BMIJ
ISSN: 2148-2586

bmij (2021) 9 (4):1582-1598
doi: https://doi.org/10.15295/bmij.v9i4.1957

Araştırma Makalesi / Research Article

Derin makine öğrenmesi metoduyla sosyal medya verilerine dayalı seçim sonucu tahmini

Election result forecast based on social media data with the deep machine learning method

İbrahim Sabuncu[1]

Eda Şen[2]

[1] Dr. Öğr. Üyesi, Yalova Üniversitesi, Yalova, Türkiye, isabuncu@yalova.edu.tr

ORCID: 0000-0001-8625-9256

[2] Öğrenci, Yalova Üniversitesi, Yalova, Türkiye, edasenn97@gmail.com

ORCID: 0000-0001-5955-4639

Sorumlu Yazar/Corresponding Author:

İbrahim Sabuncu,

Yalova Üniversitesi, Yalova, Türkiye,
isabuncu@yalova.edu.tr

Öz

Bu çalışmanın amacı, sosyal medya verilerinden yararlanarak politikacıların oy oranlarının günlük değişiminin ve seçim sonucunun tahmin edilebilirliğinin araştırılmasıdır. Bu amaçla yapılan çalışmada, 3 Kasım 2020 ABD seçimine katılan adaylar hakkında 01.07.2020- 03.11.2020 tarihleri arasında paylaşılan 20.746.834 adet tweet RapidMiner programı ile Twitter platformundan toplanmıştır. Twitter'dan toplanan verilere, Vader algoritması ile duygu analizleri yapılmıştır. Tweetler, pozitif, negatif, NPS (pozitif-negatif) ve nötr duygu kategorilerine göre gruplandırılmıştır. Duygu kategorilerine ayrılmış tweet sayıları kullanılarak, günlük oy oranlarını ve seçim sonucunu tahmin etmek için altı farklı makine öğrenmesine dayalı tahmin modeli oluşturulmuştur. Tahmin modellerinde, bağımsız değişkenler adaylar hakkında paylaşılan duygu kategorilerine göre ayrılmış günlük Twitter verisidir. Bağımlı değişkenler ise anket ve ekonomik göstergelere dayalı yapılmış, adayların günlük oy oranı tahminleridir. Tahmin modelleri 109 günlük veri ile eğitilmiştir. En doğru sonucu veren tahmin modeli Derin Makine Öğrenmesi (Deep Machine Learning) algoritması kullanılarak, seçim sonucu %1,7 hata payıyla tahmin edilebilmiştir. Bu çalışma, Twitter'daki çok çeşitli manipülasyonlara rağmen, makine öğrenmesi aracılığıyla, Twitter'ın halen politik eğilimlerin takibi ve seçim sonuçları tahmininde kullanılabilecek bir veri kaynağı olabileceğini göstermektedir.

Anahtar Kelimeler: Sosyal Medya Analitiği, Tahmin Modeli, Derin Makine Öğrenmesi, Siyasi Seçimler, Twitter

Jel Kodları: M30, C5, C13, L86

Abstract

This study aims to research the predictability of the daily variation of the vote rates of politicians and the election result by using social media data. For this purpose, 20,746,834 tweets shared between 01.07.2020 - 03.11.2020 about the candidates participating in the U.S.A. election on November 3 2020, were collected from the Twitter platform using the RapidMiner program. Sentiment analyzes were made on the data collected from Twitter by the Vader algorithm. Tweets are grouped into positive, negative, N.P.S. (positive-negative), and neutral sentiment categories. Six different machine learning-based forecast models were created to predict the daily vote rates and the election result using the number of tweets divided into sentiment categories. In forecast models, the independent variables are daily Twitter data about candidates grouped by sentiment categories. The dependent variables are the daily vote rate estimates of the candidates based on surveys and economic indicators. Forecast models are trained with 109 days of data. Using the Deep Machine Learning algorithm, the forecast model that gave the most accurate result, the election result could be predicted with a margin of error of 1.7%. This study shows that despite the wide variety of manipulations on Twitter, Twitter can still be a data source that can be used to monitor political trends and predict election results through machine learning.

Keywords: Social Media Analytics, Forecast Model, Deep Machine Learning, Political Elections, Twitter

Jel Codes: M30, C5, C13, L86

Başvuru/Submitted: 18/10/2021

Revizyon/Revised: 13/12/2021

Kabul/Accepted: 21/12/2021

Yayın/Online Published: 25/12/2021

Atıf/Citation: Sabuncu, İ., & Şen, E., Derin makine öğrenmesi metoduyla sosyal medya verilerine dayalı seçim sonucu tahmini, bmij (2021) 9 (1): 1582-1598, doi: https://doi.org/10.15295/bmij.v1i1.1111

Extended Abstract

Election result forecast based on social media data with the deep machine learning method

Literature

Recently, social media has become an essential platform for politicians to speak out to the public, making them more accessible to their voters (Conway, Kenski and Wang, 2015; Golbeck, Grimes and Rogers, 2010; Graham, Jackson and Broersma, 2016). Twitter has emerged as a standard communication tool between the leaders of the competing parties and voters. It has been emphasized that Twitter is an important platform to influence the election result and public opinion (Grover, Kar, Dwivedi and Janssen, 2019).

Social media data, especially Twitter data, are used for election monitoring and forecasting. The studies showed that social media is an essential source of information in predicting the election results (Bansal and Srivastava, 2018; Burnap, Gibson, Sloan, Southern and Williams, 2016; Castro, Kuffó and Vaca, 2017; Ceron Guzman, 2016; Conway et al., 2015; Golbeck et al., 2010; Graham et al., 2016; Grover et al., 2019; Kušen and Strembeck, 2018; Makazhanov, Rafiei and Waqar, 2014; Toker, Erdem and Özşarlak, 2017; Tumasjan, Sprenger, Sandner and Welpe, 2010; Wicaksono, Suyoto and Pranowo, 2017). Recently, it has been observed that there is a positive and significant correlation between the social media mention rate about a candidate and his vote rate (Bansal and Srivastava, 2018).

This study collected data about the 2020 U.S. Presidential elections, one of the most recent elections. The result of this election has been predicted using data collected from Twitter. Unlike the methods used in the literature, six different machine learning methods were used for prediction. According to the established forecast models, the daily vote rates of the candidates were predicted, and the candidate who would win the election was tried to be estimated.

Design and method

This study researched the predictability of the daily vote rates of the candidates participating in the election and the election results based on the data obtained from the Twitter social media platform. As a result, the following hypothesis was developed:

- Research Hypothesis H_1: There is a relationship between the number of daily (total/positive/negative/net positive/neutral) tweets posted about a candidate and the daily vote rate (polls).

More than twenty million (20,746,834) tweets shared between 01.07.2020 and 03.11.2020 about candidates in the 2020 U.S. election were collected from the Twitter platform with the RapidMiner software. However, analyses were used only 15.285.748 tweets posted about the two most popular candidates (Trump and Biden). In addition, the candidates' daily vote rates estimates based on national polls and economic indicators were also collected (Economist, 2020). These daily vote rates estimates are called daily vote rates to avoid confusion.

Sentiment Analysis of the data collected from Twitter was carried out with RapidMiner software using the Vader algorithm. Tweets were grouped according to the positive, negative, N.P.S. (positive-negative) and neutral sentiment categories. Correlation analyses were made with the IBM SPSS Statistics program to investigate the correlation between these sentiment categories and daily voting rates.

In order to predict the election result, six different machine learning-based forecast models were created by using Auto Model in the RapidMiner program. The number of tweets per day in different sentiment categories was used to forecast the daily vote rates for that day. The Deep Machine Learning model with the lowest error rate predicted actual election results.

Findings and discussion

A moderate positive correlation was found between daily voting rates (polls) and the total/positive/negative/neutral tweets for candidate Trump. Nevertheless, a moderate negative correlation was found for candidate Biden between daily voting rates (polls) and the total/ positive/negative/net positive/ neutral tweets.

The vote rates were 46% for Trump and 54% for Biden based on the November 3 daily Twitter data using the Deep Machine Learning forecast model. So, the Deep Machine Learning model could predict the election's winner by using Twitter data.

Conclusion, recommendation and limitations

In conclusion, this study shows that social media is still a valuable data source that can be used to forecast election results. In addition, it contributes to the literature as the first study to use the deep machine learning method for this purpose. It is thought that the developed forecasting model will benefit research companies that make election forecasts, relevant party management and politicians, and researchers who will work on political marketing, management and forecasting models.

Giriş

Sosyal medya, kullanıcıların fikirlerini, düşüncelerini ve ürettikleri içerikleri yayınlamalarını ve paylaşmalarını sağlayan çevrimiçi platformlardır (Ulusoy, 2012). Facebook, Instagram, Twitter gibi sosyal medya platformlarında bireyler ve kurumlar fotoğraflar, videolar ve makaleler gibi içeriklerle duygu ve düşüncelerini paylaşabilmektedirler (Kim ve Ko, 2010).

Sosyal medya, politikacıların halka seslenebilmeleri için önemli bir platform olarak kullanılmaktadır ve onları seçmenleri için daha kolay ulaşılabilir hale getirmiştir (Conway, Kenski ve Wang, 2015; Golbeck, Grimes ve Rogers, 2010; Graham, Jackson ve Broersma, 2016). Bir adayın kamuoyuna tanıtılmasında veya siyasi propaganda gibi amaçlarla sosyal medya politikacılar tarafından yaygın olarak kullanıldığı görülmektedir (Cerf, 2017; Chatfield, Reddick ve Choi, 2017; Kelly Garrett ve Weeks, 2013).

Politikacılar ve seçmenler tarafından sosyal medyanın sık kullanımı, sosyal medyadan politik eğilimler hakkında önemli miktarda verinin birikmesini sağlamıştır. Bu verileri kullanarak politik eğilimleri izleme ve seçim sonuçları tahmini gibi konularda kullanılması güncel bir araştırma konusu olarak ortaya çıkmıştır. Bu konuda yapılan çalışmalarda genelde Twitter platformu kullanılmaktadır. Nitekim politikacılar tarafından en yaygın kullanılan ve halkın siyasi söylemleri için en çok tercih ettiği platformun Twitter olduğu görülmektedir (Tumasjan, Sprenger, Sandner ve Welpe, 2010) . Dolayısıyla politik eğilimlerin takibi için en fazla veriyi barındıran Twitter platformlarından toplanan verilerle, siyasi seçim sonuçlarının tahmini çalışmaları da yapılmaktadır.

Literatürdeki çalışmaların çoğunluğu, Twitter'dan toplanan veriler kullanılarak, düşük hata oranları ile seçim sonuçlarının tahmin edilebileceğini iddia etmektedir. Ancak, son on yılda seçimlere dolaylı müdahale amacıyla Twitter'da yanlış bilgilerin kasıtlı yayılması Twitter verilerinin güvenilirliğine şüphe düşürmektedir. Twitter verileri halen politik eğilimleri doğru yansıtmakta mıdır? Twitter verileri ile seçim sonuçlarını doğru tahmin etmek halen mümkün müdür? Twitter verileri ile seçim sonuçları arasından bir bağ kurmak mümkünse bu bağı oluşturmak için en uygun tahmin modeli veya algoritması hangisidir?

Bu makalede, yukarıda ifade edilmiş soruları yanıtlamak için, 3 Kasım 2020 Amerika Başkanlık seçimi ele alınmıştır. Bu seçime katılan iki önemli başkan adayın (Donald Trump ve Joe Biden) günlük oy oranlarıyla (anket ve ekonomik verilere dayalı günlük oy oranı tahmini ile) o aday hakkında paylaşılan tweetler arasında bir ilişki olup olmadığı araştırılmıştır. Böylece politik eğilimlerdeki değişimin Twitter verilerine dayalı tahmin edilebilirliği araştırılmıştır. Ayrıca toplanan Twitter verileri ile seçim sonucu tahmin edilmeye çalışılmıştır. Bu amaçla, oluşturulan araştırma modeli metodoloji kısmında detaylı açıklanmıştır.

Literatür taraması

Sosyal medya verilerini kullanarak seçim sonuçlarını tahmin etmek amaçlı son on yılda farklı ülkelerde çeşitli çalışmalar yapılmıştır. Bu çalışmaların en eskilerinden biri de Tumasjan vd. (2010) tarafından yapılan çalışmadır. Çalışma, Twitter'ın siyasi müzakereler için bir forum olarak kullanılıp kullanılmadığını ve Twitter'daki çevrimiçi mesajların, çevrimdışı siyasi duyguları geçerli olarak yansıtıp yansıtmadığını araştırmak için Alman federal seçimi ele alınmıştır. 2009 Almanya federal seçimlerinden önce partilerden veya politikacılardan bahseden 100.000'den fazla Twitter mesajını analiz etmişlerdir. Genel olarak, Twitter'ın gerçekten de siyasi müzakereler için bir platform olarak kullanıldığı sonucuna varmışlardır.

Makazhanov, Rafiei ve Waqar (2014) tarafından, biri Kanada'da ve diğeri Pakistan'da gerçekleşen iki farklı seçimin sonuçlarını tahmin çalışması yapılmıştır. Çalışmada, Twitter verilerine dayalı olarak kullanıcıların siyasi tercihlerini tahmin etmeye çalışmışlardır. Tahmin için makine öğrenmesinden faydalanılan araştırmada SVM (*Support Vector Machine)* ve lojistik regresyon (Lgstc) modelleri kullanılmıştır. Kullanıcıların davranışlarının incelenmesinin modelleri geliştirmeye yardımcı olabileceğini ve daha doğru tercih tahminlerine yol açabileceği sonucuna varmışlardır.

Conway vd. (2015) tarafından yapılan çalışmada, Twitter ile siyasi adayların ve partilerin Twitter mesajları arasındaki ilişkiye odaklanılmıştır. Ülkenin önde gelen gazetelerinden makalelerin, başkan adaylarının kampanya tweetlerini, Cumhuriyetçi ve Demokrat partilerin tweetlerindeki konu içeriklerini analiz etmişlerdir. Dijital ortamdaki büyük verilerin bilgisayar destekli içerik analizi ve zaman serisi analizinin, medya ve politika arasındaki ilişkiyi araştırmak için kullanılabileceğini göstermişlerdir.

Bir diğer çalışmada Burnap, Gibson, Sloan, Southern ve Williams (2016), Twitter'dan toplanan tweetler ile 2015 Birleşik Krallık genel seçimleri sonucu tahmini üzerine çalışmışlardır. Twitter'dan topladıkları

verileri duygu analizi yaparak kurdukları temel tahmin modeli ile mecliste en çok koltuk sahibi olacak partiyi tahmin etmeye çalışmışlardır.

Ceron Guzman (2016) ise, Twitter'dan elde ettiği veriler ile Kolombiya 2014 Cumhurbaşkanlığı Seçim sonuçlarını tahmini üzerine çalışmıştır. Rastgele Orman algoritması kullanılmıştır. Test setindeki performanslarından, makine öğrenimi sınıflandırması için Lojistik Regresyon algoritması kullanılmıştır. %4,84 ortalama mutlak hata ile seçim sonucunu tahmin ederek başarılı olmuştur. Kolombiya'da seçim tahmini iyi sonuç verse de anket şirketlerinin tahmini daha doğru yaptıkları sonucuna varmıştır.

Graham vd. (2016) tarafından yapılan çalışmada, 2010 genel seçimlerinde İngiliz ve Hollandalı Parlamento adaylarının Twitter'ı nasıl kullandıkları karşılaştırılmaktadır. Hollandalı politikacıların Twitter'ı İngiltere'deki adaylardan daha fazla kullanma olasılıklarının daha yüksek olduğunu ve ortalama olarak İngiliz meslektaşlarına göre iki kat daha fazla tweet attığı sonucunu bulmuşlardır.

Benzer çalışma olarak Castro, Kuffó ve Vaca (2017), Twitter'dan topladıkları veriler ile 2015 Venezuela parlamento seçimlerini tahmin etmeye çalışmışlardır. Twitter'dan toplanan verilerin duygu analizi yapılmış ve ardından sosyal ağ analizi ve denetimsiz makine öğrenmesi metotları kullanılmıştır. Sonuç olarak %87,5 doğruluk oranı ile seçim sonucunu doğru tahmin edebilmişlerdir.

Türkiye'de yapılan bir çalışmada Toker, Erdem ve Özşarlak (2017) tarafından, 2015 Türkiye yerel seçimleri üzerine tahmin çalışması yapılmıştır. 1011 Twitter kullanıcısının seçim kampanyası boyunca Twitter'daki gönderileri toplanmış, rassal olarak seçilen 364 kullanıcının hesapları incelendiğinde bir kısmının politik söylemlerden kaçındığı, diğer kısmın ise daha aktif ve keskin politik tutum gösterdiği sonucuna varmışlardır. Yapılan bu çalışmada sosyal medyanın seçim sonuçlarını tahmin etmede önemli bir bilgi kaynağı olacağını vurgulamışlardır. Yaş, meslek gibi demografik verilerine bakıldığında, büyük şehirlerdeki Twitter kullanıcı sayısının, bu şehirlerin gelişmişlik düzeyi ve seçmen yoğunluğu ile uyumlu olduğunu görmüşlerdir.

Wicaksono, Suyoto ve Pranowo (2017), Twitter'dan toplanan veriler ile 2016 ABD başkanlık seçimleri üzerine bir tahmin çalışması yapmışlardır. Twitter'dan toplanan tweetlerin Binary Multinomial Naive Bayes Sınıflandırıcısı kullanılarak duygu analizlerini yapmışlardır. Doğal dil işleme (NLP) analizini kullanmışlardır. Sonuç olarak bölgelere göre kimin kazanacağını tahmin etmeye çalışmışlardır.

İncelenen başka bir çalışmada Kušen ve Strembeck (2018) tarafından, R programlama ile indirilen tweetler doğrultusunda 2016 Avusturya Başkanlık seçimini tahmin etmeye çalışmışlardır. SentiStrength algoritması ve doğal dil işleme (NLP) analizi kullanılmıştır. Duygusal tweetlerin sırasıyla beğeni, retweet ve cevap sayısı ile pozitif yönde ilişkili olduğu bulunmuştur. Ayrıca bir tweetteki hashtaglerin ve URL'lerin sayısının, beğeni ve retweet sayısı ile pozitif yönde ilişkili olduğu bulunmuştur. Bununla birlikte, yanıt sayısına gelince, URL'lerle yalnızca pozitif bir korelasyon bulunmuştur. Tweet popülaritesi açısından her iki cumhurbaşkanı adayı için retweet ile beğeni sayısı arasında güçlü bir pozitif korelasyon olduğu gözlemlenmiştir. Seçimin galibini beğeni ve retweet sayısına göre doğru tahmin edebilmişlerdir.

Farklı bir çalışmada ise Bansal ve Srivastava (2018) tarafından, Twitter Search API ile R programlama dili kullanılarak toplanan verilerden, Hibrit Konuya Dayalı Duygu Analizi (HTBSA) yönteminden yararlanarak seçimleri tahmin etmeye çalışmışlardır. Bir adayın Twitter popülaritesinin yani tweetlerde bahsedilme sayısı ile seçmenlerin oy oranı arasındaki ilişkinin anlamlı pozitif bir ilişki olduğu görülmüştür. Son zamanlarda sosyal medya ile her bir tarafın kazandığı koltuk sayısı arasında pozitif ve anlamlı bir ilişki olduğu görülmüştür. Elde ettikleri sonuca göre doğru tahminler yapıldığı sonucu elde edilmiştir.

Son olarak incelenen bir diğer çalışmada Grover, Kar, Dwivedi ve Janssen (2019), Twitter REST API kullanılarak Pyhton programlama dilinde toplanan Twitter verileri ile 2016 Amerika seçim sonuçlarını tahmin etmeye çalışmışlardır. R yazılımı ile duygu analizi yapmışlardır. Büyük miktarda metin verisinden bilgi almak için doğal dil işleme (NLP) ve metin madenciliği prensiplerini kullanmışlardır. Makine öğrenimi algoritmaları, içerik analizi ve ağ analizinden yararlanılmıştır. Twitter'da serbest metin kullanımının, bahsetme kullanımına kıyasla oylama sonuçlarıyla daha güçlü bir korelasyona sahip olduğunu göstermektedir. Twitter verileri seçim sonucunu ve kamuoyunu seçmenleri etkilemede önemli bir faktör olduğunu vurgulamışlardır.

İncelenen çalışmalar, Twitter verileri kullanılarak farklı ülkelerde seçim sonuçlarını tahmin edebilmenin mümkün olduğunu göstermektedir. Ancak, Twitter'da politik manipülasyonların yaygınlaştığına dair haberler mevcuttur (Golovchenko, Buntain, Eady, Brown ve Tucker, 2020; Jamieson, 2020; Karami, Lundy, Webb, Turner-McGrievy, McKeever B. ve McKeever R., 2021). Bu

manipülasyonlar, Twitter'daki verilerin güvenilirliğini değiştirmiş midir? Twitter halen politik eğilimleri takip ve seçim sonuçlarını tahmin için kullanılabilecek bir veri kaynağı mıdır? Tahmin için kullanılan çeşitli yöntemler vardır. Makine öğrenmesi bu yöntemler arasındadır. Ancak hangi makine öğrenmesi algoritması bu tür bir tahmin modeli için en uygundur? Bu çalışmada, belirtilen sorulara yanıt vermeye çalışarak, literatüre katkı sağlanması amaçlanmaktadır. Nitekim bu çalışmada diğer çalışmalardan farklı ve en güncel seçimlerden biri olan 2020 Amerika Başkanlık seçimleri ele alınmıştır. Bu seçimin sonucu Twitter'dan elde edilen veriler ile tahmin edilmiştir. Literatürdeki kullanılan yöntemlerden farklı olarak, Vader duygu analizi algoritması, RapidMiner Auto Model yazılımı aracılığıyla 6 farklı makine öğrenmesi kullanılmıştır. Kurulan tahmin modellerine göre adayların oy oranları tahmin edilmiş olup, seçimi kazanacak aday tahmin edilmeye çalışılmıştır.

Metodoloji ve analizler

Bu çalışmada 3 Kasım 2020 Amerika seçimleri hakkındaki Twitter sosyal medya platformundan elde edilen veriler ile seçime katılan adayların oy oranlarındaki değişimin ve seçim sonuçlarının tahmin edilebilirliği araştırılmıştır. Araştırmanın bağımsız değişkenleri Twitter'dan toplanan aşağıda açıklanmış verilerdir. Bağımlı değişkeni ise 3 Kasım 2020 gerçek seçim sonucu ile 3 kasım öncesi 109 güne ait, çok sayıdaki ulusal anket ve ekonomik göstergelere dayalı günlük oy oranlarının tahminidir (Economist, 2020). Economist sayfasından elde edilen oy oranları tahminlerinin, bu çalışma kapsamındaki Twitter verilerine dayalı oy oranı tahminleri ifadesiyle karıştırılmaması için, Economist'in seçim anketleri ve ekonomik göstergelere dayalı günlük oy oranlarının tahmini veriler, **günlük oy oranları** olarak isimlendirilmiştir.

İki farklı değişken türü arasındaki ilişki, yani bir adayla ilgili günlük oy oranları ve o aday hakkında ilgili gün paylaşılan tweet sayıları arasındaki ilişki araştırılmıştır. Adayın oy oranındaki değişimin Twitter verilerine dayalı tahmin edilebilirliği ve nihayetinde seçim günü alınacak gerçek oy oranı yani seçim sonucunun da tahmin edilebilirliği test edilmiştir.

Bağımsız değişken olarak, Twitter'da herkese açık şekilde paylaşılmış, adaylar veya seçimle ilgili anahtar kelimeleri içeren tweetler toplanmıştır. Halka açık paylaşılan bu tweetlerin toplanması için etik kurul onayı gerekmemektedir.

Toplanan tweetlerin duygu analizleri yapılarak, tweetler pozitif, negatif, NPS (pozitif-negatif tweet sayısı), nötr olarak sınıflandırılmıştır. NPS, net destekçi skoru (Net Promoter Score) için kullanılan kısaltma olup, bu çalışmada, pozitif tweet sayılarının negatif tweet sayılarından çıkarılması ile elde edilen sayıyı temsil etmektedir. Tweetlerin duygu kategorisine göre beş farklı bağımsız değişken için hipotezler tanımlanmış ve test edilmiştir. Hipotez testlerinde kullanılan değişkenler aşağıda verilmiştir:

Bağımlı değişken: $y_{ij}: i.adayın, j.\text{güne ait oy oranı tahmini}$
Bağımsız değişken: $x_{ijk}: i.aday\ hakkında, j.\text{gün}, paylaşılan\ k.kategorideki\ tweet\ sayısı$
i: 1,2 $\{1: Trump, 2: Biden\}$
j: 1..109 {1: 01.07.2020 ... 109: 02.11.2020};
k: 1..5 $\{1: Toplam, 2: Pozitif, 3: Negatif, 4: Net\ Pozitif, 5: \text{Nötr}\}$

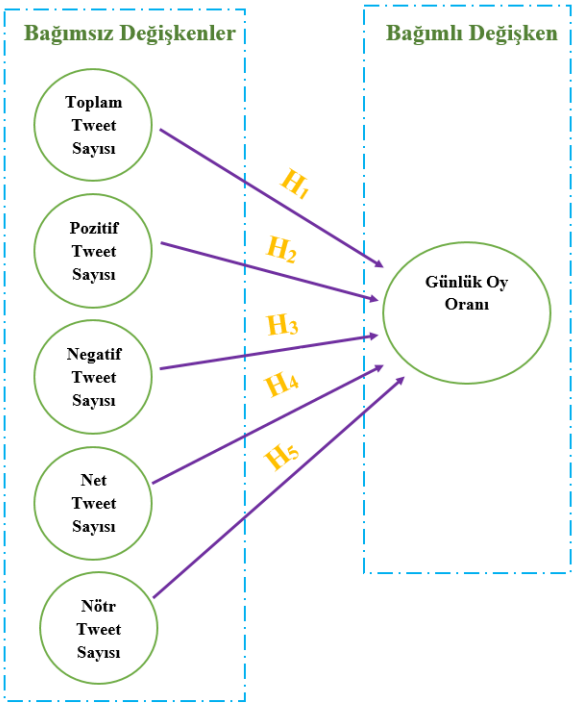
Araştırma hipotezleri ana hipotez ve alt hipotezler şeklinde aşağıda verilmiştir, (değişkenler arasında ilişki olmadığını ifade eden) sıfır hipotezlerin yazılmasına gerek görülmemiştir.

Araştırma Hipotezi H_1: Bir aday hakkında paylaşılmış olan (toplam/pozitif/negative/net pozitif/nötr) tweetlerin sayısı x_{ijk} ile o aday ile ilgili oy oranı y_{ij} arasında bir ilişki vardır.

Bağımsız değişkenlere göre alt hipotezler:

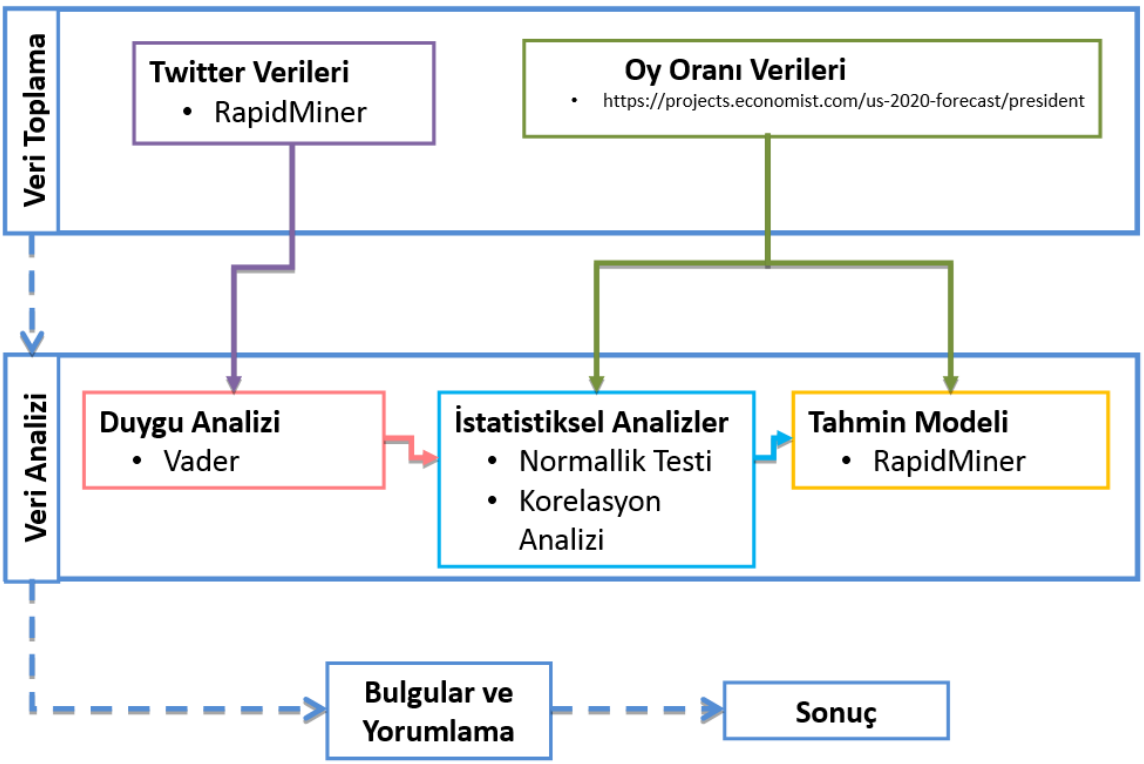
- H_{11}: Bir aday hakkında paylaşılmış olan **toplam** tweetlerin sayısı x_{ij1} ile o aday ile ilgili oy oranı y_{ij} arasında bir ilişki vardır.
- H_{12}: Bir aday hakkında paylaşılmış olan **pozitif** tweetlerin sayısı x_{ij2} ile o aday ile ilgili oy oranı y_{ij} arasında bir ilişki vardır.
- H_{13}: Bir aday hakkında paylaşılmış olan **negatif** tweetlerin sayısı x_{ij3} ile o aday ile ilgili oy oranı y_{ij} arasında bir ilişki vardır.
- H_{14}: Bir aday hakkında paylaşılmış olan **net pozitif** (pozitif – negative) tweetlerin sayısı x_{ij4} ile o aday ile ilgili oy oranı y_{ij} arasında bir ilişki vardır.
- H_{15}: Bir aday hakkında paylaşılmış olan **nötr** tweetlerin sayısı x_{ij5} ile o aday ile ilgili oy oranı y_{ij} arasında bir ilişki vardır.

Değişkenler arasındaki ilişki, Şekil 1'deki diyagramda gösterilmiştir.



Şekil 1: İlişki Hipotezleri

Hipotezlerin geçerliliğini test etmek için gerekli olan veriler (günlük oy oranları ve Twitter verileri) toplandıktan sonra, Twitter verilerine Vader algoritması duygu analizi yapılmıştır. İçerdiği duyguya bağlı olarak kategorize edilmiş tweetler ve günlük oy oraları arasındaki ilişkinin tespiti için korelasyon analizi yapılmıştır. Ardından, Twitter verileri ile günlük oy oranları için makine öğrenmesine dayalı tahmin modelleri kullanılmıştır. Makine öğrenmesi modelinde, bir adaya ait oy oranı y_{ij} adayın kendisi ve rakibi hakkında paylaşılan tweet sayılarına x_{ijk} (10 farklı değişkene) bağlı olarak tahmin edilmeye çalışılmıştır. Bahsedilen araştırma süreci Şekil 2'deki diyagramla özetlenmiştir.



Şekil 2: Araştırma Süreç Diyagramı

Veri toplama

Araştırmada bağımlı değişken günlük oy oranlarıdır. ABD'de oy oranlarının değişimini takip için ulusal seçim anketleri (oy oranı tahminleri) pek çok farklı kuruluş tarafından düzenli olarak yapılıp yayınlanmaktadır. Bu çalışmada, çok sayıdaki ulusal seçim anket verisi ve ekonomik göstergeleri kullanarak yapılan günlük oy oranları tahminlerinin (Economist, 2020) bağımlı değişken olarak kullanılması uygun görülmüştür. Nitekim bu verilerin, tek bir ulusal anket şirketinin verisine oranla, gerçek oy oranlarını ve politik eğilimdeki günlük değişimleri daha iyi temsil ettiği düşünülmektedir. Böylece, bağımlı değişken y_{ij} olarak 109 günlük (01.07.2020…02.11.2020) oy oranı tahminleri (Economist, 2020), ve 3 Kasım 2020 (110. Gün) için gerçek oy oranları (seçim sonucu) kullanılmıştır. Kavram karmaşası olmaması için anket verilerine dayalı oy oranı tahminleri, günlük oy oranları olarak isimlendirilmiştir.

Bağımsız değişken olarak kullanılacak Twitter verilerini toplamak için yaygın olarak kullanılan birçok yazılım bulunmaktadır. R programlama dili, Python programlama dili ve RapidMiner veri madenciliği programı bunlardan birkaçıdır. Bu çalışmada, RapidMiner veri madenciliği yazılımı, kodlama bilgisi gerektirmediği, diğer yazılımlara göre daha basit, kullanışlı ve ücretsiz olduğu için tercih edilmiştir.

Araştırma konusuyla ilgili Tweetler belirlemek için tweetlerin içeriğinde aranacak anahtar kelimeler belirlenmiştir. Anahtar kelimeler seçilirken, konuyla alakasız tweetleri engellemek amacıyla, kelimeler seçime aday olan partilerin Twitter hesaplarından seçilmiştir. Bu yöntemle bazı işe yarar tweetlere de ulaşılamamış olmasına rağmen, elde edilmek istenen asıl verileri bozacak konuyla alakasız birçok tweet de filtrelenmiştir.

İlk olarak, 2020 öncesi son seçimde en çok oy alan dört partiden seçime katılacak başkan ve başkan yardımcısı adayların Twitter hesapları incelenmiştir. Bu hesaplarda sıkça paylaşılan sloganlar, adayların isimleri, temsil ettikleri partilerin isimlerini ve kısaltmalarını içeren anahtar kelimeler tespit edilmiştir. Böylece doğrudan adayın ismi geçmese de adaydan bahseden, adayı destekleyen veya karşı çıkan tweetlerin de toplanması mümkün olmuştur. Ek olarak, adaylardan biriyle doğrudan ilgili olmayan ancak seçimle ilgili olan "USAelection" ve "#NovemberElection" etiketleri de (hashtag) anahtar kelimelere eklenmiştir. Sonuç olarak oluşturulan anahtar kelime listesi Tablo 1'de verilmiştir.

Tablo 1: Twitter'dan veri toplamada kullanılan anahtar kelimeleri

Parti Adı	Anahtar Kelimeler
Republican Party	#MAGA2020 @GOP Trump @POTUS @realDonaldTrump Pence @Mike_Pence @VP "Keep America Great"
Democratic Party	@DNC @TheDemocrats Biden @JoeBiden "Our best days still lie ahead" "No Malarkey!"
Libertarian Party	@LPNational "Jo Jorgersen" @Jorgensen4POTUS "Spike Cohen" @RealSpikeCohen
Green Party	@GreenPartyUS @TheGreenParty "Howie Hawkins" @HowieHawkins "Angela Walker" @AngelaNWalker

Belirlenen anahtar kelimeleri içeren tweetler RapidMiner Studio yazılımı kullanılarak, Twitter API (Kullanıcı Ara Yüzü) aracılığıyla indirilmiştir. Anahtar kelimeleri içeren, 01 Temmuz 2020 ile 03 Kasım 2020 tarihleri arasında paylaşılmış 20.746.834 adet tweet elde edilmiştir. Twitter veri setinde tweetin içerdiği metin (Text), paylaşıldığı tarih (Created-At), kim tarafından paylaşıldığı (From-User) gibi 17 farklı nitelik bilgisi bulunmaktadır. Bu veriler başka araştırmalarda da kullanılabilmesi amacıyla,

makalenin yazarı tarafından veri portalında paylaşılmıştır (Sabuncu, 2020). İndirilen tweetlerin örneği Şekil 3'te verilmiştir.

Row No.	Text	Created-At	From-User	From-User-Id	To-User	To-User-Id	Language	Source	Geo-Loc
1	@Xanthippie ...	Oct 1, 2020 3:...	Shoq	8431652	Xanthippie	18141610	en	<a href="http...	?
2	@SaysSimon...	Oct 1, 2020 3:...	Jeff	26590016	SaysSimonson	1250322626	en	<a href="http:/...	?
3	@JessicaFK...	Oct 1, 2020 3:...	Deborah Del...	28590114	JessicaFKane	142818373	en	<a href="http:/...	?
4	Why Donald ...	Oct 1, 2020 3:...	Slimer 1	7844187971...	?	-1	en	<a href="http:/...	?
5	@realDonald...	Oct 1, 2020 3:...	driver@driver	1034527691...	realDonaldTr...	25073877	en	<a href="http:/...	?
6	RT @LeahR7...	Oct 1, 2020 3:...	JohnRah	22040128	?	-1	en	<a href="http:/...	?
7	RT @COswe...	Oct 1, 2020 3:...	ProudPapa9...	1304766507...	?	-1	en	<a href="http:/...	?
8	RT @DC_Dr...	Oct 1, 2020 3:...	Stephanie Ka...	139497010	?	-1	en	<a href="http:/...	?
9	RT @Marcqu...	Oct 1, 2020 3:...	E□	746014052	?	-1	en	<a href="http:/...	?
10	RT @JoeBid...	Oct 1, 2020 3:...	Mia	2463017019	?	-1	en	<a href="http:/...	?
11	RT @PhilipR...	Oct 1, 2020 3:...	Yolanda Ran...	1101154043...	?	-1	en	<a href="http:/...	?
12	@mychloegirl...	Oct 1, 2020 3:...	I'm Ridin with ...	2760437711	mychloegirl6...	4121502046	en	<a href="http:/...	?
13	RT @kenklip...	Oct 1, 2020 3:...	FoxyBluebon...	2361222446	?	-1	en	<a href="http:/...	?
14	RT @rosapa...	Oct 1, 2020 3:...	Luis Enrique ...	919382552	?	-1	es	<a href="http:/...	?
15	RT @Michell...	Oct 1, 2020 3:...	Lily's Dad.	393164259	?	-1	en	<a href="http:/...	?
16	RT @RL9631...	Oct 1, 2020 3:...	usSteve Coxus	187022370	?	-1	en	<a href="http...	?
17	RT @realDon...	Oct 1, 2020 3:...	BBN1971	34809827	?	-1	en	<a href="http:/...	?
18	RT @Meidas...	Oct 1, 2020 3:...	David Scott	485584481	?	-1	en	<a href="http:/...	?

ExampleSet (8,615,875 examples, 0 special attributes, 17 regular attributes)

Şekil 3: Toplanan Twitter Veri Seti Örneği

Veriler dört parti adayı için toplansa da analizler 2020 öncesi son ulusal seçimlerde toplam oyların %98'ini alan iki partinin (Demokratlar ve Cumhuriyetçiler) adayları hakkındaki verilerle sınırlandırılmıştır. Böylece, içeriğindeki anahtar kelimelere göre Biden veya Trump ile ilgili olduğu tespit edilen 15.285.748 tweet ile analizler yapılmıştır. Analizlerde, veri setindeki metin (Text) ve paylaşıldığı tarih (Created-At) sütunlarındaki veriler kullanılmıştır.

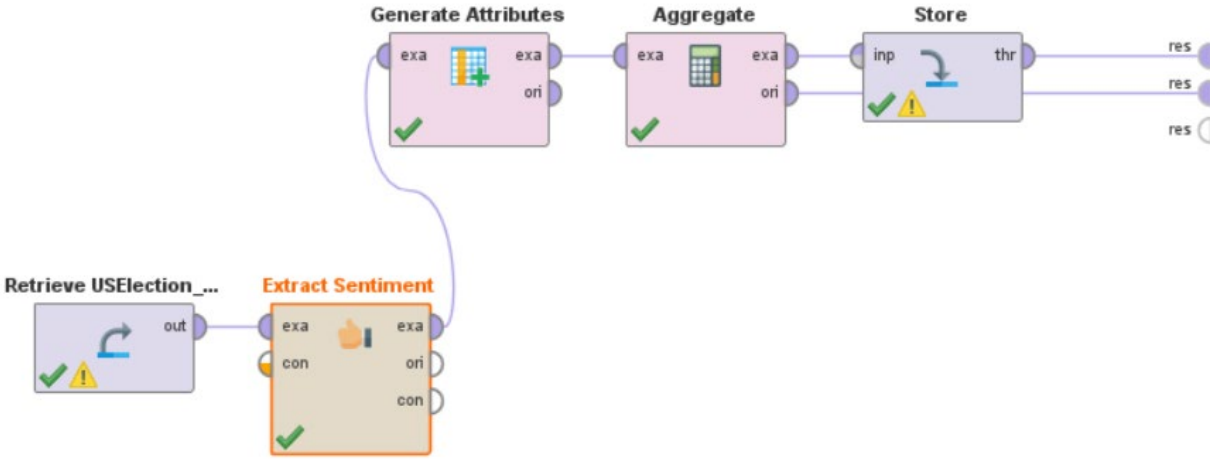
Duygu analizi

Toplanan tweetlerin duygu analizlerinin yapılabilmesi için öncelikle RapidMiner Text Processing eklenti paketinde Extract Sentiment operatörü kullanılmıştır. Bu operatör 4 farklı duygu analizi algoritması sunmaktadır. Bu algoritmalar;

- Vader: Sosyal medya içeriklerinin duygu analizleri için geliştirilmiş sözlük tabanlı bir yöntemdir ve ücretsizdir.
- Sentiwordnet: Sözlük tabanlı bir yöntemdir ve ücretsizdir.
- Aylien: Ticari bir metin madenciliği yazılımıdır. Makine öğrenmesi tabanlı duygu analizi gerçekleştirebilir ve ücretlidir.
- MeaningCloud: Makine öğrenmesi tabanlı duygu analizi ve diğer metin madenciliği işlemlerini yapan bir diğer ticari uygulamadır ve ücretlidir. Ücretsiz versiyonunda aylık sadece 10000 adet veri analiz edilmesine izin verilmektedir.

Bu çalışmada, araştırma bütçesi olmadığı için, ücretsiz iki algoritma içerisinden özellikle sosyal medya için geliştirilmiş olan Vader algoritması seçilmiştir. İndirilen tweetlerin duygu analizinde kullanılan RapidMiner modeli Şekil 4'te gösterilmiştir.

Duygu analizleri yapılan, 01.07.2020-03.11.2020 tarihleri arası paylaşılmış iki parti adayı ile ilgili tweetler, içeriğindeki anahtar kelimelere göre ilgili oldukları adaya göre ayrılmıştır. İçerdiği duyguya (pozitif, negatif, nört) ve ilgili olduğu adaya (Biden, Trump) göre sınıflandırılmış tweetlerin örneği Tablo 2'de verilmiştir. Bazı tarihlerde veri kaybı olduğu için o günlere ait tweetler analiz edilememiştir.



Şekil 4: RapidMiner Duygu Analiz Modeli

Tablo 2: Duygu analizi sonuçları

Tarih	Biden Pozitif	Biden Negatif	Biden Nötr	Trump Pozitif	Trump Negatif	Trump Nötr
1.7.2020	2199	1091	1185	8331	11046	7013
2.7.2020	4792	3282	4809	27190	29339	21982
3.7.2020	3322	2877	3375	32114	27983	25271
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
1.11.2020	26277	18658	22778	71205	59544	56260
2.11.2020	30482	20189	20280	68531	66884	56250
3.11.2020	30493	15281	27716	70643	48056	56526

Tablo 2'de verilmiş günlük tweet sayıları kullanılarak adaya ait günlük toplam (pozitif+negatif+nötr), NPS (pozitif-negatif) tweet sayıları da hesaplanmıştır. Bu verilere ilgili günler için adaylarla ilgili seçim anketi verileri ve ekonomik göstergelere dayalı günlük oy oranları (Economist, 2020) eklenmiştir. Bu verilerin örnekleri Biden ve Trump için Tablo 3 ve Tablo 4'de gösterilmektedir.

Tablo 3: Biden'a ait tweet sayıları ve günlük oy oranı verileri

Tarih	Toplam	Pozitif	Negatif	NPS	Nötr	Oy Oranı
1.7.2020	4475	2199	1091	1108	1185	54,8
2.7.2020	12883	4792	3282	1510	4809	54,9
3.7.2020	9574	3322	2877	445	3375	55
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
31.10.2020	53851	22406	13174	9232	18271	54,1
1.11.2020	67713	26277	18658	7619	22778	53,9
2.11.2020	70951	30482	20189	10293	20280	54,2

Tablo 4: Trump'a ait tweet sayıları ve günlük oy oranı verileri

Tarih	Toplam	Pozitif	Negatif	NPS	Nötr	Oy Oranı
1.7.2020	26390	8331	11046	-2715	7013	45,2
2.7.2020	78511	27190	29339	-2149	21982	45,1
3.7.2020	85368	32114	27983	4131	25271	45
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
31.10.2020	138221	54243	42996	11247	40982	45,9
1.11.2020	187009	71205	59544	11661	56260	46,1
2.11.2020	191665	68531	66884	1647	56250	45,8

Korelasyon analizi

Tahmin modellemesi yapılmadan önce, Tablo 3 ve 4'de verilmiş olan veriler arasındaki ilişki korelasyon analizi ile araştırılmıştır. Öncelikle normallik testleri yapılmıştır. Bu analizler için SPSS istatistiksel analiz programından faydalanılmıştır.

Normallik testi için kullanılan araştırma hipotezi ($H_{1n)}$: Veriler %95 güvenle normal dağılımlı değildir.

Biden'a ait verilerin normallik testi sonuçları Tablo 5'de verilmiştir. Sonuçlara göre toplam, pozitif, negatif, NPS, nötr ve oy oranı Sigma değerleri 0,05'ten küçük olduğu için H_0 reddedilir yani değişkenler normal dağılımlı değildir. Bu sebeple korelasyon incelenirken Spearman's rho değerleri dikkate alınmıştır.

Tablo 5: Biden'a ait verilerin normallik testi

	Kolmogorov-Smirnov			Shapiro-Wilk		
	İstatistik	Veri Sayısı	Anlamlılık Düzeyi (Sig.)	İstatistik	Veri Sayısı	Anlamlılık Düzeyi (Sig.)
Toplam	,186	109	,000	,861	109	,000
Pozitif	,183	109	,000	,862	109	,000
Negatif	,191	109	,000	,868	109	,000
NPS	,211	109	,000	,845	109	,000
Nötr	,190	109	,000	,852	109	,000
Oy Oranı	,107	109	,004	,968	109	,010

Biden'a ait verilerin korelasyon analizi sonuçları Tablo 6'da verilmiştir. Analizler sonucunda, günlük oy oranları ile toplam, pozitif, negatif, NPS ve nötr tweetlerin sayıları arasında 0.01 düzeyinde anlamlı negatif yönlü korelasyon olduğu tespit edilmiştir.

Tablo 6: Biden'a ait verilerin korelasyon analizi

			Oy Oranı	Toplam	Pozitif	Negatif	NPS	Nötr
Spearman' s rho	**Oy Oranı**	**Korelasyon Katsayısı**	1	-,466**	-,467**	-,463**	-,261**	-,446**
		Anlamlılık Düzeyi (Sig.)		,000	,000	,000	,006	,000
		Veri Sayısı (N)	109	109	109	109	109	109

**** Korelasyon 0,01 düzeyinde anlamlıdır.**

Trump ile ilgili verilerin normallik test sonuçları Tablo 7'de verilmiştir. Sonuçlara göre toplam, pozitif, negatif, nötr ve oy oranı normal dağılımlı değildir. Bu nedenle korelasyon incelenirken Spearman's rho değerleri dikkate alınmıştır. Sadece NPS'nin Sigma değeri ise 0,05'ten büyük olduğu için H_0 reddedilemez yani normal dağılımlıdır. Bu nedenle NPS'nin korelasyonu incelenirken Pearson değeri dikkate alınmıştır. Trump'a ait verilerin korelasyon analizleri Tablo 8'de verilmiştir. Görüldüğü üzere günlük oy oranları ile toplam, pozitif, negatif ve nötr tweetler arasında 0.01 düzeyinde anlamlı pozitif yönlü korelasyonu vardır.

Tablo 7: Trump'a ait verilerin normallik testi

	Kolmogorov-Smirnov			Shapiro-Wilk		
	İstatistik	Veri Sayısı	Anlamlılık Düzeyi (Sig.)	İstatistik	Veri Sayısı	Anlamlılık Düzeyi (Sig.)
Toplam	,229	109	,000	,867	109	,000
Pozitif	,167	109	,000	,905	109	,000
Negatif	,184	109	,000	,855	109	,000
NPS	,066	109	,200*	,970	109	,015
Nötr	,243	109	,000	,839	109	,000
Oy Oranı	,107	109	,004	,968	109	,010

Tablo 8: Trump'a ait verilerin korelasyon analizi

			Oy Oranı	Toplam	Pozitif	Negatif	NPS	Nötr
Spearman' s rho	Oy Oranı	Korelasyon Katsayısı	1	,514**	,521**	,470**		,431**
		Anlamlılık Düzeyi (Sig.)		,000	,000	,000		,000
		Veri Sayısı (N)	109	109	109	109		109
Pearson	Oy Oranı	Korelasyon Katsayısı					,172	
		Anlamlılık Düzeyi (Sig.)					,074	
		Veri Sayısı (N)					109	

**** Korelasyon 0,01 düzeyinde anlamlıdır.**

Makine öğrenmesi ile tahmin modelleri

Tablo 3 ve 4'de verilmiş olan, günlük oy oranlarını (Economist, 2020) toplam, pozitif, negatif, net pozitif (NPS) ve nötr tweet sayılarını (bağımsız değişkenler) kullanarak tahmin için 6 farklı makine öğrenmesi algoritmasından faydalanılmıştır. Makine öğrenmesi modellerinde, bir adaya ait oy oranı y_{ij} tahmin edilecek veri olarak belirlenmiştir. Adayın kendisi ve rakibi hakkında paylaşılan tweet sayıları ise x_{ijk} (10 farklı değişken) tahmin için kullanılacak bağımsız değişkenler olarak seçilmiştir.

Değişkenlerin seçilme işleminden sonra, yazılımın sunduğu altı farklı makine öğrenmesi algoritmasının hepsi seçilerek, analizler yapılmıştır. Kullanılan makine öğrenmesi algoritmaları aşağıda sıralanmıştır:

- Generalized Linear Model (RapidMiner, 2020a), geleneksek doğrusal modellerin uzantısıdır. Verilerin tahmin edilebilirliğini maksimize ederek genelleştirilmiş doğrusal modellere uyarlar.
- Deep Machine Learning Model (RapidMiner, 2020b), geri yayılım kullanılarak stokastik eğim inişiyle eğitilen çok katmanlı ileri beslemeli yapay sinir ağına dayanır. Her hesaplama düğümü, global model parametrelerinin bir kopyasını kendi yerel verileri üzerinde çoklu iş parçacıklı (eşzamansız şekilde) ile eğitir ve ağ üzerinde model ortalaması şeklinde periyodik olarak küresel modele katkıda bulunur. Deep Learning Model olarak da isimlendirilmektedir.
- Decision Tree Model (RapidMiner, 2020c), bir sınıfa bağlı değerler hakkında bir karar veya sayısal bir hedef değerin tahminini oluşturmaya çalışan bir düğüm koleksiyonudur. Her düğüm, belirli bir öznitelik için bölme kuralını temsil eder. Sınıflandırma için bu kural farklı sınıflara ait değerleri ayırır, regresyon için, seçilen parametre kriteri için hatayı en uygun şekilde azaltmak amacıyla ayırır.
- Random Forest Model (RapidMiner, 2020d), ağaç sayısı parametresiyle belirtilen, belirli sayıda rastgele ağaçlardan oluşmuştur. Bu ağaçlar girişte sağlanan ExampleSet'in önyüklemeli alt kümeleri üzerinde oluşturulur. Sınıflandırma için kural, farklı sınıflara ait değerleri ayırmaksa, regresyon içinde tahminin yaptığı hatayı azaltmak için bunları ayırır.
- Gradient Boosted Trees Model (RapidMiner, 2020e), regresyon veya sınıflandırma ağacı modellerinden oluşan bir topluluktur. Her ikisi de, kademeli olarak geliştirilen tahminler yoluyla tahmine dayalı sonuçlar elde eden ileriye yönelik öğrenme yöntemleridir.
- Support Vector Machine Model (RapidMiner, 2020f), Stefan Rueping'in destek vektör makinesi olan mySVM'nin Java uygulamasını kullanır. Bu öğrenme yöntemi hem regresyon hem de sınıflandırma için kullanılabilir. Birçok öğrenme görevi için hızlı bir algoritmadır ve iyi sonuçlar sağlar. mySVM, doğrusal veya ikinci dereceden ve hatta asimetrik kayıp işlevleriyle çalışır.

Makine öğrenmesi algoritmaları ile 109 günlük anket sonuçları ve 110. güne ait gerçek seçim sonucu ayrı ayrı tahmin edilmeye çalışılmıştır. Günlük Twitter verileri ile oy oranlarının günlük değişiminin tahmin edilebilirliği incelenmiştir. Analiz sonuçları bulgular ve tartışma kısmında verilmiştir.

Bulgular ve tartışma

Her iki parti adayı ile ilgili veriler için kurulan ilişki hipotezleri sonuçlarına göre sadece 4 nolu hipotezde Trump için kurulan hipotezin istatistiksel olarak anlamsız olduğu sonucuna varılmıştır. Hipotez testleri sonucunda elde edilen bulgular aşağıda sıralanmıştır:

H1; bir aday hakkında atılmış olan toplam tweetlerin sayısı ile o adaya verilen günlük oy oranı arasındaki ilişkiyi göstermektedir. 2 aday için de anlamlı sonuç çıkmıştır. Trump için pozitif yönlü orta düzeyde korelasyon olduğu sonucuna varılır iken, Biden için ise negatif yönlü orta düzeyde korelasyon olduğu görülmektedir. Bu durum, Trump hakkında Twitter'da bahsedilme sayısı arttıkça oy oranı arttığını ancak Biden için tersine bir etki olduğu düşünülebilir.

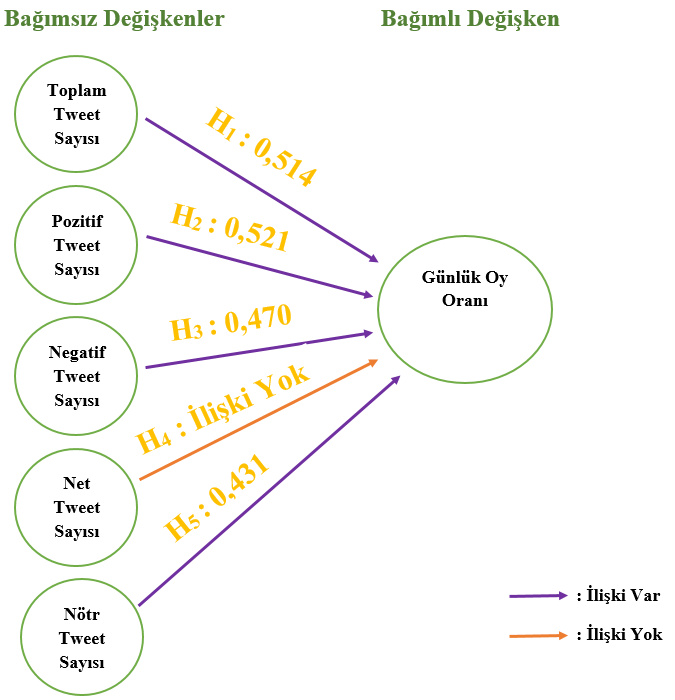
H2; bir aday hakkında atılmış olan pozitif tweetlerin sayısı ile o adaya verilen günlük oy oranı tahmini arasındaki ilişkiyi göstermektedir. Trump için pozitif yönlü orta düzeyde korelasyon, Biden için negatif yönlü orta düzeyde korelasyon olduğu sonucuna varılmıştır. Trump için elde edilen sonuç beklenen bir durumdur. Nitekim aday hakkında olumlu paylaşımların fazla olması oy oranının da fazla olması anlamına gelmektedir. Ancak Biden için beklenmedik şekilde tersine bir korelasyon görülmüştür.

H3; bir aday hakkında atılmış olan negatif tweetlerin sayısı ile o adaya verilen günlük oy oranı tahmini arasındaki ilişkiyi göstermektedir. Bu hipotez testinde, beklenildiği gibi Biden için negatif yönlü orta düzeyde korelasyon tespit edilirken, pozitif yönlü orta düzeyde korelasyon sonucu nedeniyle bu sefer Trump için beklenmedik bir durum görülmüştür.

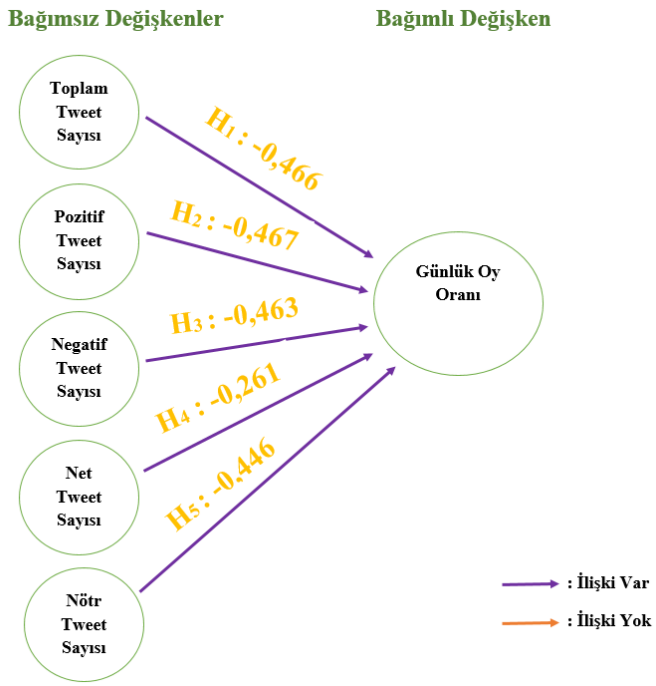
H4; bir aday hakkında atılmış olan net pozitif tweetlerin sayısı ile o adaya verilen günlük oy oranı tahmini arasındaki ilişkiyi göstermektedir. Trump için anlamlı sonuç çıkmaz iken, Biden için negatif yönlü zayıf düzeyde korelasyon olduğu sonucuna varılmıştır. Biden için elde edilen sonuç yine beklenmedik bir durum göstermektedir.

H5; bir aday hakkında atılmış olan nötr tweetlerin sayısı ile o adaya verilen günlük oy oranı tahmini arasındaki ilişkiyi göstermektedir. Trump için pozitif yönlü orta düzeyde korelasyon, Biden için negatif yönlü orta düzeyde korelasyon olduğu sonucuna varılmıştır.

Elde edilen bulgular görselleştirilerek tartışılmıştır. İlk yapılan analizler olan korelasyon analiz sonuçları Şekil 5 ve 6'da iki farklı aday için ayrı ayrı gösterilmiştir.



Şekil 5: Trump ile ilgili verilerin korelasyon analizi sonuçları



Şekil 6: Biden ile ilgili verilerin korelasyon analizi sonuçları

Trump hakkındaki Tweet sayılarının Biden'a göre çok daha fazla olmasına bakılacak olursa korelasyon analizinden sonra regresyon analizinde seçimi büyük farkla Trump'ın kazanacağı tahmin edilirdi. Ancak, anket verileri böyle bir durumun pek olası olmadığını göstermektedir. Ayrıca, korelasyon analiz sonuçlarında görülen anormallikler (pozitif tweet sayısı ile ters, negatif tweet sayısıyla doğru orantılı oy oranları) oy oranı ile tweet sayıları arasında doğrusal bir ilişki olmayabileceğini göstermektedir.

Korelasyon analizleri değişken arasındaki ilişkinin tespiti ve doğrusal regresyon analizi için önemli olsa da makine öğrenmesi algoritmaları ile, aralarında korelasyon bulunamayan değişkenler arasında bir bağ kuran tahmin modeli oluşturmak mümkündür. Bu nedenlerle, gün oy oranlarını tahmin için korelasyon sonuçlarından bağımsız olarak, makine öğrenmesi algoritmalara dayalı tahmin modeli oluşturulmuştur. Nitekim, gerçek seçim sonucunda olduğu gibi, Makine öğrenmesi tabanlı tahmin modeli oy oranlarının tweet sayısıyla doğru orantılı olmadığını ve Trump hakkında çok sayıdaki tweet'e rağmen, seçimi çok az farklı Biden'ın kazanabileceği yönünde tahmin sonucu vermiştir.

Anket sonuçlarını en iyi tahmin edebilen Derin Makine Öğrenmesi (Deep Learning) algoritması olmuştur. Altı farklı tahmin modelinin, günlük oy oranı anket sonuçları tahminindeki mutlak hata değerleri Tablo 9'da verilmiştir.

Tablo 9: Tahmin modeli hata sonuçları

Tahmin Modeli	Trump Oy Oranı Tahmini için Mutlak Hata	Biden Oy Oranı Tahmini için Mutlak Hata
Generalized Linear Model	0,233	0,233
Deep Learning*	0,205	0,209
Decision Tree	0,243	0,243
Random Forest	0,226	0,227
Gradient Boosted Trees	0,235	0,235
Support Vector Machine	0,225	0,225

En düşük hata oranına sahip Derin Makine Öğrenmesi (Deep Learning) algoritması gerçek seçim sonucunu tahmini için seçilmiştir. Gerçek seçim sonucu tahmini için, 109 günlük veri ile eğitilen derin öğrenme tahmin modeli kullanılmıştır. Gerçek seçim sonucu tahmini 2 farklı tür veri için ayrı ayrı yapılmıştır. İlk olarak Tablo 10'da verilen, sadece 3 Kasım'daki tek günlük tweet sayıları modelin girdi parametreleri olarak kullanılarak tahmin yapılmıştır. İkincisinde ise, Tablo 11'de verilen 3 Kasım'a kadar olan kümülatif değerler (109 günlük verinin toplam değerleri) kullanılarak seçim sonucu tahmin edilmiştir.

Tablo 10: 3 Kasım'daki günlük veriler

Tarih	Biden Toplam Tweet Sayısı	Biden Pozitif Tweet Sayısı	Biden Negatif Tweet Sayısı	Biden Net Tweet Sayısı	Biden Nötr Tweet Sayısı	Trump Toplam Tweet Sayısı	Trump Pozitif Tweet Sayısı	Trump Negatif Tweet Sayısı	Trump Net Tweet Sayısı	Trump Nötr Tweet Sayısı
3.11.2020	73490	30493	15281	15212	27716	175225	70643	48056	22587	56526

Tablo 11: 3 Kasım'daki kümülatif veriler

Tarih	Biden Toplam Tweet Sayısı	Biden Pozitif Tweet Sayısı	Biden Negatif Tweet Sayısı	Biden Net Tweet Sayısı	Biden Nötr Tweet Sayısı	Trump Toplam Tweet Sayısı	Trump Pozitif Tweet Sayısı	Trump Negatif Tweet Sayısı	Trump Net Tweet Sayısı	Trump Nötr Tweet Sayısı
3.11.2020	3543814	1405302	1032566	372736	1105946	11741934	4234653	4276070	-41417	3231211

Derin öğrenme tahmin modeliyle, Biden ve Trump oy oranları tahminleri ayrı ayrı yapılmıştır. Tablo 12'de 3 kasıma ait tek günlük veriler ve kümülatif veriler kullanılarak, Biden ve Trump için yapılan oy oranları tahminleri ve bu tahminlerin hata oranları gösterilmiştir.

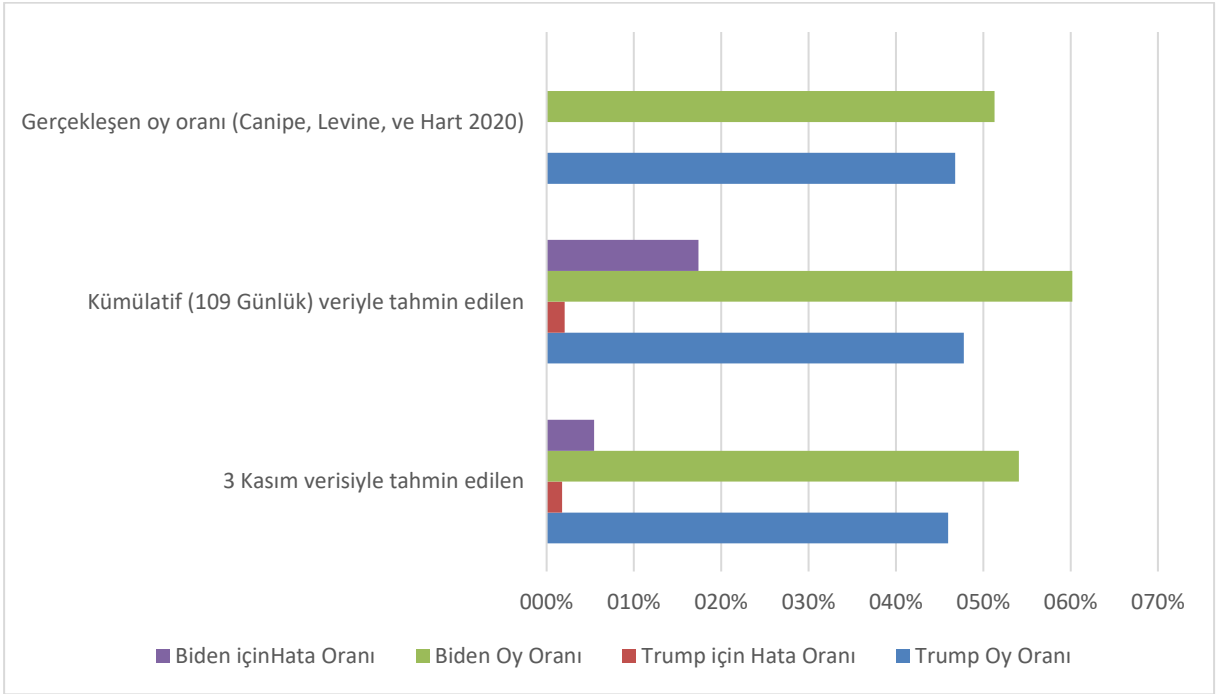
Tablo 12: Tahmin edilen ve gerçekleşen oy oranları

Değişken Adı	Trump Oy Oranı	Trump için Hata Oranı	Biden Oy Oranı	Biden için Hata Oranı
3 Kasım verisiyle tahmin edilen	45,97%	1,77%	54,09%	5,45%
Kümülatif (109 Günlük) veriyle tahmin edilen	47,77%	2,08%	60,22%	17,38%
Gerçekleşen oy oranı (Canipe, Levine ve Hart, 2020)	46,80%		51,30%	

Yapılan tahminleri hata oranının hesaplamak için aşağıdaki mutlak hata oranı formülü kullanılmıştır:

Hata Oranı = | (Gerçek değer – Tahmin değeri) / (Gerçek değer) | * 100

Tablo 12'deki veriler Şekil 7'de görselleştirilmiştir. Şekilden de anlaşılabileceği üzere 3 kasıma ait tek günlük veriler kullanılarak yapılan tahmin hem Biden hem Trump adayı için daha doğru sonuç vermiştir. Bu durumun, 109 günlük süreçte adaylarının taraftar sayısındaki yani oy oranlarındaki değişimden kaynaklanabileceği düşünülmektedir. Bu nedenle son gün elde edilen tweet sayıları, en güncel taraftar oranını yansıtabildiği için seçim sonucunun çok daha doğru tahmin edebilmesini sağladığına inanılmaktadır.



Şekil 7: Oy oranı tahmin ve gerçekleşen sonuç grafikleri

Sonuç

Değişen ve gelişen dijital ortam, sosyal medya kullanımının yoğunlaşması ve genişlemesi, bununla birlikte sosyal medyanın manipülasyon aracı olarak kullanılması, farklı ülkelerin, örneğin Rusya'nın Amerika seçimlerine müdahalesi için sosyal medyayı kullanması gibi güncel gelişmeler (Jamieson, 2020), (Golovchenko vd., 2020), (Karami vd., 2021) sosyal medyadaki verilerin güvenirliğini sorgulatmaktadır. Siyasi eğilimleri takip etmekte sosyal medyanın halen geçerli bir araç olup olmadığı araştırılması gereken bir konudur. Bu çalışmada güncel bir seçim olan 2020 Amerikan seçimleri ele alınarak, Twitter verilerinin halen seçimleri tahmin etmede ve siyasi eğilimleri belirlemede faydalı bir veri kaynağı olarak kullanıp kullanamayacağını araştırılmıştır.

Çalışma kapsamında öncelikle Twitter sosyal medya verileri ve anket verilerine dayalı günlük oy oranı tahminleri arasında korelasyon analizleri yapılmıştır. Nitekim, literatürdeki çeşitli çalışmalar (Kušen ve Strembeck, 2018), (Grover vd., 2019) Twitter verileri ile oy oranı arasında pozitif yönlü güçlü korelasyon olduğunu iddia etmektedir. Ancak, bu çalışma kapsamında yapılan korelasyon analizlerinin sonuçlarında anormallikler tespit edilmiştir. Örneğin Trump hakkındaki negatif tweet sayısıyla oy oranı arasında pozitif bir ilişki varken, Biden hakkındaki pozitif tweet sayısıyla oy oranı arasında negatif bir ilişki söz konusudur. Ayrıca eğer tweet sayısı doğrusal bir ilişki gösterseydi, Trump hakkındaki Tweet sayılarının (11.741.934) Biden hakkında paylaşılanlardan (3.543.814) üç kattan daha fazla olmasından dolayı, seçimi büyük farkla Trump'ın kazanacağı tahmin edilirdi. Bu bulgular nedeniyle, tweet sayıları ile oy oranı arasında doğrusal bir ilişki olmadığını ve regresyon analizi gibi klasik yöntemlerle oy oranı tahmin etmeninin mümkün olamayacağı sonucuna ulaşılmasını sağlamıştır. Elde edilen bu bulgular nedeniyle, geleneksel istatistiksel tahmin modelleri yerine yapay zekâ tabanlı makine öğrenmesi algoritmalarından faydalanarak seçim sonucu tahmin edilmeye çalışılmıştır.

Derin Makine Öğrenmesi yöntemi kullanılan tahmin modeli, 109 günlük Twitter verisi ve günlük oy oranları verileri ile eğitilmiştir. Ardından, gerçek seçim sonucunu tahmin için bu model kullanılmıştır. Araştırma sonucunda derin makine öğrenmesi yöntemiyle, Twitter sosyal medya verileri kullanılarak günlük oy oranlarının tahmin edilebileceği, dolayısıyla politik eğilimlerin değişiminin takibi ve seçim sonucunun tahmini için bu verilerin kullanılabileceği sonucuna ulaşılmıştır. Nitekim, Twitter verileri kullanılarak, 2020 ABD genel seçim sonuçları, Trump adayı için sadece %1,77, Biden için ise %5,45 hata oranıyla doğru şekilde tahmin edilebilmiştir. Bu sonuç Twitter verilerinin halen seçim sonuçlarının tahmini için kullanabilecek önemli bir veri kaynağı olabileceğini göstermektedir.

Tahmin modellerinde 6 farklı makine öğrenmesi algoritması kullanılmıştır. Bunların içerisinde her iki aday için de en az hata payı ile oy oranı tahmini yapabilen yöntem derin makine öğrenmesi olmuştur. Elde edilen bu sonuç, araştırmanın birincil amacı olmasa da sosyal medya verilerine dayalı politik eğilim takibi ve seçim sonuçları araştırmalarında kullanılabilecek en uygun yönteminin derin makine öğrenmesi olabileceği bulgusunu da sağlamıştır.

Sonuç olarak, bu çalışma sosyal medyanın seçim sonuçları tahmini için halen kullanılabilecek bir veri kaynağı olduğunu göstermektedir. Ayrıca, derin makine öğrenmesi metodunu bu amaçla kullanan ilk çalışma olarak literatüre katkı sağlamaktadır. Geliştirilen tahmin modeli, seçim tahmini yapan araştırma şirketlere, ilgili parti yönetimi ve politikacılara ve politik pazarlama, yönetim ve tahmin modelleri üzerinde çalışma yapacak araştırmacılara fayda sağlayacağı düşünülmektedir.

Hakem Değerlendirmesi / Peer-review:

Dış bağımsız

Externally peer-reviewed

Çıkar Çatışması / Conflict of interests:

Yazar(lar) çıkar çatışması bildirmemiştir.

The author(s) has (have) no conflict of interest to declare.

Finansal Destek / Grant Support:

Yazar(lar) bu çalışma için finansal destek almadığını beyan etmiştir.

The author(s) declared that this study has received no financial support.

Yazar Katkıları / Author Contributions:

Fikir/Kavram/Tasarım - *Idea/Concept/ Design*: **İ.S**. Veri Toplama - *Data Collection*: **İ.S.** Analiz ve/veya Yorum - *Analysis and/or Interpretation:* **İ.S.,E.Ş.**., Kaynak Taraması - *Literature Review:* **İ.S.,E.Ş.**, Makalenin Yazımı - *Writing the Article*: **İ.S.,E.Ş.**, Eleştirel İnceleme - *Critical Review*: **İ.S.,E.Ş.**, Onay - *Approval*: **İ.S.,E.Ş.**

Kaynakça / References

Bansal, B. ve Srivastava, S. (2018). On predicting elections with hybrid topic based sentiment analysis of tweets. *Procedia Computer Science*, *135*, 346–353. doi:10.1016/j.procs.2018.08.183

Burnap, P., Gibson, R., Sloan, L., Southern, R. ve Williams, M. (2016). 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Electoral Studies*, *41*, 230–233. doi:10.1016/j.electstud.2015.11.017

Canipe, C., Levine, A. J. ve Hart, S. (2020). U.S. election results. 18 Şubat 2021 tarihinde https://graphics.reuters.com/USA-ELECTION/RESULTS-LIVE-US/jbyprxelqpe/ adresinden erişildi.

Castro, R., Kuffó, L. ve Vaca, C. (2017). Back to #6D: Predicting Venezuelan states political election results through Twitter. *2017 4th International Conference on eDemocracy and eGovernment, ICEDEG 2017*, 148–153. doi:10.1109/ICEDEG.2017.7962525

Cerf, V. G. (2017). Information and misinformation on the Internet. *Commun. ACM(CACM)*, *60*(1), 9. doi:10.1145/3018809

Ceron Guzman, J. A. (2016). A Sentiment Analysis Model of Spanish Tweets.

Chatfield, A., Reddick, C. ve Choi, K. (2017). *Online Media Use of False News to Frame the 2016 Trump Presidential Campaign*. doi:10.1145/3085228.3085295

Conway, B. A., Kenski, K. ve Wang, D. (2015). The rise of Twitter in the political campaign: searching for intermedia agenda-setting effects in the presidential primary. *J. Comput. Mediat. Commun.*, *20*(4), 363–380.

Economist, T. (2020). Forecasting the US elections. 12 Aralık 2020 tarihinde https://projects.economist.com/us-2020-forecast/president adresinden erişildi.

Golbeck, J., Grimes, J. M. ve Rogers, A. (2010). Twitter use by the U.S. Congress. *J. Am. Soc. Inf. Sci. Technol.*, *61*(8), 1612–1621.

Golovchenko, Y., Buntain, C., Eady, G., Brown, M. A. ve Tucker, J. A. (2020). Cross-platform state propaganda: Russian trolls on Twitter and YouTube during the 2016 US presidential election. *The International Journal of Press/Politics*, *25*(3), 357–389.

Graham, T., Jackson, D. ve Broersma, M. (2016). New platform, old habits? Candidates use of Twitter during the 2010 British and Dutch general election campaigns. *Sage Journals*, *18*(5), 765–783.

Grover, P., Kar, A. K., Dwivedi, Y. K. ve Janssen, M. (2019). Polarization and acculturation in US Election 2016 outcomes – Can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, *145*(September), 438–460. doi:10.1016/j.techfore.2018.09.009

Jamieson, K. H. (2020). *Cyberwar: how Russian hackers and trolls helped elect a president: what we don't, can't, and do know*. Oxford University Press.

Karami, A., Lundy, M., Webb, F., Turner-McGrievy, G., McKeever, B. W. ve McKeever, R. (2021). Identifying and analyzing health-related themes in disinformation shared by conservative and liberal Russian trolls on twitter. *International journal of environmental research and public health*, *18*(4), 2159.

Kelly Garrett, R. ve Weeks, B. E. (2013). The promise and peril of real-time corrections to political misperceptions. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, (February 2013), 1047–1057. doi:10.1145/2441776.2441895

Kim, A. J. ve Ko, E. (2010). Impacts of luxury fashion brand's social media marketing on customer relationship and purchase intention. *J. Glob. Fash.Mark.*, *1*(3), 164–171.

Kušen, E. ve Strembeck, M. (2018). Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. *Online Social Networks and Media*, *5*, 37–50. doi:10.1016/j.osnem.2017.12.002

Makazhanov, A., Rafiei, D. ve Waqar, M. (2014). Predicting political preference of Twitter users. *Social Network Analysis and Mining*, *4*(1), 1–15. doi:10.1007/s13278-014-0193-5

RapidMiner. (2020a). Generalized Linear Model. 6 Kasım 2020 tarihinde https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/functions/generalized_linear_model.html adresinden erişildi.

RapidMiner. (2020b). Deep Learning. 6 Kasım 2020 tarihinde https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/neural_nets/deep_learning.html adresinden erişildi.

RapidMiner. (2020c). Decision Tree. 6 Kasım 2020 tarihinde https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_decision_tree.html adresinden erişildi.

RapidMiner. (2020d). Random Forest. 6 Kasım 2020 tarihinde https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_random_forest.html adresinden erişildi.

RapidMiner. (2020e). Gradient Boosted Trees. 6 Kasım 2020 tarihinde https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/gradient_boosted_trees.html adresinden erişildi.

RapidMiner. (2020f). Support Vector Machine. 6 Kasım 2020 tarihinde https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/support_vector_machines/support_vector_machine.html adresinden erişildi.

Sabuncu, İ. (2020). USA Nov.2020 Election 20 Mil. Tweets (With Sentiment And Party Name Labels) Dataset. 20 Kasım 2020 tarihinde https://ieee-dataport.org/open-access/usa-nov2020-election-20-mil-tweets-sentiment-and-party-name-labels-dataset adresinden erişildi.

Toker, H., Erdem, S. ve Özşarlak, P. (2017). 2015 Haziran Ve Kasım Seçimlerinde Siyasal Eğilim: Yeni Bir Kamuoyu Ölçümleme Aracı Olarak Twitter. *Erciyes İletişim Dergisi*, *5*(1), 221–234.

Tumasjan, A., Sprenger, T. O., Sandner, P. G. ve Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Fourth international AAAI conference on weblogs and social media* içinde (C. 37, ss. 455–479). Citeseer. doi:10.15581/009.37.2.455-479

Ulusoy, N. (2012). Sözlüklerdeki Sinema Sevgisi: New York'ta Beş Minare ve Çoğunluğun İnternet Sözlüklerine Yansıması. *Beta Yayıncılık, İstanbul*, 195–211.

Wicaksono, A. J., Suyoto ve Pranowo. (2017). A proposed method for predicting US presidential election by analyzing sentiment in social media. *Proceeding - 2016 2nd International Conference on Science in Information Technology, ICSITech 2016: Information Science for Green Society and Environment*, 276–280. doi:10.1109/ICSITech.2016.7852647