partinin ikili kombinasyonlarının bir sonucu olarak, analizde on parti çifti (*"partiCifti"*) vardır. Çalışmada her bir parti çifti için oluşturulan alt veri kümelerini kullanan ikili sınıflandırmalar ile kutuplaşma ölçülmüştür.

TBMM genel kurul görüşmeleri veri kümesi 2011-10-01 ve 2023-05-01 tarihleri arasındaki konuşmaları kapsamaktadır. Her bir zaman aralığı (*"zamanAraligi"*) 12 aylık sürenin başlangıç ve bitiş tarihleridir. Zaman serisi oluşturulurken, bir zaman aralığının başlangıç ve bitiş tarihi bir ay (*"birAy")* artırılmıştır.

Her sınıflandırma görevinde, bir zaman aralığı ve parti çifti için paragraflar seçilerek bir alt veri kümesi oluşturulmuştur. Bu işlem Şekil 3.3'de *"altVeriKumesiOlustur"* fonksiyonu ile gösterilmiştir. Daha sonra bu alt veri kümesi her parti için eşit sayıda paragraf olacak şekilde paragrafların özelliklerine göre dengelenmiştir.

Dengeli alt veri kümeleri Şekil 3.3'de *"dengeliAltVeriKumesiOlustur"* fonksiyonu ile elde edilmiştir. Konuşmalar, TBMM genel kurul tutanakları derleminde Genel Kurul günü (birleşim), oturum, bölüm ve konu meta verilerine sahiptir. Şekil 3.3'de konuşma içeriğini *"konusmaMetinleri"*; "*konu", "bolum", "oturum", "birlesim"* ise konuşma içeriğine ait meta verileri gösterir. Bu meta veriler, belirtilen sırayla iç içe geçmiş sıralı bir listedir ve bire çok (one-to-many) hiyerarşik yapıya sahiptir. Bir parti çifti için dengeli bir alt veri kümesi, alt hiyerarşiden üst hiyerarşiye doğru her parti için eşit sayıda konuşma seçerek oluşturulmuştur. Örneklem büyüklüğü, en az konuşmaya sahip partinin yaptığı konuşma sayısına göre belirlenmiştir. Bu prosedür sayesinde alt veri kümeleri konu, bölüm, oturum, gün ve siyasi partiye göre dengelenmiştir. Böylece, veri kümesinin dengesiz yapısına, özellikle de görüşme konularına dayalı kutuplaşma yanlılığına (bias) çözüm üretilmiştir.

TBMM Genel Kurul görüşmeleri veri kümesindeki konuşmalar oturum, bölüm ve konu meta verilerine sahiptir ancak diğer parlamento görüşmeleri veri kümeleri bu meta verilere sahip olmayabilir. Bu parlamentolarda, konuşmaları görüşme konularına göre gruplandırmak için konu modellemesi kullanılması ve veri kümesi genel kurul gününe ve konu gruplarına göre dengelenmesi çalışmada önerilmektedir.

Siyasi partilerin pozisyonlarını belirlemek için denetimli makine öğrenimi parti üyeliği sınıflandırması kullanılmıştır. On parti çifti arasındaki kutuplaşma, her biri için ikili sınıflandırma ile ölçülmüştür. Partiler arasında net bir ayrım, yüksek kutuplaşma anlamına gelir. Bu nedenle, parti üyeliği sınıflandırmasının yüksek doğruluğu yüksek kutuplaşmayı, düşük doğruluğu ise düşük kutuplaşmayı göstermektedir. Sınıflandırma, lib-lineer kernel ile lojistik regresyon algoritmalarını kullanılmıştır ve parametre ince ayarından sonra ters düzenleme parametresi olarak C=100 000 seçilmiştir. Özellik kümesi olarak bir gram, iki gram ve üç gram TF-IDF değerleri kullanılmıştır.

Tabakalı (srtratified) 4 katlı çapraz doğrulama (*"KFoldCV"*) kullanarak, bir parti çift için her bir ikili sınıflandırma görevinden dört doğruluk değeri (*"dogruluklar_cv"*) elde edilmiştir.

Tahminin kesinliğini sağlamak ve zaman serisindeki her nokta için bir güven aralığı elde etmek için dengeli veri setinin oluşturulması ve 4 katlı çapraz doğrulama ile sınıflandırma sekiz kez *(N=8)* tekrarlanmıştır. Böylece 32 adet *(n)* doğruluk değeri elde edilmiştir. Bu değerlerin ortalamasından (*"dogrulukOrtalamasi"*) zaman serilerindeki bir nokta (ay) için kutuplaşma ölçüsü elde edilmiştir.

Kutuplaşma ölçütünün güven aralığı elde edilen 32 doğruluk değerinin ortalaması ve standart sapması kullanılarak bulunmuştur. Güven aralığı (*"guvenAraligi")* algoritmadaki kutuplaşmanın üst ve alt sınırını gösterir. Kutuplaşma %95 güven seviyesinde *(Z=1.96)* değeri ile ölçülmüştür.

8 kez tekrarlanan veri seti oluşturma işleminde aynı parti çifti ve zaman noktası (ay) için her defasında farklı dengeli veri kümeleri elde edilmiştir. Bunun için konu, bölüm, oturum, genel kurul günü ve zaman aralığının çoğunluk setlerinden paragraflar seçilirken farklı tohumlama değerleri (*"tohum"*) kullanılmıştır.

Şekil 3.3'e göre yapılan sınıflandırma görevlerinin sonunda 12 aylık hareketli kutuplaşma ölçütlerini ve oluşturulan alt veri kümelerinin istatiksel verilerini içeren zaman serisi *("zamanSerileri")* elde edilmiştir. Veri kümeleri istatiklerinin (*"kutuplaşmaFarkı", "mutlakKutuplaşmaFarkı", "jaccardMesafesi")* kutuplaşma üzerine etkisi bir sonraki bölümde (Bkz. Bölüm 3.1.2) incelenecektir.

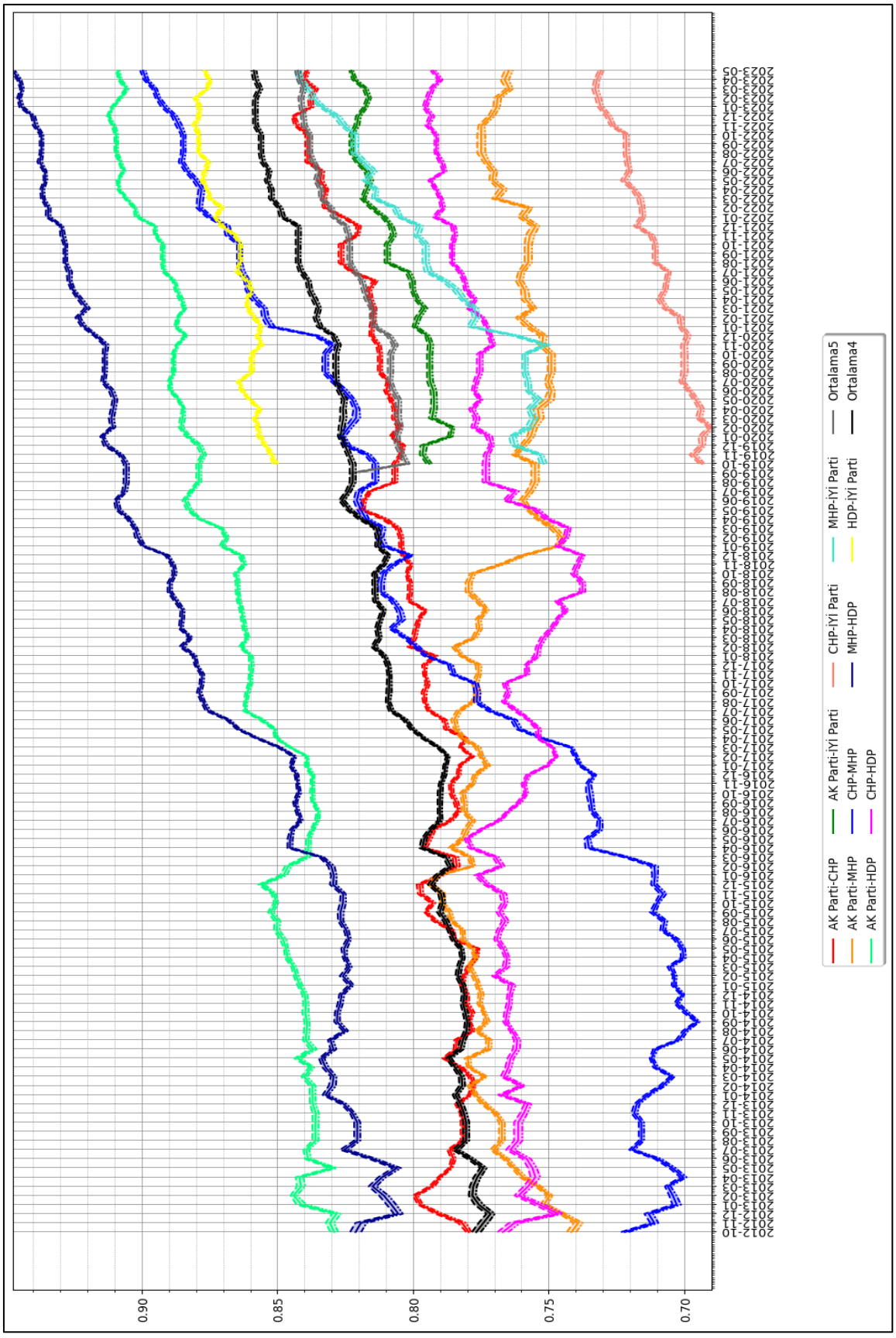
Şekil 3.4'de ikili metin sınıflandırmasından elde edilen kutuplaşma doğruluk değerleri Şekil 3.3'deki adımların milletvekili konuşmalarına uygulanması sonucu elde edilen kutuplaşma zaman serilerini göstermektedir. Kutuplaşma ölçütleri şekilde %95 güven aralığı ile birlikte gösterilmiştir.

Şekil 3.3. Parlamento görüşmelerinden parti çiftlerinin aylık kutuplaşma ölçütü algoritması

```
    Veri: Milletvekillerinin konuşmaları, [konuşmalar]
    Sonuç: Parti çiftlerinin 12 aylık hareketli sınıflandırma sonuçları, [zamanSerileri]
1   konusmaMetinleri ⊆ konu ⊆ bolum ⊆ oturum ⊆ birlesim ⊆ zamanAraligi
2   onIslenmisKonusmalar ← onIslem(konusmalar)
3   paragraflar ← paragraflaraBol(onIslenmisKonusmalar)
4   zamanSerisiBaslangici ← (2011 − 10 − 01, 2012 − 10 − 01)
5   zamanSerisiSonu ← (2022 − 05 − 01, 2023 − 05 − 01)
6   zamanSerileri ← []
7   foreach partiCifti ∈ partiCiftleri do
8       kutuplasma ← []
9       std ← []
10      boyut ← []
11      zamanAraligi ← zamanSerisiBaslangici
12      while zamanAraligi =< zamanSerisiSonu do
13          altVeriKumesi ← altVeriKumesiOlustur(zamanAraligi, partiCifti, paragraflar)
14          dogruluklar ← []
15          for k ← 1 to (N = 8) do
16              dengeliVeriSeti← dengeliVeriKumesiOlustur(partiCifti, altVeriKumesi,
                seviyeler=[konu,bolum,oturum,birlesim, zamanAraligi],tohum=k)
17              dogruluklar_cv ← KFoldCV(dengeliVeriSeti,katlama=4)
18              dogruluklar·append(doğruluklar_cv)
19          end
20          (12ay, ilkAy, sonAy)←veriKumesiBoyutlari(dengeliVeriSeti)
21          dogrulukOrtalamasi ←average(dogruluklar,axis=0)
22          dogrulukStandartSapmasi ←standartDeviation(dogruluklar,axis=0)
23          kutuplasma·append(dogrulukOrtalamasi)
24          std·append(dogrulukStandartSapmasi)
```

Şekil 3.3. (devam) Parlamento görüşmelerinden parti çiftlerinin aylık kutuplaşma ölçütü algoritması

```
25              boyut·append((12ay, ilkAy, sonAy))
26              zamanAraligi ← zamanAraligi + birAy
27          end
28          m ← lenght(kutuplasma)−1
29          n ← N· katlama
30          (guvenAraligi_j)_{j=0}^{m} ← kutuplasma_j ± Z.std_j/√n
31          (jaccardMesafesi)_{j=1}^{m} ← (boyut.ilkAy_{j−1} + boyut.sonAy_j)/(boyut.ilkAy_{j−1} + boyut.12ay_j)
32          jaccardMesafesi_0←0
33          (kutuplasmaFarki)_{j=1}^{m} ← kutuplasma_j − kutuplasma_{j−1}
34          kutuplasmaFarki_0←0
35          (mutlakKutuplasmaFarki)_{j=0}^{m} ← |kutuplasmaFarki_j|
36          olcutler←[kutuplasma, guvenAraligi, kutuplasmaFarki,
                        mutlakKutuplasmaFarkı, jaccardMesafesi]
37          zamanSerileri·append(olcutler)
38      end
```



Şekil 3.4. 12 aylık hareketli sınıflandırma sonuçları

3.1.2. Veri kümesi boyutu istatistiklerinin kutuplaşma ölçütüne etkisinin çıkarılması

Şekil 3.3'de elde edilen kutuplaşma ölçütünün ana kaynağı konuşmaların içeriğinden kaynaklanmasına karşın, veri kümesi boyutu değişkenliğinin de bunda etkisi vardır.

Veri kümesi boyutunun istatistikleri ve kutuplaşma ölçütü arasındaki ilişki korelasyon katsayısı ile tespit edilmiştir. Korelasyon, iki değişkenin birbiriyle ilişkisini gösteren istatistiksel bir ölçüdür. Korelasyonda ilişkinin derecesi korelasyon katsayısı ile ölçülür ve değeri [-1,1] aralığındadır. +1 değeri mükemmel aynı yönlü ilişkiyi, -1 değeri ise mutlak ters yönlü ilişkiyi belirtir. 0 değeri korelasyon olmadığını gösterir, bu da iki değişkenin birbirinden bağımsız olduğu anlamına gelir. Korelasyon katsayısındaki p-değeri, iki değişken arasında gerçek bir korelasyon olmadığını ve ölçülen korelasyon katsayısının şans veya rastgele hatadan kaynaklanma olasılığını gösteren istatistiksel bir ölçüdür. "α değeri" istatistiksel anlamlılık düzeyini ifade eder.

Çalışmada veri kümesi istatistikleri kutuplaşma arasındaki ilişki Pearson korelasyon katsayısı (Rodgers ve Nicewander, 1988) kullanılarak ve α=0,05 seçilerek %95 güven aralığında test edilmiştir. 0,05'ten küçük bir p-değeri istatistiksel olarak anlamlı kabul edilir ve iki değişken arasında gerçek bir korelasyon olduğu sonucuna varmak için güçlü kanıtlar olduğu anlamına gelir.

Parlamento görüşmelerinde konuşma sayısı aydan aya değişmektedir. Parlamentolar genellikle yaz aylarında -Türkiye'de Temmuz, Ağustos ve Eylül- veya seçimlerden önceki aylarda tatile girmektedir. Bu nedenle, veri kümesinde bu aylarda hiç konuşma bulunmamaktadır. Öte yandan, Türkiye'de aralık ayına denk gelen bütçe genel kurul dönemlerinde parlamento yoğun olarak çalışmakta ve çok sayıda konuşma yapılmaktadır.

12 aylık hareketli polarizasyon hesaplanırken kullanılan veri kümesi geçmiş 12 aylık süreyi kapsar. Takip eden iki ayın veri kümelerinde ilk veri kümesini 2-12nci aylarını kapsayan kısım ve sonraki veri kümesinin 1-11nci aylarını kapsayan kısım ortaktır. Takip eden aylar için veri kümelerini farkı ilk veri kümesinin birinci ayı ikinci veri kümesinin sonuncu ayından oluşur. Küme teorisinde iki kümenin benzerliği Jaccard indeksi ile ifade edilir. İki kümenin farkı ise Jaccard mesafesi ile ifade edilir.

$$J_i(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (3.1)$$

$$J_d(A,B) = 1 - J_i(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad (3.2)$$

Eş. 3.1 ve 3.2 Jaccard mesafesinin küme teorisine göre hesaplanışını göstermektedir. Çalışmada ardışık alt veri kümelerinin farkı olarak Jaccard mesafesini kullanılmıştır. Bu, birbirini takip eden bir önceki ayın ilk ayının ve bir sonraki ayın son ayının toplamının iki aydaki tüm konuşmaların toplamına oranıdır.

Veri kümelerindeki değişkenliğin sınıflandırmaya yani kutuplaşma değerine yansıması korelasyon analizi ile incelenmiştir. Zaman serilerine göre takip eden ayların kutuplaşma farkının mutlak değeri ve birbirini izleyen ayların Jaccard mesafesi arasında korelasyon vardır. Şekil 3.5'de bu değerler *"mutlakKutuplasmaFarki"* ve *"jaccardMesafesi"* olarak belirtilmiştir. Çalışmada bu ilişkinin Pearson korelasyon katsayısı %95 güvenle 0,37 olarak bulunmuştur. Ardışık aylar için ölçülen polarizasyon değerleri arasındaki farkın bir kısmının veri kümelerinin bu değişkenliğinden yani Jaccard mesafesinden kaynaklanması doğaldır. Nihai zaman serisini elde etmek için takip eden aylara ait veri kümesinin doküman sayısı farkından kaynaklana etkinin ortadan kaldırılması gerekir.

Takip eden aylar için veri kümelerinin Jaccard mesafesinin kutuplaşma farkının mutlak değerine etkisini çıkarmak için regresyon analizi kullanılmıştır. Regresyon analizinde bağımlı değişken ya da yanıt değişkeni (Y) ardışık ayların polarizasyon farkının mutlak değeri, bağımsız değişken (X) ise veri kümelerinin Jaccard mesafesidir.

Kutuplaşma farkının mutlak değeri kullanıldığı için bu değer her zaman sıfırdan büyük çıkacaktır. Bu değer kesikli hale getirilirse sayma sayıları ile ifade edilebilir.

Bağımlı değişkenin sayma sayısı olduğu yaygın iki regresyon analizi çeşidi vardır. Negatif Binom Regresyonu ve Poisson Regresyonu sayma sayılarını bağımlı değişken olarak kullanan istatiksel modellerdir.

Poisson regresyonu sayma sayısı olan bağımlı değişkenin varyans ve ortalamasının eşit olduğu durumlarda kullanılır. Fakat her zaman bu varsayım sağlanamayabilir ve bu durum

sayma sayıları için yaygın bir sorundur. Çalışmada 800 adet ölçüm için polarizasyon farkının mutlak değerinin varyansı (σ^2) $7{,}64x10^{-6}$ ve ortalaması (μ) 0,0031'dir.

Negatif Binom Regresyonu ise, yanıt değişkeni (bağımlı değişken) varyansının ortalamadan farklı olduğu durumlarda kullanılır. Çalışmada bu nedenle Negatif binom regresyonu kullanılmıştır.

$$\log(\lambda_x) = \propto + \beta X$$
$$\log(\lambda_{x+\Delta}) = \propto + \beta(X + \Delta)$$
$$\log(\lambda_{x+\Delta}) - log(\lambda_x) = \Delta\beta$$
$$\log\left(\frac{\lambda_{x+\Delta}}{\lambda_x}\right) = \Delta\beta \qquad (3.3)$$
$$\frac{\lambda_{x+\Delta}}{\lambda_x} = e^{\Delta\beta}$$
$$\lambda_x = \frac{\lambda_{x+\Delta}}{e^{\Delta\beta}}$$

Eş. 3.3 negatif biyonomiyal regresyonda bağımsız değişkendeki bir değişikliğin (Δ) yanıt değişkeni üzerindeki etkisini göstermektedir. Bağımsız değişkenin katsayısı (β), yanıt değişkeni üzerindeki çarpımsal etkiyi tanımlar. Bağımsız değişkendeki (Δ) kadarlık bir değişimin sonucu olarak elde edilen (x + Δ), yanıt değişkeninde $e^{\beta\Delta}$ kat artış veya azalışa neden olur.

Çalışmada bağımsız değişkenin (x + Δ) değeri, bağımsız değişkenin (x) değerinin karşı olgusu olarak yorumlanmıştır. Bu durumda, bağımsız değişkenin yanıt değişkenindeki etkisi $\lambda_x \leftarrow \frac{\lambda_{x+\Delta}}{e^{\Delta\beta}}$ eşitliği ile giderilebilir. Elde edilen kutuplaşma sonucu, konuşmaların içeriğinin kutuplaşma farkına katkısı olarak yorumlanmıştır.

Şekil 3.5'de kutuplaşma ölçütü *"kutuplasma"*, ardışık ayların polarizasyon farkının mutlak değeri *mutlakKutuplasmaFarki*, işaretli kutuplaşma farkı *"kutuplasmaFarkı"*, ardışık ayların veri kümelerinin Jaccard mesafesi "*jaccardMesafesi"* ile belirtilmiştir. Şekil 3.5'ye girdi olan "*zamanSerileri"* matrisi çalışmada Şekil 3.3'den elde edimiştir. Bu matris bütün parti çiftleri için kutuplaşma ölçütlerini ve veri kümesi istatistiklerini tutan matristir.

Şekil 3.5'de her parti çifti için ardışık iki ayın kutuplaşma değeri arasındaki farkın mutlak değeri alınır. Sürekli değişken olan bu fark negatif binom regresyonunda yanıt değişkeni olarak kullanılabilmesi için sayma sayısına çevrilir. Bu işlem için üs değeri (*"pwr"*) 6 olarak seçilmiştir. 13. satırda sürekli değişken 10^6 ile çarpılarak 6 haneye yuvarlama işlemi uygulanmıştır. Elde edilen sayma sayısı ("*bagimliDegiskenSaymaSayis")* regresyonda kullanılan yanıt değişkenini yani bağımlı değişkenin elemanlarını oluşturur. Regresyonda bağımsız değişken ("*bagimsizDegisken*") ardışık iki ayın veri kümelerinin Jaccard mesafesidir. Jaccard Mesafesi 0'a eşit ise regresyona dâhil edilmemiş ve bu aylarda kutuplaşmanın değişmediği varsayılmıştır.

Bir parti çifti için analizin ilk ayında, önceki ay olmadığı için Jaccard mesafesi 0'dır. Bu yüzden 22. satırda, ilk aydaki kutuplaşma değeri (*"zamanSeileri[i,kutuplasmaIndisi,0]"*) değiştirilmeden başlangıç noktası olarak atanmıştır.

19. satırdaki regresyon analizinin eğim parametresi (β) 0,0274 olarak bulunmuştur. Bir parti çifti için ardışık iki ayın kutuplaşma farkının polarizasyon farkı 25. satırda *y* ile, veri kümelerinin Jaccard mesafesi 26. satırda *x* ile gösterilmiştir. Ardışık aylardaki veri kümelerinin Jaccard mesafesinin etkisi $y_{arindirilmis} \leftarrow y/e^{\beta x}$ eşitliğine göre giderilmiştir. $y_{arindirilmis}$ arındırılmış kutuplaşma farkının mutlak değeridir. Ardışık aylardaki kutuplaşmanın yönü (negatif = azalan kutuplaşma, pozitif = artan kutuplaşma) Şekil 3.5'de *"zamanSerileri[i,kutuplasmaFarkiIndisi,j]" ile* gösterilmiştir. 29. ve 31. satırda; kutuplaşma farkının mutlak değeri yerine işaretli farkı kullanmak için $y_{arindirilmis}$, kutuplaşmanın yönüne göre pozitif veya negatif değer alır. Ardışık veri setlerinin Jaccard mesafesi 0 ise Şekil 3.5'nin 27. satırında işaretli fark sıfıra eşitlenmiştir.

32. satırda, her parti çifti için birinci aydan itibaren kutuplaşma ölçütüne işaretli kutuplaşma farkı $y_{arindirilmis}$ öz yinelemeli olarak eklenerek nihai kutuplaşma ölçütü bulunmuştur.

Her ay için parlamentodaki ortalama kutuplaşma, analizdeki parti çiftlerinin toplam kutuplaşmasının parti çifti sayısına bölünmesi bulunarak "*ortalamaKutuplasma"* elde edilmiştir.

Şekil 3.5. Ardışık alt veri kümesi farkının etkisinin kutuplaşma üzerindeki etkisinin çıkarılma algoritması

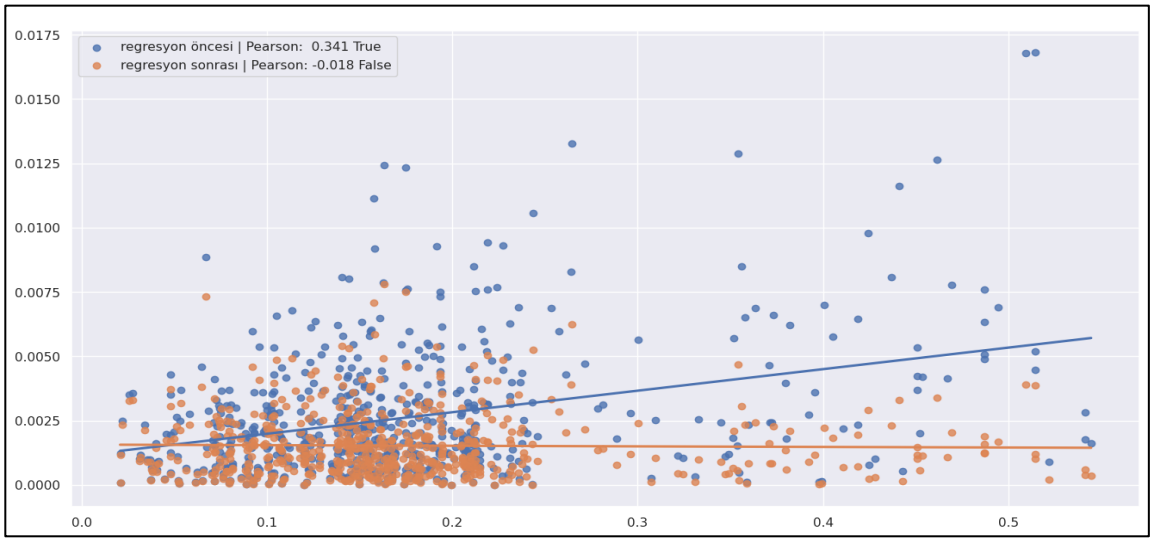
```
    Veri: 12 aylık hareketli sınıflandırma sonuçları, (zamanSerileri)
    Sonuç: Veri kümesi yapısının etkisinden arındırılmış zaman serisi
1   kutuplasmaIndis ← 0
2   kutuplasmaFarkiIndis ← 2
3   kutuplasmaMutlakFarkiIndis ← 3
4   jaccardMesafesiIndis ← 4
5   bagimliDegisken ← []
6   bagimsizDegisken ← []
7   pwr ← 6
8   for i = 0 to partiCiftiSayisi do
9       for j = 0 to aySayisi do
10          x ← zamanSerileri[i, jaccardMesafesiIndis,j]
11          y ← zamanSerileri [i, kutuplasmaMutlakFarkiIndis,j]
12          if  x > 0 then
13              bagimliDegiskenSaymaSayisi ← round(y,pwr) . 10^pwr
14              bagimliDegisken · append(bagimliDegiskenSaymaSayisi)
15              bagimsizDegisken · append(x)
16          end
17      end
18  end
19  istatikselModel ← negatifBinom(y = bagimliDegisken, x = bagimsizDegisken) /* log(
    y)= α + β x */
20  β ← istatikselModel.params[1]
21  for i = 0 to partiCiftiSayisi do
22      tempKutuplasma ← zamanSerileri[i, kutuplasmaIndis,0]
23      for j = 0 to aySayisi do
24          x ← zamanSerileri[i,dif f jaccardMesafesiIndis,j]
25          y ← zamanSerileri[i,kutuplasmaMutlakFarkiIndis,j]
26          if x=0 then
27              y_arindirilmis ←0
28          else if zamanSerileri[i, kutuplasmaFarkiIndis,j] < 0 then
29              y_arindirilmis ← −1. (y/e^βx)
30          else if zamanSerileri[i, kutuplasmaFarkiIndis,j] > 0 then
```

Şekil 3.5. (devam) Ardışık alt veri kümesi farkının etkisinin kutuplaşma üzerindeki etkisinin çıkarılma algoritması

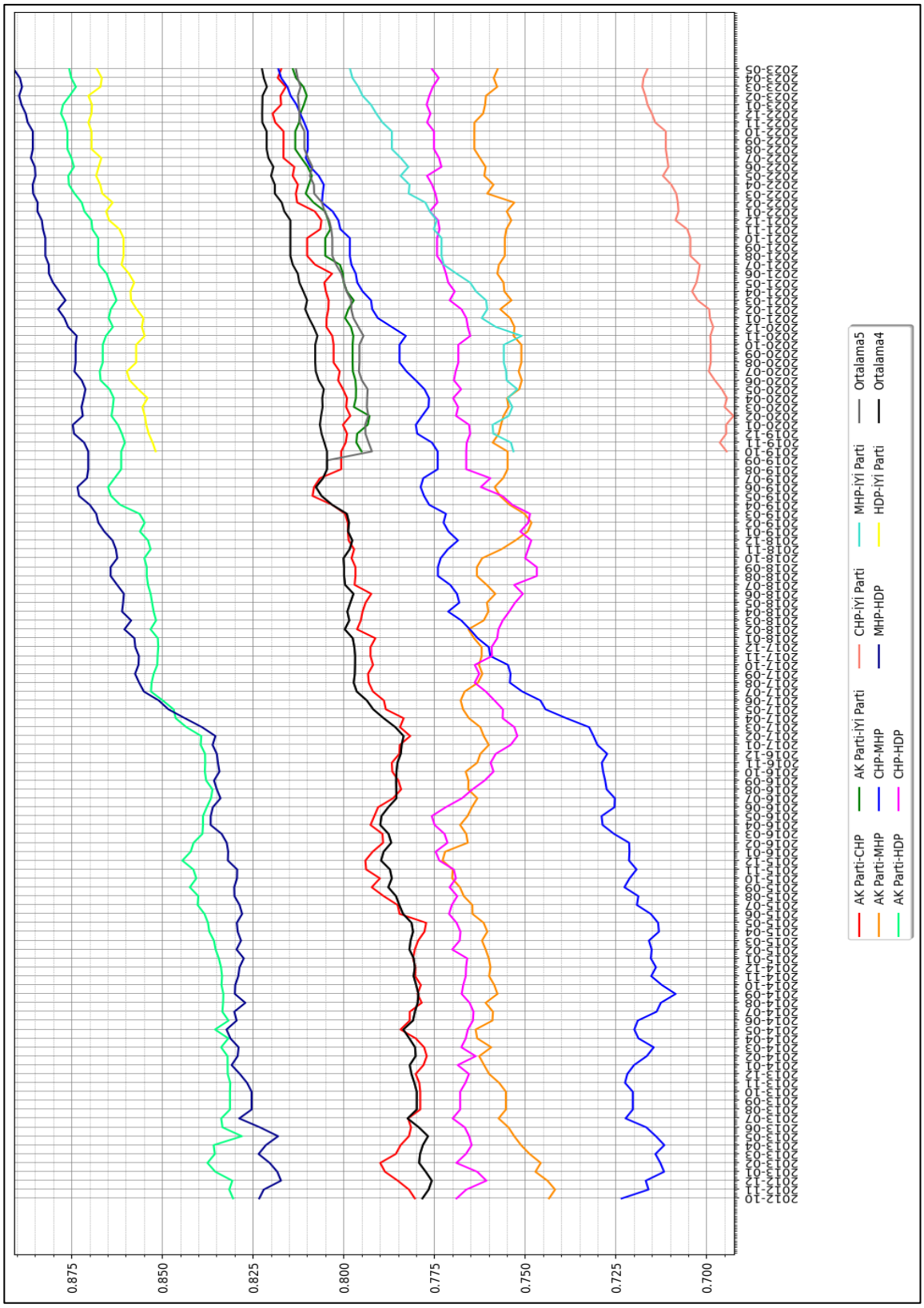
```
                    y_arindirilmis ← y/e^(βx)
                    tempKutuplasma ← tempKutuplasma + y_distilled
                    zamanSerileri[i, kutuplasmaIndis,j] ← tempKutuplasma
                    zamanSerileri[i, kutuplasmaFarkiIndis,j] ← y_arindirilmis
                    zamanSerileri[i, kutuplasmaMutlakFarkiIndis,j] ← |y_arindirilmis|
            end
    end
    ortalamaKutuplasma ← average(zamanSerileri[:,kutuplasmaIndis],axis = 1)
    zamanSerileri.append(ortalamaKutuplasma)
```

Şekil 3.6'da kutuplaşma ölçütünden aylık veri kümesi boyutu değişkenliğinin etkisinin arındırılmadan önceki ve sonraki değerler görülmektedir. Saçılım grafiğindeki korelasyon eğrileri incelendiğinde kutuplaşma mutlak farkı ile veri kümesi Jaccard mesafesi arasındaki 0,34 olan Pearson korelasyonunun giderildiği görülebilir.

Şekil 3.7'de aylık veri kümesi değişkenliğinden kaynaklanan etkiden arındırılmış 12 aylık hareketli kutuplaşma ölçütü görülmektedir Çalışmada yapılan zaman serisi analizi Şekil 3.7'deki nihai zaman serilerine göre yapılmıştır.



Şekil 3.6. Sınıflandırma sonuçlarının birinci dereceden farkı ve veri kümelerinin Jaccard mesafesi korelasyonu

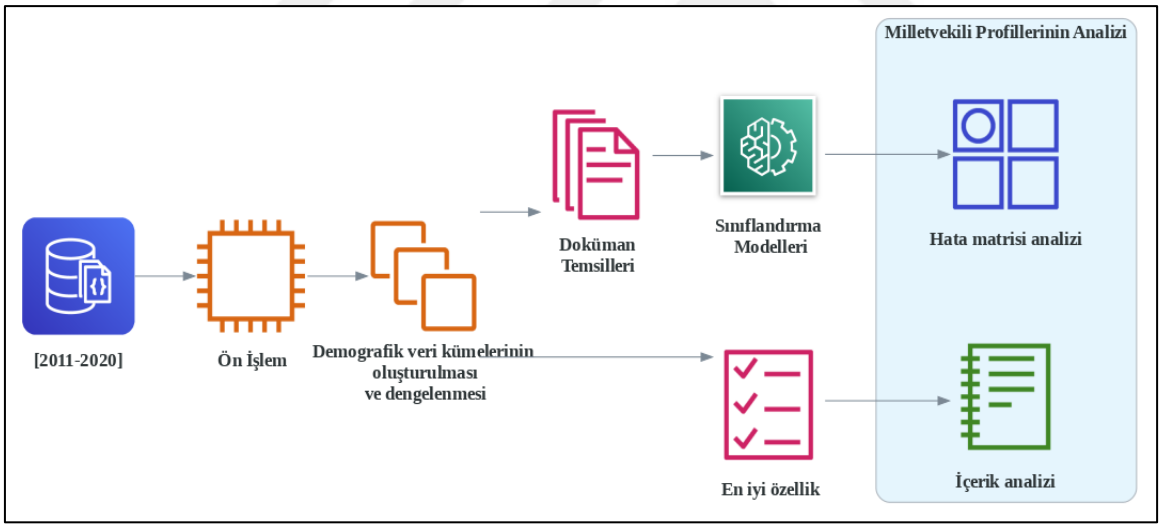


Şekil 3.7. Ardışık veri kümelerinin Jaccard mesafesinin etkisi çıkarıldıktan sonra 12 aylık hareketli kutuplaşma zaman serisi

3.2. TBMM Genel Kurul Görüşmelerinde Yazar Profili Oluşturma (YPO)

Türkiye Büyük Millet Meclisi Genel Kurul konuşmalarından milletvekillerin profillerinin incelendiği çalışmada kullanılan veri kümesi 2011-2020 yılları arasındaki 65 570 bireysel konuşmayı içermektedir.

Demografik özelliklerin tahmini çalışmalarında özellikler; karakter n-gramları, karakter frekansları, cümle uzunlukları, konuşma parçaları, noktalama işaretleri, işlev kelimeleri gibi stil tabanlı özellikler olabileceği gibi kelime n-gramları, terim frekansı-ters doküman frekansı (TF-IDF) gibi içerik tabanlı özellikler veya her ikisinin bir kombinasyonu da olabilir. Rangel, Rosso, Potthast ve Stein (2017) ve Rangel, Rosso, Montes-y-Gómez, Potthast ve Stein (2018) PAN Author Profiling yarışmalarında özellikleri ve sınıflandırıcıları değerlendirmişlerdir. İçerik tabanlı özelliklerin stil tabanlı olanlara ve klasik makine öğrenimi algoritmalarının derin öğrenme tekniklerine göre üstünlüğünü bildirmişlerdir.



Şekil 3.8. Yazar Profili Oluşturma çalışması adımları

Şekil 3.8'de milletvekillerinin demografik ve parlamento özelliklerine göre profillerinin incelendiği çalışmanın adımları görülmektedir. Konuşma metinlerinin ön işlemleri, doküman temsilleri ve sınıflandırma modelleri çalışmanın ana adımlarıdır. Analiz sonuçları hata matrisi analizi ve içerik analizi ile değerlendirilmiştir.

Çalışmada doküman temsili için TF-IDF, LDSE, Bag of Words (BoW), önceden eğitilmiş kelime vektörleri ve paragraf vektörleri (PV) kullanılmıştır. Sınıflandırıcı olarak Lojistik

Regresyon (LR), Destek Vektör Makinesi (DVM) ve İleri Beslemeli Sinir Ağı (İBSA) kullanılmıştır.

3.2.1. Yazar profili oluşturma çalışması için veri kümelerinin oluşturulması

TBMM Genel Kurul görüşmeleri veri kümesi, 2012-2020 yılları arasında genel kurul oturumlarında 1024 milletvekilinin bireysel konuşmalarının 65 570 stenografik transkriptini içermektedir. Türk Parlamentosunda, diğer dünya parlamentolarında olduğu gibi, konuşmalar gözden geçirme prosedüründen sonra tutanaklara yazılmaktadır. Yanlış yazılan kelimeler düzeltilmekte ve ciddi şekilde rahatsız edici kelimeler tutanaklardan çıkarılmaktadır. Böylece dokümanlar, parlamento tutanaklarının yazım prosedürünün bir parçası olarak zaten bu gibi ön işlemlerden geçmektedir. Konuşmalar tutanaklara yazılırken "(alkışlar)", "(gülüşmeler)", "(X parti sıralarından alkışlar)", "(gürültüler)" gibi notlar dokümanlara eklenmektedir. Milletvekillerinin sadece kendi konuşmalarını içeren dokümanları analiz etmek için ön işlemede eklenen tüm notlar temizlenmiştir. Milletvekillerinin soyadları, parti adları ve il adları dokümanda sıklıkla yer almaktadır, ancak bu kelimeler yaş, cinsiyet veya meslek gibi demografik özellikleri analiz ederken çok az anlam ifade etmektedir. Bu nedenle, bu kelimeler dokümanlardan çıkarılmıştır.

Her bir sınıflandırma görevi için 65 570 dokümandan oluşan derlemden alt veri kümeleri oluşturulmuştur. Her veri kümesi, her sınıfta azınlık sınıfının doküman sayısı kadar konuşma olacak şekilde alt-örneklem (under-sampling) altında dengelenmiştir. Böylece her sınıf eşit sayıda dokümana sahip olmuştur.

Yaş alt veri kümesindeki yaş aralığı kategorileri *40 yaş altı, 40-50 yaş arası, 50-60 yaş arası* ve *60 yaş üstüdür.* Bunlar milletvekilinin konuşma sırasındaki yaşına göre oluşturulmuştur. Eğitim ve meslek alt veri kümeleri TBMM web sitesindeki özgeçmişlerine göre oluşturulmuştur. Özgeçmişlerinde meslek bilgisi belirtilmeyen milletvekilleri için meslek, üniversitede okudukları bölüme göre etiketlenmiştir. Örneğin, hukuk fakültesi mezunları mevcut mesleklerini bilmesek de *hukuk* olarak etiketlenmiştir.

Meslek, eğitim ve parti üyeliği görevlerinde, sınırlı sayıda konuşması olan kategoriler göz ardı edilmiştir. Örneğin, meslek sınıflandırmasında eczacıları, gazetecileri, ilahiyatçıları ve diğer bazı meslekleri göz ardı edilmiş ve meslek veri kümesinde doküman sayısı 65 570'den

45 556'ya düşürülmüştür. Meslek sınıflandırması görevinde dört kategori kullanılmış ve azınlık sınıfı tıptır. Her sınıf azınlık sınıfının doküman sayısı 6873 doküman ile temsil edilmiştir. Azınlık sınıfına az örnekleme yapıldıktan sonra, meslek sınıflandırmasının alt veri kümesi 45 556'dan 27 492'ye düşürülmüştür

Tüm derlem ele alındığında doküman kelime uzunlukları 20 ile 800 arasında değişmektedir. Dokümanların ortalama uzunluğu 480 kelimedir. Çizelge 2, milletvekillerinin demografik özellikleri ve kategorilerin istatistikleri görülmektedir. N-Gram, toplam 1-gram, 2-gram ve sayısını gösterir.

Çizelge 3.1. Milletvekillerinin demografik özelliklerini sınıflandırmak için oluşturulan veri kümeleri

Milletvekilin Özelliği	Sınıflar	Milletvekili Sayısı	Konuşma Sayısı	N-Gram
Cinsiyet	Kadın	199	8937	286 000
	Erkek	820	8937	
Yaş	40 yaş altı	146	4762	304 300
	40-50	411	4762	
	50-60	464	4762	
	60 üstü	284	4762	
Eğitim	Profesör ya da Doçent	102	3807	244 500
	Doktora	61	3807	
	Yüksek Lisans	173	3807	
	Lisans	416	3807	
Meslek	Hukuk	251	6873	439 600
	Ekonomi ve Finans	129	6873	
	Tıp	110	6873	
	Mühendislik	177	6873	
Parti	AK Parti	602	10 052	693 900
	CHP	293	10 052	
	MHP	108	10 052	
	HDP	124	10 052	
Parti Durumu	İktidar	618	16 181	527 600
	Muhalefet	559	16 181	
Seçim Bölgesi	Ege	148	5068	601 500
	Karadeniz	138	5068	
	İç Anadolu	170	5068	
	Doğu Anadolu	116	5068	
	Marmara	297	5068	
	Akdeniz	157	5068	
	Güneydoğu Anadolu	133	5068	

3.2.2. Doküman temsilinde kullanılan özelikler

Metin sınıflandırmanın dönüştürme aşamasında, metin özelliklerle temsil edilir. Özelliklerin vektörel olması hesaplama işlemlerini kolaylaştırır. Çalışmada dokümanların temsili için BoW, TF-IDF, LDSE, önceden eğitilmiş Kelime Vektörleri ve Paragraf Vektörleri gibi farklı DDİ yöntemleri kullanılmıştır.

- Kelime çantası (Bag of words, BoW): Metin gösteriminde en sık kullanılan yöntemlerden biridir. BoW gösteriminde, vektörün her boyutu dokümandaki bir kelimenin sayısını gösterir.
- Terim frekansı-ters doküman frekansı (Term frequency-inverse document frequency, TF-IDF): Özellikle klasik makine öğrenimi algoritmalarında metin gösterimi için sıklıkla kullanılan bir diğer yöntem de TF-IDF'dir. TF, bir kelimenin bir dokümandaki sıklığını gösterir ve kelimelerin doküman için önemini ortaya koyar. Tüm metinlerde sık kullanılan terimler kelimenin ayırt ediciliğini azaltır. Buna bir çözüm olarak, bir terimin doküman sıklığının tersi (IDF) metin sınıflandırması için kullanılır. BoW ve TF-IDF gösteriminde, kelimelerin sırası göz ardı edilir ve bu durum anlam kaybına yol açabilir.
- Düşük boyutlu istatistiksel yerleştirme (Low dimensionality statistical embedding, LDSE): Rangel, Franco-Salvador ve Rosso (2018) bu yaklaşımı dil çeşitliliği Yazar Profili Oluşturma görevinde ayrıntılı olarak açıklamıştır. Ana felsefe, belirli bir metindeki kelimenin ağırlık dağılımının ilgili kategoriye daha yakın olması gerektiğidir. Çalışmada kelime ağırlığı için TF-IDF değeri kullanılmıştır. LDSE'de, sınıflardan her birinin ağırlıkları, bu sınıfa ait dokümanların ağırlıklarından hesaplanır. Sınıfların terim ağırlıklarının her biri, sınıfa bağlı ağırlıkların toplamı ile terimin tüm ağırlıklarının toplamı arasındaki orandır. Sınıflar için bu ağırlık metriğine sınıfa bağlı ağırlık matrisi denir. Hem eğitim hem de test setleri için dokümanların nihai temsili, sınıfların her biri için kelime ağırlıklarına ortalama, standart sapma, minimum, maksimum, olasılık ve oran gibi toplam fonksiyonları uygulanarak bulunur. Böylece, bir dokümanı birkaç sayıda ağırlıkla temsil edilir hale gelir (kategori sayısı ile fonksiyon sayısının çarpımı). Bu az sayıda ama çok açıklayıcı ağırlıklar dokümanları temsil etmek için kullanılabilir. LDSE, veri kümesindeki tüm terimleri dikkate alma ve dolayısıyla uygulanan tüm özellikleri temsil etme avantajına da sahiptir. Bunun aksine, yaygın olarak kullanılan PCA (Principal Component Analysis) veya LSA (Latent Semantic Analysis) boyut

azaltma teknikleri, daha az katkıda bulunan ve veri kümesindeki tüm özellikleri temsil etmeyen terimleri kaldırır.

Çalışmada, LDSE temsilini elde etmek için unigram, bigram ve trigram TF-IDF ağırlıkları kullanılmıştır. Dokümanların nihai temsilini elde etmek için sınıfa bağlı ağırlık matrisi üzerinde altı toplama fonksiyonu kullanılmıştır. Bu fonksiyonlar ortalama, standart sapma, minimum, maksimum, olasılık ve oran fonksiyonlarıdır.

- Önceden eğitilmiş kelime vektörleri (Pretrained word vectors): Harris (1954) anlamsal olarak benzer kelimelerin benzer bir dağılıma sahip olduğunu belirtmiştir. Kelime gömme ve kelime vektörleri, kelimeleri bir vektör uzayında anlamsal olarak temsil eder. Kelime yerleştime için en çok word2vec (Mikolov, Chen, Corrado ve Dean, 2013), GloVe (Pennington, Socher ve Manning, 2014), fastText (Bojanowski, Grave, Joulin ve Mikolov, 2017), ELMo (Peters ve diğerleri, 2018) ve BERT (Bidirectional Encoder Representations from Transformers) (Devlin, Chang, Lee ve Toutanova, 2018) modelleri kullanılır.

Mikolov ve diğerleri (2013) word2vec modelinde Yapay Sinir Ağlarını (YSA) kullanmıştır. Word2vec modeli, bir kelimenin çevresindeki kelimelerden tahmin edilmesine (Continuous Bag-Of-Words, CBOW) veya verilen bir kelimeden çevresindeki kelimelerin tahmin edilmesine (Skip-gram) dayanmaktadır. Word2vec'te YSA kelime dizisini kabul eder ve eğitimin sonunda kelime vektörleri üretir.

Çalışmada 1994'den 2023’e kadar olan Genel Kurul konuşmaları birleştirerek bir eğitim derlemi oluşturuldu. Kelime vektörlerini elde etmek için derlemi word2vec CBOW modeli ile eğitilmiştir. Eğitim derlemi nedeniyle, önceden eğitilmiş kelime vektörleri parlamento görüşmeleri için alana özgüdür. Kelime vektörlerinin kelime büyüklüğü 440.705 ve boyutluluğu 300'dür.

- BERT (Bidirectional encoder representations from transformers): Word2vec, GloVe ve fastText ile elde edilen kelime vektörleri kelimeler arasında anlamlılık taşımasına rağmen, her kelime için yalnızca tek bir global temsile sahiptir. Anlamsal kelime yerleştirme modellerindeki tek temsil, kelimenin bağlamdaki anlamını temsil etmemesine yol açar. Anlamsal kelime vektörlerinin yanı sıra ELMo ve BERT gibi

kelimeyi bağlamlarıyla birlikte temsil eden bağlamsal kelime yerleştirme modelleri mevcuttur. ELMo, çift yönlü UKSB kullanarak sol ve sağ bağlamları kodlar. BERT, UKSB yerine Transformer (Vaswani ve diğerleri, 2017) modelini kullanır. Dönüştürücüler (Transformer), dikkat mekanizmalarına sahip bir kodlayıcı-kod çözücü mimarisidir. BERT, bir dil temsil modeli oluşturmak için dönüştürücünün yalnızca kodlayıcı tarafını kullanır. Ön eğitim sırasında BERT'in tahmin görevleri Maskeli Dil Modelleme (Masked Language Model, MLM) ve Sonraki Cümle Tahminidir (Next Sentence Prediction, NSP). MLM, temsilin sol ve sağ bağlamı birleştirmesini sağlar. İnce ayar aşamasında BERT, Dönüştürücülerdeki dikkat mekanizması nedeniyle birçok alt görevi modeller.

Schweter (2020) Türkçe bir BERT modeli olan BERTurk'ü tanıtmıştır. Model, 35 GB boyutunda bir Türkçe derlem üzerinde ön eğitime tabi tutulmuştur. Özellik kestirimi görevlerinde, BERTurk'ün 128 000 kelime boyutlu versiyonuna ince ayar yapılmıştır.

- Paragraf vektörleri (PV): Paragraf vektörlerinde her vektör kelime dizisini temsil eder. Dizi bir cümle veya paragraf olabilir. Doc2vec (Le ve Mikolov, 2014) ve Skip-Thought (Kiros ve diğerleri, 2015) cümle gömme için en çok kullanılan modellerdir. Doc2vec, word2vec'in bir uzantısıdır ve bir paragraf belirtecini (paragraf vektörü) kelime vektörlerine birleştirerek tahmin görevi yapar. Yazarlar Doc2vec modelinin çıktısını Paragraf Vektörleri (PV) olarak adlandırmıştır. Bir kelime dizisi bir paragraftan rastgele örneklenir ve örneğin merkezindeki kelimeyi tahmin eder. "Paragraf belirtecinin, paragrafın mevcut bağlamında veya konusunda neyin eksik olduğunu göstermek için bellek görevi gördüğünü" belirtmişlerdir. Bu modele Paragraf Vektörlerinin Dağıtılmış Bellek Modeli (PV-DM) adı verilmiştir. Paragraf vektörünün bir başka yaklaşımı da Dağıtılmış Kelime Torbası (Distributed Bag-Of-Words, DBOW) versiyonudur. PV-DBOW bir paragraf vektörünü girdi olarak alır ve paragraftan rastgele örneklenen kelimeleri tahmin eder.

Çalışmada PV-DBOW, PV-DM ve her iki modelin birleşimi arasından en iyi doğruluk skoruna sahip olması nedeniyle PV-DBOW kullanılmıştır. Özellik vektörlerinin boyutluluğu 100 ve negatif örnekleme 5'tir.

3.2.3. Kullanılan makine öğrenmesi ve derin öğrenme modelleri

Milletvekillerinin demografik özelliklerini belirlemek için TF_ID, BoW, LDSE, önceden eğitilmiş kelime vektörleri ve paragraf vektörü doküman temsilleri ile klasik makine öğrenimi algoritmaları ve derin öğrenme teknikleri kullanılmıştır. Çalışmada model isimleri *doküman temsili_sınıflandırıcı* olarak verilmiştir.

- TF-IDF_DVM: Destek Vektör Makinaları (DVM), verileri farklı kategorilere ayırmak için farklı hiper düzlemler oluşturur. DVM, metin sınıflandırması için en iyi makine öğrenimi algoritması olarak düşünülmüştür (Joachims, 1997). Doküman gösterimi için TF-IDF değerlerini ve sınıflandırıcı olarak Destek Vektör Makineleri kullanılmıştır.
- TF-IDF_LR*:* Lojistik Regresyon (LR), karar sınırı olan doğrusal bir sınıflandırıcıdır. Bir lojistik fonksiyon, bir örneğin sınıfa ait olma olasılığını belirler. Modelde doküman gösterimi için TF-IDF değerleri ve sınıflandırıcı olarak Lojistik Regresyonu kullanılmıştır.
- LDSE_LR: dokümanların LDSE gösterimini kullanılmıştır. LDSE ağırlıkları dokümanların TF-IDF değerlerinden çıkarılmıştır. Sınıflandırıcı olarak Lojistik Regresyon kullanılmıştır.
- PV_LR: Dokümanları temsil ederken paragraf vektörleri kullanılmıştır. Doğruluk ölçütlerine göre DVM ile Lojistik Regresyon karşılaştırdıktan sonra sınıflandırıcı olarak LR algoritması seçilmiştir.
- BERT: Demografik özellik tahmini görevlerinde metin sınıflandırması için Türkçe BERT modeli BERTurk'e ince ayar yapılmıştır. BERTurk temel modeli 128k ön eğitimli temsile (https://huggingface.co/dbmdz/bert-base-turkish-128k-uncased, Son Erişim Tarihi: 2 Temmuz 2022), on iki dönüştürücü kodlayıcı (transformer encoder) bloğuna, 12 öz-dikkat başlığına (self-attention) ve 110M parametreye sahiptir.
- BoW_İBSA ve word2vec_İBSA: Doküman temsili için BoW temsilini ve word2vec kelime vektörleri kullanılmıştır. Ön eğitimli kelime vektörleri TBMM Genel Kurul görüşmelerine özgüdür. Sınıflandırıcı olarak Yapay Sinir Ağlarını kullanılmıştır. Evrişimli Sinir Ağı (ESA) (Convolution Neural Network, CNN) (Kim, 2019), Tekrarlayan Sinir Ağı (TSA) (Recurrent Neural Network, RNN) (Graves, 2013) ve İleri Beslemeli Sinir Ağı (İBSA) topolojileri ile yapılan metin sınıflandırma deneylerinden sonra, en iyi doğruluğa sahip olduğu için İBSA tercih edilmiştir. Ağın topolojisi gizli

katmanda 512 nöron, aktivasyon fonksiyonu Relu, dopout oranı 0,5 ve optimize edici yaklaşım Adam'dır.

3.2.4. En iyi özellik analizi

En iyi özellik analizi, bir derlemdeki belli bir kategoriyi en iyi temsil eden terimlerin, sözcüklerin belirlenmesi görevidir. Özellik seçilimi (feature selection), sınıflandırma görevi için en ilgili özelliklerin seçilmesi sürecidir. Bu nedenle, demografik özellik kategorilerinin en önemli terimlerini belirlemek için özellik seçimi tekniklerinden yararlanılmıştır. Metnin konusunu yansıtan en alakalı kelimeler sıfatlar ve isimlerdir. Milletvekillerinin özelliklerine göre konuların değişimini araştırmak için analizde metinler oluşturulurken metinde geçen konu kelime kökleri (isim ve sıfat) kullanılmıştır.

Çalışmada bir demografik özelliğin kategorileri içi en iyi özellikler üç adımda elde edilmiştir. İlk olarak, χ^2 (ki-kare) özellik seçimi algoritması kullanılmış ve her bir terim için χ^2 değeri için hesaplanmıştır. İkinci adım olarak TF-IDF değerlerinin kategoriler üzerindeki dağılımı hesaplanmıştır. TF-IDF dağılımını elde etmek için Rangel, Franco-Salvador ve Rosso (2018) çalışmasında sınıf bağımlı matris kullanılmıştır. Ardından her bir terim, ikinci adımda elde edilen değerlere göre belli bir kategoriye atanmıştır. TF-IDF değerlerinden terimi kategorilere atamak yanında özellik seçilimde de yararlanmak için kategori dağılımı belli bir eşik değerini geçen özellikler seçilmiştir. Çalışmada bu eşik değer %10 olarak seçilmiştir. Son adım olarak, her bir kategori için en yüksek χ^2 değerine sahip terimler sınıfların en iyi özellik (best features) olarak seçilmiştir.

$$\chi^2 = \frac{(O_i - E_i)^2}{E_i} \tag{3.4}$$

Eş 3.4'de χ^2 , i terimi için beklenen frekans değeri E_i, gözlemlenen değer ise O_i'dir.

$$\Delta = \begin{bmatrix} W_{11}\ W_{12} \ldots W_{1m}\ \delta(d_1) \\ W_{21}\ W_{22} \ldots W_{2m}\ \delta(d_2) \\ \ldots \quad \ldots \quad \ldots \quad \ldots \\ W_{n1}\ W_{n2} \ldots W_{nm}\ \delta(d_n) \end{bmatrix} \tag{3.5}$$

Eş 3.5'de Δ matrisi dokümanların tf-idf matrisidir. Matristeki her satır bir d dokümanını, her sütun sözlükteki bir t kelime terimini, W_{ij} onun tf-idf değerini ve $\delta(d_i)$ i dokümanının atanmış c sınıfını, yani bu dokümanın gerçekte demografik özelliğini temsil eder.

$$W_{(t,c)}=\frac{\sum d\in D/c=\delta(d)^{W}dt}{\sum d\in D^{W}dt} \tag{3.6}$$

Eş 3.6'da elde edilen terim ağırlığı $w_{(t,c)}$ kategori c'ye ait dokümanların ağırlıkları ile ilgili terim t için toplam ağırlık dağılımı arasındaki orandır.

$$\beta=\begin{bmatrix} w_{11}\ w_{12}\ \dots\ w_{1m} \\ w_{21}\ w_{22}\ \dots\ w_{2m} \\ \dots\quad \dots\quad \dots\ \dots \\ w_{k1}\ w_{k2}\ \dots\ w_{km} \end{bmatrix} \tag{3.7}$$

Eş 3.7'de sınıf bağımlı matris β, terimlere ait w ağırlıklarının k adet sınıfların her biri üzerindeki yüzdelik dağılımını gösterir.

Çalışmada bir kategoriyi en iyi açıklayan terimler bulunurken χ^2 değerleri büyükten küçüğe doğru sıralanmış ve her bir kategori için eğer bir terimin β matrisi içindeki değeri, söz konusu kategorinin ilk %10'luk kısmında ise seçilmiştir.

3.2.5. Başlangıç noktası

Her bir sınıflandırma görevi için, modelleri değerlendirmek üzere iki başlangıç noktası (Baseline) yöntemi kullanılmıştır.

- Temel-çoğunluk: Her zaman çoğunluk sınıfını tahmin eden istatistiksel bir taban çizgisi. Sınıflar dengeli olduğu için n sınıflandırmadaki kategori sayısını göstermek üzere bu değer (1/n)'e eşittir.
- Baseline-VecAvg: Socher ve diğerleri (2013) dokümanları temsil etmek için kelime vektörlerinin ortalamasını kullanmıştır. Demografik özelliklerin tahmini görevleri için VecAvg modelini taban çizgisi olarak kullanılmıştır. Modelde, kelime vektörleri word2vec'ten elde edilmiş ve sınıflandırıcı olarak LR kullanılmıştır.

3.3. TBMM Genel Kurul Görüşmelerinde Yakın Anlamalı Kavramlar

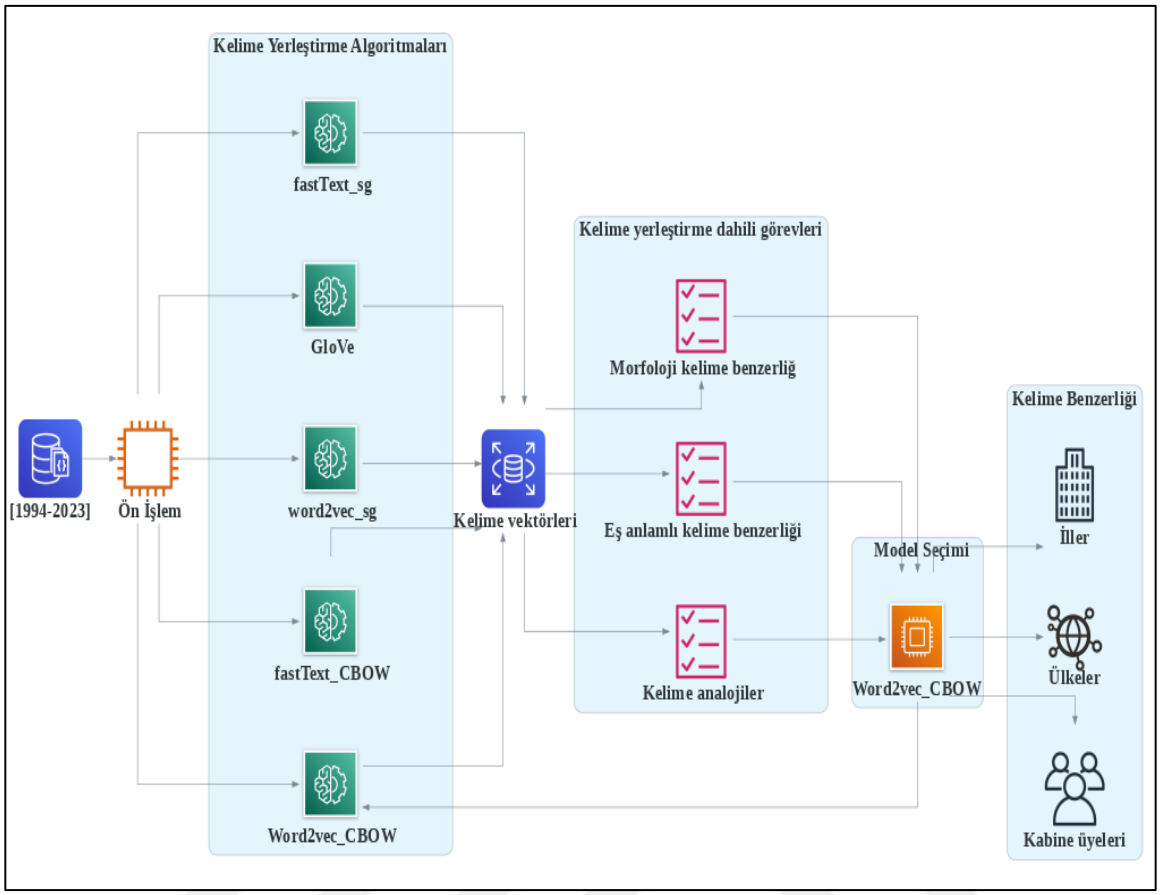
Bu çalışmada, TBMM Genel Kurul görüşme tutanaklarındaki benzer kavramların çıkarılması için kelime yerleştirme algoritmaları ile elde edilen kelime vektörleri kullanılmıştır. Çalışmada kullanılan veri kümesinin özellikleri şunlardır;

- Veri kümesi TBMM parlamento çalışma alanıyla ilgilidir ve dili Türkçedir.
- TBMM genel kurul tutanakları, konuşma dilinin doğrudan yazıya aktarılması şeklinde olduğu için Türkçenin bölgelere göre farklılığını yansıtır. Türkçede kullanılan bütün kelimeler, deyimler dokümanlarda mevcuttur.
- Tutanaklar yazıya geçirilirken TBMM tutanak uzmanları tarafından düzeltildiği için kelimelerin yanlış yazımı, kısaltılması gibi metin analizini zorlaştıran durumlar söz konusu değildir. Özellikle alan yazında çok fazla çalışmanın olduğu Twitter, Facebook gibi sosyal medya mecralarında kelimelerin kısaltılması, yazarken değiştirilmesi gibi metin analizini zorlaştıran durumlar bu veri kümesinde gözlenmez.
- TBMM Genel Kurul oturumlarında yazılı dokümanların (Tasarı, önergeler vs.) kâtip üye tarafından okunması tutanağa geçirildiği için hem konuşma dilinin hem de yazı dilinin dokümanlarda bulunması bu anlamda bir çeşitlilik oluşturur.

Birleşim, TBMM Genel Kurulunun belirli bir günde açılan toplantısıdır[1]. Oturum, bir birleşimin ara ile bölünen kısımlarından her birisidir. Bu çalışmada kullanılan veri kümesi, 01.09.1994 ve 23.04.2023 tarihleri arasındaki 3625 adet Genel Kurul birleşim tutanaklarını içerir.

Şekil 3.9 milletvekili konuşmalarından oluşturulan derlemden elde edilen kelime vektörleri ile yapılan TBMM Genel Kurul tutanaklarında yakın anlamlı kavramların çıkarılması çalışmasının ana adımları gösterilmektedir. Konuşma metinleri ön işlemden geçirildikten sonra kelime yerleştirme algoritmaları ile kelime vektörleri elde edilmiş, elde edilen vektörlerin vektör uzayındaki yakınlıklarından iller, ülkeler ve kabine üyelerinin kelime benzerliği analiz edilmiştir.

[1] https://www.tbmm.gov.tr/psozluk.htm



Şekil 3.9. TBMM Genel Kurul tutanaklarında yakın anlamlı kavramlar çıkarılmasının adımları

3.3.1. Kelime benzerliği için veri kümesinin oluşturulması

Veri kümesi oluşturulurken açık veri olarak TBMM İnternet sitesinde https://www.tbmm.gov.tr/tutanak/tutanaklar.htm adresinden yayınlanan TBMM Genel Kurul Tutanak metinlerinden yararlanılmıştır.

Veri kümesinde bulunan 2800 birleşim dosyası, Doğal Dil İşleme (DDİ) metotları uygulanarak tek bir metin dosyası haline getirilmiştir. Türkçe DDİ metotlarının uygulanması aşamasında Java tabanlı Zemberek[2] kütüphanesi kullanılmıştır. Oluşturulan dosya üzerinde gerçekleştirilen DDİ adımları ve dosyanın özellikleri şunlardır:

- Birleşim metinleri bölütlere (token) ayrılmıştır. Her bir bölüt bir kelimeyi ifade eder ve sözlükteki en küçük atomik yapıdır.

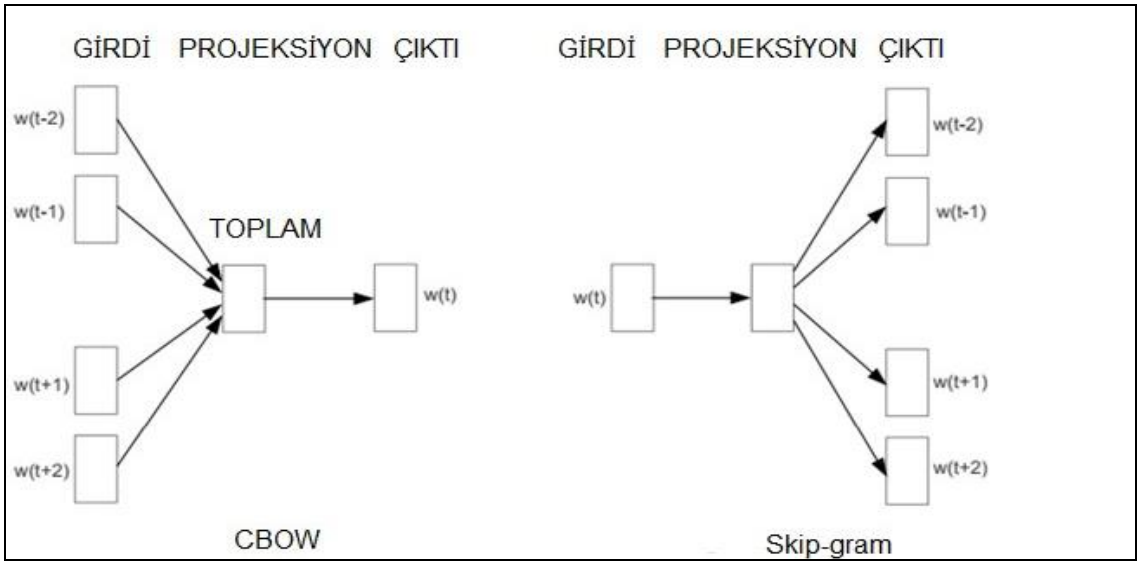
[2] https://github.com/ahmetaa/zemberek-nlp

- Zemberek kütüphanesinden yaralanarak bölüt tipleri çıkarılmıştır. Tarih, Sayı, Noktalama İşareti bölüt tipine ait olan metin analizinde kullanılmayan sözcükler filtrelenerek derlem dışında tutulmuştur.
- Sondan eklemeli bir dil olan Türkçede aynı kelime sonuna ekler alarak birden çok biçimde metinde yer alabilir. Bu çalışmada olduğu gibi sözcüklerin benzerliğinin ve analojilerin çıkarılması için aynı anlamlı fakat sentaktik olarak farklı sözcüklerin tek bir kelime vektörü ile temsil edilmesi önemlidir. Örneğin "Sepetteki elmalar çürümüştü.", ve "sepetteki elmaların rengi çok güzel." Cümlelerinde "elmalar" ve "elmaların" sözcüğü aynı anlama fakat farklı yazılışa sahiptir. Bu sözcük "elma" olarak standartlaştırılmazsa derlemde anlamsal olarak aynı sözcük onlarca farklı kelime vektörü ile temsil edilebilir. Çalışmada sözcüklerin standartlaştırılması ve aynı kavramların tek bir sözcük altında toplanması için lematizasyon kullanılmıştır. Kelime kökü kelimenin sonuna aldığı ek çıkarılarak elde edilir. Lematizasyon ise morfolojik analiz yaparak elde edilen kelimenin o dile ait sözlükte bulunmasını şart koşar. Zemberek kütüphanesinden yaralanarak her bir cümlenin dilbilimsel analizi yapılmıştır. Analiz sonucu her bir sözcük için elde edilen kelime başı (lemma) alternatiflerinden en uzunu seçilmiştir. En uzun kelime başı seçilmesindeki amaç kelimenin anlam kaybına uğramasını en aza indirgemektir.
- Bölütlerden oluşan TBMM birleşim dosyaları bir araya getirilerek art arda eklenmiş kelimelerden oluşan yeni bir metin dosyası elde edilmiştir. Bu dosyanın boyutu 718 MB'dır.

Kelime yerleştirme algoritmalar ile elde edilen kelimelerin vektör vektörleri, kelimelerin semantik anlamını içerdiği için DDİ çalışmalarında sıkça kullanılır. 1994-2023 yılları arasındaki milletvekili konuşmalarını kapsayan veri seti oluşturulduktan sonra kelime vektörlerinin çıkarılması için word2vec, fastText ve Glove kelime yerleştirme algoritmaları kullanılmıştır.

3.3.2. Word2vec[3]

Mikolov, Chen ve diğerleri (2013) word2vec modeli basitçe (Harris, 1954)'in anlam olarak benzer kelimelerin benzer bağlamlarla birlikte bulunacağı mantığına dayanır. Bu çalışmada kelime vektörlerinin oluşturulması için word2vec modeli kullanılmıştır. Word2vec modeli, CBOW (Continous Bag of Words) ve Skip-gram olarak adlandırılan iki farklı dil modelinden oluşur. Her iki modelde de bir pencere bir derlem boyunca kayar, her bir adımda yapay sinir ağı pencere içindeki kelimelerle (bağlam) eğitilir.



Şekil 3.10. Word2vec CBOW ve Skip-gram modeli

Şekil 3.10 Mikolov, Chen ve diğerleri (2013) çalışmalarındaki Şekil 1'de tanıttıkları word2vec modelinin görsel anlatımıdır. Her iki modelde de bir pencere içindeki kelimeler derlem (corpus) boyunca kayar, her bir adımda yapay sinir ağı pencere içindeki kelimelerle eğitilir. CBOW modeli derlem boyunca her bir kelimenin bu pencerenin ortasında bulunup bulunmadığını tahmin eder. CBOW modelinde, derlemdeki her bir kelime için; sinir ağına girdi olarak verilen bağlamın (w_{t-1}, w_{t-2}, w_{t+1}, w_{t+2}) merkezindeki kelime olan w_t olma ihtimali hesaplanır ve sistem doğru kelimeyi bulması için eğitilir.

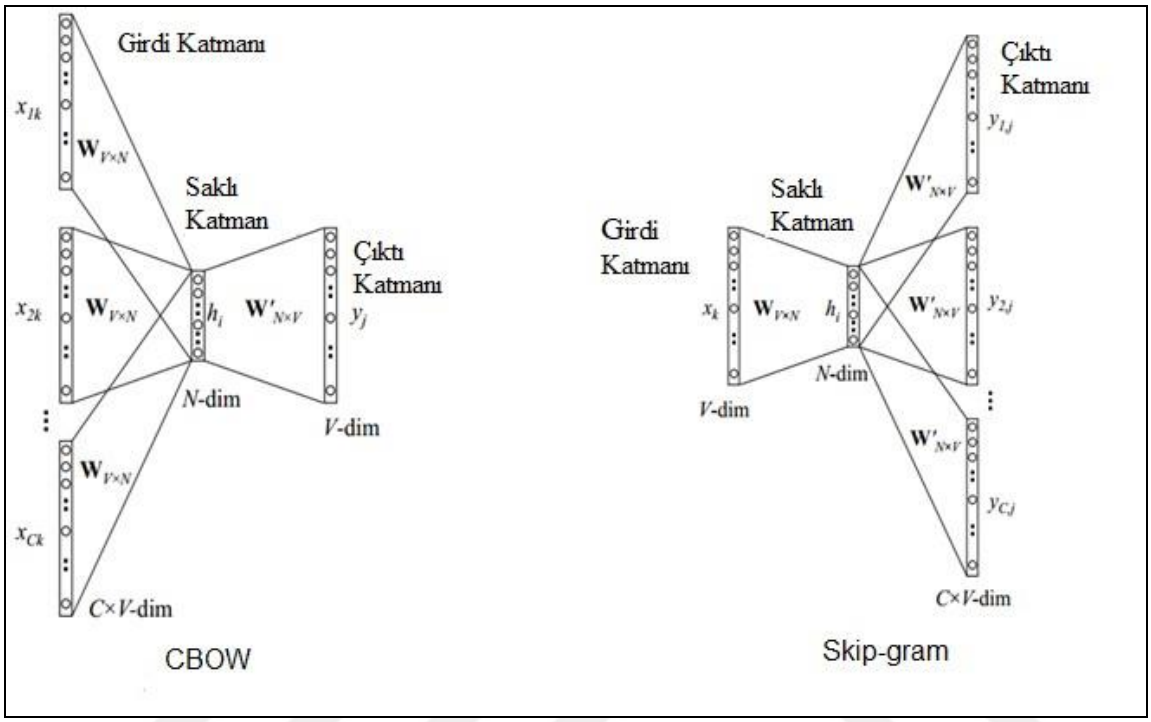
[3] https://code.google.com/archive/p/word2vec/

Skip-gram modelinde ise pencerenin ortasındaki kelimeden, etrafındaki kelimeler (bağlamı) tahmin edilmeye çalışılır. Skip-gram Modeli CBOW modelinin tersidir. Sinir ağına verilen girdi w_t vektöründen bağlam (w_{t-1}, w_{t-2}, w_{t+1}, w_{t+2}) kelimeleri tahmin edilmeye çalışılır.

Mikolov, Sutskever, Chen, Corrado ve Dean (2013) word2vec modelinde eğitim hızını artırmak için sinir ağının çıktı katmanımda negatif örneklem (negative sampling) kullanmışlardır. Negatif örneklemde merkezi kelime bağlam kelimelerinin ve rastgele seçilen N (N=4) adet kelime (w_{ns1}, w_{ns2}, w_{ns3}, $w_{ns4)}$ ile karşılaştırılır. Merkezi kelimenin rastgele seçilen kelimelerle aynı bağlamda olmadığı kabul edilir. Bu durumda tahminlerden elde edilen hata bağlam kelimeleri için 1-P(w_{t-1}) , negatif örneklem kelimeleri için P(w_{ns}) olacaktır.

Word2vec yalnızca bir tane saklı katmandan oluşan derin olmayan yapay sinir ağı yapısındadır. Şekil 3.11. (Rong, 2014)'un word2vec modelini açıkladığı çalışmasındaki Şekil 2'deki CBOW ve Skip-gram modelinin sinir ağı yapısını gösterir. Derlemin eğitimi sonunda sinir ağının ağırlık parametreleri W_{VXN}'den kelime vektörleri elde edilir. Boyutu derlemdeki sözcük sayısı (V) kadar olan vektörlerden, C adeti (pencere boyu) girdi vektörü olarak kullanılır. Pencere boyutu bağlam olarak da düşünülebilir. Bir kelimeyi temsil etmesi için isteğe bağlı olarak seçilen ve yapay sinir ağı için bir hiper-parametre olan N kelime vektörlerinin boyutunu belirler. N boyutlu bir kelime vektöründe N tane öznitelik olduğu düşünülebilir. Saklı katmanda N tane nöron vardır. Girdi katmanı ve saklı katman arasında [VxN] boyutlu ağırlık matrisi W; saklı katman ve çıktı katmanı arasında [NxV] boyutlu W' ağırlık matrisi bulunur. Çıktı katmanın sonucunda derlemdeki her bir kelimenin hedef kelime olma ihtimalini gösteren değerlerden oluşan V boyutlu vektör elde edilir. Bu vektör hedef kelime ile karşılaştırılarak sinir ağı geri yayılımlı olarak eğitilir. Girdi katmanında V boyutla temsil edilen kelime vektörleri, ağırlık matrislerinde boyutları indirgenerek, N boyutla olarak temsil edilebilir hale gelir. Yapay sinir ağının eğitilmesi sonucunda boyutu NxV olan ağırlık matrisleri elde edilir. CBOW modelde gizli katman ve çıktı katman arasındaki W' ağırlık matrisinden kelime vektörleri elde edilir. Skip-gram modelde girdi katmanı ve gizli katman arasındaki W ağırlık matrisinden kelime vektörleri elde edilir. Bu kelime vektörleri anlam benzerliğine göre vektör uzayında birbirine yakın konumlanırlar ve semantik bilgi içerirler.

TBMM Genel Kurul tutanaklarında anlamsal yakınlık elde etmek için word2vec modeli uygulanırken yapay sinir ağının eğitilmesi sonucunda çıkan kelime vektörlerinin boyutu 300'dür. Derlemde 5'den az sayıda geçen kelimeler göz ardı edilmiştir.



Şekil 3.11. Word2vec CBOW ve word2vec Skip-gram modelinin sinir ağı yapısı

3.3.3. FastText

FastText (Bojanowski, Grave, Joulin ve Mikolov, 2017) modeli kelime tabanlı word2vec modelinin karakter tabanlı varyantı olarak değerlendirilebilir. Word2vec modelinde eğitim aşamasından sonra kelime vektörleri elde edilir. Ancak eğitim aşamasında derlemde bulunmayan bir kelimenin daha sonra karşılaşılması durumunda bu kelimenin anlamsal karşılığı elde edilemez. Oysa karakter tabanlı bir modelde eğitim karakter n-gramlarla yapıldığı için, muhtemele tüm karakter varyantlarının anlamsal karşılığı vardır. Eğitim aşamasından sonra sözlüğe yeni bir kelime girmesi durumunda bu yeni kelimenin karakter n-gram karşılığı eğitim aşamasındaki karakter n-gramlar varyantlarında bulunacaktır. FastText'in bu özelliği sözlük dışı kelime (SDK) (Out-Of-Vocabulary, OOV) sorununa bir çözüm olarak beklenebilir.

Word2vec algoritması kelime tabanlı olduğu için karakter tabanlı modellere göre sözlük dışı kelimeleri ve morfolojik çeşitliliği ortaya çıkaramaz. Word2vec modelinde morfolojik olarak yakın olan kelimeler ayrı bir kelime olarak değerlendirilir. Örneğin “gönül”, “gönüllü”, “gönüllülük” kelimelerinin hepsi ayrı bir kelime olarak değerlendirilir fakat bu kelimelerdeki biçimsel benzerlik beraberinde anlamsal benzerlikte taşır. FastText karakter n-gramlar sayesinde biçimsel yakınlığı da göz önünde bulundurur. Örnekte ki kelimelerin “önl”, “önü”, “nül” gibi ortak n-gramları olacak ve bu durum bu kelimeleri biçimsel olarak ve anlamsal olarak birbirine yaklaştıracaktır. FastText bir dildeki morfolojik benzerlikten kaynaklanan anlamsal benzerliği de yakalar.

FastText’te merkezi kelimenin n-gram kombinasyonları, alt-kelime (subword) formları ve kendi formu bulunur, bağlam kelimelerinde ise sadece kelimenin kendi formu dikkate alınır.

“Günümüzün teknolojileri ile yazılım öğrenmek çok kolaydır” cümlesini ele alırsak, “yazılım” kelimesi merkezi kelime olsun n-gram kombinasyonlarında n=3 olsun. Bu durumda,

<”ya”,”yaz”,”az”,”azı”,”zı”,”zıl”,”ıl”,”ılı”,”lı”,”lım”> ,<”yazılım”> bölütleri modelde merkezi kelime ve n-gramları olarak düşünülür. Burada kelimenin sözcük formu ve karakter n-gram formu ayrı değerlendirilir. Örnekteki Türkçe karşılığı olan karakter n-gramlar ayrı bir kelime olarak değerlendirmez. ”Yaz”,”azı” gibi kelimeler sözcük olarak modelde yer aldığında <”yaz”>, <”azı”> olarak örnekteki karakter formundan ayrı ele alınır. Bağlam kelimeleri ise <”ile”>, <”öğrenmek”> gibi sözcük formuyla değerlendirilir.

G n-gramlardan oluşan bir sözlük olsun. Bir *w* sözcüğü, $G_w \subset \{1, \ldots, G\}$ *w*'de görünen n-gram kombinasyonlarının kümesidir. Her bir n-gram g ile vektör karşılığı Z_g, bağlam kelimeleri *c* ile ve sözlük *V* ile gösterilirse skor fonksiyonu Eş. 3.8’deki gibi olur. Böylece bir kelime vektörü n-gramlarının vektör değerlerinin toplamıyla temsil edilir.

$$s(w, c) = \sum_{g \in G} Z_g^T V_c \tag{3.8}$$

3.3.4. GloVe[4]

GloVe-GlobalVectors (Pennington, Socher ve Manning, 2014), en yaygın kullanılan kelime yerleştirme algoritmaları arasındadır. SVD, LSA gibi sayma ve boyut azaltma tabanlı yöntemler istatistiksel bilgileri ortaya çıkarır ancak analojiler gibi tahmine dayalı görevlerde yetersiz kalır. Buna karşılık, word2vec gibi sinir ağı ve tahmine dayalı modeller genellikle analojiler gibi tahmine dayalı görevlerde iyidir, ancak küresel istatistiksel bilgileri yeterince ortaya çıkaramaz. GloVe, hem yerel hem de küresel bilgileri birleştirerek, bu kelime yerleştirme modellerinin en iyi yönlerini birleştirmeyi hedefler.

Küresel yön, bir derlemdeki kelimelerin genel birlikte oluşum istatistiklerini ifade eder. GloVe, birlikte görünen kelime çiftlerinin birliktelik matrisini oluşturarak derlemin küresel bilgisini kullanır. Bu bilgi, dildeki sözcükler arasındaki genel anlamsal ilişkileri temsil eder.

Yerel yön ise, kelimelerin derlemdeki bağlamlarına karşılık gelir. GloVe, her bir kelimenin başka bir kelimenin bağlamında görünme olasılığına dayalı olarak birlikte oluşum sayılarını ağırlıklandırarak kelimelerin yerel bağlamını dikkate alır. Bu yerel bilgi, kelimelerin çağrışımları gibi daha incelikli anlamlarını ortaya çıkarır.

GloVe, kelimelerin bir arada bulunma istatistiklerinin küresel ve yerel yönlerini birleştirerek, hem kelimeler arasındaki geniş anlamsal ilişkileri hem de bağlamsal ince detaylarını ortaya çıkaran kelime vektörlerini oluşturur.

GloVe hem yerel bir bağlam penceresi kullanarak kelime birliktelik matrisini oluşturur, hem de koşullu olasılığı kullanarak küresel bağlamı elde eder.

Pennington, Socher ve Manning (2014) bir kelimenin belli bir karakteristiğinin ya da anlamının kelimelerinin birliktelik olasılıklarından nasıl çıkarılabileceğini bir örnekle açıklamışlardır. Buz (*ice*) ve buhar (*steam*) hedef kelimelerinin 6 milyar bölütten oluşan derlemden seçilen bağlam kelimeleri katı (*solid*), gaz (*gas*), su (*water*), biçim (*fashion*) ile ilişkisi değerlendirilmiştir.

[4] https://github.com/stanfordnlp/GloVe

Çizelge 3.2. GloVe algoritmasında kelimelerin birlikte geçme ihtimallerini gösteren değerler

Probability and Ratio	k = solid	k = gas	k = water	k = fashion
P(k | ice)	$1{,}9 \times 10^{-4}$	$6{,}6 \times 10^{-5}$	$3{,}0 \times 10^{-3}$	$1{,}7 \times 10^{-5}$
P(k | steam)	$2{,}2 \times 10^{-5}$	$7{,}8 \times 10^{-4}$	$2{,}2 \times 10^{-3}$	$1{,}8 \times 10^{-5}$
P(k | ice) / P(k | steam)	8,9	$8{,}5 \times 10^{-2}$	1,36	0,96

Buz ve *buhar* arasındaki ilişki bağlam kelimelerden her biri ile birlikte geçme olasılıklarının oranları ile bulunmuştur. Örneğin maddenin halini belirten katı kelimesi buz ile buhara göre çok daha yakından ilişkilidir ve Çizelge 3.2'in üçüncü satırındaki oran çok büyük çıkmıştır. Maddenin halini belirten gaz kelimesi ise buhar ile buza göre yakından ilişkilidir ve oran çok küçük çıkmıştır. Bağlam kelimeleri incelenen kelimelere anlamsal olarak eşit yakınlıkta veya ilgisiz ise oran bire yakın olacaktır. Su kelimesi buz ve buhar kelimesi ile anlamsal olarak yaklaşık aynı anlamda olduğu için oran bire yakındır, yine biçim kelimesi buz ve buhar ile ilgisiz olduğu için bu oran bire yakındır. Böylece kelimenin anlamı direkt kelime birliktelik olasılığı ile değil olasılıkların oranı ile temsil edilebilir. Kelime birliktelik matrisi X olsun, X_{ij} j indisindeki kelimenin i indisindeki kelimenin bağlamında (pencere) bulunma sayısı, $X_i = \sum_{k=1}^{V} X_{ik}$ ise derlem boyunca herhangi bir kelimenin i indisindeki kelimenin bağlamında bulunma sayısını gösterir. Bu durumda $P_{ij}=P(j|i)=X_{ij} / X_i$, j indisindeki kelimenin i indisindeki kelimenin bağlamı olma olasılığını verir.

Birlikte geçme ihtimallerinin oranını yukardaki örnekteki gibi buz, buhar kelime vektörlerini ve k bağlam kelime vektörünü (katı, gaz, su, biçim) girdi olarak olan Eş. 3.9'da bir F fonksiyonu ile tanımlarsak

$$F\left(w_i, w_j, \widetilde{w_k}\right) = \frac{P_{ik}}{P_{jk}} \tag{3.9}$$

$\frac{P_{ik}}{P_{jk}}$, w_i kelimesinin $\widetilde{w_k}$ bağlamında geçme olasılığının w_j kelimesinin $\widetilde{w_k}$ bağlamında geçme olasılığına oranıdır. $\frac{P_{ik}}{P_{jk}}$ oranı ile temsil edilen bilgi iki kelime vektörü ve bu kelimenin bağlam vektörünü girdi olarak alan F fonksiyonuna bağlıdır. Model F fonksiyonunun parametrelerini öğrenmek için eğitilerek olasılık oranıyla temsil edilen bilgiyi kodlar. Bu

oran ile temsil edilen bilgiyi kelime vektör uzayında göstermek için vektörlerin doğrusal yapısından dolayı çıkarma işlemi yapılabilir.

$$F\left(w_i - w_j, \widetilde{w_k}\right) = \frac{P_{ik}}{P_{jk}} \tag{3.10}$$

Eş.3.10'da eşitliğin sağ tarafı skaler büyüklükte iken sol tarafı ise vektör formundadır. Eş 3.11'de sol tarafı skalere çevirmek için vektörlerin iç çarpım özelliğinden yararlanılır.

$$F\left((w_i - w_j)^T \widetilde{w_k}\right) = \frac{P_{ik}}{P_{jk}} \tag{3.11}$$

Birlikte geçme matrisi düşünüldüğünde bir kelime ve bir bağlam kelimesi arasındaki ayrım isteğe bağlıdır ve bu iki rolü denklemde değiştirebiliriz. Tutarlılık için değiştirme işlemi sadece kelime vektörleri $w \leftrightarrow \widetilde{w_k}$ arasında değil kelimelerin derlemde geçme sıklığı $X \leftrightarrow X^T$ arasında da yapılmalıdır. Kelime sıklık matrisleri simetriktir ve F Fonksiyonun simetrik olması için benzer biçimlilik gereklidir. Eş. 3.12'de fonksiyonların homomorfizm, F(A-B)C→F(AC-BC) →F(AC)/F(AB), özelliği kullanılarak Eş. 3.13 elde edilir.

$$(F\left(\left(w_i - w_j\right)^T \widetilde{w_k}\right) = F\left(w_i^T \widetilde{w_k} - w_j^T \widetilde{w_k}\right) = \frac{F(w_i^T \widetilde{w_k})}{F\left(w_j^T \widetilde{w_k}\right)}) = \frac{P_{ik}}{P_{jk}} \tag{3.12}$$

$$F(w_i^T \widetilde{w_k}) = P_{ik} = \frac{X_{ik}}{X_i} \tag{3.13}$$

Eş 3.13'de X_i, i indeksindeki kelimenin derlemde toplam geçme sayısını, X_{ik} ise i indeksindeki kelimenin k bağlamında birlikte geçme sayısını gösterir. F fonksiyonunu üstel e fonksiyonu olarak kabul edilip eşitliğin her iki tarafının logaritması alınırsa eşitlik 3.14 elde edilir.

$$w_i^T \widetilde{w_k} = logP_{ik} = logX_{ik} - logX_i \tag{3.14}$$

X_i bağlama bağlı değildir bu yüzden yerine i kelimesi için bayes b_i terim eklenebilir, b_i terimi eklenince denklemin simetriyi koruması için bağlam kelimesi için bayes b_k terimi de eklenir.

$$w_i^T \widetilde{w_k} + b_i + \widetilde{b_k} = \log X_{ik} \quad (3.15)$$

Eş. 3.15'de en küçük kareler yöntemini kullanarak eğitim esnasında optimize edilecek hata fonksiyonu *J* ortaya çıkar.

$$J = \sum_{i,j=1}^{V} f(X_{i,j})(w_i^T \widetilde{w_j} + b_i + \widetilde{b_j} - \log X_{ij})^2 \quad (3.16)$$

Hata fonksiyonu her adımda (iterasyon) iki kelime vektörünün iç çarpımı ile birlikte geçme sayısını tahmin eder, daha sonra derlemdeki gerçekleşen birliktelik değeri X_{ij} ile olan farkı minimize etmek, 0'a yaklaştırmak için her iki kelimenin vektör değerlerini değiştirerek eğitim aşamasını gerçekleştirir. Hedef kelime *i* ve bağlam kelimesi *j* kelimesinin değerleri kelime-bağlam ve bağlam-kelime olmak üzere iki ayrı matrisde tutulur. Eğitim aşaması sonucunda yüksek boyutlu X matrisi düşük boyutlu iki adet matrise dönüştürüldüğü için bu aşama matris faktorizasyonu olarak adlandırılabilir. Bir kelimenin vektör değerleri genellikle kelime-bağlam ve bağlam-kelime matrislerindeki değerlerin toplamı ile elde edilir. Burada dikkat edilmesi gereken nokta yerel bir bağlama göre elde edilen iç çarpım değerlerinin küresel birliktelik matris değerlerine göre optimize edilmesidir. Böylece hem yerel bağlamın hem küresel istatistiklerin bilgisinin kelime vektörlerinde kodlanması sağlanır.

Eş. 3.16'da $f(X_{i,j})$ ağırlık fonksiyonudur, V ise sözlüğün boyutudur. Ağırlık fonksiyonu, sık ve seyrek kelime birlikteliklerinin önemini dengelemek için kullanılır. Sık rastlanan kelime birlikteliklere daha düşük ağırlıklar verirken, seyrek kelime birlikteliklerine daha yüksek ağırlıklar atar önemini artırır. Böylece ağırlık fonksiyonu daha seyrek ama daha anlamlı sözcükler arasındaki anlamsal ilişkileri ortaya çıkarmaya yardımcı olur. Ayrıca "ve", "veya", "ile" gibi derlemdeki kelimelerin büyük çoğunluğu ile anlamsal ilişkisi olmayan durak kelimelerini bir yerel bağlam olarak önemini azaltır. Fonksiyon seyrek kelimelere aşırı ağırlıklandırmamak için azalmayan olmalı, çok sık gerçekleşen kelimeleri aşırı ağırlıklandırmamak için görece küçük değerler üretmeli ve hata fonksiyonundaki $logX_{ij}$ ifadesinin tanımsız *log0* üretmesi yerine *f(0)=0* sonucunu üretmelidir. Bu durumda hata fonksiyonu Eş. 3.17'deki gibi davranır.

$$f(x) = \begin{cases} (x/x_{max})^{\propto}, & x < x_{max} \\ 1, & x \geq x_{max} \end{cases} \quad (3.17)$$

TBMM Genel Kurul tutanaklarında anlamsal yakınlık elde etmek için GloVe modeli, python dilinde Tensorflow[5] uygulama çatısı ile gerçekleştirilmiştir.

Uygulamada, TBMM birleşimlerinden oluşan ve DDİ adımları uygulanmış metin dosyası kullanılarak GloVe modeli ile elde çıkan kelime vektörlerinin boyutu 300'dür. Derlemde 5'den az sayıda geçen kelimeler göz ardı edilmiştir.

3.3.5. Kelime benzerliği

Kosinüs benzerliği, çok boyutlu bir uzayda iki vektör arasındaki benzerliğin bir ölçüsüdür. Kelime vektörleri düşünüldüğünde, her kelime, boyutların kelimenin farklı özelliklerine veya niteliklerine karşılık geldiği (bir derlemdeki sıklığı, bağlamı veya semantik anlamı gibi) yüksek boyutlu bir vektör olarak temsil edilir. Çalışmada kullanılan modeller 300 boyutlu kelime vektörü elde edilir. İki kelimenin benzerliği kelime vektörlerinin kosinüs benzerliği ile belirlenir. Kosinüs benzerliği Eş.3.18'de gösterildiği gibi vektörlerin iç çarpımlarının birim uzunluklarının çarpımına oranıdır.

$$Benzerlik(A,B) = \cos(\theta) = \frac{A.B}{||A||.||B||} \quad (3.18)$$

3.3.6. Kelime analojileri

Doğal dil işlemede, kelime analojisi, belirli bir kelime grubunda birbiriyle benzer ilişkili kelimeleri bulmayı ifade eder. Bir kelime çifti arasında ilişkinin başka bir kelime çifti arasında olup olmadığının incelendiği ve amacın bir kelime çifti, üçüncü bir kelime ve bu üçüncü kelime ile verilen kelime çiftiyle benzer bir ilişkiyi tamamlayan dördüncü bir kelimeyi bulmak olduğu bir dil görevi türüdür.

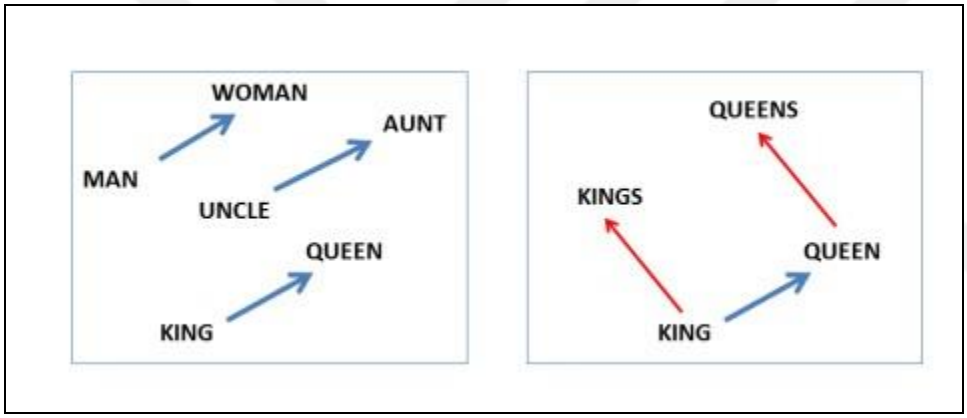
[5] https://www.tensorflow.org/

Kelime analojisi genellikle kelime vektörleri kullanılarak bulunur. Kelime vektörleri düşünüldüğünde, iki kelime arasındaki vektör farkı, bu kelimeler arasındaki ilişkinin bir temsili olarak düşünülebilir. Kelime analojisinde vektör farkları kullanılır ve temel mantık Eş. 3.19 ve Eş. 3.20'de gösterilmiştir.

$$w_A - w_B \approx w_C - w_X \tag{3.19}$$

$$w_X \approx w_C + w_B - w_A \tag{3.20}$$

"A kelimesinin B kelimesi ile olan ilişkisi C kelimesi ile X kelimesi arasında da geçerlidir." Bu örnekte A, B, C kelimesi yardımı ile X kelimesi bulunmaya çalışılır.



Şekil 3.12. Kelime analoji örneklerinin vektör uzayındaki temsilleri

Şekil 3.12, Mikolov, Yih ve Zweig (2013) çalışmasındaki Şekil 2'de gösterdikleri analoji örneklerinin vektör uzayındaki temsilleri gösterir. Kelime çiftlerinin vektör uzayında dağılımından anlamlı sonuçlar bulunabilir. Kelime vektörleri ile basit cebirsel işlem yaparak sadece kelimeler arasındaki benzerliği değil analoji adı verilen karmaşık semantik ilişkiler bulunabilir. Şekilde cinsiyet ilişkisini gösteren kelime çiftlerinde iki kelime vektörü arasındaki uzaklık eşittir. Sağda ise kelimelerin tekil çoğul ilişkisi vektörlerin arasındaki mesafeden bulunabilir. *Man (kadın)*, *woman (erkek)* ve *uncle (amca)* kelime vektörlerini vektör uzayındaki konumlarından *aunt (hala)* kelimesinin *uncle* kelimesi ile cinsiyet ilişkisi bulunabilir.

Mikolov, Yih ve Zweig (2013) çalışmalarında 2 kelime çiftinden oluşan kayıtlardan bir veri kümesi oluşturmuşlardır ve sonuçları bu veri kümesi ile değerlendirmişlerdir. *King (kral)*,

man (erkek), woman(kadın), quenn (kraliçe) kelime gruplarını kullanan (Mikolov, Yih ve Zweig, 2013), $w_{king} - w_{man} + w_{woman}$ cebirsel işleminin W_{quenn}'e en yakın sonucu verdiğini göstermiştir. Böylece kelime vektörlerinin vektör uzayında bulunduğu konumlar arasındaki mesafeye göre analojiler bulunabilmektedir.

Kelime analojileri TBMM genel kurul tutanaklarında siyasi parti ve genel başkanı olarak uyarlandığında:

Deniz Baykal Cumhuriyet Halk Partisinin (CHP) genel başkanı, Devlet Bahçeli Milliyetçi Hareket Partisinin (MHP) genel başkanıdır. "Deniz Baykal ile CHP arasındaki ilişki, Devlet Bahçeli ile X kavramı arasında vardır." Burada analojiyi ortaya çıkarmak için ilişkiyi tesis eden genel başkanlık kavramını bilmeye gerek yoktur ve bu ilişki gizlidir (latent). Kelime analojisi görevinden beklenen X kavramının karşılığı olarak MHP kavramını bulmasıdır.

3.3.7. Kelime vektörlerinde boyut indirgeme ve görselleştirme

Yüksek boyutlu kelimelerin 2 boyutlu veya 3 boyutlu vektör uzayında gösterilmesi için taşıdığı anlamı kaybetmeden boyutunun indirgenmesi gerekir. Çalışmada 300 boyutlu kelime vektörlerinin 2 boyuta indirgenerek görselleştirilmesi için t-SNE (t-Distributed Stochastic Neighbor Embedding) (Van der Maaten ve Hinton, 2008) algoritması kullanılmıştır. T-SNE algoritmasının çalışma prensibi şu şekildedir. İlk olarak, t-SNE yüksek boyutlu noktalara göre bir olasılık dağılımı oluşturur. Bu dağılımda benzer noktalar daha yüksek bir olasılığa sahiptir. Benzer bir olasılık dağılımını düşük boyutlu harita üzerinde de tanımlar ve her iki haritadaki noktaların konumlarına göre iki dağılım arasındaki Kullback-Leibler sapmasını (KL divergence) (Kullback ve Leibler, 1951) en aza indirecek şekilde optimizasyon gerçekleştirir.

Temel Bileşen Analizi (TBA) (Principal Component Analysis, PCA) (Jolliffe, 2005) kullanılmıştır. TBA vektör boyut azaltırken bilgiyi korumayı vaat eder ve yüksek boyutlu verilerin görselleştirilmesinde sıklıkla kullanılır. TBA daha karmaşık algoritmalar uygulamadan önce verilerin boyutluluğunu azaltmak için bir ön işleme adımı olarak kullanılabilir. Çalışmada kelimeler kelimeleri 2 boyutlu harita üzerinde görselleştirmek için t-SNE kullanılırken önişlem aşamasında TBA kullanılarak hem hesaplama karmaşıklığından kaçınılmış hem de boyutlar kelimelerin en önemli özellikleri seçilerek azaltılmıştır.

4. BULGULAR

TBMM Genel Kurul tutanaklarının yapay zeka tabanlı metin analizinin yapıldığı çalışmada bulgular üç ana kapsamda ele alınmıştır. Birinci bölümde siyasi partilerin 12 aylık hareketli kutuplaşması ile ilgili bulgular, ikinci bölümde milletvekillerinin demografik özellikleri ile ilgi bölümler ve üçüncü bölümde ise milletvekili konuşma metinlerinde kelime benzerliği ile ilgili bulgular araştırılmıştır.

4.1. TBMM Genel Kurul Görüşmelerinde Siyasi Parti Kutuplaşması Bulguları

TBMM Genel kurul tutanakları derleminin kapsadığı 2011 ve 2023 yılları arasında 4 siyasi parti 208-2023 yılları arasında ise 5 siyasi parti grubu vardır. Çizelge 4.1 siyasi parti adlarını, kısa adlarını ve derlemde bulundukları yasama dönemlerini ve tarihleri göstermektedir

Çizelge 4.1. TBMM'de siyasi partiler

Parti Kısa Adı	Parti Adı	Derlemde Temsil Edildiği Tarihler
AK Parti	Adalet ve Kalkınma Partisi	24,25,26,27 (2011-2023)
CHP	Cumhuriyet Halk Partisi	24,25,26,27 (2011-2023)
MHP	Milliyetçi Hareket Partisi	24,25,26,27 (2011-2023)
HDP	Halkların Demokrat Partisi	24,25,26,27 (2011-2023)
İYİ Parti	İYİ Parti	27 (Ekim 2018-2023)

İYİ Parti 25 Ekim 2017 tarihinde ağırlığını Türkiye'de milliyetçi ideolojini temsilcisi MHP'li ve ülkücü kadroların oluşturduğu kadrolarla kurulmuştur. Parti kendisini milliyetçi, demokrat ve kalkınmacı bir parti olarak tanımlar (Akşener, 2021). Saygı ve Aslan (2020) parti tüzüklerinden milliyetçilik ve din anlayışlarını karşılaştırdıkları çalışmalarında İYİ Parti'yi Türkiye'nin milliyetçi ideolojiyi benimseyen partileri arasında göstermişlerdir. MHP'nin ideolojik bağlamının Türk-İslam sentezi olduğunu, İyi Parti'nin ise milliyetçi bir ideolojiyi, tüzüğünde barındırmasına rağmen, dine dair bir görüşe rastlanmadığını belirtmişlerdir. Bu çalışmaya göre İYİ Parti'nin sağda ve MHP'ye göre daha seküler milliyetçi bir parti olduğu söylenebilir.

HDP merkezine Doğu ve Güneydoğu bölgelerindeki terör sorununu alan bir partidir. Parti Türkiye'de ideolojik olarak sol ideolojiye eğilimli olarak kabul edilir. Ayrılıkçı PKK terör

örgütü ile bağının olduğu ve örgütün siyasi kanadı olduğuna dair suçlamalar getirilen HDP bu yönüyle milliyetçi ideoloji ile en çok karşıtlığa sahip siyasi partidir. 7 Haziran 2021 tarihinde PKK ile bağı olduğu gerekçesi ile kapatma davası açılmıştır.

İlk genel başkanı Türkiye Cumhuriyeti'nin kurucusu Mustafa Kemal Atatürk olan CHP kendini Atatürkçü, laik ve solda tanımlar. 2018 yılında kabul edilen tüzüğünde kendisini "Cumhuriyet Halk Partisi, başta Kurtuluş Savaşımız olmak üzere Aydınlanma ideallerini, emek mücadelelerini, sosyal demokrasinin özgürlük, eşitlik ve dayanışma ilkelerini benimseyen çağdaş demokratik sol bir siyasal partidir." ifadesiyle tanımlamıştır (*CHP Tüzüğü*, 2018).

MHP'nin ideolojik çerçevesini Türk milliyetçiliği ve Türk-İslam sentezi oluşturur (Saygı ve Aslan, 2020). MHP Türkiye'nin üniter yapısına, toprak bütünlüğüne ve Türk kültürüne vurgu yapmaktadır. Türkiye'nin üniter yapısını bozmakla ve bölücülükle suçladığı HDP ile yüksek karşıtlığı ve kutuplaşması beklenebilir. MHP, Türkiye'de milliyetçiliği en çok temsil eden parti olarak kabul edilmektedir.

AK Parti 14 Ağustos 2001 tarihinde Recep Tayyip Erdoğan liderliğinde kurulmuştur. Kasım 2002 yılından beri tek başına iktidardadır. Çalışmada kullanılan derlemin kapsadığı tarihler (2011-2023) boyunca iktidarda AK parti hükümetleri bulunmaktadır. AK Parti ideolojik olarak sağda görülür. Parti 2023 vizyon belgesinin Muhafazakâr Demokrat kimlik bölümünde "Adalet ve Kalkınma Partisi kendisini siyasetin merkezinde konumlandıran muhafazakâr-demokrat bir kitle partisidir" (*Ak Parti 4. olağan büyük kongresi siyasi vizyonu*, 2023) ibaresi ile kimliğini muhafazakâr ve demokrat parti olarak tanımlamıştır.

Veri kümelerinin kapsadığı 2011-2023 yılları arasında Türkiye'deki siyasi parti ilişkilerini belirleyen en önemli siyasi değişim yeni hükümet sistemi ve seçim öncesi ittifaklardır.

İktidardaki AK Parti, anayasa uzlaşma komisyonuna 2012 yılında başkanlık sistemi önerisinde bulunmuştur. AK Parti'nin bu önerisi üzerinde uzlaşılamamış ve anayasayı değiştirmek için yeterli milletvekili sayısına sahip olmadığı için öneri hayata geçememiştir. MHP lideri Devlet Bahçeli Ekim-2016 ve Kasım-2016'daki partisinin TBMM'deki grup konuşmalarında mevcut sistemin fiili olarak başkanlık hukuki olarak parlamenter sistem olduğu, bu tutarsızlığın fiili tıkanıklık oluşturduğunu belirtmiştir. Bu tıkanıklığı aşmak için

ya fiili olarak parlamenter sisteme dönülmesini ya da başkanlık sisteminin hukuksal zeminin oluşturulmasını talep etmiştir. Bu konuşmaların ardından hazırlanan AK Partinin anayasa değişiklik teklifine destek vereceklerini MHP liderinin 22 Kasım 2016 tarihli meclis grup konuşmasındaki açıklamasıyla (Bahçeli, 2016) Türkiye'de yönetim değişikliğinin önü açılmıştır. Bu destek ile birlikte AK Parti ve MHP'nin birlikte anayasa değişikliği için referanduma gitmek için yeterli milletvekili sayısına sahip olmuştur. Bu nedenle çalışmada başkanlık sisteminin baskın bir şekilde ülke gündemine gelme tarihini Ekim-2016 olarak kabul edilmiştir.

Anayasa Değişikliği referandumu 16 Nisan 2017 tarihinde yapılmış ve AK Parti ve MHP'nin iş birliği ve evet oyları ile kabul edilmiştir. Anayasa değişikliğine hayır kampanyasını CHP ve HDP örgütlemiştir. Böylece Türkiye'nin hükümet sistemi kökten değişmiştir. Yeni hükümet sistemi Cumhurbaşkanlığı Hükümet Sistemi (CHS) olarak adlandırılmış ve Parlamenter Hükümet Sistemi CHS'ye dönüştürülmüştür. Bu referandum ile birlikte MHP'nin iktidar-muhalefet blokundaki konumu değişmiş, seçim ittifaklarında AK Parti ile hareket ederek muhalefet bloku yerine iktidar blokunda yer almıştır. CHP-HDP ise muhalefet blokunda kalmaya devam etmişlerdir. Bu durumun en çok CHS öncesi dönemde muhalefet blokunun merkezini oluşturan CHP ve MHP kutuplaşmasına yansıması beklenir.

Siyasi partilerin kutuplaşmasını etkileyen diğer olay ise seçim ittifaklarıdır. 13.3.2018 tarih ve 7102 sayılı Kanunla siyasi partilerin ittifak halinde seçime girmesi düzenlenmiştir. Bu durumun doğal sonucu ittifak içindeki partilerin birbirine yaklaşmasıdır. AK Parti ve MHP, 24 Haziran 2018 tarihinde yapılan milletvekili ve cumhurbaşkanlığı seçimlerine aynı ittifak (Cumhur İttifakı) içinde girmişlerdir. Cumhur İttifakı'nın karşısında ise CHP ve İYİ Partinin ana bileşenlerini oluşturduğu Millet İttifakı kurulmuştur. Bu iki karşıt ittifak 24 Haziran 2018 ve 14 Mayıs 2023 Cumhurbaşkanlığı ve milletvekili seçimlerinde resmi olarak 31 Mart 2019 yerel seçimlerinde ise metropollerde iş birliği şeklinde devam etmiştir.

Türkiye'deki siyasi partiler ideolojik durumlarının ve siyasal sistemdeki başat değişikliklerin kutuplaşmaya etkisini incelemek için siyasi parti kutuplaşmalarını kullanarak zaman serisi analizi yapılmıştır.

Kutuplaşmayı ölçerken sosyal medya verilerinde olduğu gibi parlamento görüşmelerinden aylık kesikli veri kümeleri oluşturmak mümkün değildir. Ancak 12 aylık hareketli veri

kümelerinin kullanılması, aylık kutuplaşma verileri sağlayarak zaman serisi analizinin uygulanmasına olanak tanımaktadır. Böylece aylık kutuplaşma verisi ile partilerin ilişkileri derinlemesine incelenebilmiştir.

TBMM, 27. Dönemden itibaren İYİ Parti'nin katılımı ile birlikte 5 partili olmuştur. Bu yüzden çalışmada 2011-2023 (24-27nci dönem 4 partili) arası ve 2018-2023 (27. Dönem 5 partili) arası iki ayrı analizle incelenmiştir. Metin sınıflandırmalarında hareketli veri setleri kullanıldığı için yeni dönemin tamamı ile veri setine yansıması bir yıllık süreyi gerektirir. TBMM Genel Kurul görüşmeleri ekim ayında başlar. Bu yüzden şekillerde 27. Dönemin başlangıcı bir yıl gecikme ile Ekim-2019'dan başlatılmıştır.

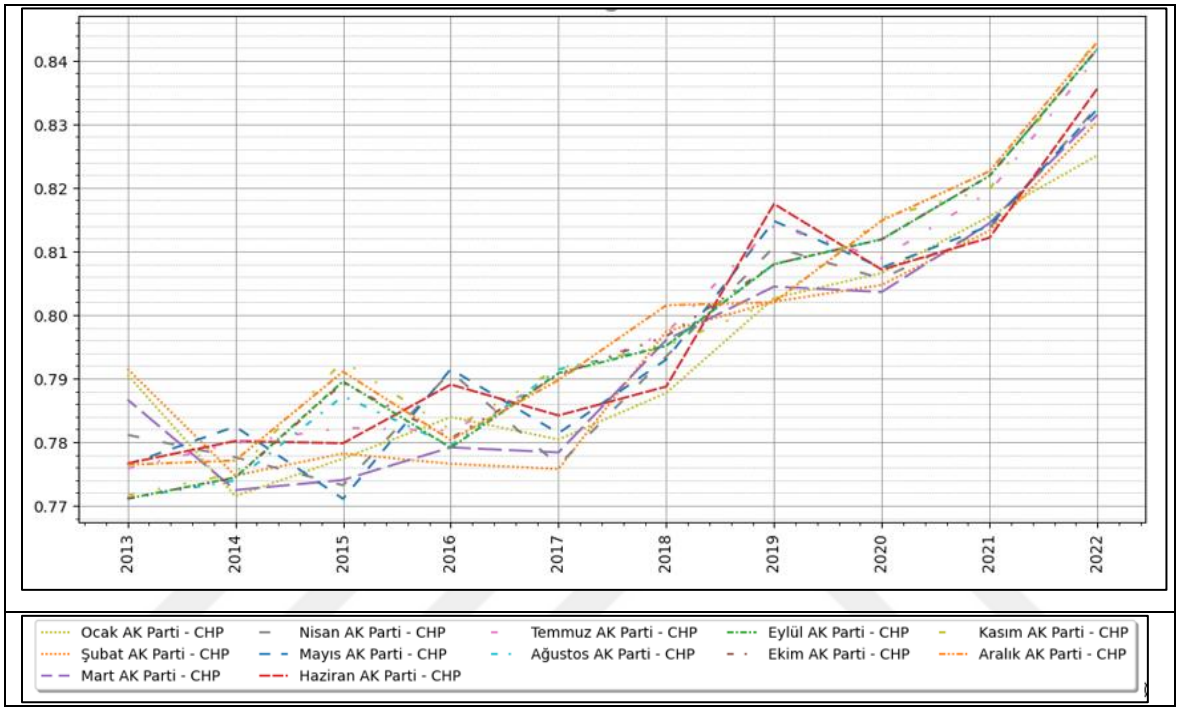
4.1.1. Bir yıllık ayrık ve 12-aylık hareketli kutuplaşma ölçütü karşılaştırması

Bir yıllık dönemi kapsayan kutuplaşma çalışmalar takvim yılını temel alarak ocak ayında başlar. Ancak, başlangıç ayının seçimi analiz sonuçlarını etkileyebilir. Oysa her aydan başlayarak 12 aylık dönem bir yıllık süreyi temsil eder. Örnek olarak Şekil 4.1'de AK Parti ve CHP arasında kutuplaşma görülmektedir. Kutuplaşma ölçütleri Şekil 3.4'deki kutuplaşma ölçütlerinden seçilmiştir. Bir yıllık kutuplaşmanın seviyesi dönemin başlangıç ayına bağlı olarak değişmektedir.

2014 ve 2015 yılında veri kümelerini Nisan ya da Mayıs ayından başlarsa kutuplaşma düşmüş, eylül ve aralık ayında yükselmiş haziran ayında sabit kalmıştır. 2019 ve 2020 yılında ise haziran ve mayıs ayında düşmüş aralık ve kasım ayında yükselmiştir. Bununla birlikte, bu ayların hepsi bir yıllık kutuplaşmayı temsil etmesi mümkündür. Her bir yılın bir nokta ile temsil edildiği ayrık kutuplaşma temsili yerine kutuplaşmanın hareketli temsilinden kazançlar elde edilebilir. Kutuplaşmanın aylık temelde ölçülmesi, bir yıllık dönem 1 veri noktası yerine 12 veri noktası ile temsil edileceği için kutuplaşma ölçütündeki bu değişkenliğe ve tutarsızlığa bir çözüm sağlar.

Kutuplaşmanın bir yıl yerine her ay için ölçülmesi, zaman içindeki eğilimler ve örüntüler hakkında daha ayrıntılı bilgiler sağlar. Sonuçları aylık olarak analiz edilerek, verilere daha uzun bir süre boyunca bakıldığında gözlemlenemeyen daha kısa vadeli eğilimleri ve değişimleri tespit etmek mümkün olur. Kutuplaşmanın yıllık yerine aylık olarak ölçülmesi,

zaman müdahalesi (time intervention), seriler arasındaki nedensellik (causality) ve tahmin (forecasting) gibi zaman serisi analizlerine olanak sağlar. Şekil 4.1'de 2013'ten 2022'ye kadar sekiz veri noktası vardır ve bu sayı zaman serisi analizleri için yeterli değildir(Hanke ve Wichern, 2013: 80). 12 aylık hareketli kutuplaşma ile bu sayı 84'e çıkar ve zaman serisi analizi mümkün olur.



Şekil 4.1. Başlangıç Aylarına göre AK Parti ve CHP arasındaki bir yıllık kutuplaşma

4.1.2. Parti çifti zaman serilerinin kutuplaşma mesafeleri

Parti çiftlerinin zaman serilerinin kutuplaşma değerleri karşılaştırılırken Dinamik Zaman Bükmesinden (DZB) (Dynamic Time Warping, DTW) elde edilen en küçük maliyet yolundan faydalanılmıştır. Elde edilen ölçüm değeri grafiklerde *dtw* ile gösterilmiştir.

Ses dalgaları, finans verisi gibi zaman serileri hiçbir zaman mükemmel şekilde senkronize olamazlar, aralarında zaman kaymaları vardır. DZB mükemmel bir şekilde senkronize olmayan iki zamansal diziyi karşılaştırmanın, aralarındaki benzerlikleri bulmanın bir yoludur. DZB örüntü tanıma, konuşma tanıma ve veri madenciliği gibi alanlarda yaygın olarak kullanılmaktadır.

Herhangi iki zaman serisi öklid mesafesi veya diğer benzer mesafeler kullanılarak zaman ekseninde noktadan noktaya karşılaştırılabilir. İlk zaman serisinin *t* zamanındaki değeri, ikinci zaman serisinin *t* zamanındaki değeri ile karşılaştırılacaktır. Bu seriler çok benzer olsalar bile serilerden birinin zaman ekseninde kayması, fazlarının farklı olması benzerliklerinin çok düşük çıkmasına sebep olur. Bu, iki zaman serisi şekil olarak çok benzer ancak zaman olarak faz dışı olsa bile çok zayıf bir karşılaştırma ve benzerlik skoru ile sonuçlanacaktır. DZB, *t* zamanındaki ilk sinyalin değerini ikinci serinin *t+1* ve *t-1* veya *t+2* ve *t-2* zamanlarındaki değerleriyle karşılaştırır. Böylece DTW, benzer şekillere ve farklı fazlara sahip sinyaller için düşük benzerlik sonucunun önüne geçer vermemesini sağlar.
Örnek olarak A, B serileri ele alınırsa:

A= [0, 0, 1, 2, 1, 0, 1, 0, 0, 2, 1, 0, 0,3,2]

B= [0, 1, 2, 3, 1, 0, 0, 0, 2, 1, 0, 0, 0,3,1]

Şekil 4.2'de örnek A ve B serilerinin zaman bükmesi olmadan ve dinamik zaman bükmesi ile eşleştirilmesi görülmektedir. DZB'de zaman kaymaları, faz değişikliği dikkate alınarak ilk serideki bir nokta eşleştirme yapılan serideki birden çok nokta ile eşleşip en düşük maliyetli eşleşmeler hesaplanarak seriler boyunca toplam eşleştirme maliyeti hesaplanır.

DZB algoritması çalışma prensibi şu şekildedir:
İlk önce her iki serinin bütün noktalarının birbiri ile uzaklıkları hesaplanarak bir maliyet matrisi oluşturulur. A ve B serileri için i ve j serilerin indisleri, (i,j) maliyet matrisindeki bir hücre olmak üzere, matrisinden bükme yolu bulurken (i_k, j_k)'ya giden yol şu şekilde hesaplanır.

$$D_{min}(i_k, j_k) = \min_{i_{k-1}, j_{k-1}} D_{min}(i_{k-1}, j_{k-1}) + d(i_k, j_k | i_{k-1}, j_{k-1}) \quad (4.1)$$

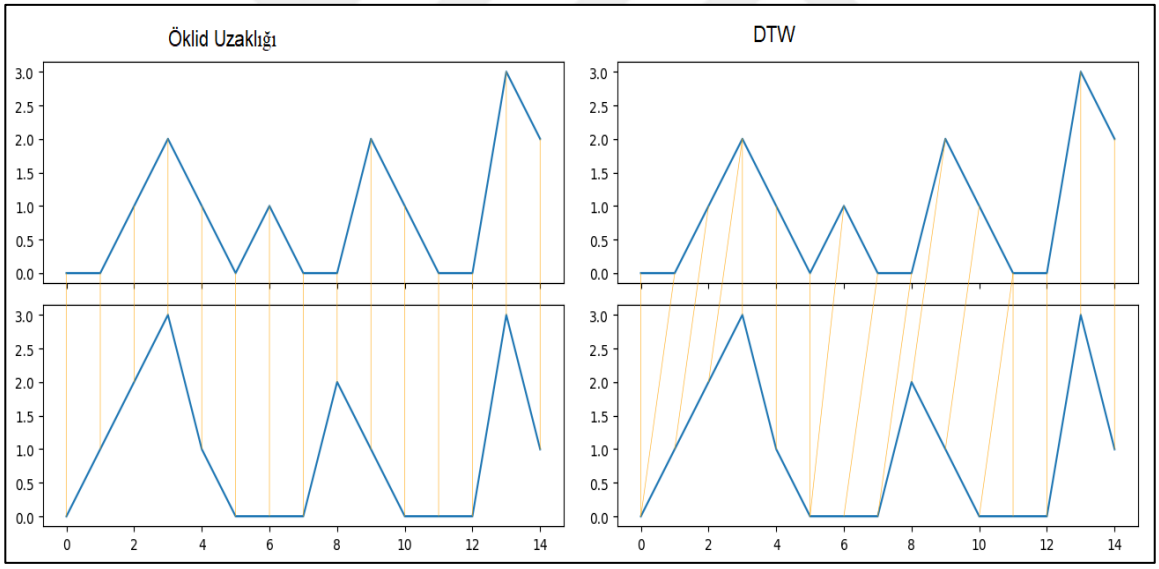
Eş. 4.1'de D öklit uzaklığını gösterir. Bütün yolun maliyeti ise $D = \sum_k d(i_k, j_k)$ ile hesaplanır.

Bükme yolu bulunurken tüm olası yollardan geçmek maliyetlidir zaman karmaşıklığına sebep olur. Olası bükme yollarının sayısını kısıtlamak için matris üzerinde dolaşırken yatay hareket [$(i, j) \rightarrow (i, j+1)$], dikey hareket [$(i, j) \rightarrow (i+1, j)$] ve diyagonal hareket [$(i, j) \rightarrow$

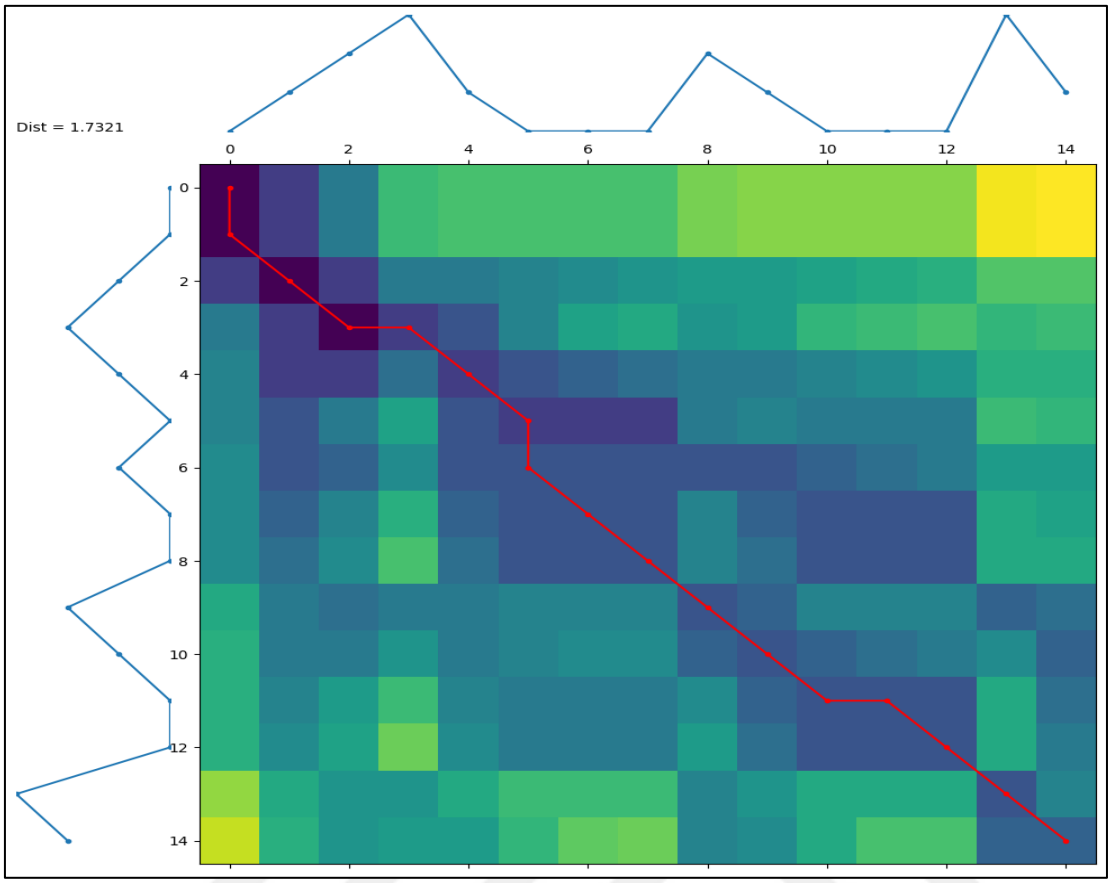
($i+1$, $j+1$)] gerçekleştirilir. Bu hareketler sayesinde maliyet matrisinin sol üst köşesinden başlanır ve sağ alt köşesine doğru her bir adımda en az maliyetli noktalar min(d(i-1,j-1), d(di, j-1), d(i-1,j)) ile seçilerek dolaşılır ve bükme yolu tamamlanır. Bükme yolunda seçilen noktaların maliyetler toplamlarının eleman sayısına bölünerek DZB uzaklığı hesaplanır.

Şekil 4.3'de örnek seriler için maliyet matrisi, bükme yolu ve hesaplanan uzaklık ölçüsü (D) gösterilmiştir. Şekildeki matrisin yatay ekseni örnek serilerdeki B serisini, dikey ekseni ise A serisinin indislerini gösterir. Bükme yolunun diyagonal olduğu yerlerde iki seride bükme olmamış, yatay ve dikey olduğu yerlerde bükme (warping) olmuştur.

Parti çiftlerinin birbirleri ile ilişkileri incelenirken benzerlik ölçüsü olarak DZB algoritması ile elde edilen mesafe kullanılmıştır. En küçük mesafe kutuplaşma arasındaki farkın en az olduğu parti çiftlerini gösterir.



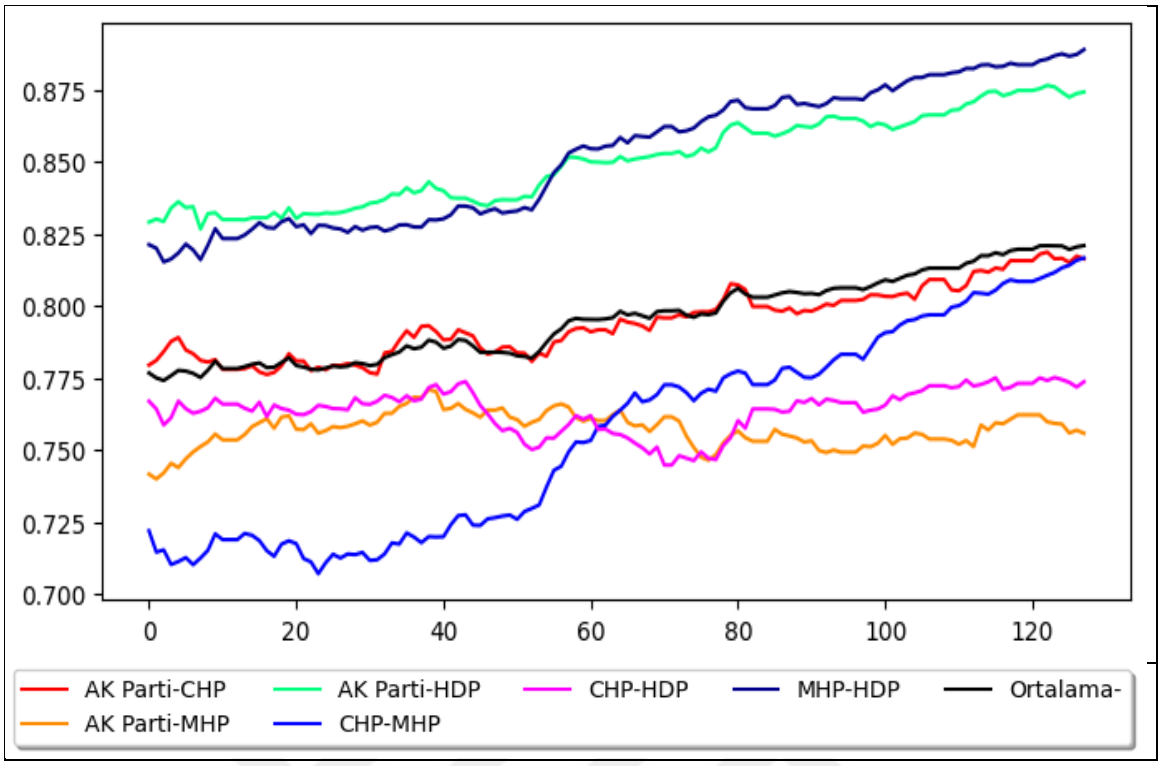
Şekil 4.2. Zaman Serilerinde Öklid uzaklığı ve DTW



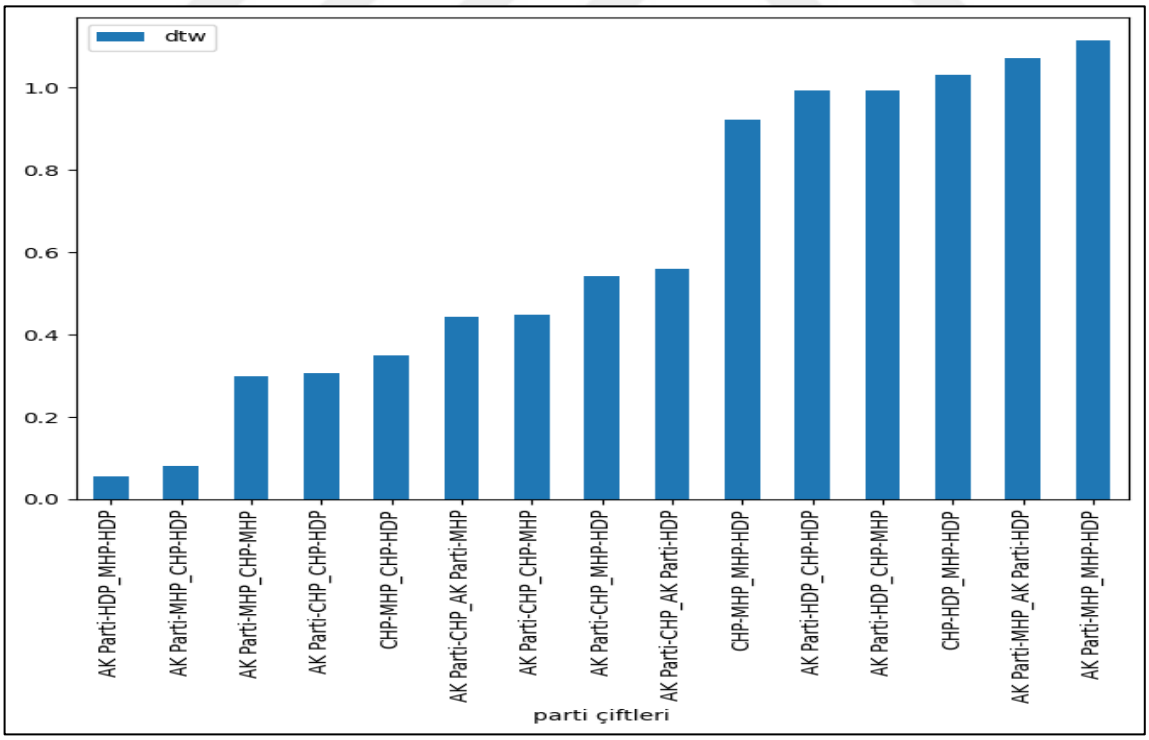
Şekil 4.3. Örnek seriler için maliyet matrisi, bükme yolu ve uzaklık

Kutuplaşma seviyelerinin incelendiği analizde uzaklık çubuk grafiklerde “dtw” etiketi ile gösterilmiştir. Görece kutuplaşma seviyelerini incelemek için bütün serilerin ilk elemanı serini bütün elemanlarından çıkarılarak serilerin 0 başlangıç noktasından başlaması sağlanmıştır. Şekil 4.4’de ve Şekil 4.7’de serilerin bu işlem öncesi ve sonrası oluşan değerleri görünmektedir. Çubuk grafiklerde görece kutuplaşma değerleri “dtw_0” ile gösterilmiştir.

Görece kutuplaşmada bütün seriler aynı noktadan (0) başladığı için ölçekleri de birbirine yaklaşmıştır. Bu yüzden görece kutuplaşma çizge grafiklerin şekil benzerliği için daha uygundur.



Şekil 4.4. 24-27. Dönemde (2011-2023) kutuplaşma seviyeleri



Şekil 4.5. 24-27. Dönemde (2011-2023) parti çiftleri arasındaki kutuplaşma seviyesi DTW mesafeleri

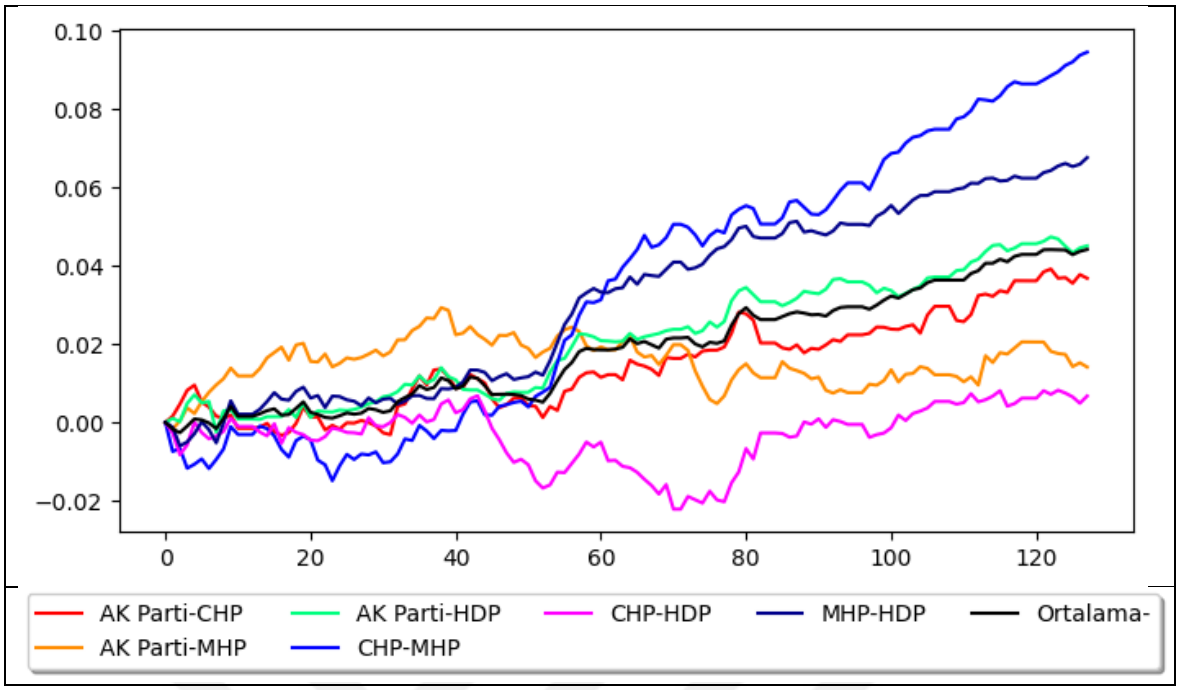
Şekil 4.4'de tüm derlem boyunca (2011-2023) TBMM'de yer alan 4 siyasi partinin oluşturduğu 6 parti çiftinin kutuplaşma seviyelerinin çizge grafiği görünmektedir. Bu çizge grafiklere göre DZB ile ölçülen mesafeler Şekil 4.5'de gösterilmiştir. Bu sonuçlara bakarak parti çiftlerinin kutuplaşma seviyelerinin derlem boyunca birbirlerine ne kadar yakın seyrettiği değerlendirilebilir.

En az kutuplaşma seviyeleri farkı (AK Parti-HDP) ve (MHP-HDP) parti çiftleri arasındadır (0,0558). AK Parti MHP'nin sağ ve milliyetçi ideolojiyi temsil ettiği düşünüldüğünde, bu durum HDP'nin sağ-milliyetçi ideoloji ile kutuplaşmasının bir sonucu olarak değerlendirilmiştir.

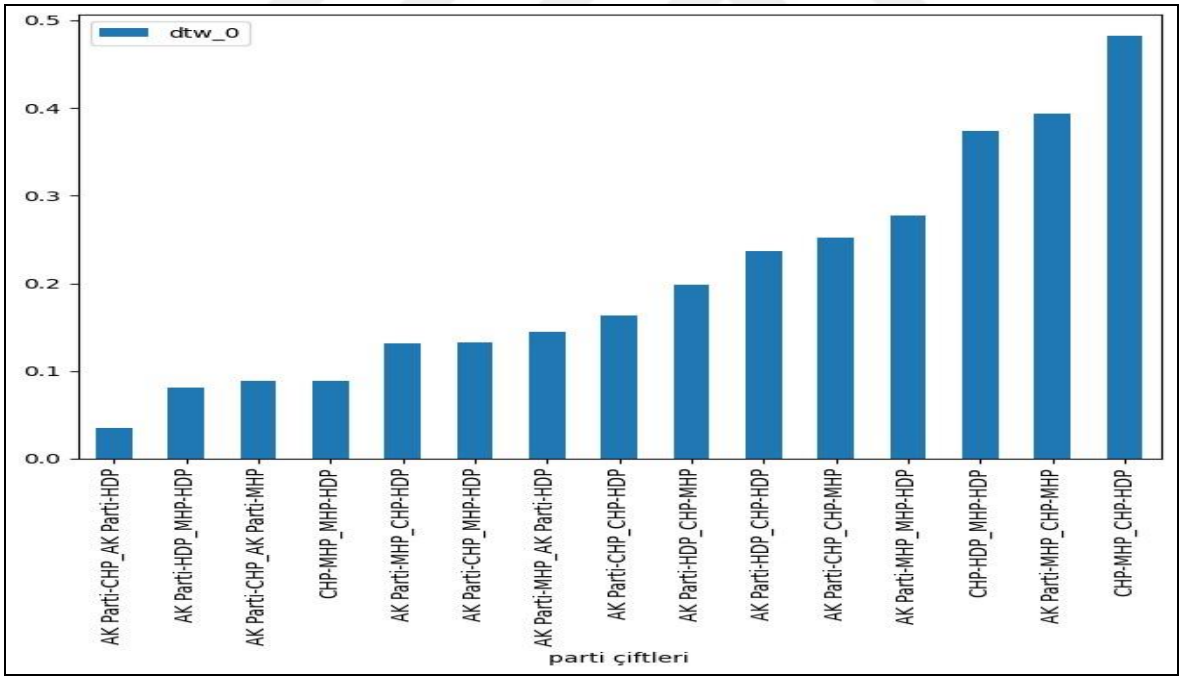
Sağ yönelimli (AK Parti-MHP) ve sol yönelimli (CHP-HDP) parti çiftlerini kutuplaşma seviyeleri özellikle Şubat-2014 ve Haziran-2019 arasında biri birine yakın seyretmiştir (Bkz. Şekil 3.7). Sağ ve sol ideolojilerin kutuplaşma seviyelerine yansımasının sonucu olarak Şekil 4.5'de bu parti çiftlerini kutuplaşma seviyeleri en küçük ikinci *dtw* uzaklığıdır (0,0817).

En yüksek kutuplaşma seviyesi farkı ise (AK Parti- MHP) ve (MHP-HDP) arasında (1,1144), en yüksek ikinci fark (1,0711) (AK Parti-MHP) ve (AK Parti-HDP) arasındadır. (AK Parti-HDP) ve (MHP-HDP) 2011-2023 yılları boyunca en yüksek kutuplaşma seviyeleri olarak seyretmiştir. Buna karşın parlamentoda bulunan 4 partiye (AK Parti, CHP, MHP, HDP) ait 6 parti çifti düşünüldüğünde 2012'de (AK Parti-MHP) kutuplaşma seviyesi dördüncü sırada iken Ocak-2017'den itibaren düşerek altıncı ve son sırada olmuştur. Böylece düşük kutuplaşma seviyeli (AK Parti-MHP) en yüksek seviyeli (AK Parti-HDP) ve (AK Parti-MHP) ile en yüksek farkı oluşturmuştur. Bu sonuçlar HDP'nin sağ ve milliyetçi ideolojilerle olan yüksek kutuplaşmasını göstermektedir.

Altı parti çiftinin her ay için kutuplaşma ortalamalarından elde edilen ve ortalamayı gösteren çizge ile *dtw* mesafesinin en az olduğu parti çifti AK Parti-CHP çiftidir (0,0268). Bu sonuca göre iktidar partisi ve muhalefetin merkezi konumundaki ana muhalefet partisinin kutuplaşma seviyesi aynı zamanda TBMM'de dört partinin oluşturduğu ortalama kutuplaşmayı en çok yansıtır.



Şekil 4.6. 24-27. Dönemde (2011-2023) görece kutuplaşma seviyeleri



Şekil 4.7. 24-27. Dönemde (2011-2023) parti çiftleri arasındaki görece kutuplaşma DTW mesafeleri

Şekil 4.6'da çizge grafiği görülen görece kutuplaşma bir parti çiftini zaman içinde kutuplaşma seviyesindeki farklılaşmayı belirtir. Bütün parti çiftlerinin başlangıç noktasının 0 olarak ele alındığı analizde parti çiftlerinin kutuplaşma seviyeleri yerine kutuplaşmalarını ne kadar artırdıklarına yani görece kutuplaşmalarına odaklanılmıştır. Şekil 4.7'de parti çiftlerinin görece kutuplaşmalarının zaman içindeki DZB mesafeleri *dtw_0* ile gösterilmiştir.

Çizgelerdeki şekil benzerliği, serilerdeki örüntüleri, zaman içinde benzer veya farklı eğilimler gösterip göstermediğini belirlemeye yardımcı olur. Parti çiftlerini kutuplaşmaları aynı noktadan yani 0'dan başladığı için ölçekleri de birbirine yaklaşmıştır. Aynı ölçekteki çizgelerin DZB mesafeleri şekil benzerliğini daha iyi yansıtır. Örneğin CHP-HDP ve MHP-HDP kutuplaşması şekil olarak benzer olmasına karşın kutuplaşma seviyeleri farklı ölçekte seyreder. Derlem boyunca CHP-MHP 0,72-0,81 aralığında MHP-HDP 0,82-0,89 aralığında bir kutuplaşmaya sahiptir. Kutuplaşma seviyeleri arasındaki bu büyük farkın sonucu olarak Şekil 4.5'de *dtw_0* mesafesi 15 ölçüm arasında en yüksek üçüncü ölçüm çıkmıştır (1,0319). Kutuplaşma seviyelerini 0 noktasından başlatmak ölçek farkını azaltır ve serileri aynı ölçeğe yaklaştırır. Şekil 4.7'de (CHP-MHP) - (MHP-HDP) parti çiftlerinin görece kutuplaşma seviyesi farkı (0,89) ile en düşük dördüncü farktır. Bu mesafe aynı zamanda en düşük ikinci (AK Parti-HDP_MHP-HDP, 0,0814) ve üçüncüye (AK Parti-CHP_AK Parti-MHP, 0,0889) çok yakın çıkmıştır.

CHP-MHP ve MHP-HDP'nin derlem boyunca kutuplaşma seviyelerinin yüksek farkına rağmen görece kutuplaşma farkının düşük çıkması şekil benzerliğini gösterir. Bu iki parti çifti benzer örüntüye ve benzer eğilimlere sahiptir. MHP'nin Ekim 2016'da Ak Parti'ye hükümet değişikliği referandumunda destek vereceğini açıklaması ile birlikte muhalefet blokundan kopması ve iktidar blokuna yaklaşması sonucu bu iki parti ile kutuplaşmasını benzer şekilde artırmış, sonuç olarak bu seriler benzer eğilimlere ve şekle sahip olmuştur.

Görece kutuplaşma siyasi partilerin kutuplaşma hızlarını gözlemlemeye yarar. CHP-MHP görece kutuplaşmasını en çok artıran, en hızlı kutuplaşan partidir. Kutuplaşma seviyesi olarak Ekim-2016 ve Mart-2017 arasında en düşük parti çifti iken derlemin sonunda, Mayıs-2023'de en yüksek üçüncü parti çiftidir (Bkz. Şekil 3.4).

Derlem boyunca (AK Parti-CHP) ve (AK Parti-HDP) görece kutuplaşma farkının en az olduğu partidir (0,0349). CHP ve HDP'nin iktidara karşı muhalefet durumunun hiç

değişmemesi ve bu iki partinin sol ideolojiye yakınlığı AK Parti ile kutuplaşma eğilimlerini birbirine yaklaştırmıştır.

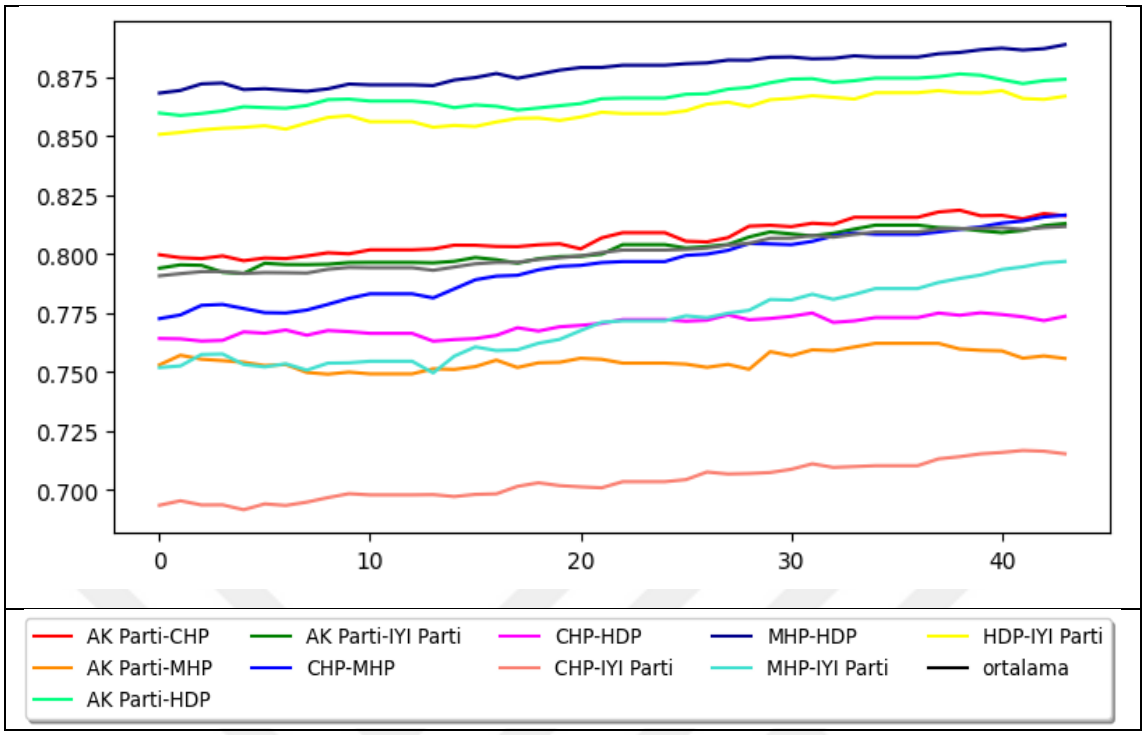
(AK Parti-HDP) ve (MHP-HDP) arasında en düşük ikinci görece kutuplaşma farkı vardır (0,0814). Bu iki parti derlem boyunca aynı zamanda en yüksek kutuplaşma seviyesine sahiptir. HDP'nin sağ milliyetçi ideoloji ile yüksek kutuplaşması Şekil 4.5'deki kutuplaşma seviyesinin yanında Şekil 4.7'deki şekil benzerliğinin ölçüldüğü kutuplaşma eğilimlerine de yansımıştır.

En düşük üçüncü kutuplaşma seviyesi (AK Parti CHP) ve (AK Parti-MHP) arasındadır (0,0889). Ak Parti-CHP arasındaki kutuplaşmanın seyri çok değişen hızda artarken Ocak-2016'dan itibaren AK Parti-MHP görece kutuplaşması azalmış ve bu iki serinin kutuplaşması birbirine yaklaşmış ve Haziran-2018'de çapraz kesmiştir. Bu durumun sonucu olarak görece kutuplaşmanın *dtw_0* farkı düşük çıkmıştır. Bu iki parti çiftinin görece kutuplaşmasının paralel seyretmemesi ve çapraz kesmesi MHP'nin muhalefetteki durumunun değişmesinin ve AK Partiye yaklaşmasının bir sonucu olarak değerlendirilmiştir.

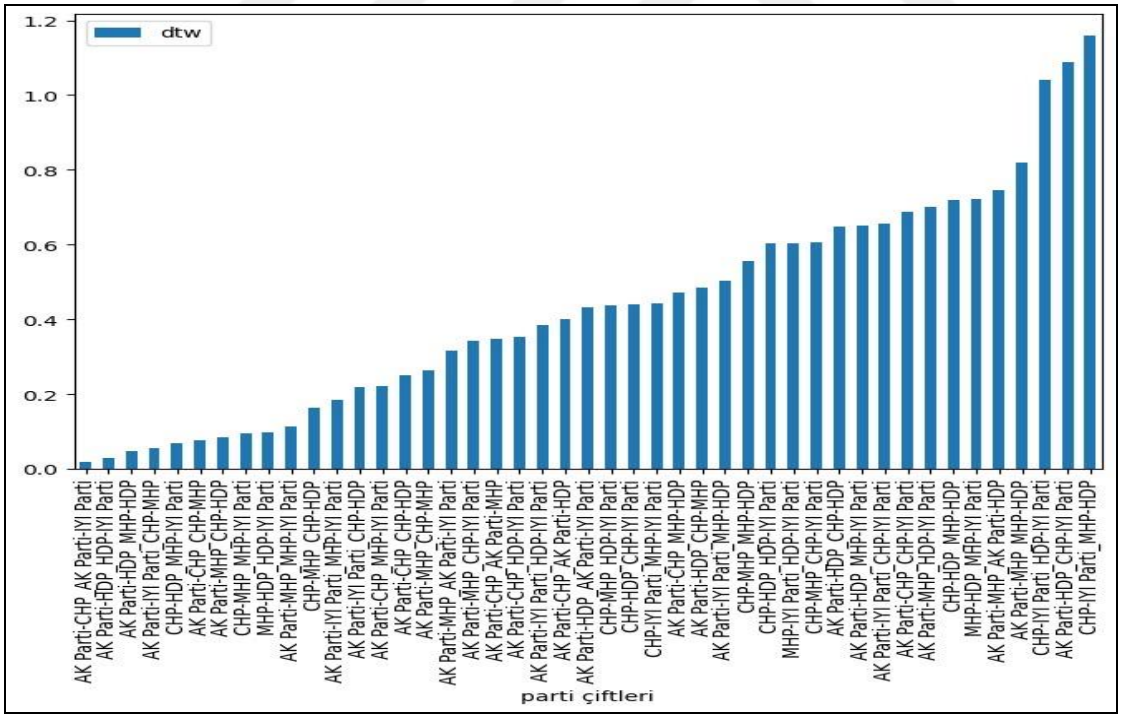
(CHP-MHP) arasındaki görece kutuplaşmanı sürekli artması ve en yüksek olarak seyretmesi (CHP-HDP) arasındaki görece kutuplaşmanın değişkenliği düşük seyretmesi sonucu bu iki parti çiftinin görece kutuplaşma farkı en yüksek parti çifti (0,4824) olmasını sağlamıştır. (MHP-HDP) görece kutuplaşmasının (CHP-MHP) parti çifti ile benzer eğilimde olması sebebi ile (MHP-HDP) ve (CHP-HDP) görece kutuplaşması farkı en düşük üçüncü parti çiftidir (0,3741).

(CHP-MHP) arasındaki görece kutuplaşma Ocak-2016'dan itibaren sürekli artmış, (AK Parti-MHP) arasındaki kutuplaşma ise Ocak-2016'dan itibaren azalmıştır. Bu iki zıt yönlü hareket iki parti çiftinin kutuplaşma farkını sürekli artırarak 2023 yılı sonunda toplam olarak en yüksek ikinci görece kutuplaşma farkı (0, 3933) olarak gerçekleşmesini sağlamıştır.

AK Parti-HDP görece kutuplaşma olarak ortalama görece kutuplaşmaya en yakın *dtw_0* mesafesine sahip parti çiftidir (0,0164).



Şekil 4.8. 27. Dönemde (2018-2023) kutuplaşma seviyeleri



Şekil 4.9. 27. Dönemde (2018-2023) parti çiftleri arasındaki kutuplaşma seviyesi DTW mesafeleri

27. dönemde (2018-2023) İYİ Partinin parlamentoya katılması ile analiz 5 partili hale gelmiştir. 27. Dönemin beş partili olması ve derlemdeki son dönemi içermesi nedeni ile bu dönem ayrıca incelenmiştir.

Şekil 4.8'de 27. dönemdeki 5 siyasi partinin oluşturduğu parti çiftlerinin kutuplaşma seviyelerinin çizge grafiği görünmektedir. Bu serilerin DZB mesafeleri Şekil 4.9'da *dtw* ile gösterilmiştir.

Kutuplaşma seviyelerine göre AK Parti-CHP ve AK Parti-İYİ Parti en küçük *dtw* mesafesine sahip parti çiftleridir (0,0179). 27. Dönemde iktidardaki AK Parti, muhalefetin oluşturduğu seçim ittifakının (Millet İttifakı) iki büyük partisi ile benzer kutuplaşma seviyeleri oluşturmuştur. Bu sonuç seçim ittifaklarının kutuplaşma seviyelerine etkisini gösterir.

27. Dönemde sağ, milliyetçi ideolojinin kutuplaşma seviyelerine etkisi bu ideoloji ile çatışma halindeki HDP'nin üç sağ milliyetçi parti; AK Parti, MHP ve İYİ Parti ile oluşturduğu kutuplaşma seviyelerinden gözlemlenebilir. HDP ile kutuplaşma söz konusu olduğunda bu üç partinin çok benzer davrandığı ve ideolojik köklerinden dolayı HDP'ye karşı benzer kutuplaşma seviyeleri oluşturduğu gözlemlenmiştir. 27. Dönem boyunca MHP-HDP, AK Parti-HDP ve İYİ Parti-HDP en yüksek kutuplaşma seviyelerini göstermiştir.

CHP, İYİ Parti ve HDP muhalefet blokunu oluşturduğu düşünülebilir. HDP ve İYİ Parti muhalefet rolünde olmalarına rağmen oluşturdukları yüksek kutuplaşma seviyesi bu parti çiftleri için milliyetçi ideolojinin iktidar-muhalefet statüsünü baskıladığını gösterir. Şekil 4.9'da (AK Parti-HDP) -(HDP-İYİ Parti) en düşük ikinci mesafe (0,0299), (AK Parti-HDP) (MHP-HDP) en düşük üçüncü mesafedir (0,0489) ve kutuplaşma seviyeleri benzer seyreder.

(AK Parti-İYİ Parti) ve (CHP-MHP) parti kutuplaşma seviyelerinin 0,056 *dtw* mesafesi ile benzer seyretmesi Cumhur ittifakı üyelerinin (AK Parti, MHP) Millet İttifakı üyeleri (CHP, İYİ Parti) kutuplaşmalarının benzerliğinden kaynaklandığı değerlendirilmiştir. Sonuçlara göre seçim ittifaklarının etkisi bu parti çiftlerinin kutuplaşma seviyelerine yansımıştır.

CHP ve İYİ Parti 27. Dönemde muhalefet blokunun partileridir ve birlikte seçim ittifakı kurmuşlardır. Bu yakınlaşmanın sonucu olarak 27. dönem boyunca kutuplaşma seviyesinin en düşük seyrettiği parti çiftidir. Derlemde (CHP-İYİ Parti) çifti HDP'nin sağ milliyetçi

partilerle oluşturduğu kutuplaşma seviyeleri ile en yüksek farkı oluşturmuştur. (CHP-İYİ Parti) ve (MHP-HDP) 1,1587 ile en yüksek kutuplaşma farkına sahiptir. (CHP-İYİ Parti) (AK Parti-HDP) *dtw* mesafesi 1,0874 ile en yüksek ikinci, (HDP- İYİ Parti) 1.0405 ile en yüksek üçüncü kutuplaşma seviyesine sahiptir.

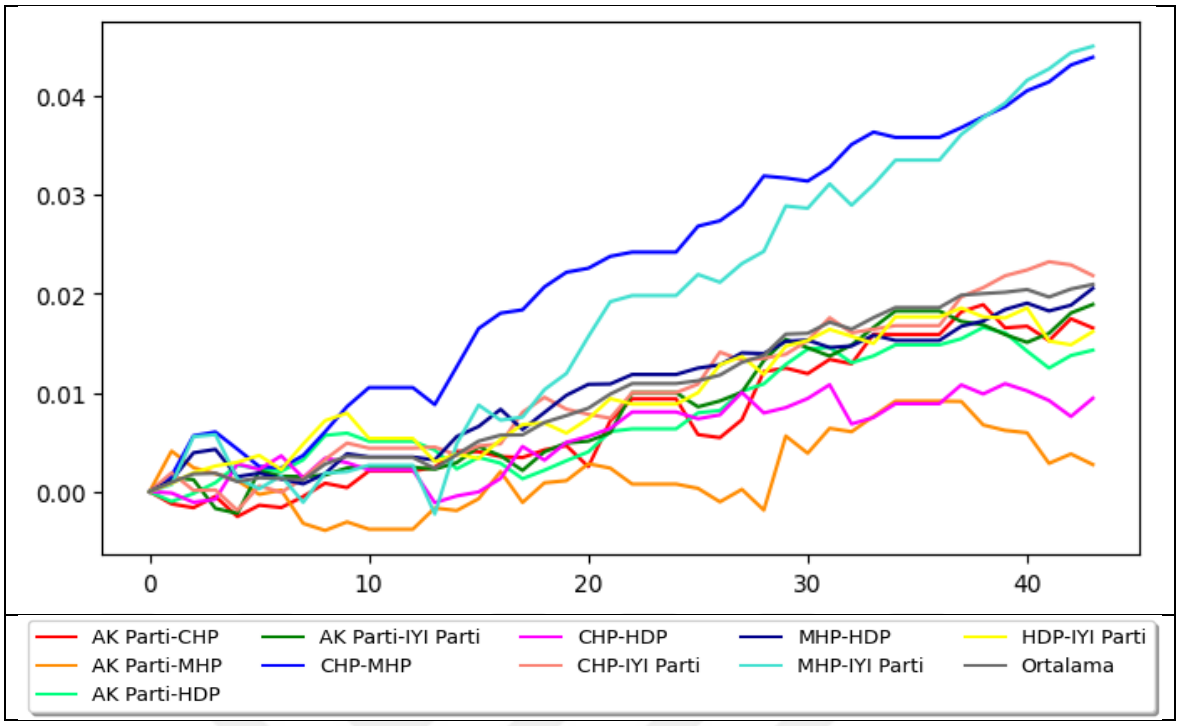
Kutuplaşma seviyeleri dikkate alındığında bütün partilerin kutuplaşma ortalaması ile en benzer eğilimi AK Parti-İYİ Parti (0,0088) ve AK Parti-CHP (0,026) gösterir. Bu parti çiftlerinin kutuplaşması TBMM'deki ortalama kutuplaşmanın da göstergesidir denilebilir.

Şekil 4.10'da parti çiftlerinin görece kutuplaşma seviyeleri, Şekil 11'de parti çiftlerinin birbirleri ile oluşturdukları DZB mesafeleri *dtw_0* olarak gösterilmiştir. Bu mesafeler aynı zamanda eğrilerin şekil benzerliğini, eğilim benzerliğini de gösterir.

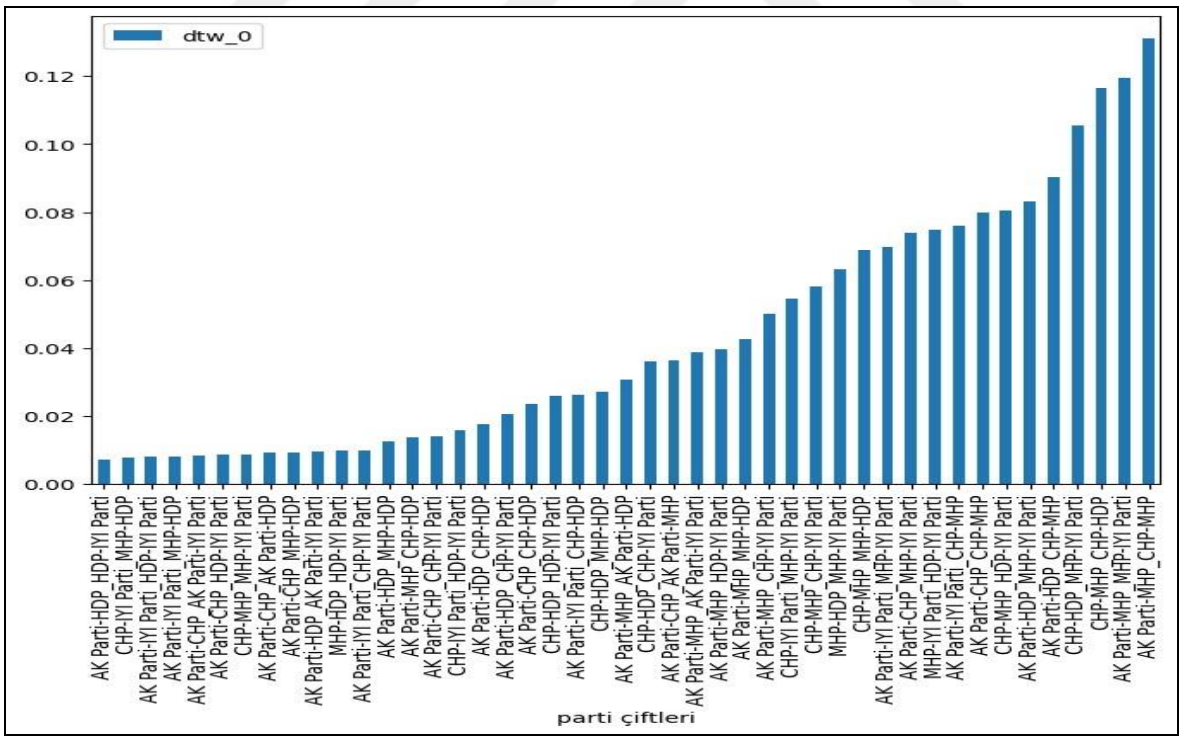
Şekil 4.10'da CHP-İYİ Parti, AK Parti-HDP, MHP-HDP, HDP-İYİ Parti ve AK Parti-İYİ Parti çiftleri benzer eğilimi gösterdiği için bu parti çiftlerinin oluşturduğu *dtw_0* mesafeleri düşük ve birbirine yakın çıkmıştır. Bu partilerin kutuplaşma hızları birbirine çok benzerdir. En düşük mesafe 0, 0072 ile (AK Parti-HDP) (HDP-İYİ Parti) arasındadır. Bütün *dtw_0* mesafelerinin düşük olduğu bu gruptaki en yüksek mesafe ise 0,1 ile CHP-İYİ Parti ve AK Parti-İYİ Parti arasındadır.

En çok görece kutuplaşma farkı oluşturan parti çifti (CHP-MHP)'dir. Bu iki partinin 27. Dönem öncesi muhalefet blokunu oluştururken 27. Dönem ile birlikte MHP'nin iktidar partisi AK Parti ile ittifak kurması, CHP'nin ise muhalefet partisi olarak devam etmesi görece kutuplaşma seviyelerine yansımıştır. (AK Parti-MHP) iktidar blokunun iki üyesi olarak en düşük görece kutuplaşma seviyesine sahiptir. Bu iki durumun sonucu olarak (CHP-MHP) ve (AK Parti-MHP) 0,131 ile en yüksek görece kutuplaşma farkına sahiptir.

(MHP-İYİ Parti) görece kutuplaşması, bu iki milliyetçi partinin ideolojik yakınlıklarına rağmen MHP'nin iktidar blokunun İYİ Partinin ise muhalefet blokunun üyesi olmasının sonucu olarak özellikle Kasım-2021'den itibaren artmıştır ve dönem sonunda en yüksek farka ulaşmıştır. 27. dönem sonunda ise en yüksek ikinci kutuplaşma seviyesine sahiptir. AK Parti-MHP ise görece kutuplaşma seviyesi en düşük seyreden parti çiftidir. Bu durumun sonucu ise (MHP-İYİ Parti) ve (AK Parti-MHP) en düşük ikinci görece kutuplaşmaya sahip (0.1196) olmuştur.



Şekil 4.10. 27. Dönemde (2018-2023) görece kutuplaşma seviyeleri



Şekil 4.11. 27. Dönemde (2018-2023) parti çiftleri arasındaki görece kutuplaşma DTW mesafeleri

4.1.3. Parti çiftleri arasında Granger nedensellik analizi

Granger nedenselliği (Granger, 1969), bir zaman serisinin başka bir zaman serisi değişkenini tahmin etme gücünü değerlendirmek için kullanılan istatistiksel bir yöntemdir.

Örnek olarak A ve B olmak üzere iki zaman serisi ele alındığında, A'nın geçmiş değerleri, B'nin gelecek değerlerinin tahminine yardımcı oluyorsa A ve B arasında Granger nedenselliği olduğu kabul edilir. Nedenselliğin yönü A'dan B'ye doğrudur ve *A Granger-neden B* olarak ifade edilir. Granger nedensellik iki zaman serisi değişkeni arasında öngörülebilirlik anlamında istatiksel kanıt sağlar, sonuçları kesinlik (deterministik) belirten bir neden sonuç ilişkisi olarak değerlendirmemek gerekir.

Çalışmada iki zaman serisi arasındaki tahmin gücünü ölçerken basit regresyon analizi (Ordinary Leats Squares, OLS) kullanılmıştır ve statmodel kütüphanesinden yararlanılmıştır. İki serinin Granger nedenselliği 1ve 2 ay bekleme (lag) değeri için test edilmiştir

Durağan zaman serilerinde ortalama ve varyans zaman içinde değişmez. Bu özellik, model parametrelerinin tutarlı bir şekilde tahmin edilmesini sağlar. Nedensellik ilişkileri hakkında güvenilir çıkarımlar yapılması için zaman serisi değişkenlerinin durağan olması önemlidir. Augmented Dickey-Fuller (ADF) testi, bir zaman serisinin durağan mı yoksa durağan olmayan mı olduğunu belirlemeye yardımcı olan istatistiksel bir hipotez testidir.

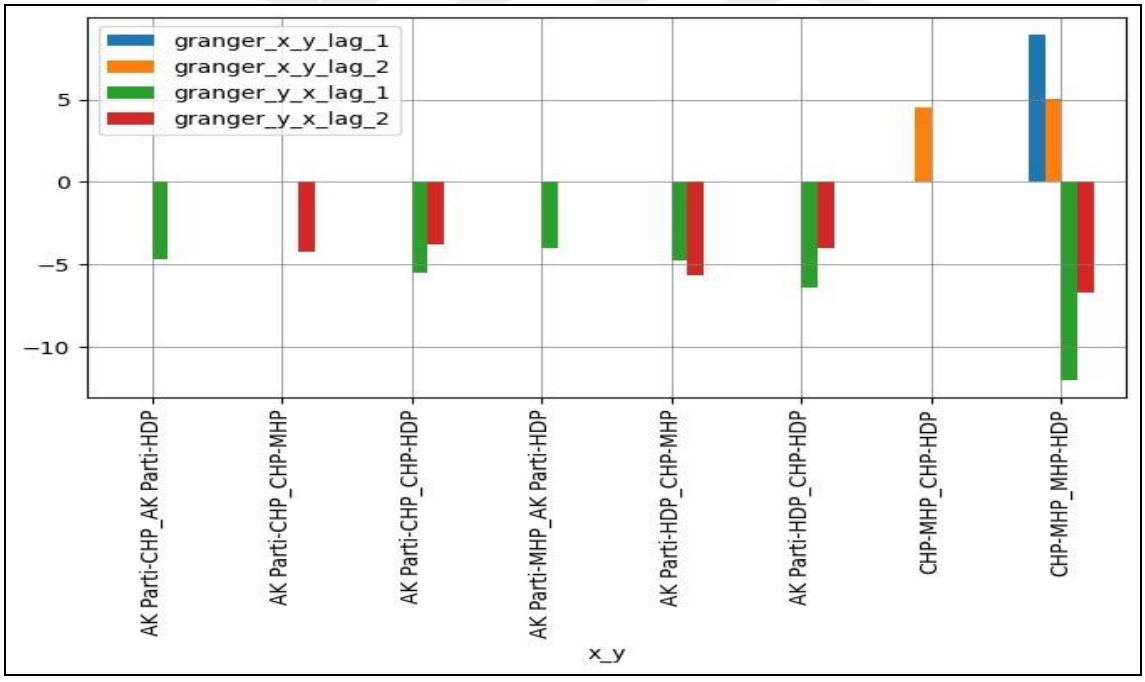
Çalışmada seriler ADF testi ile durağan olmadığı belirlendikten sonra birinci dereceden fark alma işlemi uygulanmıştır. Fark alma işleminden sonra tekrar ADF testi uygulanmış ve serilerin durağan hale geldiği görülmüştür. Böylece Granger nedensellik testinin durağan seriler gereksinimi sağlanmıştır.

Granger testin nedensellik anlamlılığı F-Test ile ölçülmüştür. F-istatistik değeri Granger nedenselliğin gücünü verir. İstatiksel anlamlılığın güven aralığı %95 ($p<0,05$)'dir.

Şekil 4.12 2011-2023 yılları arasındaki dört partinin (AK Parti, CHP, MHP, HDP) oluşturduğu parti çiftlerine ait zaman serilerinin Granger nedensellik sonuçlarını gösterir. Şekilde birinci parti çifti x, ikinci parti çifti y, zaman gecikmesinin ay olarak değeri lag

olarak ifade edilmiştir. Şekildeki değerler Granger nedensellik testinin F-istatik değerleridir. F-istatistik değerleri daima pozitiftir Şekildeki pozitif ve negatif değerler %95 güven aralığına göre istatiksel olarak anlamlı Granger nedenselliğinin yönünü gösterir. Görselleştirmeyi kolaylaştırmak için bu şekilde gösterilmiştir. Şekildeki parti çiftleri isimleri baz alındığında; pozitif değerler ikinci serinin birinci serinin Granger nedeni olduğunu, negatif değerler ise birinci serinin ikinci serinin Granger nedeni olduğunu gösterir.

Granger nedenselliği sonuçları yorumlanırken bir zaman serisi değişkeninin (X) diğer bir (Y) serisi üzerindeki tahmin gücünü ölçen istatistiksel bir kavram olduğuna, kesin bir nedensellik belirtmediğine dikkat edilmelidir. Granger X arttığında kesin olarak Y'nin de artacağı ya da azalacağı anlamına gelmez. Çalışmada nedensellik sonuçları yorumlanırken kutuplaşma ifadesi yüksek kutuplaşmayı belirtmemiş, parti çiftleri arasındaki ilişkiyi belirtmiştir.



Şekil 4.12. 24-27. Dönemde (2011-2023) parti çiftleri arasında Granger nedenselliği

Şekil 4.12'de Granger nedenselliğin en fazla olduğu parti çifti CHP-MHP ve MHP-HDP'dir. Granger nedenselliği bir ay ve iki ay gecikme ile istatiksel olarak anlamlıdır ve yönü çift yönlüdür. Bu iki parti çiftinin kutuplaşmasının bir ay ve iki ay gecikme ile birbirini etkilediği söylenebilir. Bu parti çiftleri yüksek kutuplaşma seviyeleri farkına (Bkz. Şekil 4.4) rağmen görece kutuplaşma benzerliği farkının düşük çıkması (Bkz. Şekil 4.7) şekil benzerliğine veya

eğilim benzerliğine işaret eder. Granger nedensellik testi bu durumu doğrulamıştır. CHP-MHP ve MHP-HDP kutuplaşmaları çift yönlü olarak birbirini en çok etkileyen serilerdir.

Ak Parti-HDP ve CHP-HDP arasındaki Granger nedenselliğin yönü AK Parti-HDP'den CHP-HDP'ye doğrudur. AK Parti-HDP kutuplaşmasının bir ay ve iki ay gecikme ile CHP-HDP kutuplaşmasını etkilemesi HDP ve CHP'nin derlem boyunca aynı durumda, muhalefet rolünde olması, parti kutuplaşma seviye farklarının yüksekliği (Bkz. Şekil 4.4), görece kutuplaşma seviye farklarının yüksekliği (Bkz. Şekil 4.7) düşünüldüğünde beklenmeyen bir durumdur.

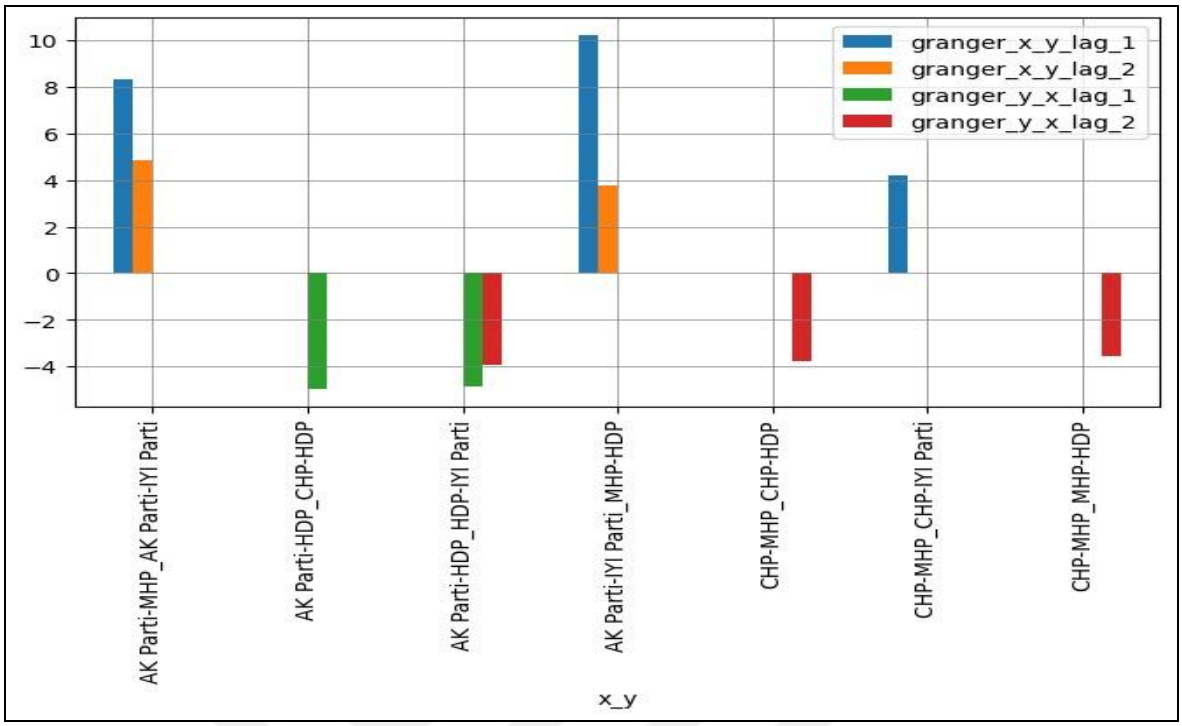
AK Parti-CHP parti çifti aynı şekilde CHP-HDP ile Granger nedensellik oluşturmuştur. CHP-HDP kutuplaşmasının seyri bu iki partinin AK Parti ile kutuplaşmasından aynı şekilde etkilendiği sonucuna varılabilir. İki partinin AK Partiye muhalefet olma durumları derlem boyunca değişmediği için bu sonuç beklenebilir. CHP_HDP iki ay gecikme ile CHP-MHP'nin Granger nedenidir. CHP'nin HDP ile kutuplaşma seviye seyrinin MHP ile kutuplaşma seviyesini de etkilediği Granger nedensellik teorisine göre söylenebilir.

AK Parti-HDP ile CHP-MHP arasındaki Granger nedenselliği bir ay ve iki ay gecikme ile AK-Parti-HDP'den CHP-MHP'ye yönündedir. İdeolojik olarak Ak Parti ve MHP'nin sağ yönelimli, CHP ve HDP'nin sol yönelimli partiler olduğu düşünüldüğünde, bu sonuç siyasi partilerin ideolojik konumlarının Granger nedenselliğe yansıdığını gösterir.

AK Parti-CHP kutuplaşması AK Parti HDP kutuplaşmasının bir ay gecikme ile Granger nedenidir. AK Partinin merkez soldaki CHP ile kutuplaşması sol yönelimli HDP ile kutuplaşmasını etkilemesi ideolojinin sonuçlar üzerine etkisini gösteren diğer bir bulgudur.

AK Parti-CHP kutuplaşması ise iki ay gecikme ile CHP-MHP kutuplaşmasının Granger nedenidir. Bu sonuç merkez sol ideolojideki CHP'nin milliyetçi muhafazakâr iki sağ parti ile oluşturduğu parti çiftlerinin kutuplaşma seyrinin üzerindeki etkiyi göstermektedir.

AK Parti-MHP ve AK Parti-HDP arasındaki birinci seriden ikinci seriye doğru bir aylık gecikme ile Granger nedenselliği vardır. Şekil 4.12'den AK Parti'nin MHP ile ilişkisinin HDP ile olan ilişkisini etkilediği sonucuna varılabilir.



Şekil 4.13. 27. Dönemde (2018-2023) parti çiftleri arasında Granger nedenselliği

Şekil 4.13'de 27. dönemde %95 güvenle istatiksel olarak anlamlı Granger nedensellik olan parti çiftleri görünmektedir.

Şekil 4.12'deki sonuçlara benzer şekilde Şekil 4.13'de CHP_MHP'den MHP_HDP'ye yönlü 2 ay gecikmeli Granger nedenselliği vardır. Derlemin kapsamı 27. Döneme daraltıldığında Şekil 4.12'deki CHP_HDP ve CHP_MHP arasındaki Granger nedenselliğin yönü değişmiş ve CHP_MHP'den CHP-HDP'ye doğru gerçekleşmiştir. AK Parti-HDP ve CHP-HDP arasındaki Granger nedenselliğ bir ay gecikme ile Şekil 4.12'deki nedensellikle aynı yönlü gerçekleşmiştir.

AK Parti-İYİ Parti ve MHP_HDP çiftlerinin kutuplaşma seviyeleri MHP-HDP'den AK Parti-İYİ Partiye yönlü bir ve iki ay gecikmeli Granger nedenselliği oluşturmuştur.

AK Parti-İYİ Parti kutuplaşması AK Parti-MHP kutuplaşmasının bir ve iki ay gecikme ile Granger nedenidir. Sonuçlardan AK Partinin milliyetçi ideolojiye sahip İYİ Parti ile kutuplaşması diğer bir milliyetçi ideolojideki MHP ile kutuplaşmasını etkilediği söylenebilir.

AK Parti HDP kutuplaşması bir ay ve iki ay gecikme ile HDP-İYİ Parti kutuplaşmasının Granger nedenidir. HDP sağ milliyetçi ideolojiye sahip AK Parti ile kutuplaşması yakın ideolojiye sahip İYİ Parti ile kutuplaşmasını tahmin etmede yardımcı olur. Bu sonuç ideolojinin Granger nedenselliğe yansıdığı diğer bir ölçümdür.

CHP'nin İYİ Parti ile kutuplaşması MHP ile kutuplaşmasının Granger nedenidir. Merkez sol ideolojinin milliyetçi ideolojilerle kutuplaşmasının bu partilerle ilişkisini tahmin etmeye yardımcı olduğu sonucu çıkarılabilir.

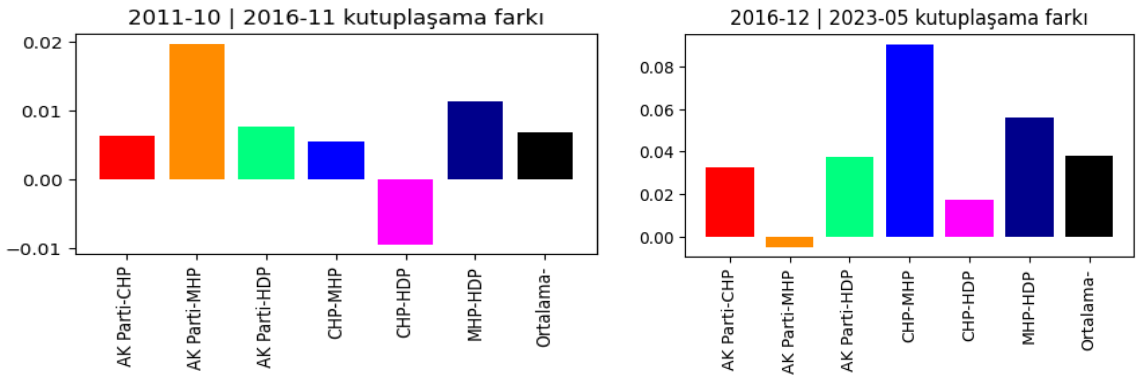
Siyasi parti kutuplaşma zaman serilerinde yapılan zaman serisi analizi ideolojilerin ve seçim ittifaklarının Granger nedensellik üzerinde anlamlı etkisinin olduğunu göstermiştir.

4.1.4. Hükümet sistemi değişikliğinin parti kutuplaşmalarına etkisi

22 Kasım-2016 tarihli TBMM'deki MHP grup toplantısı MHP lideri Devlet Bahçeli'nin anayasa değişikliği işin AK Partiye destek vereceğini açıkladığı tarihtir. Bu yüzden Kasım 2016 MHP'nin muhalefet statüsünün değişmeye başladığı tarih olarak ele alınabilir. MHP AK Parti arasında iş birliği Nisan-2017 anayasa değişikliği referandumuna evet oyu vererek ve takip eden seçimlerde aynı ittifak içinde yer alarak devam etmiştir. Muhalefet blokunda değişimin özellikle CHP-MHP, MHP-HDP ve AK Parti-MHP kutuplaşma seviyesine yansıması beklenir.

Şekil 3.7'den çıkarılacak sonuca göre bloklaşmanın ve seçim ittifaklarının oluşturduğu yeni durumun kutuplaşmaya hemen yansımadığı görülür. Bunun kullanılan yöntemden kaynaklanan sebebi kutuplaşmanın ayrık değil 12-aylık hareketli ölçülmesidir. Takip eden her ay için bir aylık doküman analize girmiş, bir yıllık kutuplaşma ölçülürken yeni bloklaşmanın kutuplaşma sonuçlarına yansıması tedrici olmuştur.

Kasım 2017 iktidar muhalefet bloklaşmasının yeniden belirlendiği tarih olarak ele alınırsa bu tarih öncesi ve sonrası iki zaman aralığının başında ve sonunda oluşan kutuplaşma farkları Şekil 4.14'de gösterilmiştir.



Şekil 4.14. Ekim 2012 – Ekim 2016 ve Kasım 2016 – Mayıs 2023 arasında kutuplaşma farkı

Bu iki zaman dönemini karşılaştırılırken ortalama kutuplaşmanın arttığı göz önünde bulundurulmalıdır. Ortalama kutuplaşma ilk dönemde 0,0068 artmışken ikinci dönemde 0,0382 artmıştır.

İlk dönemde kutuplaşması en çok artan çift AK Parti-MHP'dir. Bu parti çifti ikinci dönemde kutuplaşması azalan tek çifttir. 4 partinin oluşturduğu parti çiftlerinden AK Parti-MHP birinci dönemde en düşük ikinci ya da üçüncü parti çifti iken ikinci dönemde en düşük parti çifti olmuştur. İkinci dönemin sonunda CHP-MHP kutuplaşmasının çok altında kalmıştır.

CHP-MHP kutuplaşması ikinci dönemde en hızlı artan ve en büyük farkı (0,0906) oluşturan parti çiftidir. Bu fark ilk dönem için 0,0054'tür. Bu ölçüm MHP ve CHP'nin TBMM'de farklılaşmasının derecesini gösterir.

MHP-HDP ise birinci dönem 0,0113'lük bir fark oluşturmuşken ikinci dönem kutuplaşmasını artırarak 0,056'lık bir fark oluşturmuştur. MHP-HDP ikinci dönemde MHP-CHP'den sora en büyük kutuplaşma farkının olduğu parti çiftidir.

İki dönemdeki farklılaşmanın sonucunda CHP-MHP kutuplaşması Ekim-2012'den Kasım-2017'ye kadar en düşük kutuplaşma seviyesi olarak kalmıştır. CHP-MHP kutuplaşmasının hızı Aralık-2017, MHP-HDP kutuplaşmasının hızı ise Şubat 2017'de artmıştır. MHP-HDP kutuplaşması Ekim-2012-Nisan-2014 arasında AK Parti-HDP kutuplaşmasını yakın seyredip hemen altında yer alırken Nisan-2014 tarihinden itibaren artarak Nisan-2014 Mayıs-2023 arasında kadar en yüksek kutuplaşma seviyesi olarak kalmıştır.

CHP-HDP arasındaki kutuplaşma farkı birinci dönem için 0,0094 aşağı yönlü iken ikinci dönem 0,0175 yukarı yönlüdür. Bu kutuplaşma CHP-HDP'nin TBMM'deki genel kutuplaşma seviyesinin artmasından ve CHP'nin milliyetçi ideolojinin yükselişinden etkilenmesinin bir sonucu olarak yorumlanmıştır.

AK Parti ve CHP'nin iktidar ve muhalefet olma durumu ve birbirlerine karşı tavırlarında bir değişiklik olmamıştır. İki dönem karşılaştırıldığında ilk dönemki fark 0.0063, ikinci dönemki fark 0,0325'dir. TBMM'deki ortalama kutuplaşmanın artışı doğrultusunda bu parti çifti de kutuplaşmasını artırmıştır. Derlem boyunca ortalama kutuplaşmaya en yakın seyreden parti çifti bu durumu her iki dönem için de korumuştur.

4.2. TBMM Genel Kurul Görüşmelerinde Yazar Profili Oluşturma Bulguları

Çalışmada, demografik özelliklerin tahmininde, modelleri değerlendirmek için çapraz doğrulama kullanılmıştır. İlk olarak, dengesiz derlemden rastgele iki ayrı dengeli alt veri kümesi oluşturulmuştur. Ardından, bu iki alt veri kümesi üzerinde 4 kat (4-fold) çapraz doğrulama ile metin sınıflandırması gerçekleştirilmiştir. Modelleri karşılaştırırken doğruluk değerlerinin ortalamasını kullanılmıştır. Derin öğrenme modellerinde, çapraz doğrulamanın her bir katı bir doğrulama ve test setine bölünmüştür. Böylece, doğrulama kümesi %12,5, test kümesi %12,5 ve eğitim kümesi %75'tir. Üst parametreleri ayarlarken klasik makine öğrenimi algoritmalarına ızgara araması ve derin öğrenme algoritmalarına rastgele arama uygulanmıştır. Klasik makine öğrenimi algoritmalarının bu deneysel ortamı milletvekillerinin demografik özelliklerinin ikili analizinin gerçekleştirildiği Bölüm 4.8'de LR ile de kullanılmıştır.

Model çiftleri arasındaki anlamlılık düzeyi (McNemar, 1947) testi ile α anlamlılık eşiği ($\alpha = 0.05$) için belirlenmiştir. McNemar testi, istatistikte ikili karşılaştırmalar için kullanılır. Test için oluşturulan olumsallık çizelgesinin (contingency table) nihai değerleri, çapraz doğrulamanın her bir katının olumsallık çizelgesi değerlerinin ortalaması alınarak hesaplanmıştır.

Deney düzeneğinin tekrarlanabilirliği, her bir sınıflandırma görevi için aynı alt veri kümeleri ve üst parametreler seçilerek sağlanmıştır.

Sınıflar arasındaki ilişkiler, tüm özellik tahmin görevleri için hata matrisi üzerinde hata analizi ile incelenmiştir. Yanlış sınıflandırılmış bir örneğin tahmin edilen kategoriye yakın olduğu varsayılmıştır.

Çizelge 4.2. Konuşmaların demografik ve siyasi özelliklerine göre sınıflandırma doğrulukları

Model	Cinsiyet	Yaş	Eğitim	Meslek	Parti	Parti Durumu	Seçim Bölgesi
TF-IDF(kelime)_LR	0,81	0,52	0,60	0,66	0,84	0,92	0,54
TF-IDF(kelime_karakter)_LR	0,82	0,52	0,60	0,67	0,84	0,92	0,54
TF-IDF(kelime)_DVM	0,81	0,52	0,60	0,65	0,84	0,92	0,53
TF-IDF(kelime_karakter)_DVM	0,82	0,52	0,60	0,66	0,84	0,92	0,53
LDSE_LR	0,75	0,44	0,52	0,61	0,74	0,88	0,49
PV_LR	0,78	0,46	0,54	0,55	0,79	0,80	0,19
BERT	0,77	0,43	0,52	0,56	0,80	0,91	0,46
BoW_ İBSA	0,77	0,44	0,53	0,58	0,59	0,87	0,46
word2vec_ İBSA	0,68	0,32	0,38	0,38	0,26	0,78	0,30
Baseline-VecAvg	0,72	0,38	0,40	0,50	0,64	0,81	0,30
Baseline-majority	0,50	0,25	0,25	0,25	0,25	0,50	0,14

Çizelge 4.2 milletvekillerinin demografik ve parlamento özellikleri tahmin görevlerinde kullanılan modellerden elde edilen doğruluk değerlerini göstermektedir. TF-IDF özelliklerine sahip klasik makine öğrenimi modelleri, tüm özellik tahmin görevlerinde en yüksek doğrulukları vermiştir. TF-IDF_LR ve TF-IDF_DVM modellerinde, içerik tabanlı (kelime n-gramları) ve stil tabanlı (karakter n-gramları) özelliklerin kombinasyonu, doğruluklara ve McNemar testine göre cinsiyet ve meslek demografik özelliklerinde sınıflandırma doğruluğunu artırmıştır. Yaş, eğitim, parti, bölge ve parti statüsü özellikleri hem TF-IDF(kelime) hem de TF-IDF(kelime_karakter) özellikleri için aynı kalmıştır.

DVM ve LR yedi sınıflandırmanın beşinde aynı doğruluğa sahiptir: Cinsiyet için %82, yaş için %52, eğitim için %60, parti için %84 ve parti statüsü için %92. TF-IDF_LR sadece meslek (%67) ve seçim bölgesinde (%54) TF-IDF_DVM'den daha yüksek bir doğruluğa sahiptir. McNemar testi uygulandığında bu iki model arasında sadece seçim bölgesi özelliğinde anlamlı fark oluşmuştur.

Sadece BERT, klasik makine öğrenimi modelleriyle tek bir görevde, parti statüsü üyeliğinde rekabet edebilmiştir. BERT, milletvekillerinin parti durumunu TF-IDF_LR ve TF-

IDF_DVM'den %1 daha düşük (%91) doğrulukla tahmin etmiştir. McNemar testine göre her iki modelde de aynı hata oranına sahiptir. BERT parti aidiyetinde %80 ile klasik makine öğrenimi algoritmalarından sonra en yüksek performansa sahip olmuştur.

LDSE, çalışmada kullanılan başlangıç noktası değerlerin üzerinde doğruluğa sahip olmasına rağmen, TF-IDF özelliklerini kullanan klasik makine öğrenmesi modellerinin altında performans göstermiştir. Yine de LDSE, meslek (%61) ve bölge (%49) sınıflandırma görevlerinde TF-IDF_LR ve TF-IDF_DVM'den sonra en başarılı sonucu vermiştir. LDSE ayrıca parça durumu sınıflandırmasında %88 doğrulukla BERT'i takip etmiştir.

PV_LR, cinsiyet demografik özelliğinde %78 ile doğruluk değerleri açısından en iyi ikinci model olmuştur. Ayrıca parti sınıflandırmasında BERT'ten %1 daha düşük doğruluğa (%79) sahipti. McNemar testine göre BERT ve PV_LR arasında anlamlı bir fark vardır. Öte yandan PV, seçim bölgesi özelliğinde %19 ile genelleme yapamamıştır. Baseline-VecAvg'den (%30) daha düşük doğruluğa sahiptir ve çoğunluk taban çizgisine (%14) yakındır.

BoW_ İBSA, word2vec_ İBSA hiçbir görevde temel değerin üzerinde bir performans elde edememiştir. Parti üyeliğinde (%26) bile çoğunluk taban çizgisine yakın bir doğrulukla genelleme yapamamıştır.

BoW_ İBSA, parti üyeliğinde taban çizgisinin altında doğruluğa sahipti. İBSA başka bir derin öğrenme modeli olan BERT ile karşılaştırıldığında, BoW_ İBSA cinsiyet (%77) ve bölgede (%46) aynı doğruluğu, yaşta (%44), eğitimde (%53) ve meslekte (%58) ikinci en yüksek doğruluğu vermiştir. BoW_ İBSA ve BERT, McNemar testi uygulandığında cinsiyet demografik özelliğinde anlamlı bir fark göstermemiştir.

TF-IDF (kelime ve karakter) tabanlı modeller, cinsiyet ve meslek için TF-IDF (kelime) modellerinden daha yüksek doğruluğa sahipti. Karakter n-gramı yazarın üslubuna dair özellik (stylometry) olarak değerlendirilmiştir. Bu sonuçlardan Yazarın üslubunun konuşmaların ayırt edilmesinde etkili olduğu sonucuna varılabilir.

En yüksek doğruluk değerine sahip demografik özellik tahmini görevinin konuşmalarda en çok öne çıkan özellik olduğu söylenebilir. Çizelge 4.2, milletvekillerinin konuşmalarına

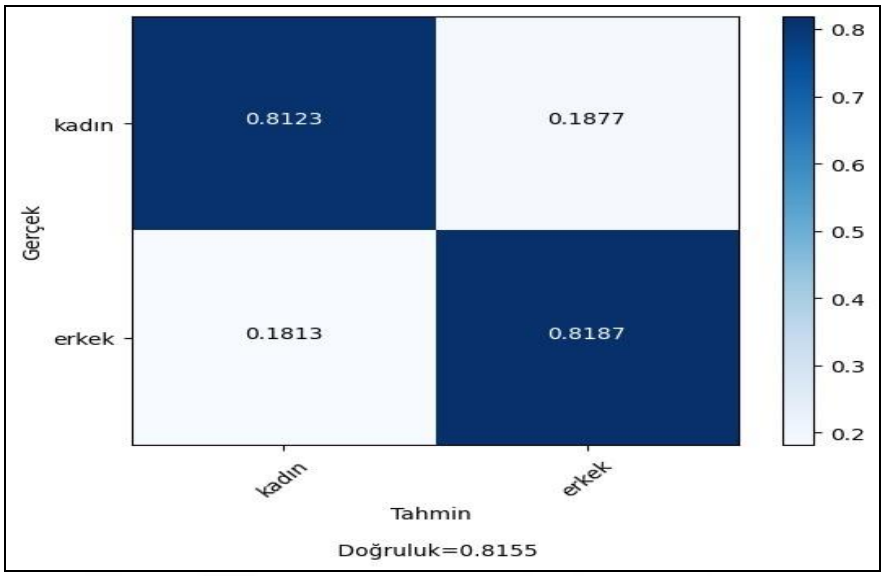
çoğunlukla parti statülerini ve parti aidiyetlerini yansıttıklarını göstermektedir. Cinsiyet konuşmalarda öne çıkan diğer bir özelliktir.

4.2.1. Cinsiyet tahmini

Cinsiyetler arasındaki karışıklık (confusion) Şekil 4.15'de gösterilmiştir. Hata matrisine göre, cinsiyet erkek milletvekilleri için (%81,87) kadın milletvekillerine (%81,23) kıyasla sadece 0,0064'lük bir farkla daha tahmin edilebilirdir. Çalışmanın sonuçları, TBMM'de erkek ve kadın milletvekillerinin konuşmalarının yüksek oranda (%82 doğrulukla) ayırt edilebilir olduğunu göstermektedir.

Kadın cinayetleri, kadına ve çocuğa yönelik şiddet, Şekil 4.16'daki kadın milletvekillerinin terimlerinden çıkan en belirgin bulgulardır. Kadına yönelik şiddet, kadın cinayeti, çocuk istismarı, kavga şiddeti, taciz tecavüz, çocuk cinsel, erkek şiddeti ve şiddete maruz kalma gibi terimler ülkede kadınlara karşı işlenen ciddi suçların göstergesidir. Toplumsal cinsiyetin anaakımlaştırılması, kadınlar ve erkekler arasında eşitliğin sağlanmasına yönelik bir stratejidir (Mazey, 2002). Toplumsal cinsiyetin anaakımlaştırılması ile ilgili terimler de Şekil 4.15'de listelenmiştir. Sözcükler kadın hakları, kadın istihdamı, toplumsal cinsiyet, toplumsal cinsiyet eşitliği, kadın işletmeleri, erkek egemen, kadın örgütlenmesi ve kadın siyaseti gibi toplumsal cinsiyet anaakımlaştırmasının sözcükleridir.

Erkek milletvekillerinin konuşmalarından çıkarılan kelimeler ise konular bakımından çeşitlilik göstermektedir. Polemik, NATO, fiyat, istihbarat teşkilatı, konu, elektrik parası, banka tarımı, rüşvet yolsuzluğu, kamulaştırma ve arazi dekarı erkeklerin terimleri arasında yer almıştır. Tarım, ekonomi ve finans, ordu ve istihbarat alanları Türkiye Büyük Millet Meclisi'ndeki (TBMM) erkek milletvekillerinin ilgi alanlarıdır.



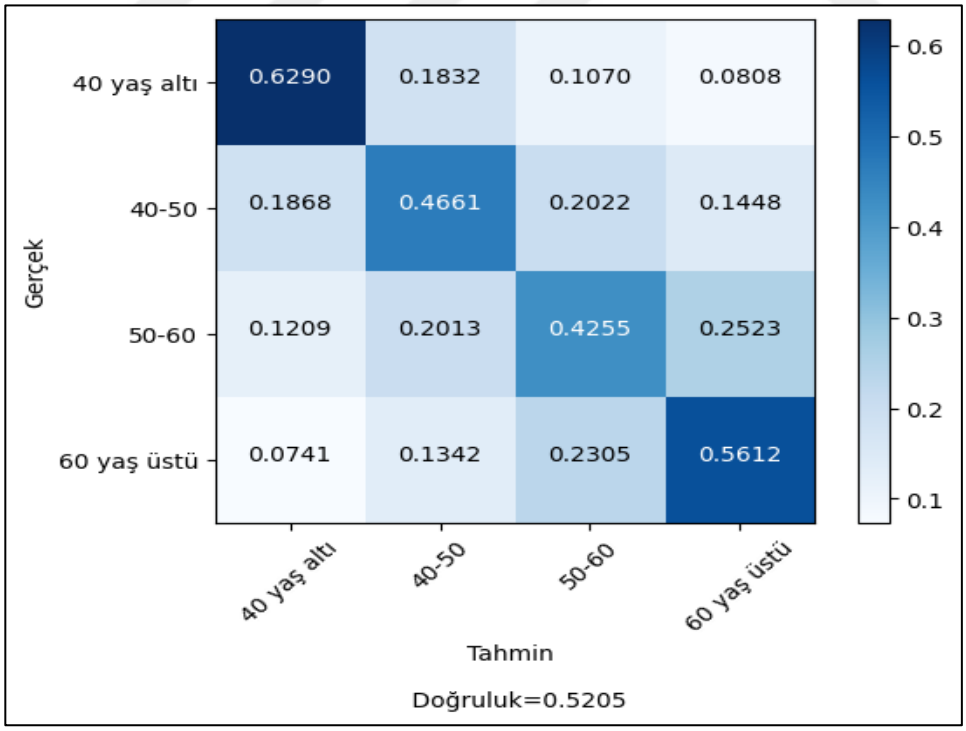
Şekil 4.15. TF-IDF(kelime_karakter)_LR ile cinsiyet sınıflandırması için hata matrisi

Şekil 4.16. Cinsiyet sınıflandırması için en iyi terimler

4.2.2. Yaş tahmini

Yaş demografik özelliğinde, yaş aralığı sınıfları doğal olarak bir sıra halinde sıralanır. Şekil 4.17'de hata matrisi yaş aralıklarının sırasını yansıtmaktadır. Birbirine yakın yaş aralıkları beklendiği gibi önemli oranlarda bir araya gelmektedir ve bu durum bir sınıflandırma görevinin doğrulanması olarak düşünülebilir. 40 yaş altı grubunun tahmin edilmesi daha kolay iken (%62,91), 50-60 (%42,55) en zor olanıdır. Uçlardaki yaş grupları, 40 yaş altı (%62,91) ve 60 yaş üstü (%56,12), iç yaş grupları olan 40-50 (%46,61) ve 50-60'a (%42,55) göre daha tahmin edilebilirdir. Hata matrisi, genç ve yaşlı milletvekillerinin konuşmalarının orta yaş gruplarındakilere göre daha ayırt edici olduğunu göstermektedir.

Şekil 4.18'de yaş kategorilerine göre özelliklerin dağılımında herhangi bir örüntü bulanamamıştır. Ancak tüm kategoriler ve özellikler dikkate alındığında en genç kategoriye ait terimlerin özellik öneminin daha yüksek olduğu gözlemlenmiştir. Bu durum, 40 yaş altı kategorisinin konuşma içeriğinin yaş durumunu daha fazla yansıttığı şeklinde yorumlanabilir.



Şekil 4.17. TF-IDF(kelime)_LR ile yaş sınıflandırması için hata matrisi

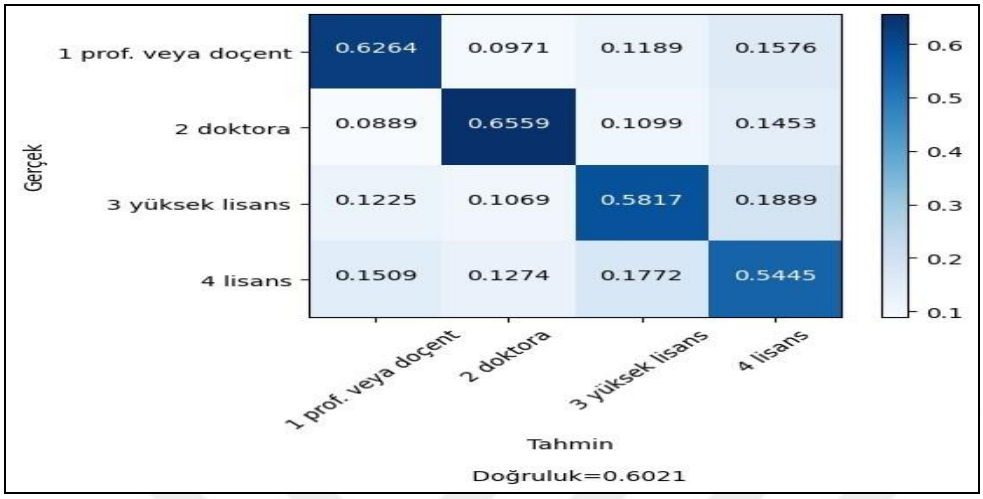
Şekil 4.18. Yaş sınıflandırması için en iyi terimler

4.2.3. Eğitim tahmini

Eğitim durumları birbirini takip ettiğinden, Şekil 4.19'daki hata matrisinde birbirine yakın eğitim durumlarının birbirine yakınsaması beklenebilir. Ancak bu durumla ilgili yaş sınıflandırmasında olduğu gibi belirgin bir durum gözlemlenmiştir. Örneğin, "doçent veya profesör", "doktora" (%09,71) ve "yüksek lisans "a (%11,89) kıyasla "lisans "a (%15,76) daha yakındır. "Profesörler veya Doçentler" lisans öğrencilerine ders vermek için yüksek lisans veya doktora öğrencilerinden çok daha fazla zaman harcamaktadır. "Doçent veya profesör" ile "lisans" arasındaki yakınlığın nedeni bu olabilir. “Doktora” ve “doçent veya profesör” %65,59 ve %62,65 doğruluk oranlarıyla kolayca tespit edilebilmektedir. Buna karşılık, “lisans” %54,45'lik bir oranla tanımlanması zor bir terimdir.

Yükseköğretimle ilgili terimler Şekil 4.20'de “Doçent veya Profesör” kategorisinde ortaya çıkmıştır. Öğretim, doçent, öğretim görevlisi, yükseköğretim, fakülte, öğretim elemanı, kurs, üniversite, vakıf üniversitesi, rektör, doktora, kurs, doktora, okul eğitimi ve eğitim direktörü “Doçent veya Prof.” kategorisinin en iyi özellikleri arasındadır. Bun kategori, veri kümesindeki en yüksek eğitim statüsünü temsil etmektedir. Bu duruma uygun olarak, yüksek

eğitim kurumları ve konuları çoğunlukta olmuştur. “Doktora”, “yüksek lisans” ve “lisans” kategorilerinin terimleri belirli bir konuya odaklanmamıştır.



Şekil 4.19. TF-IDF(kelime)_LR ile eğitim sınıflandırması için hata matrisi

Şekil 4.20. Eğitim sınıflandırması için en iyi terimler

4.2.4. Meslek tahmini

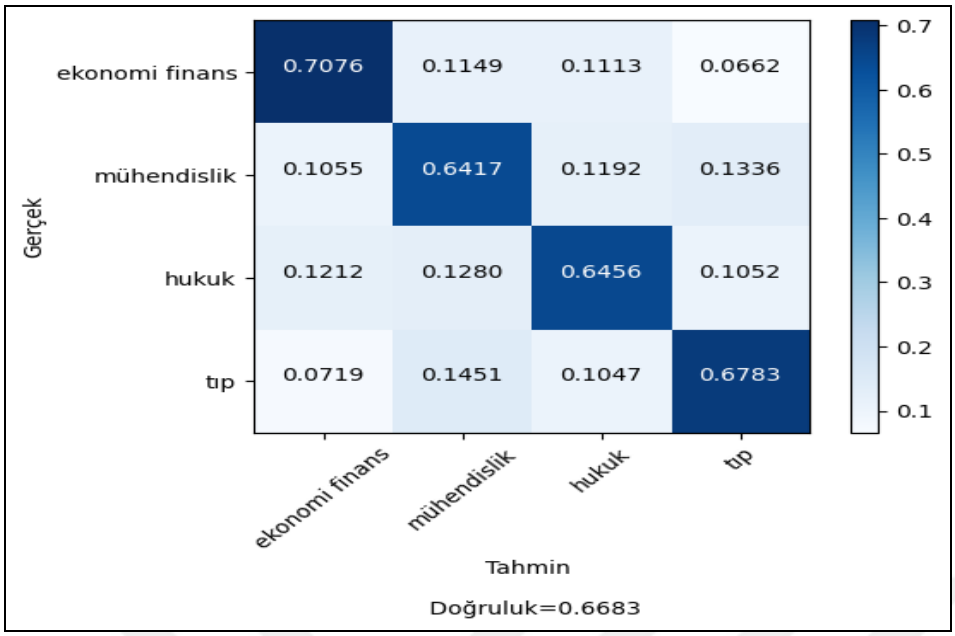
Şekil 4.21 meslek sınıflandırması için hata matrisini göstermektedir. Ekonomi ve finans (%70,76) tanımlanması en kolay meslekler olarak görünmektedir. Buna karşılık, mühendislik (%64,16) ve hukuk (%64,56) en zor örneklerdir. Sonuçlara göre, en yüksek hata değeri (%14,51) tıptan mühendisliğe doğrudur. Bu da bu iki mesleğin TBMM'de daha yakın olduğunu göstermektedir. Her iki meslek kategorisinin de pozitif bilimlere yatkın olması bu yakınlığın bir nedeni olabilir. En düşük hata oranının (%6,62) ekonomi ve maliyeden tıbba doğru olması, bu iki kategorinin kendi aralarında çok ayırt edilebilir olduğunu göstermektedir.

Şekil 4.22'de mesleki sınıflandırmada en çok öne çıkan terimler incelendiğinde, milletvekillerinin konuşmalarında kendi mesleki alanlarını yansıttıkları görülmektedir. Ekonomi, maliye, vergi, finans ve bütçeyi içeren terimler ekonomi ve finansın içeriğini diğer kategorilerden kesin bir şekilde ayırmaktadır. Kategorinin en etkili özellikleri büyüme, banka, az gelişmişlik, gelir, finans, enflasyon ve vergidir.

Mühendislik dalları, mühendislik kategorisinin terimlerini belirlemiştir. Orman mühendisliği (orman, orman yangını, orman teşkilatı, orman bölgesi, köy ormanı), ziraat mühendisliği (patates, dekar, dekar arazi, tohum, pamuk barajı, dünya çiftçisi, mandalina, litre mazot ve kır çiftçisi) ve inşaat mühendisliği (inşaat mühendisi, yapı denetim) en çok öne çıkan terimler oldu. Mega watt, jeoloji ve gıda gibi diğer mühendislik dallarıyla ilgili terimler de ortaya çıkmıştır. Tüm mühendislik dallarını ilgilendiren tek bir terim (meslek odası) bulunmaktadır.

Yargı süreci, yargı kurumları, mahkemeler, ceza hukuku, avukatlık mesleği ve barolarla ilgili terimler hukuk kategorisinin önde gelen özelliklerini oluşturmaktadır. Yargı, hüküm, yargılama, suç, ceza, infaz, avukat ve baro hukukun öne çıkan terimleridir.

Sağlık hizmetleri, sağlıkta dönüşüm, sağlık alanı çalışanları, sağlık çalışanlarının maruz kaldığı şiddet gibi konular tıp kategorisinin öne çıkan terimleridir. Doktor, hekim, hasta, sağlık, ambulans, medikal, ilaç, kanser ve koruyucu sağlık ise tıbbın özellikleridir.



Şekil 4.21. TF-IDF(kelime_karakter)_LR ile meslek sınıflandırması için hata matrisi

Şekil 4.22. Meslek sınıflandırması için en iyi terimler

4.2.5. Seçim bölgesi tahmini

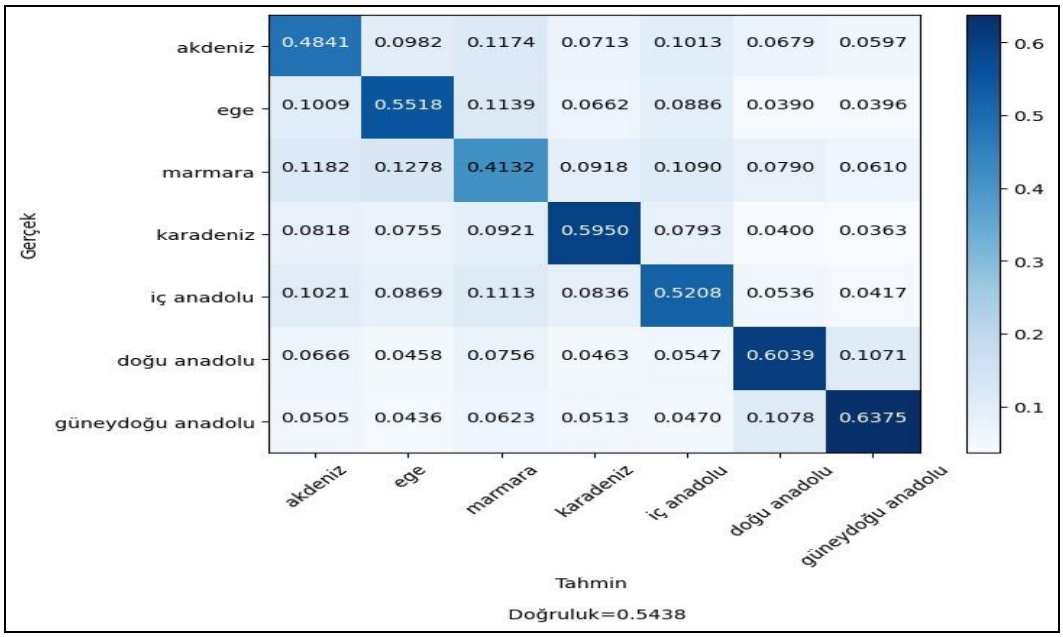
1941'de düzenlenen ilk Coğrafya Kongresi'nde iller konum, iklim, topografya, nüfus, tarımsal çeşitlilik ve benzeri özelliklerine göre yedi coğrafi bölgeye ayrılmıştır (Ertek, 2011). Çalışmada seçim bölgesi, bir milletvekilinin temsilcisi olduğu ilin coğrafi bölgesini tanımlamaktadır. Sınıflandırma görevinde kategorileri elde etmek için milletvekillerinin seksen bir seçim ilini yedi seçim coğrafi bölgesine dönüştürülmüştür. Bu sınıflar Ege, Karadeniz, İç Anadolu, Doğu Anadolu, Marmara, Akdeniz ve Güneydoğu Anadolu'dur.

Şekil 4.23'deki hata matrisine göre, Doğu Anadolu (%63,70), Güneydoğu Anadolu (%60,40) ve Karadeniz (%59,50) tanımlanması en az zor olan bölgelerken, Marmara (%41,32) ve Akdeniz (%48,41) en zor bölgelerdir. En yüksek hata oranı Marmara'dan Ege'ye (%12,78), en düşük hata oranı ise %3,63 ile Karadeniz'den Doğu Anadolu'ya aittir. Bir bölgeden diğer bölgelere yapılan hatalar incelendiğinde, bölgeler arasında bir ilişki tespit edilebilmektedir. Ege, Marmara için en yüksek hataya sahipken (%11,39), Marmara da Ege için en yüksek hataya sahiptir (%12,78). Aynı durum Doğu Anadolu ve Güneydoğu Anadolu için de geçerlidir. Doğu Anadolu için en yüksek hata %10,78 ile Güneydoğu Anadolu'da, Doğu Anadolu için en yüksek hata ise Güneydoğu Anadolu'dadır (%10,71). Bu hatalar, Ege ile Marmara ve Doğu Anadolu ile Güneydoğu Anadolu arasındaki coğrafi, kültürel veya ekonomik yakınlığın sonuçlarını gösterir. Karadeniz, Doğu Anadolu ile sınırdaş olmasına rağmen, Şekil 4.23'de en uzak bölge çiftlerinden biri olarak ortaya çıkmaktadır

Seçim bölgesi sınıflandırmasında bölgelerin ekonomik yapılarına ait ürünler, özellikle de tarım ürünleri en önemli özellikler arasında yer almaktadır. Şekil 4.24'de bölgesel konuşma tarzından ziyade bölgesel konularla ilgili özellikler öne çıkmaktadır.

Akdeniz Bölgesi için narenciye, pamuk, sebze, meyve, domates, tarım, çiftçilik ve çiftçi en önemli özelliklerdir. Ülkenin turizm bölgesi olarak turizm de en önemli özellikler arasındadır. Zeytin, Ege Bölgesi'nin ana tarım ürünüdür ve her iki bölgede de en önemli özellikler arasındadır. Üzüm de en etkili özelliklerden biridir. Jeotermal, bölgenin coğrafyası gereği zengin sıcak yeraltı su kaynakları nedeniyle bölge için tanımlayıcı bir terimdir ve en önemli özellikler arasındadır. Ülkenin en büyük metropol kenti İstanbul'un Marmara Bölgesi'nde olması nedeniyle kent, konut, kentsel dönüşüm, imar, emlak gibi kent yaşamına ilişkin terimler en önemli özellikler arasında yer alır. Bölgenin yaygın rüzgârı olan lodos da

bölge için önemli bir özelliktir. Karadeniz Bölgesi için çay, fındık ve kömür en önemli özellikler arasındadır. Doğu Anadolu Bölgesi'nde Doğu Ekspresi, dadaş (bölgenin Erzurum ilinin halkına verilen isim) ve kayısı en güçlü terimler arasındadır. Güneydoğu Anadolu'da ise GAP barajı, Antep fıstığı ve Dicle nehri en önemli özellikler arasında yer almaktadır. Güneydoğu ve Doğu Anadolu bölgelerinde bölgesel sorun terördür. Cezaevi, vahşet ve IŞİD gibi güvenlik ve terörle ilgili özellikler bölgeler için öne çıkan terimlerdir.



Şekil 4.23. TF-IDF(kelime)_LR ile seçim bölgesi sınıflandırması için hata matrisi

Şekil 4.24. Seçim bölgesi sınıflandırması için en iyi terimler

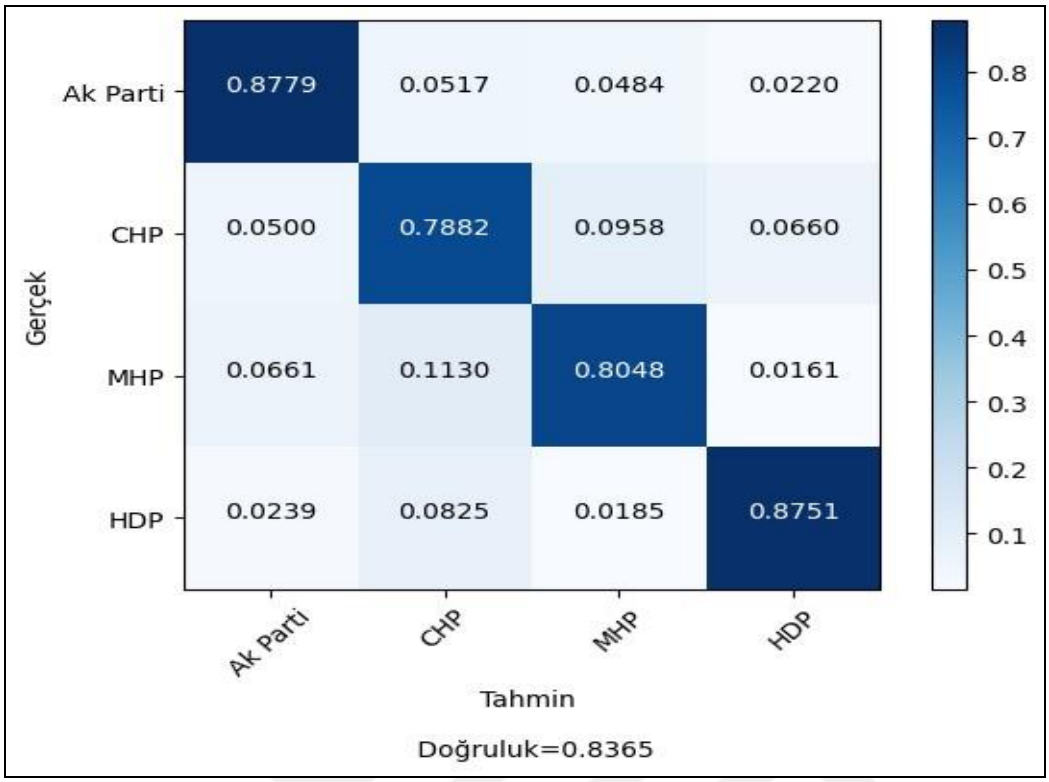
4.2.6. Parti aidiyeti tahmini

Konuşmalar, 2011- 2020 arası mecliste bulunan dört partinin milletvekillerine aittir. AK Parti 2002 yılından bu yana Türkiye'de iktidar partisidir. Muhafazakâr-sağ bir parti olarak tanımlanabilir. Cumhuriyet Halk Partisi (CHP) bu dönemde ana muhalefet partisidir. Türkiye'nin sosyal-demokrat sol partisidir. Milliyetçi Hareket Partisi (MHP) ülkede sağ eğilimli milliyetçi bir partidir. Halkların Demokratik Partisi (HDP) ülkenin sol eğilimli Kürt yanlısı bir partisidir. HDP, Türkiye'nin Güneydoğu ve Doğu Anadolu bölgelerini konu alan dar kapsamlı bir parti olarak düşünülebilir ve Türk milliyetçisi MHP ile kutuplaşmaktadır.

Şekil 4.25'de MHP'den HDP'ye hata (%1,61) ve HDP'den MHP'ye hata incelendiğinde, bu partilerin milletvekillerinin birbirlerine en uzak partiler olduğu görülmektedir. Türkiye'nin gerçek siyasi hayatı da aralarındaki kutuplaşma nedeniyle bu durumu doğrulamaktadır. AK Parti (%87,79) ve HDP (%87,51) konuşmalardan en çok tahmin edilebilen, CHP (%78,82) ise en az tahmin edilebilen partilerdir.

Şekil 4.26'daki parti üyeliği görevinde en iyi özellikleri incelendiğinde, partilerin ideolojik kökenleri ve statüleriyle ilgili özellikler bulunmuştur. Parti sınıflandırmasında, iktidardaki AK Parti, bakanımız, başbakanımız, cumhurbaşkanımız ve hükümetimiz gibi bakan, başbakan veya cumhurbaşkanı ile birinci çoğul şahıs için iyelik eki olan ayırt edici kelimelere sahiptir. Bu sözcükler aynı zamanda bakan, başbakan, cumhurbaşkanı ve hükümet gibi iyelik eki almayan muhalefet partileri için de ayırt edici sözcüklerdir.

Ana muhalefet partisi CHP, laikliği vurgulamaktadır ve Türkiye Cumhuriyeti'nin kurucusu Mustafa Kemal Atatürk tarafından kurulmuştur. Laiklik ve Atatürk ayırt edici özellikler arasındadır. Partinin sosyal demokrat sol kökenleri, ülkede daha çok solcular tarafından kullanılan emekçi, yurttaş ve grev gibi özelliklerinde de mevcuttur. MHP'nin en iyi özellikleri milliyetçi ideolojisini yansıtmasıdır. Ülkü, Türkiye'deki milliyetçilerin hedefini ima eden özel bir anlama sahiptir. Ülkücü, Türk Milleti, büyük Türk Milleti, ülkücülerin efsanevi liderine atıfta bulunan Başbuğ, şehit ve gazi MHP için en önemli özelliklerdir. Kürtçe, savaş, esir, barış, kadın ve eğitimde anadil HDP'yi tanımlayan temel özelliklerdir. Yurttaş ve emekçi gibi sol terimler de partinin partinin ideolojisini yansıtan en iyi özellikler arasında çıkmıştır.



Şekil 4.25. TF-IDF(kelime)_DVM ile parti aidiyeti sınıflandırması için hata matrisi

Şekil 4.26. Parti aidiyeti sınıflandırması için en iyi terimler

4.2.7. Parti durumu tahmini

Şekil 4.27, iktidarın (%91,98) muhalefete (%90,13) kıyasla daha öngörülebilir olduğunu göstermektedir. 92'lik doğruluk oranı, milletvekillerinin parti statülerinin Türk parlamentosundaki konuşmalarına göre yüksek oranda ayırt edilebilir olduğunu göstermektedir.

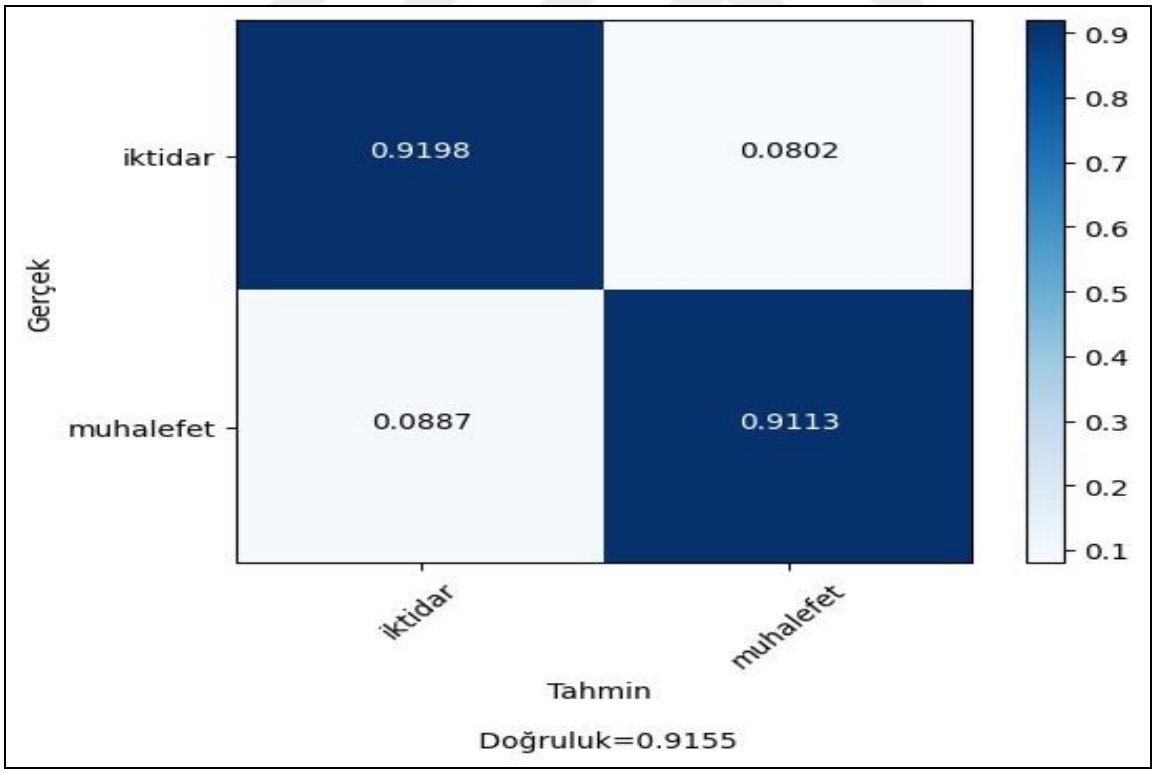
Şekil 4.28'de parti statüsüne göre konuların değişimin araştırmak için milletvekillerinin konuşmalarının en etkili özelliklerine göre hükümet ve muhalefet partileri analiz edilmiştir. Parti üyeliği sınıflandırmasındaki aynı örüntü parti statüsü üyeliğinde de mevcuttur. İktidar, bakanımız, başbakanımız veya cumhurbaşkanımız gibi birinci çoğul şahıs için iyelik eki alan ayırt edici kelimelere sahipken, muhalefet bu terimlere iyelik eki almadan sahiptir.

İktidar milletvekillerinin konuşmalarındaki en etkili özellikler; millet, terör, terör örgütü, medeniyet, birlikte, hizmet, güçlü, şehit, vatan millet, dünya, takdir, dost, sayın cumhurbaşkanı, gayret, gönül, tesis, destek, proje, duygu, başarı, eleştiri, birlik ve beraberliktir, imkân, suçlama, temenni, sonuç, soy, gurur, katkı, milli irade, millet otoritesi, fetih, kahraman, istikamet, masum, reform, şükran, hedef, üslup, ihraç, hain, karar, gelecek, istiklal, lider, etkili olarak çıkmıştır. Muhalefet için ise en sol terimlerden biri olan yurttaş muhalefet milletvekillerinin en önemli terimidir. İşçi, çiftçi, esnaf ve köylü de muhalefet partilerinin önem verdiği kelimeler arasındadır. Muhalefet partilerinin daha çok ekonomik sorunlara vurgu yaptığı görülmektedir. Para, vergi, taşeron, fatura, grev, sendika, ücret, fabrika, mazot, ithalat, geçim ekonomik terimlerdir. İhracat iktidarın, ithalat ise muhalefetin terimleri arasında yer almaktadır. TBMM'de ihracatın ekonomik başarı ya da bağımsızlık, ithalatın ise ekonomik başarısızlık göstergesi olarak kabul edildiği sonucu çıkarılabilir. Muhalefette iktidar milletvekillerine kıyasla çok daha fazla olumsuz anlamlı terim bulunmaktadır. Bunlardan bazıları işsiz, yoksul, zam, kira ve açlık gibi ekonomi ile ilgilidir. Muhalefetin olumsuz dil kullanımı diğer terimlerde de görülmektedir. Sorun, yandaş, mağduriyet, cinayet, yolsuzluk, rüşvet, baskı, sefil, intihar, yasak ve vahim de muhalefetin olumsuz anlamlı terimleridir.

İktidar ve ana muhalefet partisi değişmediği için ideolojinin dili bulgulara yansımış görünmektedir. İktidar partisi sağ muhafazakâr AK Parti, ana muhalefet partisi ise merkez sol, sosyal demokrat CHP'dir. Hükümet milletvekilleri çoğunlukla eşanlamlı kelimelerden

daha geleneksel, muhafazakâr kelimeleri tercih etmişlerdir. Çalışmada kullanılan terim çiftleri, (tercih edilen muhafazakâr terim-yaygın olarak kullanılan eşanlamlı terim) olarak gösterilirse, iktidar milletvekillerinin en etkili özellikleri (medeniyet-uygarlık), (mana-anlam), (gönül-kalp), (imkân-olanak) gibi geleneksel ya da muhafazakâr kelimelerdir. Muhalefet milletvekilleri ise eşanlamlı kelimelerden daha az muhafazakâr olanları tercih etmişlerdir. Muhalefet milletvekilleri (cevap-yanıt), (imkân-olanak) terimleri dikkate alındığında yanıt ve olanak terimlerini cevap ve imkân terimlerine tercih etmişlerdir.

Terör, terör örgütü ve şehit terimleri, TBMM'de hükümet için en öne çıkan konulardan birinin terör olduğunu göstermektedir. İktidar milletvekillerinin konuşmalarında dost, beraber, birlik ve beraberlik, güçlü, kahraman, yürek, başarı, reform, etkili gibi olumlu kelimeler çoğunluktadır. Ancak diğer yandan en olumsuz kavramlardan biri olan vatan haini bu durumun istisnası olarak iktidar milletvekilleri için en etkili özellikler arasında yer almaktadır.



Şekil 4.27. TF-IDF(kelime)_DVM ile parti durumu sınıflandırması için hata matrisi

Şekil 4.28. Parti durumu sınıflandırması için en iyi terimler

4.2.8. Milletvekillerin demografik özelliklerinin ikili analizi

Bir demografik özelliğin diğer bir demografik özellik üzerindeki etkisini dört tahmin görevi aracılığıyla incelenmiştir. İlk iki ikili analizde her yaş grubu içinde cinsiyet ve eğitim durumu sınıflandırması yapılmıştır. Üçüncü ikili analizde ise her bir siyasi parti içindeki milletvekillerinin cinsiyetini tahmin edilmiştir. Dördüncü ve son ikili analizde, ikili analiz için her bir eğitim durumu grubu içinde parti üyeliğini sınıflandırılmıştır.

Sınıflandırma görevlerinde Bölüm 3.2.3'deki deneysel kurulum kullanılmıştır. Sınıflandırmalar unigram, bigram ve trigram TF-IDF özellikleri ile Lojistik Regresyon algoritması ile yapılmıştır. Çizelge 4-7 metin sınıflandırması görevlerinin doğruluklarını göstermektedir.

Çizelge 4.3-4.6'daki konuşma sayıları, konuşmaları yapan milletvekili sayıları ve sınıflandırma başarım değerleri arasında herhangi bir ilişki bulunamamıştır. Bu durum tahminin konuşma içeriklerinden kaynaklandığı ve sınıflandırmada konuşma sayıları, konuşmaları yapan milletvekili sayılarının bir önyargı (bayes) oluşturmadığını gösterir.

Çizelge 4.3. Yaş gruplarının meslek sınıflandırması üzerindeki etkisi

Yaş Grubu	Milletvekili Sayısı	Konuşma Sayısı	Meslek Tahmini
40 yaş altı	89	1560	0,79
40-50	258	4800	0,69
50-60	310	11560	0,71
60 yaş üstü	178	5040	0,70

Çizelge 4.3'e göre, 40 yaş altı 40-50 yaş grubundan %10, 60 yaş üstünden %9 ve 50-60 yaş grubundan %8 daha yüksektir. Bu sonuçlar, en genç yaş grubunun daha büyük yaş gruplarına göre mesleklerini daha fazla yansıttığını göstermektedir. Buna ek olarak, 50-60, 60 yaş üstü ve 40-50 yaş gruplarının konuşma içerikleri %71, %70 ve %69 oranlarında sınıflandırılmıştır. Bu yaş grupları arasındaki fark 50-60 grubu kadar önemli olmamıştır.

Çizelge 4.4. Eğitim durumunun parti üyeliği sınıflandırması üzerindeki etkisi

Eğitim	Milletvekili Sayısı	Konuşma Sayısı	Parti Tahmini
Profesör ya da Doçent	102	3996	0,87
Doktora	71	464	0,72
Yüksek Lisans	167	6464	0,89
Lisans	503	18452	0,85

Çizelge 4.4 her bir meslek grubundaki parti üyeliği tahminlerini göstermektedir. Çizelgeye göre, parti üyeliği yüksek lisans mezunları (%89) için en kesin üyeliktir. Doçent veya profesör (%87) milletvekilleri de parti üyeliklerini güçlü bir şekilde göstermektedir. Doktoralı (%72) milletvekillerinin parti üyeliğini tahmin etmek diğerlerine kıyasla kolay değildir.

Çizelge 4.5. Yaşın cinsiyet sınıflaması üzerindeki etkisi

Yaş Grubu	Milletvekili Sayısı	Konuşma Sayısı	Cinsiyet Tahmini
40 yaş altı	139	3160	0,85
40-50	420	6444	0,81
50-60	470	6964	0,82
60 yaş üstü	171	1302	0,84

Çizelge 4.5'de cinsiyet özelliği tahmininde 40 yaş altı %85, 60 yaş üstü %84, 50-60 %82 ve 40-50 %81 doğruluk oranına sahiptir. En yaşlı ve en genç yaş grupları, orta yaş gruplarına göre cinsiyet özelliklerini daha fazla yansıtmaktadır.

Çizelge 4.6. Parti aidiyetinin cinsiyet sınıflandırması üzerindeki etkisi

Siyasi Parti	Milletvekili Sayısı	Konuşma Sayısı	Cinsiyet Tahmini
AK Parti	484	4054	0,80
CHP	276	6176	0,78
MHP	85	1124	0,90
HDP	118	6198	0,89

Cinsiyet, Çizelge 4.6'da parti aidiyeti göz önüne alındığında MHP ve HDP için daha öngörülebilirdir. Türk milliyetçisi, sağcı MHP ve Kürt yanlısı, solcu HDP en çok çatışma yaşayan iki partidir. Öte yandan, AK Parti ve CHP hem sağ hem de sol kanadın daha merkez partileridir. MHP (%90) ve HDP (%89) cinsiyet demografik özelliğinde AK Parti (%80) ve CHP'den (%78) daha yüksek doğruluk oranına sahiptir. Çalışmanın bu bulguları, merkeze yakın partilerde cinsiyetin TBMM'ye göre daha az anlamlı olduğunu göstermektedir.

4.2.9. Demografik ve siyasi özelliklerin genel değerlendirmesi

Çalışmada, milletvekillerinin yedi özelliği için özellik tahmin görevleri ile TBMM Genel Kurul görüşmeleri analiz edilmiştir. Derin öğrenme yöntemleri, yeni ve trend yaklaşımlar oldukları için metin sınıflandırma için en iyi performans gösteren yöntemler olarak düşünülebilir. Ancak, özellik tahmini görevlerinde, n-gram TF-IDF doküman temsilini kullanan DVM ve LR, derin öğrenme yöntemleri olan BoW_ İBSA ve BERT'ten daha yüksek doğruluk vermiştir. Mevcut çalışmada, BERT'in parti statüsü üyeliğindeki başarısı bu sonuca bir istisna olarak kabul edilebilir. BERT, parti statüsü eğiliminde %91 doğruluk oranına sahiptir.

Çalışmada klasik makine öğrenmesi yöntemleri daha başarılı olsa da klasik makine öğrenmesi ve derin öğrenme yöntemleri karşılaştırıldığında başarım değişken olabilmektedir. PAN 2019'da İngilizce ve İspanyolca Twitter verileri üzerinde cinsiyet profillemesi araştırılmıştır. İngilizcede en iyi sonuç (%84,32) n-gramlar ve Lojistik Regresyon ile elde edilmiştir (Valencia, Adorno, Rhodes ve Pineda, 2019). İspanyolcada,

Pizarro (2019) kelime n-gramları, karakter n-gramları ve DVM ile en iyi sonucu (%81,72) elde etmiştir. PAN 2019'da klasik makine öğrenmesi algoritmaları arasında en çok kullanılan DVM olmuştur. Çok az katılımcı ESA, TSA ve İBSA gibi derin öğrenme tekniklerini kullanmıştır. Sonuçlara göre, klasik makine öğrenmesi yaklaşımları derin öğrenme yaklaşımlarından daha yüksek doğruluk elde etmiştir (Rangel ve Rosso, 2019). PAN 2018'in genel bakışına göre (Rangel ve diğerleri, 2018), paylaşılan görevlerden biri cinsiyet tanımlamadır. Cinsiyet tahmini metin tabanlı bir yaklaşıma, görüntü tabanlı bir yaklaşıma ve her iki yaklaşımın bir kombinasyonuna bağlıdır. Yalnızca metin özellikleriyle ilgili olarak katılımcılar klasik makine öğrenimi ve derin öğrenme yaklaşımlarını kullanmışlardır. PAN 2018'in metin tabanlı alt görevinde en iyi sonuçlar, farklı n-gram türleri ile DVM ve Lojistik Regresyon gibi klasik makine öğrenimi algoritmalarının kombinasyonlarıyla elde edilmiştir. Veenhoven, Snijders, van der Hall ve van Noord (2018) derin öğrenme yaklaşımlarının başarısını göz önünde bulunduran tek modeli, İspanyolca için ikinci en yüksek doğruluk değeri %80,36 ile önceden eğitilmiş kelime yerleştirme ile bi-LSTM 'e aittir. PAN 2017'de Rangel ve diğerleri (2017), paylaşılan görevler arasında dört farklı dilde toplanan Twitter yazarlarından cinsiyet ve dil çeşitliliğinin belirlenmesi yer almaktadır (Arapça, İngilizce, Portekizce ve İspanyolca). Cinsiyet belirlemede Arapça, İngilizce ve İspanyolca için en başarılı yaklaşımlar klasik makine öğrenmesi modellerine dayanırken, derin öğrenme teknikleri Portekizcede en yüksek sonuca sahiptir. Miura, M. Taniguchi, T. Taniguchi ve Ohkuma (2017) kelime ve karakter gömme ile ESA, TSA, dikkat mekanizması, maksimum havuzlama (pooling) katmanı ve tam bağlı katman kullanırken Portekizcede en iyi sonucu elde etmiştir. PAN 2021'de Rangel, De la Peña Sarracén, Chulvi, Fersini ve Rosso (2021), katılımcılar Twitter kullanıcısının nefret söylemi yayıp yaymadığını tespit etmişlerdir. En başarılı modeller derin öğrenme tekniklerine aittir. İngilizcede en iyi sonucu (%75) Dukic ve Krzic (2021) BERT ile İspanyolcada en iyi sonucu (%85) Siino, Di Nuovo, Tinnirello ve La Cascia (2021) ESA ile elde etmişlerdir.

LDSE'nin başarısı özellik tahmin görevine göre değişkenlik gösterebilir. LDSE, dil çeşitliliği tanımlamada diğer tüm yaklaşımlardan daha iyi performans göstermiştir, ancak cinsiyet tanımlamada dil çeşitliliği kadar umut verici değildir (Rangel ve diğerleri, 2017). LDSE_LR modeli ile özellik tahmini görevleri için de değişken sonuçlar elde edilmiştir. Cinsiyet belirleme ve parti statüsü belirleme için çok umut vericidir, ancak diğer görevler için TF-IDF_DVM ve TF-IDF_LR ile uyumlu değildir.

Bu çalışmanın bulguları bazı kısıtlamalar ışığında değerlendirilmelidir. En az doküman sayısı sınıfına örnekleme (under-sampling) önemli veri kaybına yol açmaktadır. Örneğin, meslek demografik özelliğinde 45.556 doküman 27.492'ye düşürülmüştür. Eksik örnekleme veri kaybına yol açsa dahi, tüm kategorilerin en iyi özellik analizine ve hata analizine eşit katılımını sağlamak için bu yöntem uygulanmıştır.

Veri kısıtlılığı nedeniyle bazı kategorileri veri kümesinden çıkarılmıştır. Örneğin, ön lisans ve lise kategorileri eğitim durumu sınıflandırmasına dâhil edilmemiştir. Aynı nedenle meslek demografik özelliğinde de eczacı, öğretmen, siyaset bilimci ve diğer meslekleri kullanılmamıştır. Parti üyeliği sınıflandırmasına, derlemin kapsadığı dönemde TBMM'de bulunmadığı için İYİ Parti analizlere dâhil edilmemiştir.

Bir YPO görevinde analize dâhil olan yazar sayısı veya bir yazarın diğer yazarlara kıyasla çok fazla dokümanda yer alması, YPO görevini Yazarlık Atfetme (Authorship Attribution, AA) görevine dönüştürebilir. Parlamenter alanda, siyasi parti sözcüleri çok sayıda söz almaktadır. Örneğin, meslek demografik özelliğinde, tıp alanındaki bir milletvekilinin 1102 konuşması bulunurken, kategoride yalnızca bir konuşması olan milletvekilleri de vardır. Eksik örnekleme sürecinde, sınırlamanın etkisini azaltmak için bu durum dikkate alınmıştır. Yerine koymadan örnekleme (sampling-without-replacement) kullanarak, azınlık sınıfının doküman sayısına ulaşana kadar her milletvekili için bir konuşmayı öz yinelemeli olarak veri kümesine eklenmiştir. Böylece mümkün olduğunca çok milletvekilinin konuşmaları alt veri kümelerine dâhil edilmiştir.

Blaxill ve Beelen (2016) Westminster'da kadınların temsilini araştırmıştır. Ayrıca erkek ve kadın milletvekillerinin konuşmalarında bu terimin önemini incelemişlerdir. Kadın milletvekillerinin kullandıkları dil karşılaştırıldığında, Westminster ve Türkiye Büyük Millet Meclisi'nde sağlık terimlerinin eğitim terimlerinden çok daha ön planda olduğu görülmüştür. Ayrıca her iki parlamentodaki kadın milletvekillerinin de eğitim ve sağlık alanını vurgulamışlardır. Cinsiyete dayalı terimler, aile, çocuk ve bakım alanları her iki parlamento için de büyük önem taşımaktadır. Bu çalışmanın sonuçlarına göre, kadın cinayetleri Türkiye'de önemli bir sorundur. Kadın milletvekilleri yoğun olarak TBMM kadın hakları ve çocuk haklarından bahsetmişlerdir. Westminster'da kadın cinayetleri kavramı öne çıkmamıştır. Her iki parlamentoda da erkek milletvekillerinin ilgisi benzerlik göstermektedir. Her iki parlamentoda da nükleer, faiz, kuvvet, şirket ve toprak gibi ortak

terimler bulunmaktadır. Ekonomi, finans, dış politika ve ordu, TBMM ve Westminster'da erkek milletvekillerini kadınlardan ayıran alanlardır. Tarımla ilgili terimler TBMM'de Westminster'a göre daha fazla önem taşımaktadır.

Naderi ve Hirst (2018) parlamento oturumlarının sözlü soru dönemini analiz ederek Kanada parlamentosundaki itibar kurtarma dilini tespit etmiştir. Muhalefet üyeleri tarafından sorulan soruları itibar tehdidi, hükümet milletvekilleri tarafından sorulan soruları ise tehdit içermeyen dostane sorular olarak kabul etmişlerdir. Analizlerinde Dilbilimsel Sorgulama ve Kelime Sayımı (Linguistic Inquiry and Word Count, LIWC)(Tausczik ve Pennebaker, 2010) kullanmışlardır. İtibar tehditlerinde (muhalefet) öfke ve olumsuz duyguların, itibar dışı tehditlerde (hükümet) ise olumlu duyguların daha fazla kullanıldığını bulmuşlardır. Çalışmada TBMM'deki en iyi özellik analizinden benzer bulguları elde edilmiştir. Bu sonuçlar parlamentolarda zaten beklenen ve çalışmada elde edilen bulgularla uyumlu sonuçlardır. Güçlü, gururlu ve başarılı olmak TBMM'de iktidar partisinin öne çıkan özellikleridir ve başarıyı ima etmektedir. İşsizlik, açlık, yoksulluk ise muhalefetin en önemli özellikleri arasındadır. Bu sonuca göre mevcut çalışmanın bulguları Naderi ve Hirst'in çalışmasıyla uyumludur.

Dahllöf (2012) İsveç parlamentosunda yaş ve parti üyeliğinin cinsiyet üzerindeki etkisini araştırmıştır. Cinsiyet, yaşlı grupta genç gruba göre daha öngörülebilirdir. Ayrıca, sağ görüşlü milletvekillerinin cinsiyetini belirlemenin sol görüşlü milletvekillerine göre daha kolay olduğunu bulmuştur. Daha yaşlı grupların ve sol görüşlü politikacıların cinsiyete daha eşit kelime kullanımına sahip olduğunu öne sürmüştür. Dahllöf'ün çalışmasında olduğu gibi, daha düşük doğruluk oranının milletvekillerinin daha cinsiyet eşitlikçi bir dil kullandığını gösterebileceği ifade edilebilir. Öte yandan, farklı bir bakış açısıyla ifade edildiğinde baskın kişilik özelliklerinden biri olan cinsiyetin konuşmalarında yeterince ortaya çıkmadığı şeklinde de yorumlanabilir. Kanada parlamentosunun aksine, TBMM'deki en genç yaş grubu için cinsiyet daha tahmin edilebilir çıkmıştır. Parti üyeliği ele alındığında daha merkezdeki partiler (AK Parti ve CHP) TBMM'de daha düşük cinsiyet öngörülebilirliğine sahipken, bu durum Kanada parlamentosundaki sol kanat grubu için geçerlidir.

4.3. TBMM Genel Kurul Görüşmelerinde Yakın Anlamlı Kavramlar Bulguları

TBMM Genel Kurul tutanaklarından yakın anlamlı kavramlar ve analojilerin araştırıldığı çalışmada kelime benzerliği ve kelime analojilerinden yararlanılmıştır. Çalışmada kelime benzerliği ve kelime analojileri beş farklı model ile ölçülmüştür. Bu modeller GloVe, word2vec CBOW, word2vec skip-gram, fastText CBOW ve fastText skip-gram modelleridir.

Kelime yerleştirme algoritmaları eğitilirken bağlamın uzunluğu yani pencere boyu 5, elde edilen kelime vektörlerinin boyutu ise 300 olarak seçilmiştir. 2 veya 3 gibi daha küçük bir pencere boyutu, daha fazla yerel kelime birlikteliğini yakalarken, 10 veya 15 gibi daha büyük bir pencere boyutu, daha fazla genel birlikteliği yakalayacaktır. Kelime benzerliği ve analojilerin ölçülmesi düşünüldüğünde, pencere boyutunun 5 olarak seçilmesi ile küçük pencere boyunun sebep olacağı yerel bağlamın yakalaması için çok küçük olması ve büyük pencere boyunun sebep olacağı çok fazla ilgisiz kelimenin dikkate alınarak önemli bilgilerin göz ardı edilmesinin önüne geçilmiştir.

Kelime vektörlerinin boyutu düşünüldüğünde her bir boyut kelimenin bir karakteristiğini veya anlamlılığını temsil eder. Kelime vektörlerinin boyutu yüksek seçilerek kelimelerin daha fazla anlamlılık taşıması ve kelimeler arasında daha detaylı ilişkilerin ortaya çıkarılması sağlanabilir. Ancak, daha yüksek bir boyutluluk seçerken dikkate alınması gereken bazı durumlar da vardır. Bunlardan birincisi, boyut arttıkça modelin eğitilmesi ve kullanılması hesaplama açısından daha maliyetli hale gelebilir. Bunun yanında, daha fazla sayıda boyutla, kelime vektörlerini yorumlamak veya görselleştirmek daha zor hale gelebilir. Çalışmada hesaplama karmaşıklığı, veri görselleştirme, kelime benzerliği ve analoji başarım değerleri göz önünde bulundurularak, çalışmada kelime vektörlerinin boyutu 300 olarak seçilmiştir.

Çizelge 4.7. Kelime yerleştirme algoritma parametreleri ve istatistikleri

	GloVe	word2vec	word2vec_sg	fastText	fastText_sg
Pencere Boyu	5	5	5	5	5
Vektörü Boyutu	300	300	300	300	300
Sözcük Sayısı	218000	67800	67800	67800	67800

Çizelge 4.7.'de kullanılan kelime yerleştirme modelleri ve eğitim sonucu ortaya çıkan sözcük sayısı gösterilmiştir. Çalışmada word2vec tabanlı algoritmalar 67800, GloVe algoritmasında ise 218000 tekil sözcük oluşmuştur.

Kelime vektörlerinin kalitesi benzerlik ve analoji gibi dâhili (intrinsic) görevlerinden elde edilen başarım değerleri ile değerlendirilir. Benzerlik ve analoji görevlerinden elde edilen başarımın, bu kelime vektörlerini kullanan sınıflandırma, varlık ismi tanıma gibi hârici (extrinsic) görevlere de yansıması beklenir. Ancak Schnabel ve diğerleri (2015) dâhili ve hârici görevlerden elde edilen sonuçların her zaman uyuşmadığı sonuçlar göstermiştir. Kelime yerleştirme algoritmaları karşılaştırıldığında kelime benzerliği ve analojiler gibi dâhili görevler için de başarım farklılık gösterdiği görülmüştür (Gladkova ve diğerleri, 2016). Bu sonuçlar her görevin kelime yerleştirmelerinin farklı bir yönünden başarım elde ettiğini göstermiştir. Bu durumda bir görevde başarılı olan bir kelime yerleştirme metodunun başka bir görevde de en başarılı metot olacağını garantilemez.

Kelime benzerliği yakın anlamlılığı ve olumlu bir ilişkiyi çağrıştırmasına rağmen kelimelerin aynı bağlamda geçme sıklığını ve ihtimalini gösterir. Bunun sonucunda çoğunlukla yakın anlamlı kelimeler benzer çıksa da zıt anlamlı ya da olumsuz ilişkiye sahip kelimelerin benzerliklerinin yüksek çıktığı örnekler de vardır. Örneğin Rusya-Ukrayna savaşı bu iki ülkenin aynı bağlamda geçme sayısını ve olasılığını artırmıştır. Bunun sonucun olarak derlemden elde edilen kelime vektörlerine göre kelime benzerlikleri yüksektir. Türkçedeki deyimlerde ve kalıplarda bir kelime zıt anlamlısı ile birlikte sıklıkla kullanılır. "Sıcak demeden soğuk demeden bekledik.", "Yaz kış çalıştılar.", "Üç aşağı beş yukarı anlaştık." cümleleri bu kullanımlara örnektir. Derlemde sıcak kelimesine benzerliği en yüksek kelime zıt anlamlısı soğuk, yaz kelimesi için kış, iyi için kötü, aşağı için yukarı kelimesidir.

Eş anlamlı kelimeler aynı anlamı taşıyan kelimelerdir ve anlamsal benzerlikleri nerdeyse birbirleriyle eşittir. TBMM genel kurul tutanaklarında anlamsal kelime benzerliğini ölçmek için Türk Dil Kurumu sözlüğünde eş anlamlı kelimeler kullanılmıştır. Veri kümesi 169 adet kelime çiftinden oluşur. Kelime yerleştirme algoritmalarının bir kelimenin eş anlamlısını o kelimeye en yakın kelime olarak bulması beklenir. Bir vektör uzayında iki kelime vektörünün birbirine yakınlığı kosinüs benzerliği ile ölçülmüştür. Bu yakınlık aynı zamanda anlamsal olarak benzerliği de gösterir.

Her algoritma farklı bir vektör uzayı oluşturduğu için kelimelerin kosinüs benzerlik değerleri algoritma içinde anlamlıdır, algoritmalar arasında bu değerler standart oluşturmaz. Kelime yerleştirme algoritmaları arasında bir karşılaştırma yapılırken kosinüs benzerlik değerleri yerine aranan kelimelerin kaçıncı sırada benzerlik gösterdiği ölçülmüştür. Sonuçlardaki sıra değeri N, yakınlığı ölçülmek istene sözcüğün (Kelime2) ilgili sözcüğünün (Kelime1) en yakın ilk N (1, 2, …, N) kelimesi arasında olduğunu gösterir.

4.3.1. Kelimelerin biçimsel yapısına göre kelime benzerliği

Kelimelerin biçimsel yapısına göre benzerliğini incelemek için biçimsel yakınlıklarına göre iki veri kümesi oluşturulmuştur.

Birinci veri kümesi anlamsal olarak yakın ama biçimsel olarak farklı olan 167 adet kelimeden oluşur. Bu veri kümesine ait örnekler Çizelge 4.8.'de görülmektedir. İkinci veri kümesi ise hem anlamsal hem de biçimsel olarak yakın 108 kelimeden oluşur . Çizelge 4.9 bu veri kümesine ait örnek kelimeleri göstermektedir.

Çizelge 4.8 ve Çizelge 4.9'da word2vec CBOW modeli w2v_CBOW, word2vec skip-gram modeli w2v_sg, FastText CBOW modeli fastText_CBOW, fastText skip-gram modeli fastText_sg olarak gösterilmiştir.

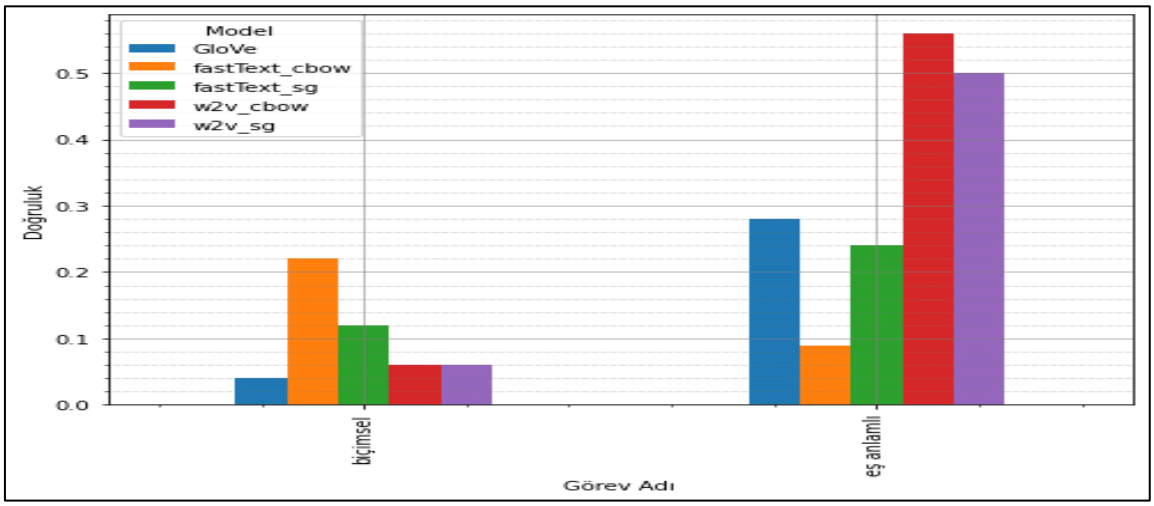
Çalışmada ilgili algoritmalar tarafından Kelime1 sütunundaki kelimeye en yakın 10 adet kelime bulunmuş, bu kelimenin Kelime2 sütunundaki eş anlamlı karşılığının yakın anlamlı kelimeler listesinde kaçıncı sırada bulunduğu ilgili algoritma sütununda gösterilmiştir. Ø sembolü eş anlamlı kelimenin algoritma tarafından en yakın 10 kelime arasında bulunamadığını gösterir. Örneğin *armağan* kelimesinin eş anlamlı karşılığı *hediye* GloVe tarafından en yakın birinci (1), w2v_cbow tarafından en yakın ikinci w2v_sg tarafından en yakın birinci, fastText_sg tarafından en yakın altıncı kelime olarak bulunmuştur. FastText_cbow ise *hediye* kelimesini en yakın 10 kelime arasında bulamamıştır.

Çizelge 4.8. Eş anlamlı kelimeler için kelime benzerliği örnekleri

Kelime1	Kelime2	w2v_cbow	w2v_sg	fastText _cbow	fastText_sg	GloVe
abartı	mübalağa	Ø	1	Ø	Ø	9
ama	fakat	1	1	Ø	Ø	1
anı	hatıra	1	1	Ø	1	1
anlam	mana	1	1	1	1	1
armağan	hediye	2	1	Ø	6	1
arzu	istek	Ø	3	Ø	Ø	1
asır	yüzyıl	1	1	1	1	1
ayrıcalık	imtiyaz	1	1	10	4	2
başvuru	müracaat	1	1	5	1	1
bereket	bolluk	2	2	Ø	10	1

Çizelge 4.9. Kelimelerin biçimsel ve anlamsal benzerlik örnekleri

Kelime1	Kelime2	w2v_cbow	w2v_sg	fastText _cbow	fastText_sg	GloVe
acemilik	acemi	Ø	Ø	Ø	1	Ø
avukatlık	avukat	1	1	2	5	1
bilgisayarcı	bilgisayar	Ø	Ø	3	1	Ø
bilgisayarlı	bilgisayar	Ø	10	4	1	Ø
dağcı	dağ	Ø	Ø	1	Ø	Ø
dansçı	dans	Ø	Ø	9	1	Ø
devletçilik	devletçi	Ø	7	4	1	5
dindar	din	Ø	Ø	1	Ø	Ø
etçil	et	Ø	Ø	1	Ø	Ø
evcil	Ev	Ø	Ø	1	Ø	Ø



Şekil 4.29. Eş anlamlı kelimeler ve türemiş kelimelerin kelime yerleştirme algoritmaları ile ölçülen benzerlikleri

Şekil 4.29'da biçimsel ve anlamsal olarak benzer kelimeler ve eş anlamlı kelimeler için kelime yerleştirme algoritmaların sonuçları karşılaştırılmıştır. Şekil 4.29'da sıra değeri (N) 1'e eşittir yani kelime2 kelime1'in en yakın birinci kelimesi ise kelimeler benzer kabul edilmiştir.

Eş anlamlı kelimelerin sonuçlarına göre fastText algoritmalarının başarımı kelime benzerliği bulma konusunda diğer algoritmaların altında kalmıştır. FastText CBOW %0,9, fastText skip-gram ise %24 başarım elde etmiştir. GloVe algoritmasının başarım değeri %24'dür. word2vec modelinin biçimsel olarak farklı kelimelerin anlamsal benzerliğini bulmada diğer modellere göre daha başarılıdır. word2vec CBOW modeli %56, word2vec skip-gram modeli ise %50 sonuç elde etmiştir. Word2vec modelleri karşılaştırıldığında ise CBOW modeli daha başarılı bulunmuştur.

FastText eğitim esnasında tahmin edilecek merkezi kelime olarak kelime, alt-kelime (sub-words) ve kelime n-gramlarını kullandığı için kelimelerin morfolojik (biçimsel) benzerliklerini bulmakta daha başarılı olması beklenir. Çalışmada biçimsel olarak birbirine yakın kelimeleri ölçmek için yapım eki ile isimden isim olarak türeyen kelimeler kullanılmıştır. Örneğin *dansçı* kelimesi *dans* isim kelimesinden *ci-cı* ekini olarak türemiştir. Bu iki kelime biçimsel olarak yakın olduğu gibi aynı alana (gösteri, eğlence) ait iki kelime olması nedeni ile anlamsal olarak da yakındır.

Şekil 4.28'de fastText algoritmaları kelime vektörlerini, alt kelime (subwords) ve karakter n-gramları ile tahmin yaparak elde ettiği için kelimeler arasında biçimsel yakınlık bulunması halinde daha başarılı sonuçlar vereceğini göstermiştir. Biçimsel ve anlamsal olarak benzer kelimelerin test edildiği görevde fastText CBOW %22, fastText skip-gram modeli ise %12 elde etmiştir. Biçimsel yakınlık için word2vec modelleri %0,06, GloVe ise %0,04 başarım elde etmiştir. TBMM Genel Kurul tutanakları veri kümesinden elde edilen sonuçlar ve Türkçenin biçimsel olarak zengin bir dil olduğu göz önünde bulundurularak eğer benzerliği ölçülecek kelimeler biçimsel olarak da birbirine yakın ise fastText modeli önerilir.

Çizelge 4.10'da kelime yerleştirme algoritmaların morfolojik durumuna göre elde ettiği sonuçlar biçimsel olarak zengin, yarı zengin ve yalın olmak üzere üç örnek kelime ile gösterilmiştir. Oluşturduğumuz veri kümelerinde benzerlik başarımları en yüksek olan iki

algoritma olan fastText CBOW ve word2vec CBOW algoritmalarından elde edilen bu kelimelere en yakın anlamlı 20 kelime incelenmiştir.

Çizelge 4.10. Morfolojilerine göre benzerlik örnekleri

Ağırlık		Gönüllülük		Element	
w2v_cbow	fastText_cbow	w2v_cbow	fastText_cbow	w2v_cbow	fastText_cbow
önem	ağırlıkta	gönüllük	gönüllük	mineral	elemeğinin
ehemmiyet	sağırlık	mukimlik	alçakgönüllülük	oksit	moment
hız	ağırlıktaki	hasbilik	gönüllü	uranyum	kement
yoğunluk	ağırlığı	gönüllü	gönüllüce	bentonit	elemeği
patlak	ağırlı	usüller	alçakgönüllü	barit	segment
sebebiyet	ağırlıklı	usül	gönüllerdeki	antimon	elemeğiyle
yaygınlık	ağırlığındaki	usülleri	gönülsüz	toryum	elem
ivme	önem	numaraly	gönüllerindeki	ponza	ınvestment
çekidüzen	tırlık	gaye	gönüldaş	sodyum	sediment
salık	hıdırlık	mütekabiliyet	sevecenlik	stronsiyum	hidroksiklorokin
taviz	yoğunluk	prime	barışseverlik	zeolit	kolemanit
ödün	yaygınlık	düstur	hamiyetperverlik	jips	hidrojen
ektazminat	ehemmiyet	prensip	yardımseverlik	magnezyum	uranyum
destek	ağırlaştırılmalı	sistemler	severlik	krom	granit
popülarite	ağırlaşma	karşılıklılık	yurtseverlik	kobalt	hidroksit
poz	ağırlaştırma	profesyonellik	özbenlik	madensel	hidrojeoloji
görünürlük	yoğunlukta	devamlılık	güçlülük	manyezit	toryum
start	ağırlaştırmış	felsefe	örgütlülük	potasyum	metalürjik
öncelik	ağırlaştıran	dayandırılma	prenslik	feldispat	sülfürik
veriştirmek	maliyetetkinlik	çoğulculuk	hasbilik	feldspat	hidrojeolojik

FastText CBOW tarafından bulunan *ağırlık* sözcüğüne en yakın on sözcük görülmektedir. *Ağırlık* kelimesi ağır-ağırlık olarak 1 defa türemiştir ve biçimsel olarak yarı zengin olarak kabul edilebilir. Birinci yakınlıktaki *ağırlıkta* kelimesi biçimsel olarak ve anlamsal olarak yakın olmasına karşın ikinci sıradaki *sağırlık* kelimesinin biçimsel olarak yakındır ama anlamsal olarak ilgisizdir. Eğer sözcük sözlükteki çok sayıda kelime ile biçimsel benzerlik gösteriyorsa ve anlamsal yakınlık ölçülüyorsa, biçimsel benzerliğin anlamsal yakınlığı baskıladığı ve ilgisiz sonuçlar çıkardığı durumlar oluşabilir. Bu durum fastText için bir dezavantaj oluşturur.

Ağırlık kelimesinin word2vec tarafından bulunan yakın anlamlı kelimeleri gösterir. FastText ile karşılaştırıldığında yalnızca anlamsal yakınlığa odaklandığı görünmektedir. Ağırlık

kelimesini model *önem, ehemmiyet* olarak değerlendirmiştir. *Ağırlıklı* gibi hem anlamsal hem biçimsel yakınlığa sahip olan kelimeyi ise hiç bulamamıştır. FastText ise *ağırlıklı* kelimesini altıncı sırada bulmuştur. Bu durum word2vec'e göre bir avantajıdır. Kelimenin yaygın olarak kullanılan anlamı olan *önemi* ise ancak sekizinci sırada bulunabilmiştir.

Gönüllülük kelimesi biçimsel olarak zengindir. Gönül, gönüllü, gönüllülük olarak birden çok defa türetilmiş anlamları Türkçede sıklıkla kullanılır. Bu kelimenin fasText tarafından bulunan benzer kelimeleri incelendiğinde fastText'in bu anlamları başarı ile bulduğu gözlenir. 'gönüllük', 'alçakgönüllülük', 'gönüllü', 'gönüllüce', 'alçakgönüllü', 'gönüllerdeki', 'gönülsüz' gibi biçimsel ve anlamsal benzerlik yanında 'barışseverlik', 'hamiyetperverlik', 'yardımseverlik' gibi anlamsal olarak yakın kelimeleri de bulmuştur. Word2vec ise hem biçimsel hem anlamsal yakın olan kelimelerden yalnızca 'gönüllük' ve 'gönüllü' kelimelerini bulmuştur. Anlamsal olarak yakın olan sadece *hasbilik* kelimesi bulunmuştur.

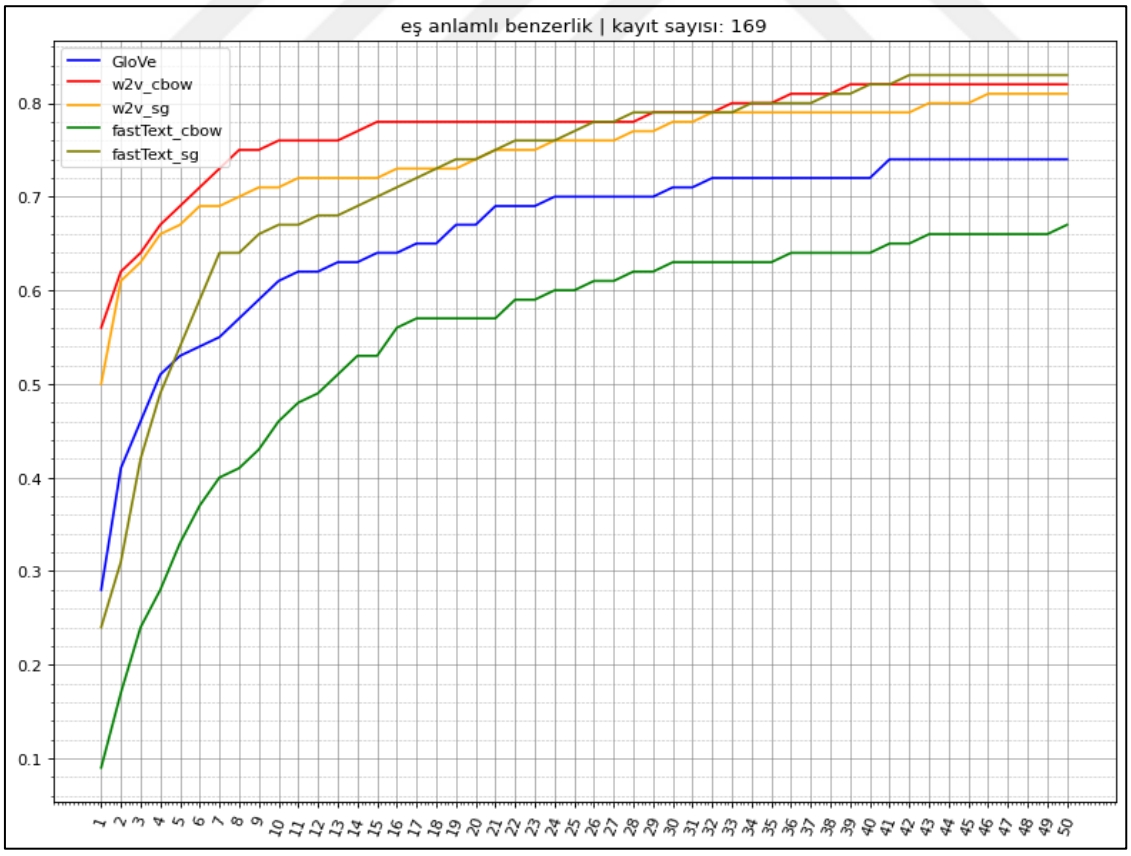
Element sözcüğü Türkçeye İngilizceden girmiştir ve bu kelimeden türetilen kelime yoktur. Dolayısı ile biçimsel olarak zengin değildir. Bu kelimenin fastText tarafından bulunan 'sediment', 'hidroksiklorokin', 'kolemanit', 'hidrojen' gibi ilişkili kelimeler ancak dokuzuncu kelimeden itibaren ortaya çıkmıştır.

Element sözcüğünün word2vec tarafından bulunan yakın anlamlı ya da ilgili sözcükleri incelendiğinde 'mineral', 'oksit', 'uranyum', 'bentonit', 'barit', 'antimon', 'toryum', 'ponza', 'sodyum', 'stronsiyum', 'zeolit' sözcüklerin en yakın ilk kelimeden itibaren bulunduğu gözlemlenmiştir.

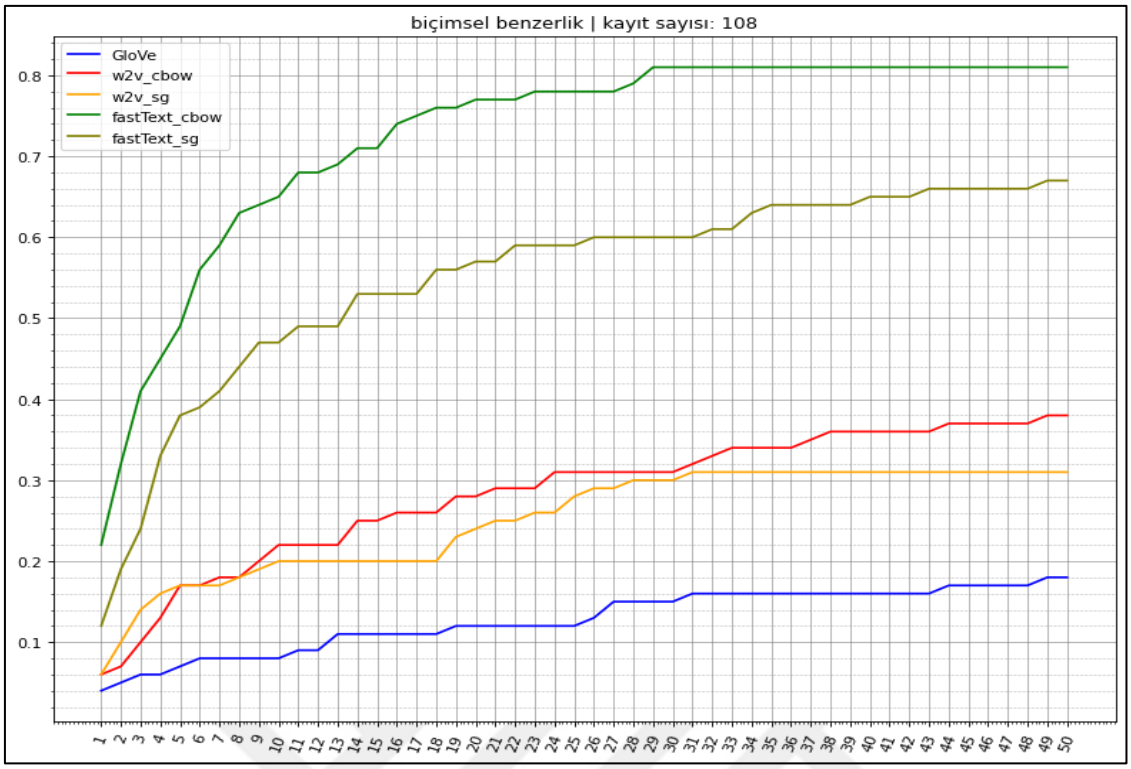
Morfolojik zenginliğine göre kelimelerin yakın anlamlılığının bulunması göz önünde bulundurulursa karakter n-gramlara dayalı fastText modeli morfolojik ve anlamsal olarak yakın kelimeler başarı ile bulurken, anlamsal olarak yakın ama morfolojik olarak farklı kelimeleri yakınlık sırası (N) yüksek tutulursa bulabilir. FastText ayrıca morfolojik olarak yakın ama anlamsal olarak ilgisiz kelimelere de yüksek benzerlik değeri gösterebilmektedir. *Element* kelimesine yakın olarak anlamsal olarak ilgisiz *elemeği, elemeği, elemeğiyle* kelimesi bulunmuştur. Kelime tabanlı olan word2vec ve GloVe modelleri ise biçimsel benzerlik şartı olmadan anlamsal yakınlığı ve ilgiyi yakalar.

FastText'in morfolojik benzerliği olmayan kelimeleri ancak yüksek sıra (N) için bulabildiği üç farklı morfolojik durum içinde gözlemlenmiştir. Bu durum tüm veri kümesi için nasıl gerçekleştiğini gözlemlemek için en yüksek 50 kelime yakınlığı için kelime yerleştirme algoritmalarının ürettiği sonuçlar Şekil 4.30'da görülmektedir. Eş anlamlı (biçimsel olarak farklı anlamsal olarak yakın) kelimelerde fastText skip-gram'ın 5. sıradan itibaren GloVe modelinden, 18. Sıradan itibaren word2vec skip-gram modelinden, 25. sıradan itibaren ise word2vec CBOW modelinden daha iyi sonuç vermeye başlamıştır. Bu durum kelime yakınlığı derecesinin öneminin daha az olduğu görevler için karakter tabanlı fastText skip-gram algoritmasının kelime tabanlı GloVe ve word2vec modellerine tercih edilebileceğini göstermektedir.

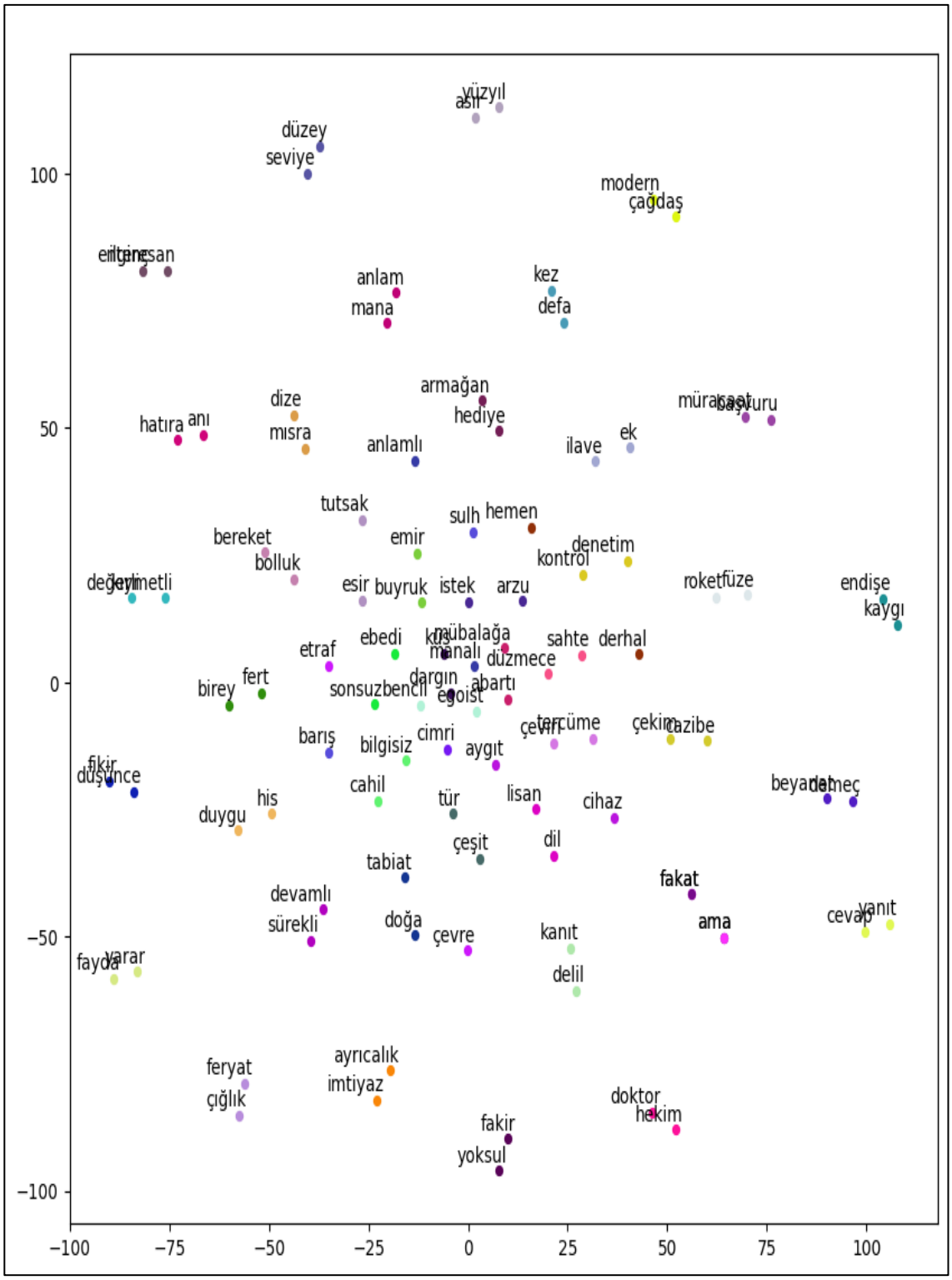
Hem biçimsel hem anlamsal yakınlığın bulunduğu yapım eki ile isimden isim olarak türemiş sözcükler için ise karakter tabanlı fastText algoritmalarının kelime tabanlı GloVe ve word2vec'e göre daha başarılı olduğu Şekil 4.31'de görülmektedir. Şekilde en yüksek başarımı fastText CBOW algoritması 29. sıra (N=29) için %81 başarım ile elde etmiştir.



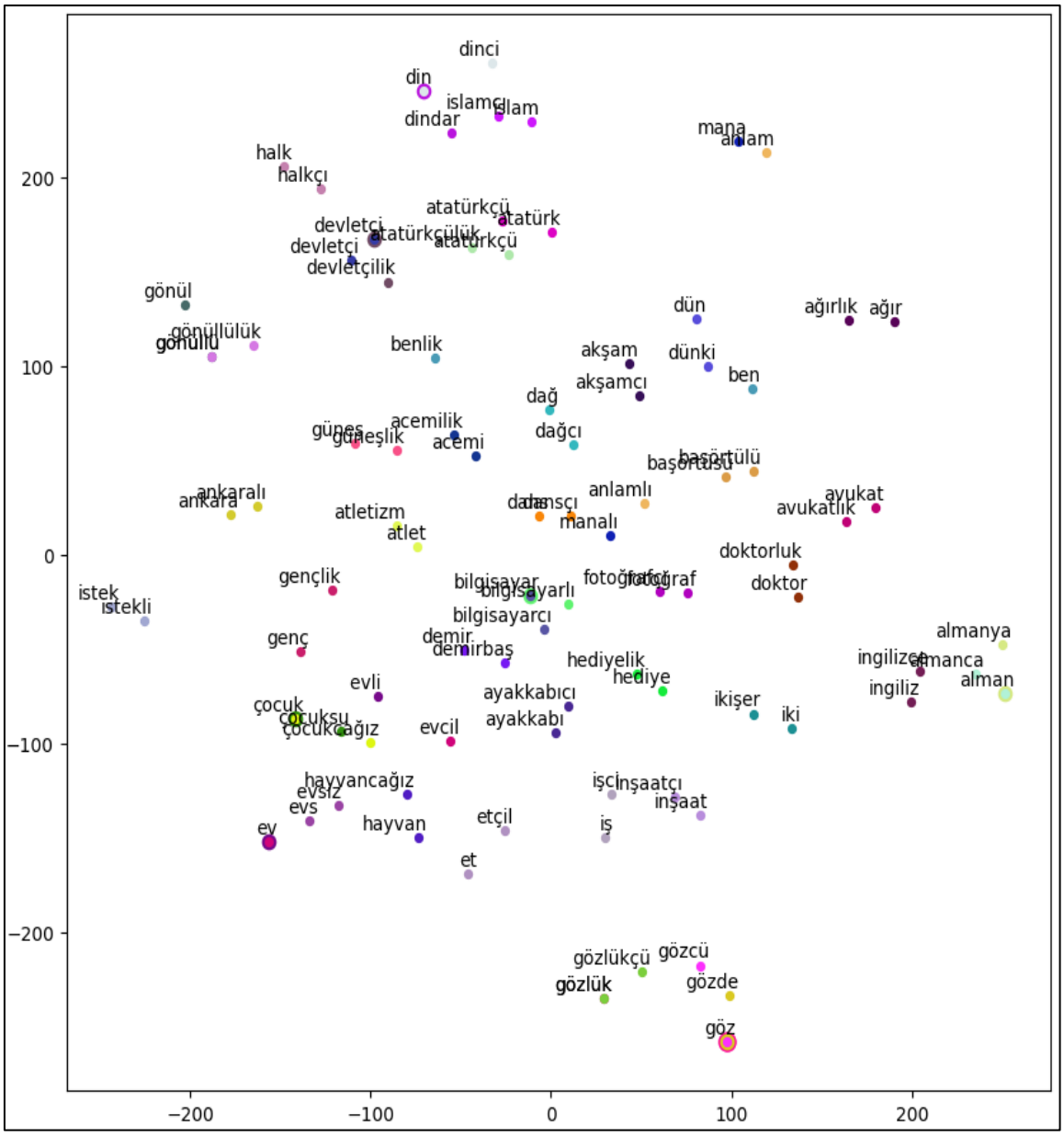
Şekil 4.30. Eş anlamlı kelimelerin ilk 50 sıra için başarımı



Şekil 4.31. Biçimsel ve anlamsal olarak benzer kelimelerin ilk 50 sıra için başarımı



Şekil 4.32. Eş anlamlı kelimelerin iki boyutlu kelime vektör uzayında gösterimi



Şekil 4.33. Biçimsel ve anlamsal olarak benzer kelimelerin iki boyutlu kelime vektör uzayında gösterimi

Şekil 4.32 ve Şekil4.33'de eş anlamlı ve biçimsel olarak zengin çekim ekli kelimelerin kelime yerleştirme algoritmaları tarafından bulunan kelime vektörleri iki boyutlu vektör uzayında gösterilmektedir. T-SNE kullanılarak yapılan görselleştirmelerde word2vec CBOW algoritmasının eş anlamlı kelimeleri fastText CBOW algoritmasının ise çekim ekli kelimeleri benzerliklerine göre birbirlerine yaklaştırdıkları görünmektedir. Şekil 4.32'de en yakın anlamlı eş anlamlı kelime çiftleri *'asır - yüzyıl', 'demeç - beyanat', 'anlam - mana', 'düzey - seviye', 'ama - fakat', 'enteresan - ilginç', 'cevap - yanıt'* olarak bulunmuştur. Şekil 4.32'de *'etraf - çevre', 'derhal - hemen', 'cimri - pinti', 'cihaz - aygıt', 'barış - sulh', 'dargın - küs', 'arzu - istek'* kelimeleri ise eş anlamlı kelime grubunda birbirine en uzak kelimelerdir.

Şekil4.33’de en yakın anlamlı eş anlamlı kelime çiftleri *'bilgisayarlı - bilgisayar', 'fotoğrafçı - fotoğraf', 'ayakkabıcı - ayakkabı', 'atatürkçülük - atatürkçü', 'bilgisayarcı - bilgisayar', 'devletçi - devletçi'* olarak bulunmuştur. En uzak kelimeler ise *'benlik - ben', 'ağırlık - ağır', 'evcil - ev', 'gözlük - göz', 'anlam - anlamlı', 'dinci - din'* kelimeleridir. Şekil 4.33 incelendiğinde fastText CBOW *ingiliz, İngilizce, alman, almanca, almanya* gibi kelimeleri bir araya getirerek ülke kavramı için bir küme oluşturmuştur. Aynı şekilde *devletçi, devletçilik, Atatürkçülük, Atatürk, Atatürkçü, halkçı* gibi kelimeleri bir araya getirerek ideolojiyi kapsayan bir kavram kümesi oluştuğu görünmektedir.

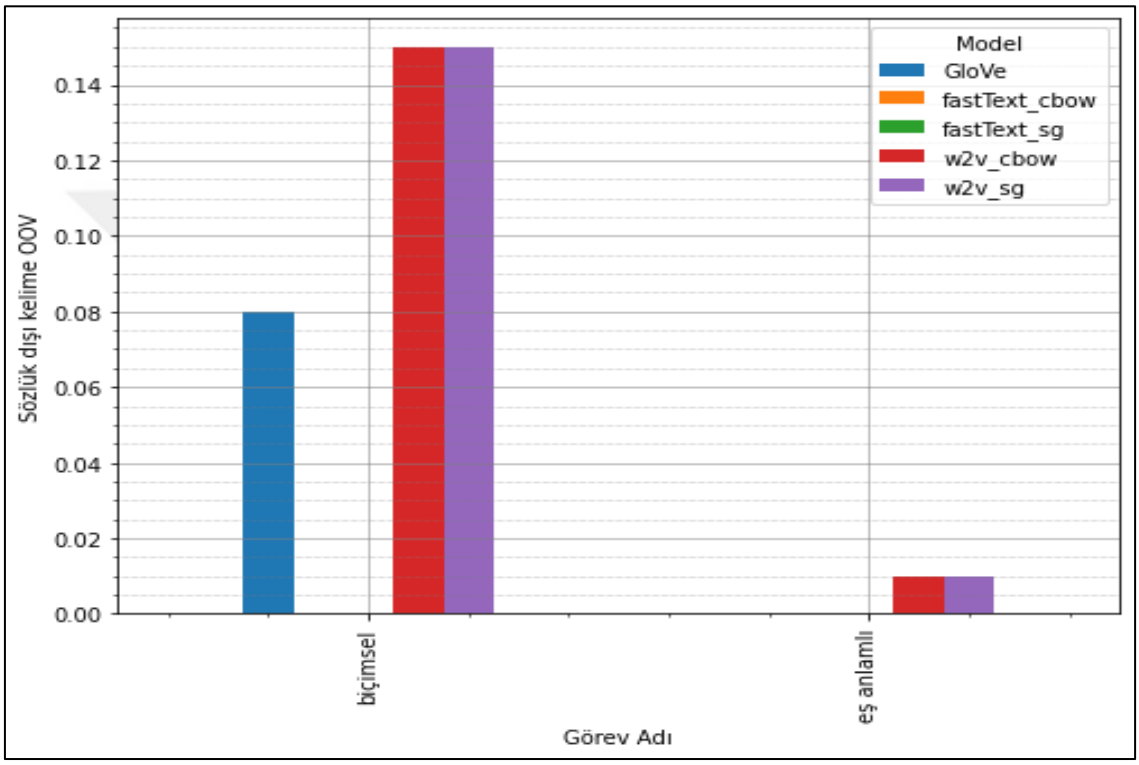
4.3.2. Sözlük dışı veya nadir kelimeler

Sözlük dışı kelime (SDK) (out of vocabulary, OOV) çalışılan dilde bulunan ama çalışılan derlemin sözlüğünde bulunmayan kelimeleri ifade eder. Kelime yerleştirmede FastText SDK için iyi bir çözüm olabilir. Bunun nedeni FastText'in tam kelimelerin yanı sıra karakter n-gramları gibi alt kelime birimleri için de kelime vektörleri üretebilmesidir. Bu, kelimelerin morfolojik varyasyonları hakkında da bilgi çıkarımını sağlar. Böylelikle bilinmeyen veya nadir bulunan kelimelerin anlamsal yapılarını elde etmek için kullanılabilir.

Bir sözlük dışı kelime ile karşılaşıldığında, FastText modeli bu sözcüğü alt sözcük birimlerine ayırabilir ve bu alt sözcüklerin her biri için yerleştirme (embedding) oluşturabilir. Sözlük dışı kelime için nihai kelime vektörü daha sonra alt kelime birimleri için gömmelerin toplamı veya ortalaması olarak hesaplanabilir. Bu yaklaşım, özellikle zengin morfolojiye sahip dillerde sözlük dışı kelime sorununa bir çözüm sunar.

Türkçe sondan eklemeli dil olması sebebiyle morfolojik çeşitliliği fazladır ve fastText nadir bulunan veya sözlük dışı kelimeler için kullanılabilir. Şekil 4.34’de biçimsel ve eş anlamlı kelime benzerliği veri kümesinde kelime yerleştirme algoritmaları tarafından sözlükte bulunamayan SDK kelime sayısının veri kümesindeki kelime sayısına oranı gösterilmiştir. Biçimsel veri kümesi için word2vec %15, GloVe %8 SDK sorununa sebep olurken fastText bütün kelimeler için kelime vektörü oluşturmuştur. Aynı sorun morfolojik olarak zengin olmayan kelimelerden oluşan eş anlamlı veri kümesinde word2vec için %1 olarak ölçülmüş, GloVe ve fastText için ise ölçülmemiştir.

Veri önişleme aşamasında derlem oluşturulurken kelimelerin yüzeysel biçimleri (surface form) yerine en uzun kelime başları (lemma) kullanılmıştır. Bu işlem ek alarak farklılaşan kelimelerin ortak kelime başı ile temsil edilmesi sağlar, böylelikle morfolojik çeşitlilik de azalmış olur. Bu durum oluşabilecek SDK etkisinin azaltır. Kelime kökleri veya kelime başları olmadan kelimelerin yüzeysel biçimleriyle yapılan metin analizinde fastText'in SDK sebebiyle ortaya çıkan olumlu etkisinin artacağı varsayılabilir.



Şekil 4.34. Kelime yerleştirme algoritması tarafından bulunamayan kelimeler

4.3.3. TBMM Genel Kurul görüşmelerinde kelime analojileri

TBMM genel kurul tutanaklarında kelime analojilerini incelemek için literatürde sıklıkla kullanılan ülke-başkent kelime analojisinin yanında parlamento alanına özgü analoji veri kümeleri oluşturulmuştur. Parlamento alanına özgü ilk kelime analojisi görevi siyasi parti genel başkanı ve siyasi parti kelime çiftleri arasındaki ilişkidir. Çalışmada ikinci kelime analojisi görevi parlamento görüşmelerinde siyasi partiler adına konuşan parti grup başkanvekili ve siyasi parti çiftleri arasındaki ilişkiyi inceler. Üçüncü analoji görevinde ise kabine üyeleri ve bakanlık yaptığı alanı temsil eden kelime çiftlerinden oluşan görevdir.

Çizelge 4.11. Siyasi partiler ve genel başkanları

Kelime1	Klm2	Kelime3	Klm4	w2v _cbow	w2v _sg	Fast Text _cbow	Fast Text _sg	Glo Ve
tansuçiller	dyp	denizbaykal	chp	1	2	10	4	1
receptayyiperdoğan	akparti	kemalkılıçdaroğlu	chp	1	1	4	1	1
denizbaykal	chp	mustafabülentecevit	dsp	3	1	15	8	4
devletbahçeli	mhp	mithatsancar	hdp	4	3	18	3	5
pervinbuldan	hdp	meralakşener	iyiparti	3	1	7	2	3
mithatsancar	hdp	ahmetmesutyılmaz	anap	2	4	22	8	4
denizbaykal	chp	receptayyiperdoğan	akparti	2	1	13	1	1
necmettinerbakan	rp	pervinbuldan	hdp	4	1	28	2	5
mustafabülentecevit	dsp	devletbahçeli	mhp	1	1	10	1	1
devletbahçeli	mhp	kemalkılıçdaroğlu	chp	1	1	2	1	1

Çizelge 4.11 siyasi partiler ve genel başkanları veri kümesine ait örnek kayıtları ve sonuçları gösterir. $w_X \approx w_{kelime\_3} + w_{kelime\_2} - w_{kelime\_1}$ eşitliği ile elde edilen w_X kelime vektörüne en yakın kelimeler listesindeki kelime4'ün sırası Çizelge 4.11'deki algoritmalar sütununda gösterilmiştir. Çizelgede 100 değeri yakın kelimeler listesinde bulunmadığını gösterir. w_X kelime vektörüne en yakın kelimeler bulunurken girdi kelime vektörleri ($w_{kelime\_1}, w_{kelime\_2}, w_{kelime\_3}$) göz ardı edilmiştir(Mikolov ve diğerleri, 2013).

Kelime analojileri, kelimelerin yüksek boyutlu vektör uzayında matematiksel gösterimleri olan kelime yerleştirmelerin veya kelime vektörlerinin kalitesini değerlendirmek için bir ölçüt olarak kullanılabilir. Buradaki fikir, eğer kelime yerleştirme kelimeler arasındaki anlamsal ilişkileri yakalamada başarılı ise, kelime analojilerinde de iyi performans göstermeleri beklenir.

Şekil 4.35'de kelime analojileri yakalamada en başarılı olunan görev genel başkan-siyasi parti kelime çiftidir. 73 adet kayıt üzerinden yapılan ölçümde word2vec skip-gram %55

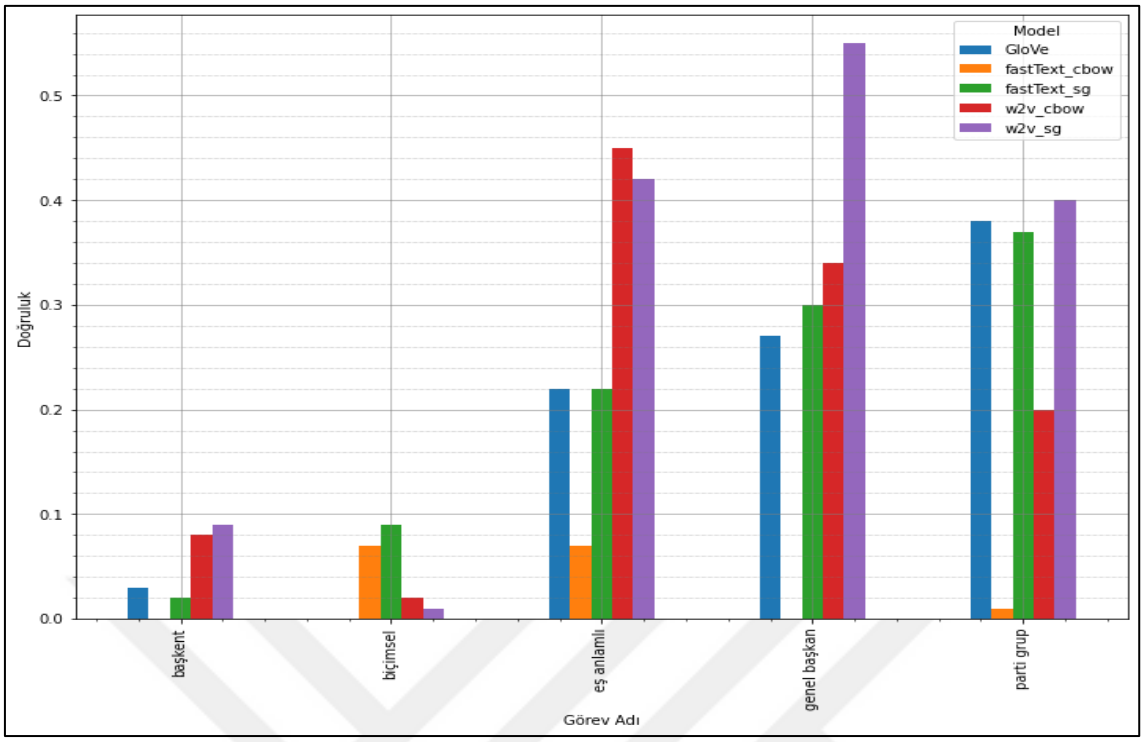
başarım elde etmiştir. İkinci en başarılı model word2cev CBOW (%34) ile arasında %20 kadar başarım farkı vardır. FastText skip-gram %30, GloVe ise %27 başarım elde etmiş, fastText CBOW ise hiçbir kelime analojisi bulamamıştır.

Siyasi parti grup başkanvekilleri ve siyasi partiler çiftlerinden oluşan analoji veri kümelerinde word2vec skip-gram modeli %40 başarım elde etmiştir. GloVe %38 ile ikinci en başarılı model iken fastText skip-gram %37 başarım göstermiştir. Bu görevde, bir pencere boyunda ortadaki kelimenin etrafındaki kelimeleri veya alt-kelimeleri tahmin etmeye dayalı skip-gram modeller bir kelimenin yanındaki kelimeleri tahmin etmeye dayalı CBOW modellerinden daha başarılı olmuştur. FastText CBOW analojilerin ancak %1'ini bulabilirken word2vec CBOW'in başarımı %20'dir.

500 kayıt üzerinde eş anlamlı kelime çiftlerinden oluşan analoji görevinde is word2vec modelleri diğer modellerden daha başarılıdır. Word2vec CBOW %45, word2vec skip-gram %42 başarım elde etmiştir. FastText CBOW bu görevde %7 ile en az başarım elde eden modeldir. GloVe ve fastText skip-gram ise %22 başarım elde etmiştir.

Yapım ekleri ile isimden isim olarak türemiş biçimsel analoji veri kümesi 500 kelimeden oluşmaktadır. Biçimsel veri kümesinde başarım diğer görevlerin altında kalmıştır. Bu görevde fastTextin morfolojik çeşitliliği yakalamasından dolayı diğer modellerin üstünde bir başarım elde etmiştir. En yüksek başarım %9 ile fastText skip-gram, bir sonraki yüksek başarım %7 ile fastText CBOW modelidir. GloVe hiçbir analoji bulamaz iken word2vec CBOW %2, word2vec skip-gram ise %1 başarım ile çok az başarım değerleri elde etmiştir. Word2vec ve GloVe modelleri biçimsel veri kümesinde analoji görevinde başarısız olduğu FastText modellerinin ise doğruluk değerlerinin çok düşük kaldığı söylenebilir.

Ülke-başkent kelime çiftlerinden oluşan ve 500 adet kayıtın olduğu veri kümesinde word2vec skip-gram %9, word2vec CBOW %8, GloVe %3 başarım elde etmiştir. FastText skip-gram analojilerin ancak %2'sini bulabilirken fastText CBOW modeli ise kelime analojisi bulamamıştır. Sonuçlar incelendiğinde TBMM genel kurul tutanaklarında ülkeler ve başkentlerin birbirine uzaklığın ülkeden ülkeye değişkenlik gösterdiği ve bir analoji oluşturacak ilişkiye sahip olmadığı değerlendirilebilir.



Şekil 4.35. Kelime yerleştirme algoritmalarının kelime analojileri başarımı

TBMM Genel Kurul tutanaklarında analojilerin incelendiği görevlerde sonuçlar arasında farklılıklar gözlenmiştir. Gladkova ve diğerleri (2016) kelime analojisi probleminde başarımın farklı dilsel ilişkiler için büyük ölçüde değişkenlik gösterdiğini belirtmişlerdir. Bu yüzden başarım aranacak analojinin biçimsel, anlamsal, alttür (hiponimi) veya üsttür (hipernimi) gibi durumlarına göre değişebilir.

4.3.4. İllerin kelime benzerliği

Çalışmada illerin, ülkelerin ve kabine üyelerinin kelime benzerliği için word2vec CBOW kelime yerleştirme algoritmasından elde edilen kelime vektörleri kullanılmıştır. Word2vec CBOW algoritması seçilirken eş anlamlı kelimeler ile yapılan kelime benzerlik görevi başarımları dikkate alınmıştır (Bkz. Şekil 4.29 ve Şekil 4.30).

TBMM Genel Kurul Tutanaklarında 81 ilin birbirleri ile benzerliklerini ölçerek vektör uzayında nasıl dağılım gösterdikleri incelenmiştir. Şekil 4.36 en yüksek benzerlik gösteren ilk 50 il ikilisini, Şekil 4.37 ise il adlarına ait 300 boyutlu kelime vektörlerinin t-sne algoritması ile 2 boyuta indirgendikten sonra oluşturulan saçılım grafiğini gösterir.

Beklenildiği gibi genel olarak sınır komşusu olan veya coğrafi olarak yakın olan illerin benzerliği yüksektir. Fakat bu durumuna istisna oluşturan durumlar da vardır.

İstanbul'un en yakın benzerlik gösterdiği beş il *'ankara', 'izmir', 'bursa', 'kocaeli', 'antalya', 'trabzon', 'eskişehir'* olarak çıkmıştır. Ankara'nın en yüksek benzerlik gösterdiği iller *'istanbul', 'izmir', 'eskişehir', 'kayseri', 'diyarbakır', 'adana', 'bursa'* illeridir. İzmir'in benzerlik gösterdiği iller ise *'istanbul', 'antalya', 'bursa', 'adana', 'muğla', 'mersin', 'ankara'* olarak çıkmıştır. Bu sonuçlar birlikte değerlendirildiğinde şehirlerin ekonomik ve sosyal durumları da benzerliklerini artırmaktadır. Türkiye'nin en büyük üç metropolü İstanbul, Ankara ve İzmir'in birbirleri ile benzerlikleri yüksek çıkmış Şekil 36'daki iki boyutlu vektör uzayında metropoller olarak adlandırılabilecek ayrı bir kümede gösterilmiştir.

Şehirler tarihlerinde paylaştıkları ortak olaylar ile de birbirine yaklaşabilir ve kelime benzerlikleri yüksek çıkabilir. Sivas'ın en benzer iller listesinde Çorum birinci, Çorum'un en benzer iller listesinde ise Sivas birinci sıradadır. Bu iki il sınır komşusu değildir. İki il arasındaki yüksek benzerlik değerlerinin her iki ilin tarihlerinde yer alan benzer olaylardan kaynaklanabilir. 1980 Mayıs-Temmuz aylarında Çorum'da alevi vatandaşları hedef alan olaylar olmuştur. 2 Temmuz 1992'de Sivas'ta çoğunluğu Alevi 33 aydın ile 2 otel çalışanının hayatlarını kaybetmeleri ile sonuçlanan Madımak oteli saldırısı olmuştur. Çorum ve Sivas kelime vektörleri arasındaki yüksek benzerliğin sebebinin bu iki olay olduğu çıkarımını doğrulamak için Türkiye tarihinde Alevileri hedef alan üçüncü bir olay ile karşılaştırılmıştır. Maraş olayları; 19 Aralık ile 26 Aralık 1978'de Kahramanmaraş'ta meydana gelen Alevileri hedef alan olaylardır. Kahramanmaraş kelimesinin kısaltılmış şekli olan *maraş* kelimesinin en yüksek benzerlik gösterdiği iller sırası ile ('sivas'- 0,5469), ('kahramanmaraş'-0,5147), ('antep'-0,4890), ('çorum'-0,4742) ve ('hatay'-0,4442)'dır. Maraş kelimesi kendisinin unvan almış hali *'Kahramanmaraş'* komşuları *Antep* ve *Hatay* ile yakın benzerlik göstermesi beklenebilir. Fakat Çorum ile olan yüksek benzerliği ve Sivas ile olan benzerliğinin diğer komşularının çok üstünde çıkması Çorum, Sivas ve Maraş'ın kelime benzerliğinin, bu üç ilde de Alevi yurttaşları hedef alan olaylar sonucu yüksek çıktığını doğrulamıştır.

Kocaeli ve Düzce sınır komşusu değildir. Aralarındaki yüksek benzerlik değerleri şehirlerin paylaştığı ortak olayların kelime benzerliğine olan etkisine diğer bir örnektir. 17 Ağustos 1999 yılında merkez üssü Kocaeli Gölcük ilçesi olan ve 18 373 kişinin yaşamını yitirdiği Türkiye tarihinin en büyük depremlerinde birisi yaşanmıştır. Bu depremden 88 gün sonra 12

Kasım 1999'da ise Bolu'nun Düzce ilçesinde gerçekleşen 845 kişinin yaşamını yitirdiği deprem olmuştur. Ölü sayıları Temmuz 2010 ayına ait *Deprem Riskinin Araştırılarak Deprem Yönetiminde Alınması Gereken Önlemlerin Belirlenmesi Amacıyla Kurulan Meclis Araştırması Komisyonu* raporundan alınmıştır. Bu depremden sonra Düzce il statüsüne geçmiştir. Düzce ve Gölcük depremi hem tarihlerinin yakınlığı ve çok şiddetli depremler olması sebebiyle birlikte anılırlar. Kocaeli kelimesinin en fazla yakınlık gösterdiği illerde ('düzce'-0,6299), ('bursa- 0,6280), ('sakarya'- 0,6085) ve ('balıkesir', 0.5973)'dir. Düzce kelimesine en benzer iller ('bolu'-0,6878), ('kocaeli'-0,6299), ('sakarya'-0.5989) ve ('elazığ', 0.54909)'dır. Düzce kelimesi komşuluğu ve daha önce bağlı olduğu il olması nedeni ile Bolu ile benzerlik göstermesi beklenir. İkici en yüksek benzerliği Kocaeli ile göstermiştir. Sonuçlara göre *Elazığ* ile olan yakınlığının da 2020 Elazığ depremi ile ilgili olduğu değerlendirilmiştir. Kocaeli ve Düzce kelimesini benzerlik değerlerini doğrulamak için Gölcük kelimesi kullanılabilir. Gölcük kelimesinin iller ile gösterdiği benzerliklere göre; 'düzce', 'bolu', 'kocaeli', 'sakarya', 'yalova', 'elazığ' sıralaması oluşur. Bu benzerlik değerleri sonucunda Kocaeli ve Düzce kelime vektörlerinin anlamsal olarak yaklaşmasının bu illerin paylaştığı ortak deprem geçmişi ile ilgili olduğu değerlendirilmiştir.

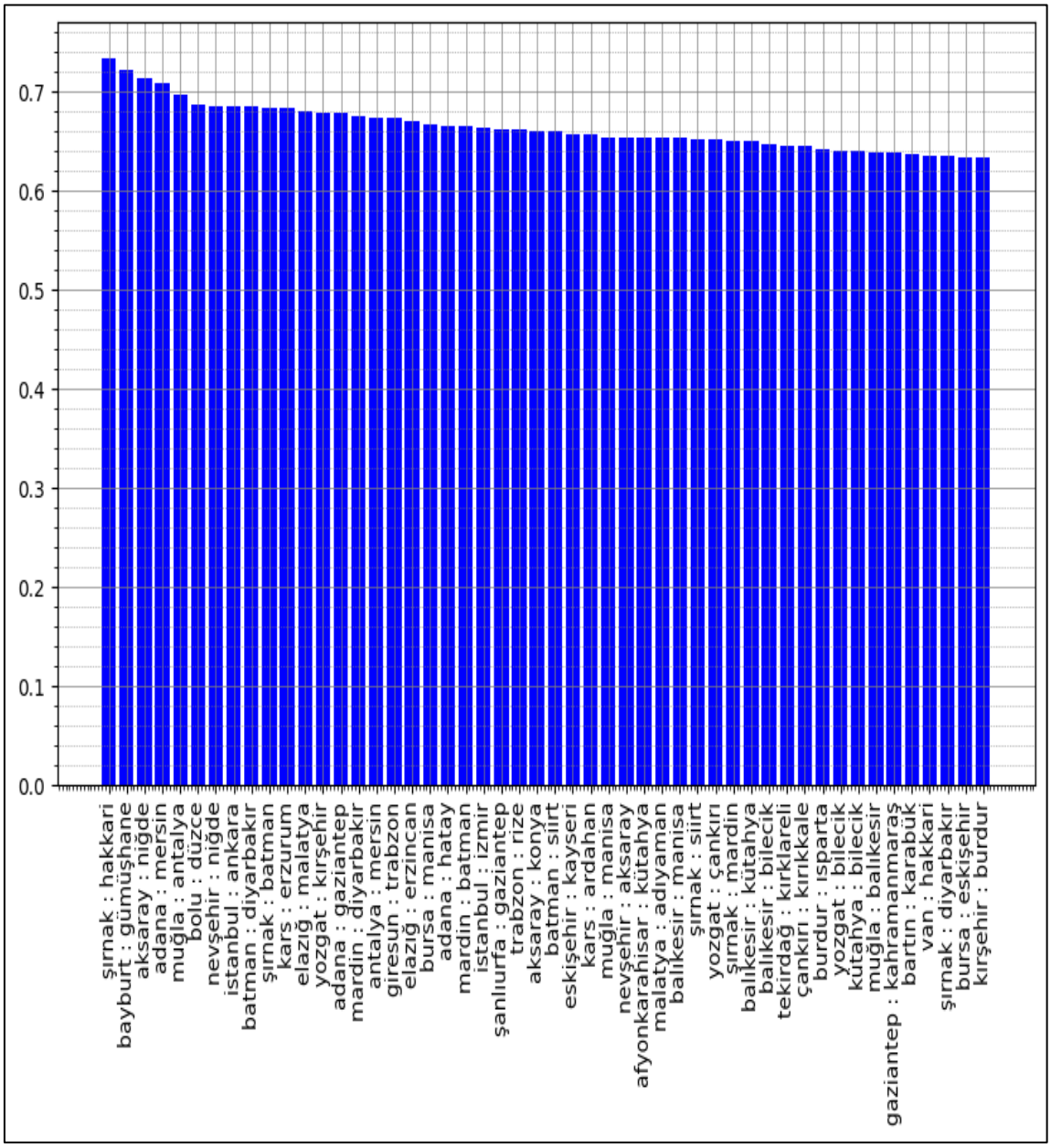
Kelime benzerliği en düşük olan il ikililerinde Ordu kelimesinin bulunduğu ikililer çoğunluktadır ve diğer iller ile oluşturduğu benzerliklerin ortalaması 0,1266 en düşük ildir. Bu durum 'ordu' kelimesinin bir il adından çok Türk Silahlı Kuvvetlerini temsil etmesidir. Derlemde *ordu* kelimesinin en yakınlık gösterdiği kelimeler yakınlıklarına göre *'ordusu', 'komutasındaki', 'müfreze', 'asker', 'kahraman', 'tsk', 'donanmak', 'ordumuz', 'kumandan', 'tugay', 'kuvvet', 'öso', 'kuvayımilliye', 'muhafız'* şeklindedir. Türkçede il adından başka bir anlamı olan diğer bir kelime ise aydın kelimesidir. Derlemde *aydın* kelimesinin yakın anlamlı kelimelerinden (*'aydınlar', 'ilhan', 'nezir', 'münevver', 'türkali')* il adından daha çok *münevver, entelektüel* anlamına karşılık geldiği görülmüştür. Aydın diğer iller ile benzerliklerinin ortalaması 0,2928 en düşük üçüncü ildir. Türk Dil Kurumu sözlüğüne göre uşak kelimesinin 1. Çocuk, 2. Herhangi bir bölgenin halkından olan erkek, 3. Erkek hizmetçi anlamı vardır. Uşak kelimesine yakın anlamlı kelime incelendiğinde bu kelimenin derlemde il adını temsil ettiği görülür. Uşak kelimesi ile en yakın anlamlı kelimeler *'afyon', 'kütahya', 'tekirdağ', 'afyonkarahisar', 'balıkesir'* olarak sıralanmıştır. Bir diğer Türkçede anlamı olan *tokat* kelimesinin yakın anlamlı kelimelerinden (*'zile', 'niksar', 'erbaa', 'reşadiye', 'turhal', 'sille', 'tekme', 'artova', 'almus'*) ise yerleşim adına karşılık geldiği gözlemlenmiştir. Diğer illerle benzerliklerin ortalaması en düşük üçüncü (0,2653) il olan Çanakkale kelimesinin en

çok benzerlik gösterdiği kelimeler (*'malazgirt', 'dumlupınar', 'zafertepe', 'çanakkaledeki', 'sarıkamış', 'galiçya', 'kutülamare', 'arıburnu', 'tabya', 'miryokefalon', 'anzak', 'zaferi', 'seddülbahir', 'mohaç'*) şeklindedir. Bu kelimeler Türk tarihindeki diğer savaşlar ve Çanakkale savaşı ile ilgili kelimeleri belirtir. Çanakkale kelimesinin derlemde il adından daha çok Çanakkale Savaşını temsil etmesi nedeni ile diğer illerle düşük benzerlik gösterdiği değerlendirilmiştir. Çanakkale kelimesine en yüksek benzerliğe sahip il olan Sakarya ile benzerliği Sakarya'nın Türk kurtuluş savaşının en önemli muharebesinin geçtiği yer olması ve muharebenin Sakarya Savaşı olarak adlandırılması sonucu olduğu değerlendirilmiştir.

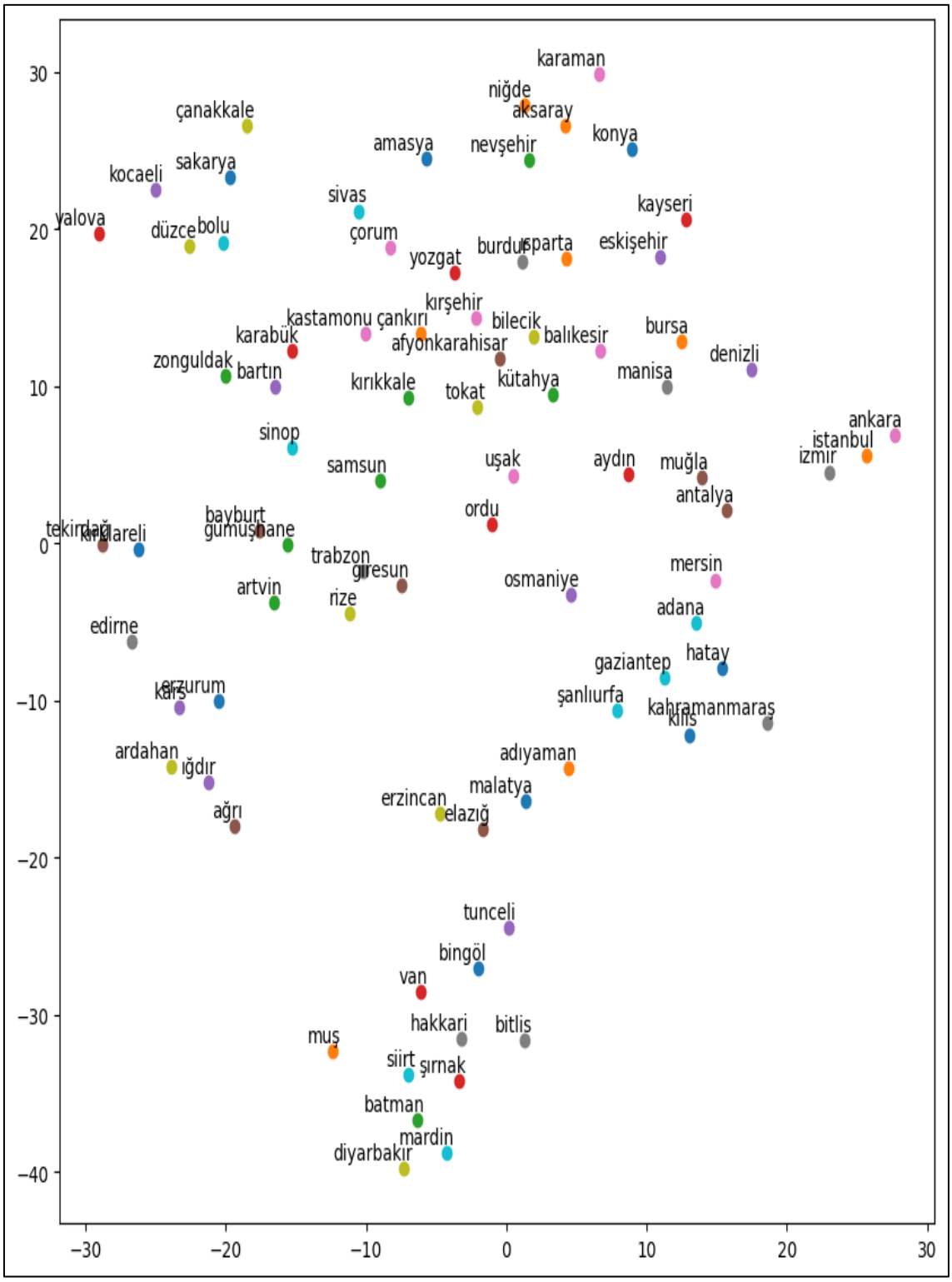
Türkçede il adlarının birlikte kullanıldığı kalıplar da illerin kelime benzerliklerinin yüksek çıkmasına neden olabilir. Edirne ilinin en yakın kelime benzerlikleri ('kars', 0.6297), ('tekirdağ', 0.5979), ('kırklareli', 0.5861), Kars ilinin en yakın kelime benzerlikleri ise ('erzurum', 0.6836), ('ardahan', 0.6568), ('edirne', 0.6297) olarak çıkmıştır. Komşuluk olarak en uzak il ikillerinden birinin birbirinin anlam olarak en yakın illeri olarak çıkmasının nedeni Türkiye'nin tümünden ya da bir ucundan diğer ucundan bahsedilirken kullanılan "Edirne'den Kars'a kadar" kalıbının siyasiler tarafından sıklıkla kullanılmasından kaynaklandığı değerlendirilmiştir.

Kelime vektörleri görselleştirilmek için 2 boyuta ya da 3 boyuta indirgenir. Bu durumda bilgi kaybı ya da yanlış bilgilere sebep olabilir. Şekil 4.37'de 2 boyuta indirgenmiş kelime vektörleri 2 boyutlu vektör uzayında gösterilmiştir. Türkçedeki "Edirne'den Kars'a kadar" kalıbından dolayı Edirne ve Kars kelimesinin birbirine benzer çıktığı belirtilmişti. Kelime vektörleri 2 boyuta indirgenirken bu kelimelerin dâhil oldukları kümeler de birbirine yaklaştırılmıştır. Sonuçta 300 boyutlu kelime vektörlerinin benzerliklerine göre oluşan Kars'ın dâhil olduğu *'erzurum', 'ardahan', 'edirne', 'ığdır', 'ağrı', 'gümüşhane'* kümesi ile Edirne'nin dâhil olduğu 'kars', 'tekirdağ', 'kırklareli' kümesi birbirine yaklaşmış, bu iki kümeye dâhil olan iller 300 boyutlu vektörlerde uzak olduğu halde 2 boyutlu vektörlerde birbirine yakın çıkmışlardır. Edirne'ye yakın olan 300 boyutlu Tekirdağ kelime vektörünün benzerlikleri *'kırklareli', 'edirne', 'bursa', 'bilecik'* iken 2 boyutlu Tekirdağ kelimesinin benzerlikleri sıra ile *'gümüşhane', 'kırklareli', 'bayburt', 'trabzon', 'artvin', 'edirne'* şeklindedir. Gümüşhane 300 boyutlu kelime vektörlerine göre Tekirdağ'a en yakın anlamlı 50. il iken 2 boyutlu kelime vektörlerine göre birinci en yakın ildir. Kars ise en yakın 47. il iken 2 boyutlu vektörlerde 10. en yakın il olmuştur. Bu sonuçlara göre 2 boyutlu kelime

vektörlerinin görselleştirmesinde oluşan kavram kümeleri kendi içinde tutarlı iken kümeler arasındaki görselleştirme yanıltıcı olabilir ve tutarsızlıklar oluşabilir.



Şekil 4.36. İller arasında kelime benzerlikleri



Şekil 4.37. İllerin 2 boyutlu kelime vektör uzayında görünümü

4.3.5. Ülkelerin kelime benzerliği

Ülkelerin kelime benzerliklerinin incelemek için 148 ülkeden 11 325 adet ülke ikilisi oluşturularak bu ikililer arasındaki kelime benzerliği ölçülmüştür. Çalışmada kelime n-gramları yerine bir kelimelik bölütler kullanıldığı için iki kelimeden oluşan ülke isimlerinde iki kelimeden ülkeyi ayırt edici olan tercih edilmiştir. Örneğin Yeni Zelanda için "zelanda" bölütü kullanılmıştır. Şekil 4.38 en yüksek benzerliğe sahip 50 kelime ikilisini Şekil 4.38 ise 300 boyutlu kelime vektörlerinin 2 boyutlu düzlemdeki dağılımını göstermektedir.

En yüksek benzerliğe sahip 50 ülkenin gösterildiği Şekil 4.38 komşu ülkelerin ya da aynı coğrafi bölgede bulunan ülkelerin benzerliklerinin arttığını göstermektedir. Şekilde Litvanya-Letonya, Kırgızistan-Özbekistan, Mozambik-Burkina, Brezilya-Arjantin gibi farklı kıtalardan ve bölgelerden ülke ikililerinin benzerlikleri yüksektir.

Ülkelerin kelime benzerliğinden çıkan sonuçlardan birisi de aralarında uzun süreli anlaşmazlığın ya da sıcak savaşın olduğu ülkelerin yüksek kelime benzerliğine sahip olduğudur. Sorun yaşayan ülkeler genellikle komşu ülkelerdir ve komşuluk ilişkisinden dolayı benzerliklerinin yüksek çıkması beklenen bir durumdur. Ancak bu ülkeler diğer komşuluk ilişkisine sahip oldukları ülkelere göre aralarında daha yüksek benzerlik değerleri taşırlar. Bu durum savaşa ya da soruna konu ülkelerin isimlerinin ikili halde kullanılmasından kaynaklanmaktadır. İsrail-Filistin sorunu, Rusya-Ukrayna Savaşı, Azerbaycan-Ermenistan anlaşmazlığı gibi kullanımlar bu kelime ikililerin yerel bağlamda birlikte geçme olasılığını artırmakta ve bu durum kelime benzerlik sonuçlarına yansımaktadır.

İsrail kelimesine en yakın anlamlı kelimeleri ('filistin', 0,6203), ('ermenistan', 0,6025), ('suriye', 0,5693), ('rusya', 0,5578), ('abd', 0,5534) ve Filistin kelimesinin en yakın anlamlı kelimeleri ('israil', 0,6203), ('keşmir', 0,5197), ('suriye', 0,511), ('myanmar', 0,5107) kelimeleridir. Bu iki ülkelerin yaşadıkları sorundan dolayı benzerlik değerlerinin artması sonuçlara yansımıştır. Bu sonuçlarda ki diğer bir ortak nokta ise Filistin'de olduğu gibi Keşmir ve Myanmar'da da Müslüman nüfusun kimliklerinden dolayı egemen devletler ile sorun yaşamasıdır.

Ermenistan'ın en yakın benzerlik gösterdiği ülkeler ('azerbaycan', 0,6693), ('israil', 0,6025), ('rusya', 0,5781), ('gürcistan', 0.5633), ('yunanistan', 0,5573), Azerbaycan'ın ise ('ermenistan', 0,6693), ('özbekistan', 0,5918), ('türkmenistan', 0,5894), ('gürcistan', 0,5796), ('kırgızistan', 0,578), ('kazakistan', 0,5532), ('rusya', 0,5465) olarak çıkmıştır. Ermenistan komşuları Rusya ile 0,58, Gürcistan ile 0,56 benzerliğe sahipken Azerbaycan ile 0,67 gibi çok daha yüksek benzerliğe sahiptir. Bu şablon Azerbaycan için de geçerlidir. Azerbaycan ve Ermenistan'ın benzerliklerinin diğer komşu ülke benzerliklerinin çok üstünde çıkması bu ülkenin aralarındaki komşuluk ilişkisinden çok anlaşmazlıktan dolayı derlemde sıklıkla aynı bağlamda geçtikleri şeklinde değerlendirilmiştir.

Rusya-Ukrayna Savaşı bu iki ülkeyi vektör uzayında birbirlerine yaklaştırmıştır. Ukrayna'nın benzer ülkeler sıralaması ('rusya', 0,7038), ('gürcistan', 0,575), ('belarus', 0,5725), Rusya'nın sıralaması ise ('ukrayna', 0,7038), ('iran', 0,6892), ('gürcistan', 0,6004), ('yunanistan', 0,5821), ('ermenistan', 0,5781) olarak çıkmıştır. Bu iki ülkenin birbirleriyle olan benzerlik değerlerinin yüksekliği ve bu değerin diğer komşularla olan benzerliklerden farkı aralarındaki yüksek benzerlik değerinin kaynağının Rusya-Ukrayna Savaşı ve anlaşmazlığıdır.

Kamboçya, benzerlik gösterdiği ülkeler listesi coğrafi konumuna göre en çok çeşitliliğe sahip ülkelerdir. Benzerlik sırasına göre ülkeler ('tayland', 0,775), ('nikaragua', 0,7677), ('nijerya', 0,7566), ('guatemala', 0,7467), ('bangladeş', 0,7443), ('zambiya', 0,7424), ('endonezya', 0,7414), ('bolivya', 0,741), ('nepal', 0,7403), ('mozambik', 0,7324), ('şili', 0,7304) şeklindedir. Komşuluk gösterdiği ülkeler bulunmak ile beraber Asya, Afrika, Güney ve Orta Amerika kıtasından ülkeleri ile de yüksek benzerliğe sahip olmasının ekonomik ve sosyal benzerliğinden kaynaklandığı değerlendirilmiştir.

Aralarındaki uzaklığa rağmen birbirine çok yakın çıkan iki ülke Kanada ve Avustralya'dır. Kanada'nın en benzer ülkeleri ('avustralya', 0,7538), ('izlanda', 0,6819), ('ingiltere', 0,6704), ('norveç', 0,66) Avustralya'nın benzer ülkeleri ise ('Kanada', 0,7538), ('izlanda', 0,6459), ('zelanda', 0,6372), ('hindistan', 0,6366), ('ingiltere', 0,6151), ('norveç', 0,6033) olarak sıralanmışlardır. Bu iki ülke birçok ortak özelliğe sahiptir. Her iki ülke de eski İngiliz kolonileridir ve İngiliz Milletler Topluluğu üyesidir. Her iki ülkenin dili de İngilizcedir. Her ikisi de çok kültürlü ve çeşitlilik arz eden kozmopolit toplumlardır, önemli ölçüde göçmen nüfusa sahiptirler. Her iki ülke de güçlü ekonomilere sahiptir ve büyük ekonomilerden

oluşan G20 grubunun üyesidir. Bu iki ülkenin kelime benzerliğini yüksek çıkması kelime yerleştirme algoritmalarının ülkelerin benzerliğini coğrafi komşuluk dışında kültürel, sosyal ve ekonomik durumlarına göre de oluşturduğunu gösterir.

Şekil 4.38'de ülkeler benzerliklerine göre kavram kümeleri oluşturmuşlardır. Nüfuzlarına göre en başat Avrupa ülkeleri İngiltere, Almanya, Fransa, İngiltere, İspanya ve İtalya bir küme oluşturmuş, bu kümeye yakın nüfus olarak orta büyüklükte refah düzeyi olarak gelişmiş Avrupa ülkeleri Danimarka, Belçika, Hollanda, İsviçre, Avusturya, Norveç, İsveç ise başka bir kavram kümesi oluşturmuşlardır.

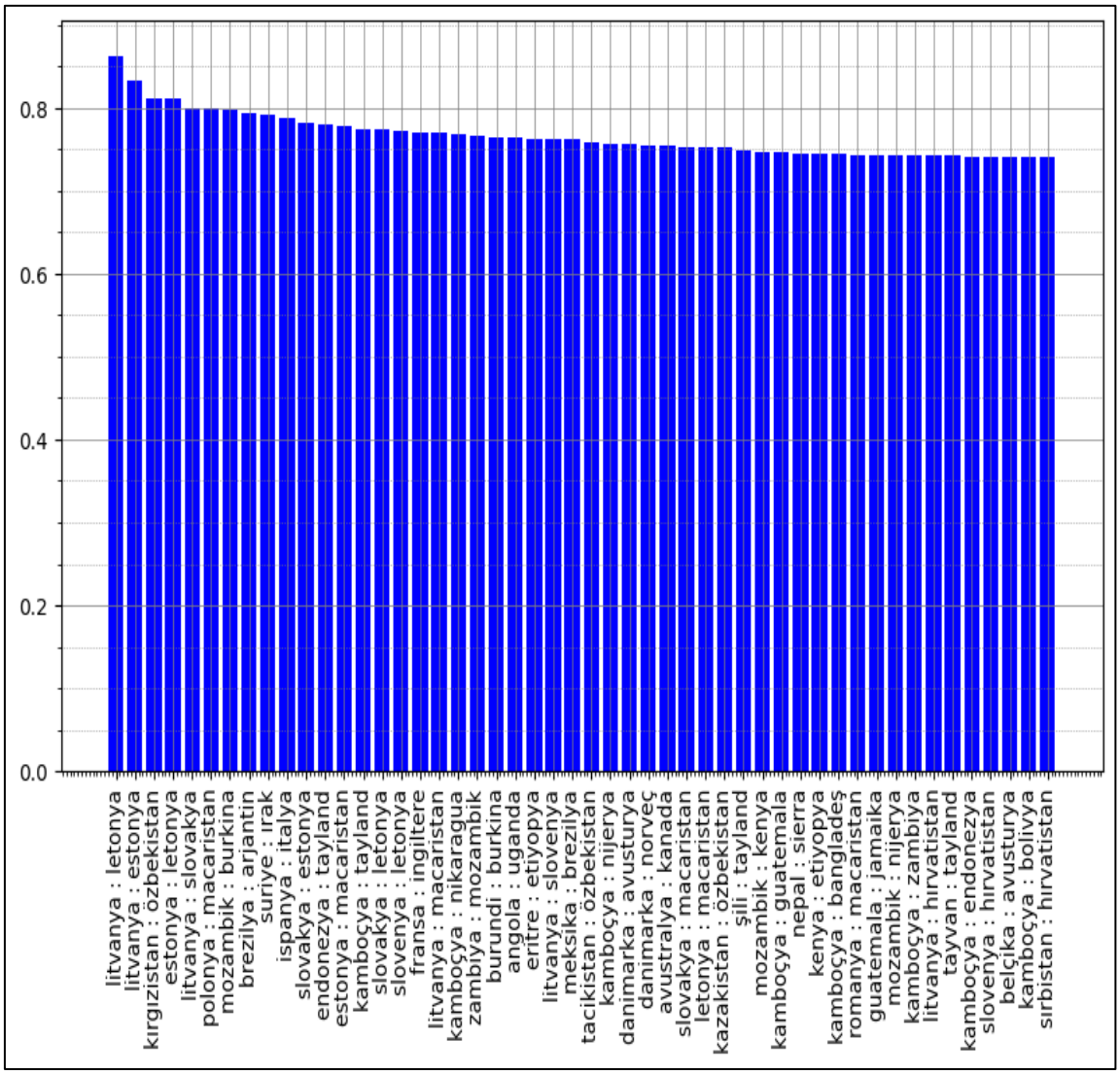
Özbekistan, Tacikistan, Kırgızistan, Türkmenistan, Kazakistan'ın oluşturduğu kümesi ise Türk dünyası kavram kümesi olarak değerlendirilmiştir.

Kuveyt, Bahreyn, Birleşik Arap Emirlikleri (Emirlik) ise Arap Dünyası Kümesini oluşturmuşlardır. Brezilya, Arjantin ve Meksika Şekil 4.38'de bir araya gelerek ayrı bir küme oluşturmuşlardır.

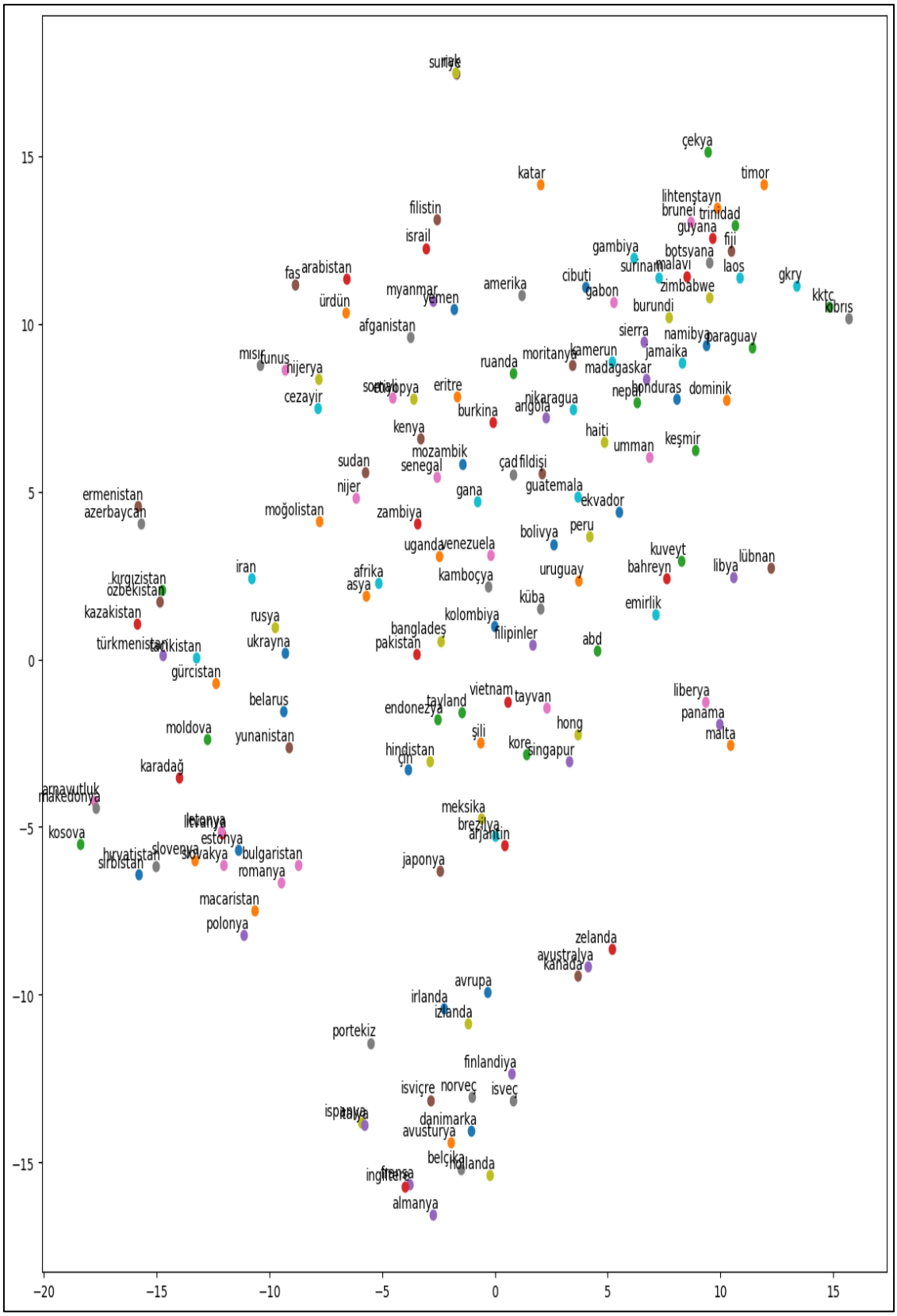
Kıbrıs adasını temsil eden Kıbrıs, Güney Kıbrıs Rum Kesimi (GKRY), Kuzey Kıbrıs Türk Cumhuriyeti (KKTC),'nin birbirleriyle benzerlikleri yüksektir ve Şekil 4.39'da ayrı bir kümeyi oluşturmuşlardır.

Irak ve Suriye, Letonya ve Litvanya, İspanya ve İtalya, Arnavutluk ve Makedonya, İngiltere ve Fransa kelime benzerliklerinin yüksek olması ve 2 boyutlu vektör uzayında çok yakın konumlanmaları nedeni ile Şekil 4.39'da görselleştirme zorlaşmış ve nerdeyse birbirlerinin üstünde çıkmışlardır.

Şekil 4.38'de Singapur, Hong-Kong, Tayvan, Vietnam, Tayland, Endonezya, Filipinler ise Güneydoğu Asya Ülkeleri kümesini oluşturmuşlardır. Çin, Hindistan ve Bangladeş Pakistan ikililerinden oluşan kümeler Güneydoğu Asya Ülkeleri kümesine yakındır.



Şekil 4.38. Ülkeler arasında kelime benzerlikleri



Şekil 4.39. Ülkelerin 2 boyutlu kelime vektör uzayında konumları

4.3.6. Kabine üyelerinin kelime benzerliği

Çalışmada kelime vektörleri ve kelime benzerliği kullanılarak 1994 ile 2023 yılları arasında hükümetlerde görev yapan kabine üyelerinin birbirleri ile ve bakanlık yaptığı konu ile ilişkileri kelime benzerliği kullanılarak incelenmiştir. Derlemde bakan isimleri adı ve soyadlarının birleşiminden oluşan bölütlerle temsil edilmiştir.

Derlem 1994-2023 yılını kapsadığı için kabine üyeleri bu tarihlerde görevde olan 52.-66. Türkiye Cumhuriyeti hükümetleri kabine üyelerinden ve başkanlarından oluşur. 52-65. Hükümetler arasında devlet bakanları ve başbakan yardımcıları, ilgili olduğu görev alanı ile ilgili sonradan bakanlık ihdas edilmişse, ilgili bakanlık ile ilişkilendirmişlerdir. Örneğin 61. Hükümet ile birlikte Aile ve Sosyal Politikalar Bakanlığı, Gençlik ve Spor Bakanlığı kurulmuştur. Daha önceki hükümetlerde bu bakanlıklar devlet bakanları tarafından yürütülmüştür. Görev alanı aile olan Devlet Bakanaları Işılay Saygın ve Nimet Çubukçu Baş çalışmada aile bakanlığını temsil eden *aile* kavramı ile ilişkilendirilmiştir. Derlemde devlet bakanlarının ve başbakan yardımcılarının görev alanları çeşitlilik gösterdiği ve belli bir kavrama atanamadıkları için kullanılmamışlardır. Bakanların isimleri sonradan değişse bile derlemde bulunan haliyle kullanılmıştır. 66. Hükümetlerdeki Hazine ve Maliye Bakanlarının görev alanları önceki dönem hükümetlerinin ekonomi ve maliye bakanlarının görevlerini kapsadığı için *ekonomi* ve *maliye* kavramları ile ilişkilendirilmişlerdir. Çizelge 4.12 bakanlıkları ve bakanlıkların adından dolayı ilişkili olduğu kelimeyi veya kavramı göstermektedir.

Çizelge 4.12. Bakanlıklar ve ilişkili oldukları kavramlar

Bakanlık	Kavram	Bakanlık	Kavram
başbakan	başbakan	kültür bakanı	kültür
adalet bakanı	adalet	kültür ve turizm bakanı	kültür, turizm
aile ve sosyal politikalar bakanı	aile	maliye bakanı	maliye
aile, çalışma ve sosyal hizmetler bakanı	aile, çalışma	milli eğitim bakanı	eğitim
bayındırlık ve iskan bakanı	bayındırlık, şehircilik	milli savunma bakanı	savunma

Çizelge 4.12. (devam) Bakanlıklar ve ilişkili oldukları kavramlar

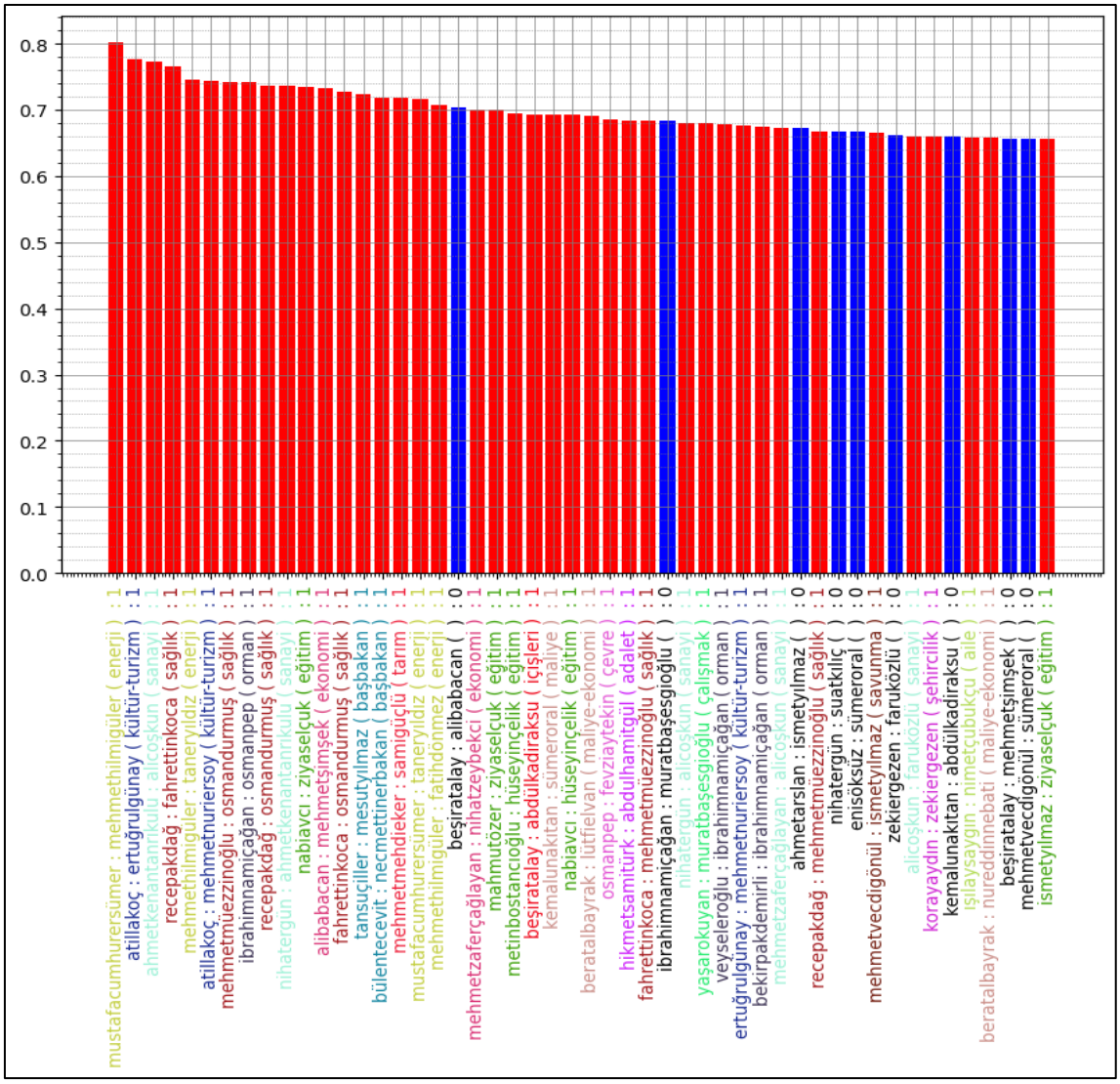
bilim, sanayi ve teknoloji bakanı	sanayi	orman bakanı	orman
çalışma ve sosyal güvenlik bakanı	çalışma	orman ve su işleri bakanı	orman
çevre bakanı	çevre	sağlık bakanı	sağlık
çevre ve orman bakanı	çevre, orman	sanayi ve ticaret bakanı	sanayi
çevre ve şehircilik bakanı	çevre, şehircilik	sanayi ve teknoloji bakanı	sanayi
dışişleri bakanı	dışişleri	tarım ve köyişleri bakanı	tarım
ekonomi bakanı	ekonomi	tarım ve orman bakanı	tarım, orman
enerji ve tabii kaynaklar bakanı	enerji	turizm bakanı	turizm
gençlik ve spor bakanı	spor	ticaret bakanı	ticaret
gıda, tarım ve hayvancılık bakanı	tarım	ulaştırma bakanı	ulaştırma
gümrük ve ticaret bakanı	ticaret	ulaştırma, denizcilik ve haberleşme bakanı	ulaştırma
içişleri bakanı	içişleri	ulaştırma ve altyapı bakanı	ulaştırma

Aynı kabine görevi yürüten bakanların kelime vektörü uzayında benzerlik etkilerini araştırmak için bakanların birbirleriyle olan kelime benzerlikleri ölçüldü. 82 adet kabine üyesinden tekrarlanmayan ikili kombinasyon kullanılarak $C_r^n = C_2^{82} = 3320$ adet bakan çifti oluşturuldu. 3320 çiftten 221 tanesinin konularına göre ortak kabine üyelikleri bulunmaktadır.

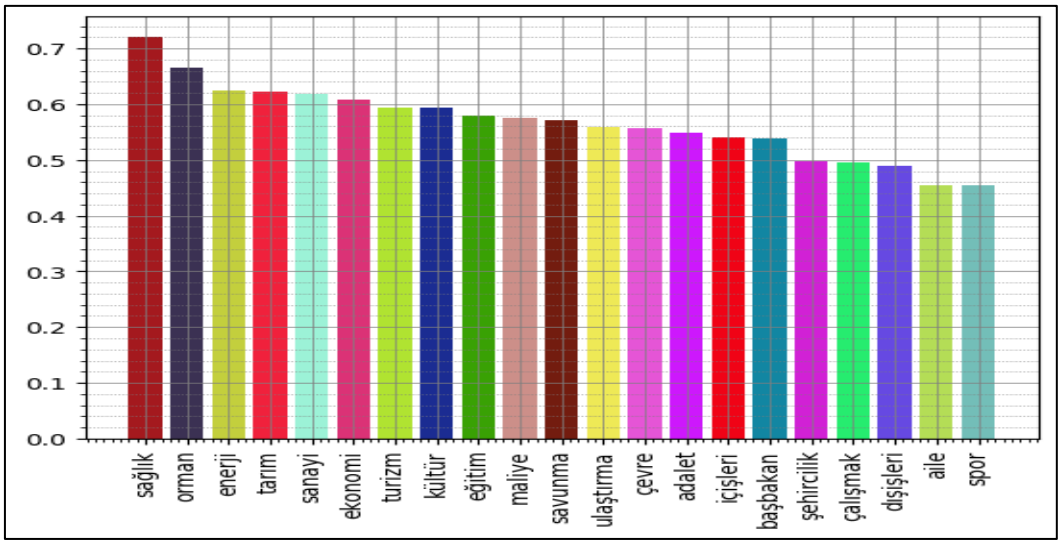
Aynı kabine görevine sahip ve sahip olmayan bakan çiftlerinin kelime benzerliğine göre karşılaştırmak için istatiksel iki örneklem ile bağımsız t-testi kullanıldı. T-testi, iki grubunun ortalamalarının birbirinden önemli ölçüde farklı olup olmadığını belirlemek için kullanılan istatistiksel bir testtir.

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}} = \frac{0{,}43 - 0{,}56}{\sqrt{\left(\frac{0{,}1^2}{3169}\right) + \left(\frac{0{,}1^2}{234}\right)}} = -18{,}69 \qquad (4.2)$$

Eş 4.2'de t, t-testi sonucu oluşan değeri, $\overline{x_1}$ aynı görevi bulunmayan kabine üyesi ikililerinin benzerlik ortalamasını (0,43), s_1 standart sapmasını (0,10), n_1 sayısını (3169) gösterir. Aynı göstergeler ortak görev yapan kabine üyesi ikilileri için $\overline{x_2}$ (0,56), s_2 (0,10) ve n_2 234'tür. 2 numaralı eşitliğe göre T-testinin p değeri (2,101e-50) 0,05'den küçüktür. Sonuç olarak kelime benzerliği açısından aynı kabine görevi yapan üyeler arasında yapmayanlara göre istatistiksel olarak anlamlı bir fark olduğunu gösterir. Bu sonuç, ortalamalar arasındaki farkın %95 güvenle şansa bağlı olmadığını, aksine gruplar arasında gerçek bir fark olduğunu göstermektedir.



Şekil 4.40. Kabine üyelerinin benzerlik değerleri



Şekil 4.41. Aynı görevi yapan kabine üyelerinin ortalama benzerlikleri

Şekil 4.40'da derlemde 82 kabine üyesinden oluşturulan 3320 bakan çiftinin kelime benzerliklerine göre sıralandığında en yüksek kelime benzerliğine sahip 50 bakan çiftinin çubuk grafiği görülmektedir. Kırmızı çubuklar aynı görevde bulunmuş kabine üyelerini mavi çubuklar farklı görevde bulunmuş kabine üyelerini gösterir. En yüksek kelime benzerliği Cumhur Ersümer ile Hilmi Güler arasındadır. Yüksek kelime benzerliğine sahip ikililer 'mustafacumhurersümer - mehmethilmigüler (enerji, 0,8019)', 'atillakoç : ertuğrulgünay (kültür-turizm, 0,7769)', 'ahmetkenantanrıkulu : alicoşkun (sanayi, 0,7728)', 'recepakdağ : fahrettinkoca (sağlık, 0,766)', 'mehmethilmigüler : taneryıldız (enerji, 0.7458)', 'atillakoç : mehmetnuriersoy (kültür-turizm, 0,7439)' olarak çıkmıştır.

Şekil 4.41'de derlem boyunca (1994-2023) yılları arasında aynı konuda bakanlık yapmış kabine üyelerinin benzerliklerinin ortalamaları konulara göre görülmektedir. Örneğin bu yıllarda sağlık bakanlarının toplam benzerliğinin bakan sayısına bölümü 0,72'dir. Şekil 4.40 ve Şekil 4.41'de sonuçlar incelendiğinde sağlık, orman ve enerji bakanlarının kendi aralarında kelime benzerliklerinin fazla olduğu görülebilir.

Aynı görevi yapmadığı halde yüksek benzerliğe sahip ikililer 'beşiratalay : alibabacan (0,7045)', 'ibrahimnamiçağan : muratbaşesgioğlu (0,6829)', 'ahmetarslan : ismetyılmaz (0,6719)', 'nihatergün : suatkılıç (0,6667)', 'enisöksüz : sümeroral (0,6665)', 'zekiergezen : faruközlü (0,661)', 'kemalunakıtan : abdülkadiraksu (0,6594)', 'beşiratalay : mehmetşimşek (0,6572)' şeklindedir. Sonuçlara göre bakanların görev aldığı kabineler incelendiğinde,

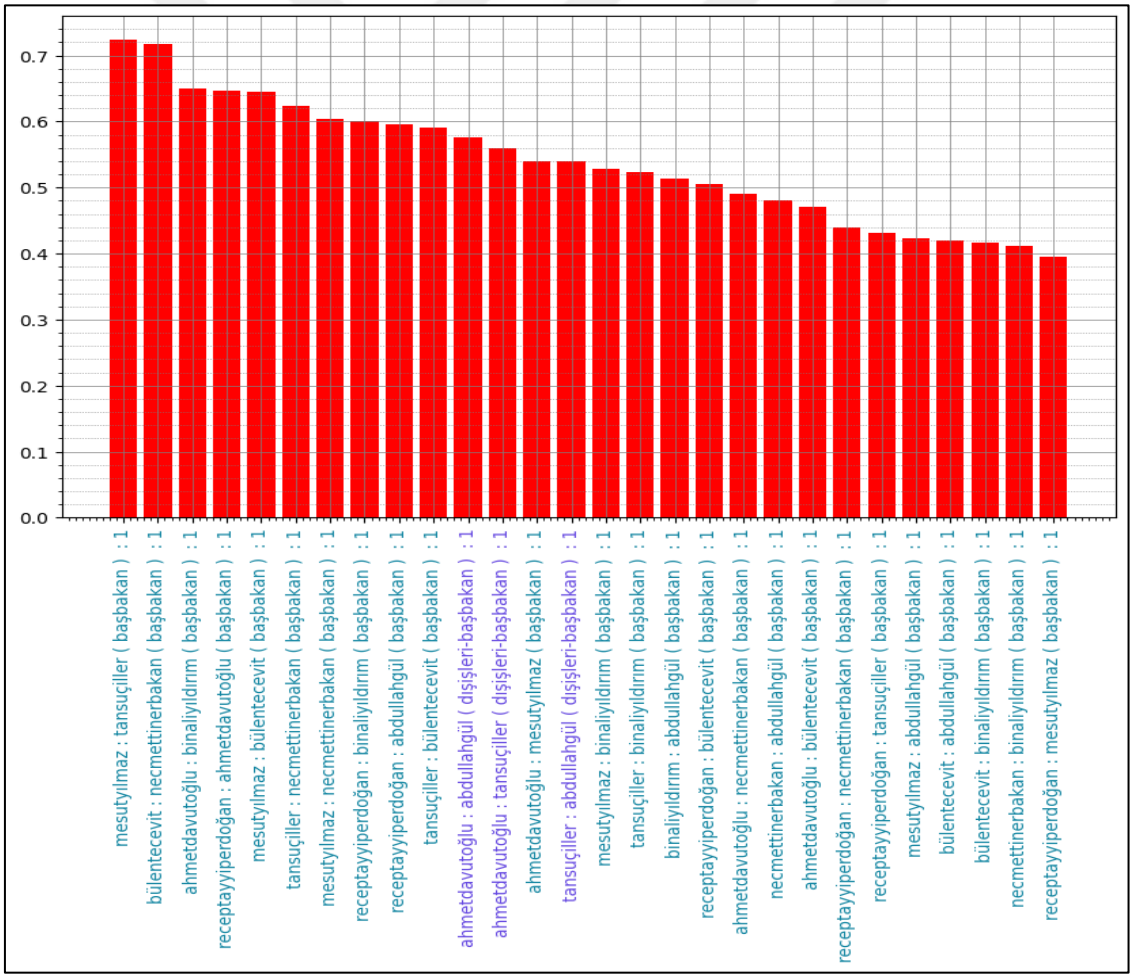
ikililerin aynı kabinede görev alan bakanlar olduğu bulunmuştur. Bu durum aynı görevi yapan bakanların yanında aynı kabinede görev almanın da kabine üyelerinin benzerliklerini artırdığını göstermiştir.

Aynı görevi yaptığı halde düşük benzerlik değerlerine sahip ikililer 'mevlütçavuşoğlu : tansuçiller (dışişleri, 0,3409)', 'faruknafızözak : mustafademir (şehircilik, 0,3274)', 'faruknafızözak : mehmetözhaseki (şehircilik, 0,3265)', 'zehrazümrütselçuk : ışılaysaygın (aile, 0,3191)', 'faruknafızözak : erdoğanbayraktar (şehircilik, 0,3141)' olarak çıkmıştır. Faruk Özak 59. Hükümette Bayındırlık ve İskân Bakanı, 60. Hükümette ise Spordan sorumlu Devlet Bakanlığı yapmıştır. Tansu Çiller ise 52. hükümette başbakan, 54. hükümette ise dışişleri bakanıdır. Zehra Zümrüt Selçuk ise 66. Hükümette Aile, Çalışma ve Sosyal Hizmetler Bakanlığı yapmıştır, yani aile bakanlığı ve çalışma bakanlığını aynı anda yürütmüştür. Benzerlik değerlerinden kişilerin birden çok konuda görev almalarının aynı görevi aldığı diğer bakanlarla kelime benzerliğinin azalmasına neden olduğu değerlendirilmiştir.

Birden farklı kabine görevi yapan bakanların kelime benzerliğine etkisi Binali Yıldırım'ın kelime benzerlik değerlerinden görülebilir. Binali Yıldırım 65. hükümetin başbakanıdır, 58-61. ve 64. Hükümette ulaştırma bakanlığı yapmıştır. Binali Yıldırım ayrıca 12 Temmuz 2018- 19 Şubat 2019 tarihleri arasında TBMM Başkanlığı yapmıştır. Çok uzun süre ulaştırma bakanlığı yaptığı için Binali Yıldırım'ı temsil eden kelime vektörü vektör uzayında ulaştırma bakanları ve başbakanlar ile yakın çıkmıştır. Binali Yıldırım'ın en yakın kelime benzerlik değerleri ('ahmetdavutoğlu', 0,6491), ('ismetyılmaz', 0,6154), ('receptayyiperdoğan', 0,6), ('ahmetarslan', 0,5953), ('lütfielvan', 0,5606) şeklindedir. Binali Yıldırım, Ahmet Davutoğlu ve Recep Tayyip Erdoğan ile Başbakanlık, İsmet Yılmaz ile TBMM Başkanlığı, Ahmet Arslan ve Lütfi Elvan ile Ulaştırma Bakanlığı görevi nedeni ile yüksek kelime benzerliğine sahiptir.

Şekil 4.42'de hükümet başkanlarının benzerlikleri incelendiğinde birbirine yakın tarihlerde başbakanlık yapmış liderlerin daha yüksek kelime benzerliğine sahip oldukları gözlemlenmiştir. Kelime benzerliğini değerlendirirken dikkat edilmesi gereken husus word2vec kelime benzerliği algoritmasının bağlam-tabanlı olduğudur. Bu nedenle kelime benzerliği söz konusu kavramların benzer bağlamlarda sıklıkla beraber kullanıldığını gösterir. Siyasi liderler söz konusu olduğunda yüksek kelime benzerliğine sahip kişilerin

birbiriyle aynı doğrultuda siyaset yaptıklarını, benzer fikirde olduklarını göstermez, bunun yerine aynı bağlamda kullanıldıklarını gösterir. Çalışmada 52. Hükümetin başkanı Tansu Çiller ile 53 ve 55. hükümetin Başbakanı Mesut Yılmaz siyasi hayatları boyunca en çok anlaşmazlığa sahip liderler iken derlemde en yüksek benzerlikle ölçülmüştür. Bu iki lider merkez sağı yani aynı siyasal tabanı temsil etmesi ve aynı dönemde genel başkanlık yapmaları nedeni ile aynı bağlamda sıklıkla geçmişlerdir. Benzerliği artıran bir diğer etkende yakın zamanda aynı görevi yürüten kişilerin ortak konulara yani benzer bağlamlara konu olmasıdır. Siyasi, ekonomik ve sosyal gündem zaman içinde değişir. Bunun doğal bir sonucu ise olarak aynı tarihte veya birbirine yakın tarihte görev alan kişilerin bağlamlarının benzemesidir. Şekil 4.42'deki benzerlik değerleri bu durumu doğrulamıştır. Birbirine yakın tarihte görev alan başbakan ikililerinin kelime benzerliği daha fazladır.



Şekil 4.42. Hükümet başkanı ikililerinin benzerlik değerleri

Şekil 4.43'de 300 boyutlu kelime vektörlerini t-sne algoritması 2 boyuta indirgendikten sonra ki vektör uzayındaki dağılımı görünmektedir. Saçılım grafiğinde noktaların renkleri Şekil 4.41'deki çubuk grafikteki bakanlık renklerine göre renklendirilmiştir. Birden çok bakanlıkta bulunan kişi saçılım grafiğinde birden çok renk ile görselleştirilmiştir. Şekilde ortak görev alan bakanların 2 boyutlu vektör uzayında birbirine yaklaşarak kavram kümeleri oluşturmuşlardır.

Metin Bostancıoğlu, Hüseyin Çelik, Nabi Avcı, İsmet Yılmaz, Ziya Selçuk birbirlerine yaklaşarak eğitim kavram kümesini oluşturmuşlardır.

Ekonomi ve maliye bakanlıkları birlikte değerlendirildiğinde altı bakan bir ekonomi kümesi oluşturmuşlardır. Bu kümenin kendi içinde iki alt küme oluşturduğu gözlemlenmiştir. Ali Babacan, Mehmet Şimşek ve Naci Ağbal oluşan bir grubu, Berat Albayrak, Lütfi Elvan ve Nureddin Nebati ikinci bir grubu oluşturmuşlardır. Bu bakanların görev yaptıkları tarihler dikkate alındığında bu farklılaşmaların görev yaptıkları tarihlere göre oluştuğu anlaşılır. Maliye Bakanlarından oluşan Sümer Oral ve Kemal Unakıtan ise bu kümeye uzak ayrı bir ikili oluşturmuşlardır.

Hikmet Sami Türk, Bekir Bozdağ ve Abdulhamit Gül Şekil 4.43'de adalet bakanlığını temsil eden bir küme oluşturmuşlardır.

Çalışma Bakanı Faruk Çelik, Sanayi Bakanı Nihat Ergün, Spor Bakanı Suat Kılıç ve Adalet Bakanı Sadullah Ergin ortak kabine görevleri olmamasına rağmen bir kavram kümesi oluşturmuştur. Sadullah Ergin 2003-2009, Faruk Çelik 2002-2007, Suat Çelik 2009-2011 yıllarında TBMM'de AK Parti grup başkanvekilliği yapmışlardır. Bu kişilerin yakın tarihlerde yaptıkları grup başkanvekilliği nedeni ile kelime benzerliklerinin yüksek çıktığı değerlendirilmiştir.

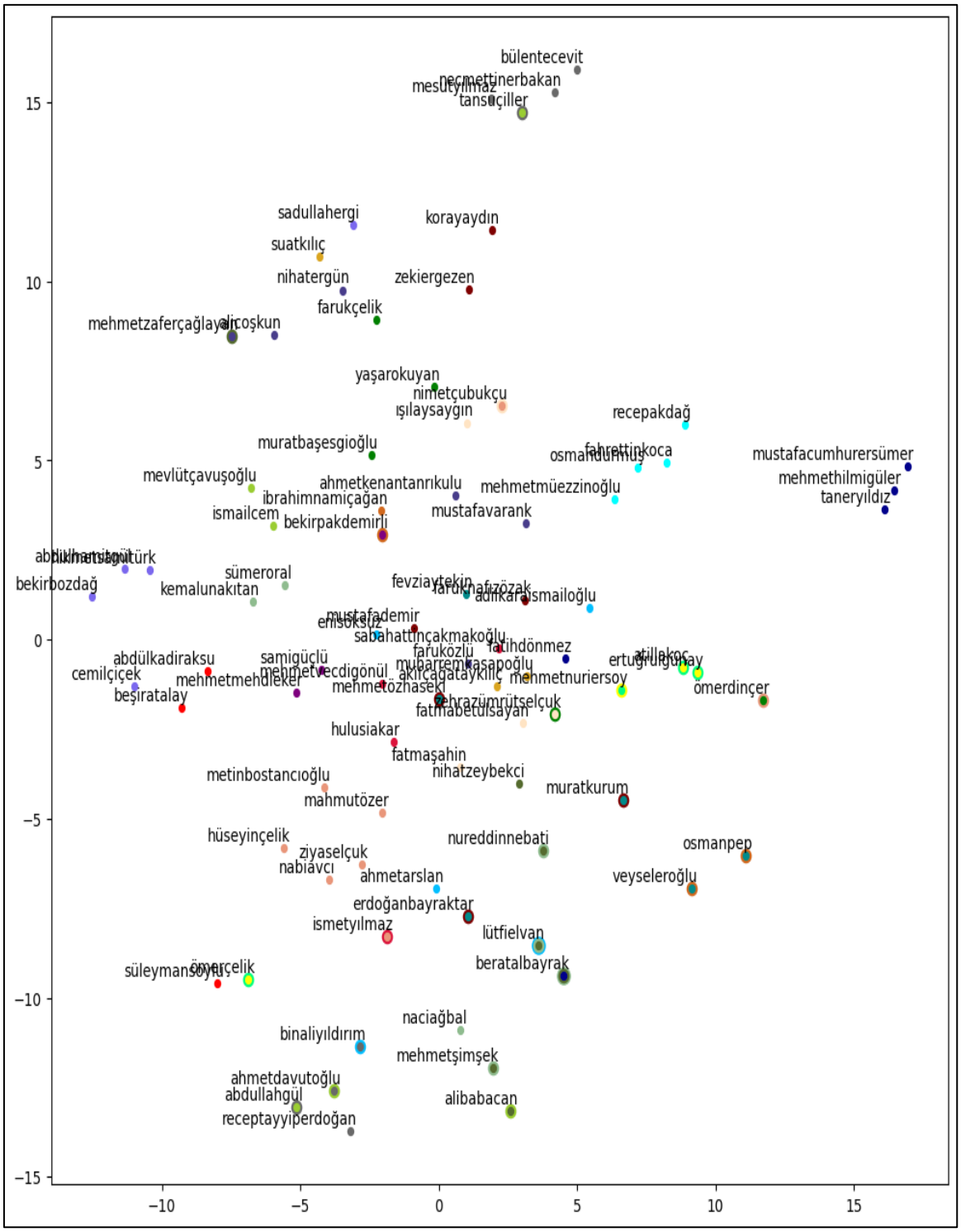
Metin Bostancıoğlu, Hüseyin Çelik, Nabi Avcı, Ziya Selçuk ve Mahmut Özer birbirlerine yaklaşarak eğitim kümesini oluşturmuşlardır.

AK Parti hükümetlerinin başkanları Recep Tayyip Erdoğan, Abdullah Gül, Ahmet Davutoğlu ve Binali Yıldırım bir kavram kümesi oluşturmuştur. AK parti öncesi hükümet başkanları Necmettin Erbakan, Tansu Çiller, Mesut Yılmaz ve Bülent Ecevit ise ayrı bir

başbakanlar kavram kümesi oluşturmuştur. Dışişleri Bakanları İsmail Cem ve Mevlüt Çavuşoğlu birbirine yakın ikililerdir.

Şekil 4.43'de Osman Durmuş, Recep Akdağ, Mehmet Müezzinoğlu ve Fahrettin Koca ise sağlık kümesini, Mustafa Cumhur Ersümer, Mehmet Hilmi Güler, Taner Yıldı ve Fatih Sönmez enerji kümesini oluşturur.

Çalışmada kabine üyelerinin benzerlikleri değerlendirilirken kabine üyesinin farklı hükümetlerde farklı bakanlık görevini yerine getirmesi, grup başkanvekilliği gibi parti görevlerinin olması vektör uzayında aynı görevi yerine getirdiği kişilerle yaklaştırır. Kabine üyeliğinde bulunduğu süre, aynı görevi yapan diğer kabine üyeleri ile aynı bağlamı paylaşma sıklığını etkiler. Bu koşullar kelime benzerliğini etkileyen fakat kontrol edilemeyen durumlardır ve çalışma için bir sınırlılık oluşturmuştur.



Şekil 4.43. Kabine üyelerinin 2 boyutlu kelime vektör uzayında konumları

5. SONUÇ VE ÖNERİLER

Parlamentoların ana işlevleri yasama, denetim ve temsildir. Bu işlevlere ilişkin işlerin görüşülüp tartışıldığı ve nihai olarak karara bağlandığı en üst karar organı genel kurul oturumlarıdır. Genel kurul oturumlarının transkript edilip metine dönüştürülmesi ile yıllar boyunca binlerce sayfayı bulan büyük metin verisi oluşur.

Çalışmada TBMM Genel Kurul tutanaklarındaki milletvekili konuşmalarından oluşan büyük metin verisinin hesaplamalı analizi yapılmıştır. Klasik makine öğrenmesi, derin öğrenme, Doğal Dil İşleme yöntemleri kullanılarak ham metin verisinden anlamlı bilgiler çıkarılmıştır.

Parlamento verilerinden kutuplaşmanın veya parti aidiyetlerinin ölçüldüğü çalışmalar şimdiye kadar takvim yılını esas almışlar ve her yıl bir veri noktası ile temsil edilmiştir. Bu yaklaşımda yıl içindeki kutuplaşmayı incelemek mümkün değildir. Bununla birlikte yıllık kutuplaşma ölçümleri veri noktalarının azlığından dolayı zaman serisi analizine imkân vermez.

Yoğun olarak verilerin üretildiği sosyal medya verileriyle bir yıldan daha az periyotları kapsayan veri kümeleri oluşturulabilir. Ancak parlamentoların çalışma şekli, üretilen konuşma sayıları göz önünde bulundurulduğunda bir ayı kapsayan veri kümelerini oluşturmak mümkün değildir. Parlamento görüşmelerinde kutuplaşma çalışmalarında oluşan bu soruna çözüm olarak her ay için geçmiş bir yıllık veri kümeleri ve metin sınıflandırması ile elde edilen 12 aylık hareketli kutuplaşma ölçütü kullanılmıştır. Çalışmada, kullanılan veri kümelerinin boyut olarak değişkenliğinin metin sınıflandırması üzerindeki etkisi arındırılarak nihai kutuplaşma ölçütü elde edilmiştir. Kullanılan bu yöntem ile bir yıl on iki veri noktası ile temsil edilebilmiş ve aylık olarak siyasi partilerin ilişkileri incelenebilmiştir.

TBMM'de siyasi partilerin kutuplaşması iki farklı zaman serisi ile incelenmiştir. 2011 ve 2023 arası dört parti, AK Parti, CHP, MHP ve HDP'nin oluşturduğu altı parti çiftine ait zaman serileri birinci analizi oluşturmuştur. İkinci analizde ise TBMM'de 27. Dönem (Ekim-2018, Mayıs-2023) İYİ Parti'nin katılımı ile beş parti ve on parti çifti ile incelenmiştir.

Her iki analizde HDP muhafazakâr sağ AK Parti ve milliyetçi sağ MHP ve İYİ Parti ile en yüksek kutuplaşma ölçütlerini oluşturmuştur. Birinci analizde DZB ile ölçülen kutuplaşma farkı en az olan iki parti çifti AK Parti-HDP iken ikinci analizde HDP'nin bu üç parti ile oluşturduğu parti çiftleridir. Bu sonuçlar milliyetçi ve sağ ideolojinin, HDP ile oluşan kutuplaşmanın ana etkeni olduğu değerlendirilmiştir ve ideolojilerin kutuplaşmaya etkisini göstermektedir.

27. dönemde oluşan 12 aylık kutuplaşma sonuçları ise seçim ittifaklarının kutuplaşma üzerindeki etkilerini göstermektedir. Millet ittifakının iki ana partisi CHP ve İYİ Parti kutuplaşma seviyesi dönem boyunca en az olan partidir. Cumhur ittifakının iki partisi AK Parti ve MHP ise dönem boyunca en küçük ikinci kutuplaşma seviyesine sahiptir.

12 aylık hareketle kutuplaşma sayesinde Granger nedensellik analizi yapmak mümkün olmuştur. 2011-2023 yılları temel alındığında AK Parti-CHP parti çifti AK Parti-HDP parti çiftinin kutuplaşmasının Granger nedenidir. Bu sonuç kutuplaşmanın ilk önce iktidar partisi ve ana muhalefet partisi arasında başladığını gösterir.

Çalışma sonucu olarak parlamento görüşmelerinde 12 aylık hareketli kutuplaşma ölçütünün kullanılmasının yıllar içinde gerçekleşen kutuplaşmanın detaylı analizinde kullanılabileceği, bu sayede kutuplaşma eğilimlerinin ve zaman serisi analizinin gerçekleştirilebileceği gösterilmiştir.

Demografik özelliklerin tahmini çalışmaları İngilizce dilinde sosyal medya verileri kullanılarak kapsamlı bir şekilde incelenmiştir; ancak Türkçe dilinde parlamento verileri kullanılarak yapılan çalışmalar sınırlıdır. Bu çalışmada, YPO alanını çeşitlendireceği düşünülen bir TBMM Genel Kurul görüşmeleri derlemi tanıtılmıştır. YPO çalışması meclis görüşmelerinde cinsiyet, yaş, eğitim durumu, meslek, seçim bölgesi, parti aidiyeti, parti durumu (iktidar/muhalefet) olmak üzere yedi özellik tahmini görevi ile gerçekleştirilmiştir.

Derin öğrenme yöntemlerinin, DDİ görevlerinin çoğunda yeni yaklaşımlar olduğu için klasik makine öğrenmesi algoritmalarına göre üstünlüğe sahip olduğu düşünülebilir. Ancak bu çalışmada klasik makine öğrenmesi algoritmaları İBSA ve BERT gibi derin öğrenme tekniklerinden daha iyi performans göstermiştir. Modellerin doğrulukları karşılaştırıldığında, TF-IDF_DVM ve TF-IDF_LR tüm özellik tahmini görevleri için en

yüksek doğruluklara sahiptir. DVM ve LR algoritmalarında stil tabanlı özellikler (karakter n-gramları) ve içerik tabanlı özelliklerin (kelime n-gramları) kombinasyonunu da kullanılmıştır. Yalnızca cinsiyet ve meslek tahmini görevleri doğruluklarını artırmıştır. Yazarın üslubunun sadece cinsiyet ve meslek tahmini üzerinde etkili olduğu söylenebilir.

Cinsiyet tahmininde elde edilen %82'lik doğruluk oranı, milletvekilinin cinsiyetinin konuşmaların içeriği ve tarzı üzerinde etkili olduğunu göstermektedir. Milletvekillerinin yaş aralığı %52 doğrulukla belirlenebilmektedir. Konuşma dili ve dokümanın içeriği milletvekillerinin yaşını yansıtmaktadır.

Çalışmada milletvekillerinin eğitim düzeyi %60 doğrulukla tahmin edilmiştir. Yaş ve eğitim seviyesi tahmininde sınıfların sırası hata matrisinde incelenmiştir. Yaş sınıflandırmasında yakın yaş aralıkları önemli oranlarda bir araya gelmektedir. Birbirine yakın yaş gruplarının konuşma tarzı ve içeriğinin birbirleri takip ettiği görülmüştür. Eğitim seviyesi sınıflandırmasında ise sınıfların sıralamasında bazı tutarsızlıklar olduğu gözlemlenmiştir: Örneğin, *doçent veya profesör*, *doktora* veya *yüksek lisansa* göre *lisansa* daha yakındır. Meslek tahmini görevindeki %67'lik doğruluk oranı, milletvekillerinin meslekleri ile konuşmaları arasında bir ilişki olduğunu göstermektedir.

Meslek sınıflandırmasındaki kategorilerin tahmin edilebilirlik değerleri, *mühendislik* ve *hukuk* için düşük *ekonomi ve finans* için ise yüksek olduğunu göstermektedir.

Milletvekillerinin seçim bölgeleri TBMM konuşmalarında %54 doğrulukla tespit edilebilmektedir. Sınıf karışıklıklarına (confussion) göre, Ege ve Marmara arasında ve Doğu Anadolu ve Güneydoğu Anadolu bölgeleri arasında bir yakınlık söz konusudur.

Parti üyeliği (%84) ve parti statüsü üyeliği (%92) doğruluk değerleri, milletvekillerin konuşmaları ve parlamenter aidiyetleri arasındaki yakın ilişkiyi göstermektedir. Bu doğruluklara göre, özellik tahmin görevlerinin en yüksek iki doğruluğu demografik alanlardan ziyade parlamenter alana aittir.

Milletvekillerinin bir özelliğinin diğeri üzerindeki etkisi için ikili sınıflandırma analizi kullanılmıştır. Bu analizde bir özelliğin her bir kategorisine ait konuşmalardan oluşan veri kümesinde başka bir özellik tahmini görevi yapılmıştır. Genç milletvekilleri, diğer yaş

gruplarına kıyasla mesleklerini daha fazla yansıtmaktadır. Genç milletvekilleri aynı zamanda cinsiyet özelliklerini de daha net gösterme eğilimindedir. Parti üyeliği ve cinsiyetin ikili sınıflandırmasında, merkez partilerin milletvekilleri konuşmalarında daha fazla cinsiyet eşitlikçi kelime kullanmaktadır. Eğitim durumları arasında yapılan analizde, parti üyeliğinin lisansüstü eğitim düzeyindeki milletvekilleri tarafından belirleyici bir şekilde yansıtıldığı görülmektedir.

Çalışmada metinlerin içerik analizi için konu kelimelerinden (sıfat ve isim) en terimler analizi yapılmıştır. Milletvekili konuşma içeriklerinde toplumsal cinsiyet konusunun somut sonuçları vardır. Kadın cinayetleri ile kadına ve çocuğa yönelik şiddet, milletvekillerinin konuşmalarından çıkarılabilecek en rahatsız edici sorunlardır. Meslek sınıflandırması, milletvekillerinin TBMM'de konuşurken meslekleriyle ilgili terimleri kullanmayı tercih ettiklerini göstermektedir.

Coğrafi bölgelerin sosyal ve ekonomik yapısı ve bölgesel sorunlar, seçim bölgesi sınıflandırmasının en güçlü terimlerinden çıkarılabilmektedir. Partilerin ideolojileri ve iktidar ya da muhalefet olma durumlarıyla ilgili terimler, milletvekillerinin parti aidiyetlerini ayırt etmek için önemlidir. Konu kelimeleri analizine göre iktidar milletvekilleri konuşmalarında daha geleneksel kelimeler kullanmaktadır.

Çalışmadan elde edilen kapsamlı deneysel değerlendirmeler: (1) klasik makine öğrenimi yöntemlerinin, TBMM Genel Kurul görüşmeleri kullanırken özellik tahmini görevlerinde derin öğrenme tekniklerinden daha başarılı olduğunu, (2) milletvekillerinin demografik özelliklerinden çok parlamenter özelliklerini gösterdiğini ve (3) kadına yönelik şiddet ve toplumsal cinsiyetin ana akımlaştırılmasının Türkiye'de önemli bir konu olduğunu göstermektedir.

TBMM Genel Kurulunda birbiri ile yakın anlamlı kelimeler ise kelime yerleştirme algoritmaları ile elde edilen kelime vektörlerinin benzerlik değerleri ile incelenmiştir. Kelime vektörlerinin kalitesi benzerlik ve analoji gibi dâhili (intrinsic) görevlerinden elde edilen başarım değerleri ile değerlendirilir.

TBMM Genel Kurul Tutanakları derleminde ortak morfolojiye sahip ek alarak türemiş kelime ikililerin benzerliğini bulmada fastText, GloVe ve word2vec’den daha yüksek

başarım elde etmiştir. Sondan eklemeli bir dil olarak Türkçe morfolojik olarak zengin bir dildir. FastText eğitim aşamasında alt-kelimeleri kullanması ve karakter tabanlı olması aynı köke sahip kelimelerin ek alarak türemiş bütün sürümlerini yakalamasını sağlar. FastText'in bu özelliği derlemde karşılaşılabilecek sözlük dışı kelime sorununa çözüm üretir. FastText eşanlamlılar ve türemiş kelimelerden oluşan veri kümesinde bütün kelimeler için kelime vektörleri oluşturarak, word2vec ve GloVe algoritmalarına göre sözlük dışı kelime sorunu için en iyi çözüm olduğun göstermiştir.

Benzer morfolojiye sahip olamayan eş anlamlı kelimelerin benzerliğinde is word2vec CBOW modeli diğer modellere göre daha iyi bir performans göstermiştir. Kelime analojilerinde için ise siyasi parti genel başkanları ile parti adları, siyasi parti grup başkanları ve parti adları arasındaki analojiler araştırılmıştır. Kelime analojilerini ortaya çıkarmada word2vec skip-gram modeli word2vec CBOW fastText ve GloVe'dan daha başarılıdır.

TBMM genel kurul tutanakları derleminde ülkeler, iller ve kabine üyelerinin isimleri morfolojik olarak farklıdırlar. Bu bölütler için kelime vektörleri karşılaştırılan 5 kelime yerleştirme algoritması tarafından da oluşturulmuştur ve sözlük dışı kelime sorunu gözlemlenmemiştir. Bu sebeple iller, ülkeler ve kabine üyelerinin kelime benzerliklerinin incelendiği çalışmada sözlük dışı kelime sorununa en iyi çözüm olarak çıkan fastText yerine kelime benzerliği ve analojilerde en yüksek başarımın elde edildiği word2vec modeli tercih edilmiştir.

Ülkeler ve illerin vektör uzayındaki yakınlıklarını en çok etkileyen etmen coğrafi komşuluk ilişkileridir. Komşuluk ilişkileri yanında il veya ülkelerin paylaştıkları deprem, afet, sosyal ve siyasi olaylar, savaş, ekonomik ve kültürel yakınlık gibi durumlar da aynı bağlam içinde geçmelerini sağlar. Örneğin çalışmada İstanbul ve Ankara, İstanbul ve İzmir coğrafi komşuluk ilişkisine sahip olmamalarına rağmen Türkiye'nin metropolleri olarak en yüksek kelime benzerliğine sahip iller arasındadır.

Kabine üyelerinin kelime benzerlikleri incelendiğinde ise aynı kabine görevini yerine getiren üyelerin kelime benzerliklerini arttığı görülmüştür. Bakanlıklara göre ortalama benzerlik değerleri karşılaştırıldığında Sağlık Bakanlarının benzerlikleri yüksektir. Sağlık ile ilgili terimlerin çokluğu ve bu terimlerin diğer bakanlıklardan ayrılması, ayrıca sağlık bakanlığının isminin bütün hükümetler boyunca değişmemesinin ortalama benzerlik

değerini artıran bir durumdur. Kelime benzerliği sonuçlarına göre aynı tarihli ya da yakın tarihli hükümetlerin kabine üyelerinin kelime benzerliğinin yüksektir.

TBMM Genel kurul tutanaklarını veri olarak kullanan çalışma; yapay zeka tabanlı yöntemlerle analizi ile ham metin verisinden değerli bilgiye ulaşmanın imkânını üç ayrı konuda yapılan analiz ile göstermiştir.

KAYNAKLAR

Abercrombie, G., Batista-Navarro, R. (2019). Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. *Journal of Computational Social Science*, 1-26.

Agun, H. V., Yilmazel, S., Yilmazel, O. (2017). *Effects of language processing in Turkish authorship attribution.* 2017 IEEE International Conference on Big Data (Big Data), 1876-1881.

Altszyler, E., Sigman, M., Ribeiro, S., Slezak, D. F. (2016). Comparative study of LSA vs Word2vec embeddings in small corpora: A case study in dreams database. *arXiv preprint* 178-187

Amasyalı, M. F., Diri, B. (2006). Automatic Turkish text categorization in terms of author, genre and gender. *Natural Language Processing and Information Systems: 11th International Conference on Applications of Natural Language to Information Systems, NLDB 2006, Klagenfurt, Austria, May 31-June 2, 2006. Proceedings 11*, 221-226.

Bahçeli, D. (2016). *Devlet Bahçeli, TBMM Grup Toplantısı Konuşması. 22 Kasım 2016*. URL: https://www.mhp.org.tr/htmldocs/genel_baskan/konusma/4154/index.html / Son Erişim Tarihi: 18.07.2023.

Baker, C. F., Fillmore, C. J., Lowe, J. B. (1998). The berkeley framenet project. *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Bartle, A., Zheng, J. (2015). Gender Classification with Deep Learning. *Stanfordcs, 224d Course Project Report*, 1-7.

Bengio, Y., Ducharme, R., Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, *13*.

Berndt, D. J., Clifford, J. (1994). Using dynamic time warping to find patterns in time series. *KDD workshop*, *10*(16), 359-370.

Bevendorff, J., Ghanem, B., Giachanou, A., Kestemont, M., Manjavacas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Wiegmann, M., Zangerle, E. (2020). Shared Tasks on Authorship Analysis at PAN 2020. *Advances in Information Retrieval*, 12036, 508-516.

Blaxill, L., Beelen, K. (2016). A feminized language of democracy? The representation of women at Westminster since 1945. *Twentieth Century British History*, 27(3), 412-449.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5, 135-146.

Boulis, C., Ostendorf, M. (2005). A quantitative analysis of lexical differences between genders in telephone conversations. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '*05, 435-442.

CHP Tüzüğü. (2018). URL: https://content.chp.org.tr/file/chp_tuzuk_10_03_2018.pdf / Son Erişim Tarihi: 18.07.2023.

Ciot, M., Sonderegger, M., Ruths, D. (2013). Gender inference of Twitter users in non-English contexts. *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1136-1145.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Collobert, R., Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th international conference on Machine learning*, 160-167.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F. (2011). Predicting the political alignment of twitter users. *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, 192-199.

Consortium, B. N. C. (2007). British national corpus. *Oxford Text Archive Core Collection*.

Dahllöf, M. (2012). Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches—A comparative study of classifiability. *Literary and linguistic computing*, 27(2), 139-153.

Deniz, A., Kiziloz, H. E. (2017). Effects of various preprocessing techniques to Turkish text categorization using n-gram features. *2017 International Conference on Computer Science and Engineering (UBMK)*, 655-660.

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

Dukic, D., Krzic, A. S. (2021). Detection of Hate Speech Spreaders with BERT. *CLEF (Working Notes)*, 1910-1919.

Dunn, J., Argamon, S., Rasooli, A., Kumar, G. (2016). Profile-based authorship analysis. *Digital Scholarship in the Humanities*, 31(4), 689-710.

Dwi Prasetyo, N., Hauff, C. (2015). Twitter-based election prediction in the developing world. *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, 149-158.

Dzieciątko, M. (2019). Application of text analytics to analyze emotions in the speeches. *Information Technology in Biomedicine: Proceedings 6th International Conference, ITIB'2018, Kamień Śląski, Poland, June* 18–20, 2018 6, 525-536.

Elmanarelbouanani, S., Kassou, I. (2014). Authorship Analysis Studies: A Survey. *International Journal of Computer Applications*, *86*(12), 22-29.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., De Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., De Macedo, L. D., Navarretta, C., Luxardo, G., Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57(1), 415-448.

Ertan, G., Çarkoğlu, A., Aytaç, S. E. (2022). Cognitive political networks: A structural approach to measure political polarization in multiparty systems. *Social Networks*, 68, 118-126.

Ertek, T. (2011). Hasan Âli Yücel ve Birinci Coğrafya Kongresi (1941). *Türk Coğrafya Dergisi*, 57, 11-19.

Eskişar, G. M. K., Çöltekin, Ç. (2022). Emotions Running High? A Synopsis of the State of Turkish Politics through the ParlaMint Corpus. *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, 61-70.

Estival, D., Gaustad, T., Pham, S. B., Radford, W., Hutchinson, B. (2007). Author profiling for English emails. *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics,* 263, 272.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., Smith, N. A. (2014). Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:*1411.4166.

Fatima, M., Anwar, S., Naveed, A., Arshad, W., Nawab, R. M. A., Iqbal, M., Masood, A. (2018). Multilingual SMS-based author profiling: Data and methods. *Natural Language Engineering*, 24(5), 695-724.

Fatima, M., Hasan, K., Anwar, S., Nawab, R. M. A. (2017). Multilingual author profiling on Facebook. *Information Processing & Management*, 53(4), 886-904.

Flores, A. M., Pavan, M. C., Paraboni, I. (2022). User profiling and satisfaction inference in public information access services. *Journal of Intelligent Information Systems*, 1-23.

Fraccaroli, N., Giovannini, A. (2020). *Central Banks in Parliaments: A Text Analysis of the Parliamentary Hearings of the Bank of England, the European Central Bank and the Federal Reserve* (SSRN Scholarly Paper 3646000).

Frid-Nielsen, S. S. (2018). Human rights or security? Positions on asylum in European Parliament speeches. *European union politics*, 19(2), 344-362.

Ganitkevitch, J., Van Durme, B., Callison-Burch, C. (2013). PPDB: The paraphrase database. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 758-764.

García-Díaz, J. A., Colomo-Palacios, R., Valencia-García, R. (2022). Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Generation Computer Systems*, 130, 59-74.

Garimella, V. R. K., Weber, I. (2017). A long-term analysis of polarization on Twitter. *Eleventh international AAAI conference on web and social media*.

Gentzkow, M., Shapiro, J., Taddy, M., diğerleri. (2016). *Measuring polarization in high-dimensional data: Method and application to congressional speech*.

Gladkova, A., Drozd, A., Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. *Proceedings of the NAACL Student Research Workshop*, 8-15.

Goet, N. D. (2019). Measuring polarization with text analysis: Evidence from the UK House of Commons, 1811–2015. *Political Analysis*, 27(4), 518-539.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424-438.

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:*1308.0850.

Grimmer, J., Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3), 267-297.

Hanke, J. E., Wichern, D. (2013). *Business Forecasting: Pearson New International Edition*. Pearson Higher Ed.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146-162.

Herring, S. C., Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, *10*(4), 439-459.

Hirst, G., Riabinin, Y., Graham, J., Boizot-Roche, M., Morris, C. (2014). Text to Ideology or Text to Party Status? *From text to political positions: Text analysis across disciplines*, 55, 93-15.

Hix, S., ve Høyland, B. (2013). Empowerment of the European parliament. *Annual Review of Political Science*, 16, 171-189.

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, 28(2), 87-106.

Howard, J., Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:*1801.06146.

Høyland, B., Godbout, J.-F., Lapponi, E., Velldal, E. (2014). Predicting party affiliations from European Parliament debates. *Proceedings of the acl 2014 workshop on language technologies and computational social science*, 56-60.

Hu, X., Cai, Z., Wiemer-Hastings, P., Graesser, A. C., McNamara, D. S. (2007). Strengths, limitations, and extensions of LSA. İçinde *Handbook of latent semantic analysis* (ss. 413-438). Psychology Press.

İnternet: Akşener, M. (2021). Mülakat, PolitikYol.com. *PolitikYol.com Haber Sitesi*. URL: https://www.politikyol.com/meral-aksener-politikyola-konustu-turkiye-bu-sistemi-birakin-2023u-bu-hafta-sonuna-kadar-bile-tasiyamaz / Son Erişim Tarihi: 18.07.2023.

İnternet: *Ak Parti 4. Olağan büyük kongresi siyasi vizyonu*. (2023). URL: https://www.akparti.org.tr/media/272148/2023-vizyonu.pdf / Son Erişim Tarihi: 18.07.2023.

Iqbal, F., Binsalleeh, H., Fung, B. C. M., Debbabi, M. (2010). Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation, 7*(1-2), 56-64.

Janssen, A., Murachver, T. (2004). The Relationship between Gender and Topic in Gender-Preferential Language Use. *Written Communication*, 21(4), 344-367.

Joachims, T. (1997). Categorization with support vector machines: Learning with many relevant features. *machine learning: ECML-98 10 th European Conference on Machine Learning*, 137-142.

Jolliffe, I. (2005). Principal component analysis. *Encyclopedia of statistics in behavioral science*.

Kaati, L., Lundeqvist, E., Shrestha, A., Svensson, M. (2017). Author profiling in the wild. *2017 European Intelligence and Security Informatics Conference (EISIC)*, 155-158.

Kapočiūtė-Dzikienė, J., Utka, A., Šarkutė, L. (2015). Authorship attribution and author profiling of Lithuanian literary texts. *The 5th Workshop on Balto-Slavic Natural Language Processing*, 96-105.

Kaynar, O., Aydın, Z., Görmez, Y. (2017). Sentiment analizinde öznitelik düşürme yöntemlerinin oto kodlayıcılı derin öğrenme makinaları ile karşılaştırılması. *Bilişim Teknolojileri Dergisi*, 10(3), 319-326.

Kim, Y. (2019). Convolutional neural networks for sentence classification. ArXiv 2014. *arXiv preprint arXiv:*1408.5882.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S. (2015). Skip-thought vectors. *Advances in neural information processing systems*, 28.

Koppel, M., Argamon, S., Shimoni, A. R. (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4), 401-412.

Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., Can, F. (2006). Chat mining for gender prediction. *Advances in Information Systems: 4th International Conference, ADVIS 2006, Izmir, Turkey, October 18-20, 2006. Proceedings* 4, 274-283.

Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79-86.

Landauer, T. K., Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), 211.

Lapponi, E., Søyland, M. G., Velldal, E., Oepen, S. (2018). The Talk of Norway: A richly annotated corpus of the Norwegian parliament, 1998–2016. *Language Resources and Evaluation*, 52(3), 873-893.

Lauderdale, B. E., Herzog, A. (2016). Measuring political positions from legislative speech. *Political Analysis*, 24(3), 374-394.

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

Levy, O., Goldberg, Y., Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, *3*, 211-225.

Li, J., Li, J., Fu, X., Masud, M. A., Huang, J. Z. (2016). Learning distributed word representation with multi-contextual mixed embedding. *Knowledge-Based Systems*, 106, 220-230.

Lim, W.-Y., Goh, J., Thing, V. L. L. (2013). Content-centric age and gender profiling. *Proceedings of the Notebook for PAN at CLEF*, 130-138.

Lin, J. (2007). *Automatic author profiling of online chat logs*. Naval Postgraduate School Monterey CA.

Mazey, S. (2002). Gender mainstreaming strategies in the EU: Delivering on an agenda? *Feminist Legal Studies*, 10, 227-240.

McCleary, R., Hay, R. A., Meidinger, E. E., McDowall, D. (1980). *Applied time series analysis for the social sciences*. Sage Publications Beverly Hills, CA.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153-157.

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

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

Mikolov, T., Yih, W., Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 746-751.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 39-41.

Miura, Y., Taniguchi, T., Taniguchi, M., Ohkuma, T. (2017). Author Profiling with Word+ Character Neural Attention Network. *CLEF (Working notes)*.

Morales, A. J., Borondo, J., Losada, J. C., Benito, R. M. (2015). Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3), 033114.

Mukherjee, A., Liu, B. (2010). Improving Gender Classification of Blog Authors. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 207-217.

Müller-Hansen, F., Callaghan, M. W., Lee, Y. T., Leipprand, A., Flachsland, C., Minx, J. C. (2021). Who cares about coal? Analyzing 70 years of German parliamentary debates on coal with dynamic topic modeling. *Energy Research & Social Science*, 72, 101869.

Naderi, N., Hirst, G. (2018). Using context to identify the language of face-saving. *Proceedings of the 5th Workshop on Argument Mining*, 111-120.

Naili, M., Chaibi, A. H., Ghezala, H. H. B. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia computer science*, 112, 340-349.

Nanni, F., Glavas, G., Ponzetto, S. P., Stuckenschmidt, H. (2019). Political text scaling meets computational semantics. *arXiv preprint arXiv:*1904.06217.

Newman, M. L., Groom, C. J., Handelman, L. D., Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse processes*, 45(3), 211-236.

Nguyen, D., Smith, N. A., Rose, C. (2011). Author age prediction from text using linear regression. *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, 115-123.

Odell, E. (2022). *Hansard Speeches Version 3* [R]. URL: https://github.com/evanodell/hansard-data3 / Son Erişim Tarihi: 18.07.2023.

Parlparse. (2023). [Python]. mySociety. URL: https://github.com/mysociety/parlparse / Son Erişim Tarihi: 18.07.2023.

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

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018). Deep contextualized word representations. ArXiv 2018. *arXiv preprint arXiv:*1802.05365, 12.

Peterson, A., Spirling, A. (2018). Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. *Political Analysis*, *26*(1), 120-128.

Pizarro, J. (2019). Using N-grams to detect Bots on Twitter. *CLEF (Working Notes)*.

Przybyla, P., Teisseyre, P. (2014). Analysing utterances in polish parliament to predict speaker's background. *Journal of quantitative linguistics*, *21*(4), 350-376.

Rangel, F., Franco-Salvador, M., Rosso, P. (2018). A low dimensionality representation for language variety identification. *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part II* 17, 156-169.

Rangel, F., Peña-Sarracén, G. L. de la, Chulvi-Ferriols, M. A., Fersini, E., Rosso, P. (2021). Profiling hate speech spreaders on twitter task at PAN 2021. *Proceedings of the Working Notes of CLEF 2021, Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st to 24th, 2021*, 1772-1789.

Rangel, F., Rosso, P. (2019). Overview of the 7th author profiling task at PAN 2019: Bots and gender profiling in twitter. *Working notes papers of the CLEF 2019 evaluation labs*, 2380, 1-7.

Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W. (2014). Overview of the 2nd author profiling task at pan 2014. *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014*, 1-30.

Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B. (2018). Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter. *Working notes papers of the CLEF*, 1-38.

Rangel, F., Rosso, P., Potthast, M., Stein, B. (2017). Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working notes papers of the CLEF*, 48.

Reddy, T. R., Vardhan, B. V., Reddy, P. V. (2016). A survey on authorship profiling techniques. *International Journal of Applied Engineering Research*, 11(5), 3092-3102.

Rheault, L., Beelen, K., Cochrane, C., Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, 11(12), e0168843.

Rodgers, J. L., Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1), 59-66.

Rong, X. (2014). Word2vec parameter learning explained. *arXiv preprint arXiv*:1411.2738.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, 12(2-3), 140-157.

Sakamoto, T., Takikawa, H. (2017). Cross-national measurement of polarization in political discourse: Analyzing floor debate in the US the Japanese legislatures. *2017 IEEE international conference on big data (Big Data)*, 3104-3110.

Salton, G., Wong, A., Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620.

Saygı, H., Aslan, C. (2020). Türkiye'de Partilerin Milliyetçilik ve Din Anlayışları. *Liberal Düşünce Dergisi*, 25(99), 79-104.

Schweter, S. (2020). *BERTurk—BERT models for Turkish* (1.0.0) [Software]. Zenodo.

Siino, M., Di Nuovo, E., Tinnirello, I., La Cascia, M. (2021). Detection of hate speech spreaders using convolutional neural networks. *CLEF (Working Notes)*, 2126-2136.

Slapin, J. B., Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), 705-722.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., 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*, 1631-1642.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538-556.

Stamatatos, E., Fakotakis, N., Kokkinakis, G. (2001). Computer-Based Authorship Attribution Without Lexical Measures. *Computers and the Humanities*, 35(2), 193-214.

Tausczik, Y. R., Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1), 24-54.

Torcal, M., Santana, A., Carty, E., ve Comellas, J. M. (2020). Political and affective polarisation in a democracy in crisis: The E-Dem panel survey dataset (Spain, 2018–2019). *Data in brief*, 32, 106059.

Türkmen, H. İ., Diri, B., Biricik, G., Doğan, R. (2011). Demographic information classification exploiting spoken language. *2011 IEEE 19th Signal Processing and Communications Applications Conference (SIU)*, 13-16.

Valencia, A. I. V., Adorno, H. G., Rhodes, C. S., Pineda, G. F. (2019). Bots and gender identification based on stylometry of tweet minimal structure and n-grams model. *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum, Lugano, Switzerland*, 2380.

Van der Maaten, L., Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \Lukasz, Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Veenhoven, R., Snijders, S., van der Hall, D., van Noord, R. (2018). Using translated data to improve deep learning author profiling models. *Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, 2125.

Yilmaz, K. E., Abul, O. (2018). Inferring political alignments of Twitter users. *2018 International Symposium on Networks, Computers and Communications (ISNCC)*, 1-6.

Yu, B. (2014). Language and gender in congressional speech. *Literary and Linguistic Computing*, 29(1), 118-132.

Yu, B., Kaufmann, S., Diermeier, D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1), 33-48.

Zheng, R., Li, J., Chen, H., Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3), 378-393.

Gazili olmak ayrıcalıktır