SİYASET BİLİMİNDE OTOMATİK METİN ANALİZİ YÖNTEMLERİ VE UYGULAMA ALANLARI

Betül AYDOĞAN ÜNAL[1]

Atıf: Aydoğan Ünal, B. (2023). Siyaset biliminde otomatik metin analizi yöntemleri ve uygulama alanları. *Hitit Sosyal Bilimler Dergisi, 16(1),* 190-208. doi: 10.17218/hititsbd.1260739

Özet: Otomatik metin analizi, büyük boyuttaki metin verilerini daha önce mümkün olmayan yollarla analiz etme yeteneği sayesinde siyaset biliminde hızla büyüyen bir alan haline gelmiştir. Ancak, metinsel verileri analiz etmek için pek çok farklı yöntemin bulunması, araştırmacıların araştırma soruları ve verileri için en uygun yaklaşımı belirleme sürecini zorlaştırmaktadır. Bu makale, siyasi olguları incelemek için kullanılan farklı otomatik metin analizi yöntemleri arasından basit istatistiksel analizler, denetimli/denetimsiz makine öğrenmesi, dağılımsal semantik modeller ve kelime gömme yöntemlerini ele alarak araştırmacılara kapsamlı bir kaynak sunmayı amaçlamaktadır. Basit sıklık dağılımlarının hesaplanması ve benzerlik/uzaklık ölçümlerinin kullanımı gibi temel yöntemlerin yanı sıra daha gelişmiş yöntemlerin temel varsayımları, ürettiği çıktılar, güçlü ve zayıf yönleri karşılaştırmalı olarak ele alınmaktadır. Bu çalışma, bu yöntemlerin siyaset bilimine katkı sağlama potansiyelini vurgulamakla birlikte uygulama alanlarından örnekler sunmaktadır.

Anahtar Kelimeler: *Otomatik Metin Analizi, Siyaset Bilimi, Büyük Veri, Makine Öğrenmesi, Araştırma Yöntemleri*

Automated Text Analysis Methods and Application Areas in Political Science

Citation: Aydoğan Ünal, B. (2023). Automated text analysis methods and application areas in political science. *Hitit Journal of Social Sciences, 16(1),* 190-208. doi: 10.17218/hititsbd.1260739

Abstract: Automated text analysis has become a rapidly growing field in political science due to its ability to analyze large-scale textual data in ways that were not previously possible. However, because there are many different methods available for analyzing textual data, it can be difficult for researchers to choose the most appropriate approach for their research questions and data. This article provides a general overview of the use of statistical summaries, supervised and unsupervised machine learning, distributional semantic models, and word embedding methods for examining political phenomena. It compares the data requirements, outputs produced, basic assumptions, advantages, and disadvantages of not only basic methods such as calculating simple frequency distributions and similarity/distance measurements but also more advanced methods. While emphasizing the potential contribution of these methods to political science, this study provides examples from application areas.

Keywords: *Automated Text Analysis, Political Science, Big Data, Machine Learning, Research Methods*

1. GİRİŞ

Dil, insanların birbirleriyle duygu ve düşüncelerini paylaşmak için kullandıkları araçtır. Dil, ayrıca insanların sosyal grup ve topluluklar oluşturmalarına olanak sağlar. Aynı zamanda dil, insanların çevrelerindeki dünyayı anlamalarına ve bilgileri işlemelerine yardımcı olmada önemli bir rol oynar. İnsanlar dili kullanarak çevrelerini ve yaşamlarında olup bitenleri anlamlandırabilirler (Gee, 2018). Tüm bu iletişimin aktarıldığı metinlerin analiz edilebilmesi, anlamlarının anlaşılabilmesi ve buna dayanarak sonuçlar çıkarılabilmesi siyaset bilimi açısından

İnceleme Makalesi / Review Article
[1] Dr. Öğr. Üyesi, Ege Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Uluslararası İlişkiler Bölümü, betul.aydogan@ege.edu.tr | http://orcid.org/0000-0003-2371-0921 | https://ror.org/02eaafc18
Asst. Prof. Dr., Ege University, Faculty of Economics and Administration Sciences, Department of International Relations, betul.aydogan@ege.edu.tr | http://orcid.org/0000-0003-2371-0921 | https://ror.org/02eaafc18

oldukça değerlidir. Siyaset biliminde araştırmalar, etrafımızdaki dünya hakkında bilgi üretmek için büyük ölçüde metinlere dayanmaktadır (Grimmer ve Stewart, 2013). Tarihsel olarak bakıldığında metinsel verileri analiz etmede kullanılan ilk yöntem, metinlerin araştırmacılar tarafından yorumlanmasına dayanan nitel metin analizidir (Wesley, 2014). Bu yöntemsel yaklaşımda, metinlerin onları oluşturan kişilerin tutumlarının veya fikirlerinin bir ürünü olduğu ve bu nedenle kişilerin görüşleri hakkında bir bilgi kaynağı olarak analiz edilebileceği kabul edilmektedir. Bunun için, araştırmacılar metinlerde neyin, nasıl ve hangi bağlamda söylendiğini analiz ederek metni daha derinden anlamayı amaçlarlar. Genel olarak amaç, incelenen kişi veya kuruluşun söylemlerinin bağlamıyla ilişkili daha iyi anlaşılmasını sağlamaktır. Son yıllarda metin madenciliği araçlarının geliştirilmesinden kaynaklı olarak hem erişilebilen verinin boyutu her geçen gün artmakta hem de büyük metin verilerini analiz etme süreci bilgisayar desteğiyle otomatikleşerek daha verimli bir hale gelmektedir (Benoit, 2020). Tüm bu gelişmeler, araştırmacıların daha önce mümkün olandan daha büyük çaplı ve daha nesnel analizler yapmasına olanak sağlamaktadır (Gül ve Nizam, 2021).

Siyaset bilimi alanında otomatik metin analizi yöntemleri, büyük miktarlardaki metinsel veriden anlamlı sonuçlar çıkarılması için kullanılan teknikleri nitelemektedir (Budge ve Pennings, 2007). İstatistiksel ve matematiksel yöntemlerle, metnin çeşitli yönlerden ölçülmesi, içindeki eğilimlerin ve ilişkilerin ortaya çıkarılması amaçlanmaktadır. Aynı zamanda bu yöntemler metinleri belirli temalara, duygulara ve ideolojilere göre sınıflandırılmak için de kullanılabilmektedir (Schoonvelde ve diğerleri, 2019). Bu yöntemlerle, bireylerin, grupların ve kurumların görüş ve tutumlarının analizi için ampirik bir temel sağlanmaktadır. Monroe ve Schrodt (2008, s.351), metni "siyasi davranışın en yaygın ve kesinlikle en kalıcı ürünü" olarak nitelemektedir. Siyaset bilimi araştırmalarında büyük miktarda metin verisini analiz etmek için hesaplama yöntemlerinin geliştirilmesine ve uygulanmasına önemli katkıları bulunan Kenneth Benoit'a (2020) göre; politikacıların davranışları onların ne hissettikleri hakkında bilgi verir, ancak politikacıların gerçek duygularını bilmek açısından onların ne söyledikleri daha önemlidir. Siyaset biliminde otomatik metin analizinde, resmî belgeler, tutanaklar, gazete makaleleri, raporlar, blog yazıları ve sosyal medya iletileri sıklıkla kullanılan metin türleridir. Ses ve video kayıtlarının transkriptleri gibi diğer türler de metne dönüştürülerek (Atalay ve Çelik, 2017) diğer metin tabanlı içerikler gibi analiz edilebilir.

Bu makalenin amacı, siyaset bilimi alanında en sık kullanılan otomatik metin analizi yöntemlerine genel bir bakış sağlamak ve bu yöntemlerin metinlerden anlam çıkarmak için nasıl kullanılabileceğini literatürden örneklerle açıklamaktır. Böylece siyaset bilimi araştırmacıları ve öğrencileri için otomatik metin analizi yöntemleri dünyasına bir giriş sağlamak ve onları kendi araştırma projelerini tasarlarken bilinçli kararlar vermeleri için gerekli bilgilerle donatmak hedeflenmektedir. Siyaset bilimi için önemli katkılar sunma potansiyeli taşıyan otomatik metin analizi yöntemleri üzerine olan Türkçe yayın konusunda önemli bir ihtiyaç mevcuttur[2]. Bu makale, siyaset bilimi ve sosyal bilimler alanındaki araştırmaların bu yöntemlerin daha yaygın bir şekilde kullanılabilir hale getirilmesi amacına yönelik olarak literatüre katkı sunmayı hedeflemektedir. Yeni analiz yöntemleri hakkında bilgilerin Türkçe olarak kullanıma sunulması, siyaset bilimi ve sosyal bilimler alanında Türkçe yayın yapan araştırmacılar, öğrenciler ve

[2] Uluslararası İlişkiler metodolojisi açısından Hatipoğlu ve diğerleri (2022) tarafından yazılan "Otomatik Metin Analizi ve Uluslararası İlişkiler" adlı kitap bölümü Türkçe kaynak olarak önemli bir katkı sunmaktadır. Muhasebe alanında Özyiğit (2022) "Muhasebe Alanına Güncel Yaklaşımlar: Metin Madenciliği" adlı makalesiyle metin analizinin uygulamasına dair Türkçe kaynak sunmaktadır. Halkla İlişkiler alanında ise, Özoran (2022) "Bir Halkla İlişkiler Aracı olarak Twitter: Dünya Sağlık Örgütü Paylaşımlarının İçerik Analizi ve Metin Madenciliği ile İncelenmesi" adlı makalesiyle metin madenciliği yöntemlerinin alandaki uygulamasını açıklamaktadır.

uygulamacılar için yöntemsel bilgi alanını ilerletmeye yardımcı olabilme potansiyeline sahiptir. Böylece, yeni anlayışların ve bakış açılarının ortaya çıkmasını sağlanarak alanın büyümesine ve gelişmesine katkıda bulunulabilecektir.

Çalışmada ilk olarak, metinsel veri kavramına ve siyaset bilimi alanında metinsel veri kaynaklarına kısa bir giriş yapılacaktır. Daha sonra, bu alanda kullanılan başlıca otomatik metin analizi yöntemleri literatürden örneklerle tanıtılacaktır. Bu makalede, sosyal bilimlerde otomatik metin analizi yöntemleri alanında son dönemde kullanılmaya başlanan ve Van Loon (2022) tarafından "aile" olarak nitelendirilen gruplama benimsenmektedir. İlk aile olan basit istatistiksel analizler, frekans analizi gibi belirli kelimelerin ne sıklıkta görüldüğünü sayarak metni temsil etmektedir. İkinci aile olan makine öğrenmesiyle belge yapısı analizi, sözcüklerin birlikte olma istatistiklerinden anahtar sözcükleri veya temaları çıkarsamaktadır. Üçüncü aile olan dağılımsal semantik benzerlik analizi, kelimelerin anlamını sayısallaştırmakta ve metinleri bu tür anlamların koleksiyonları olarak temsil edilmektedir. Tartışma bölümünde, bu yöntemlerin siyaset bilimi araştırmaları için uygulama potansiyelleri karşılaştırmalı olarak ele alınacaktır.

2. OTOMATİK METİN ANALİZİ YÖNTEMLERİ

Dijitalleşmenin etkisiyle resmi tutanaklar, basın bildirileri, haber makaleleri ve sosyal medya paylaşımları gibi siyaset biliminde analiz edilebilecek metinsel veri kaynaklarında büyük oranda artış yaşanmıştır. Otomatik metin analizi açısından veri çeşitliliği genel hatlarıyla yapılandırılmış ve yapılandırılmamış olarak ikiye ayrılır (Kılıç ve diğerleri, 2019). Aralarındaki temel fark; yapılandırılmış metin verileri analiz etmeyi kolaylaştıran tutarlı bir formata sahipken, yapılandırılmamış metin verileri ise düzensiz ve karmaşıktır. Yapılandırılmış metin verilerinin işlenmesi, veri içinde arama, sıralama ve filtreleme yapılması genellikle kolaydır. Önceden tanımlanmış bir yapıya veya biçime sahip olmayan yapılandırılmamış metin verileri, günlük dil, kısaltmalar ve yazım hataları içerebilir. Bu sebeple, bu metinlerden anlamlı sonuçlar elde etmek için daha fazla işlem gerekmektedir (Konşuk Ünlü, 2022).

Siyaset bilimi için başlıca metinsel veri kaynakları şunlardır:

- *Meclis tutanakları, yasama belgeleri ve hükümet raporları:* Bu tür metinler, bir metin belgesinde konuşmacı, tarih ve konum gibi bilgileri içerdiği ve çoğunlukla belirli bir formatı takip ettiği için genel olarak yapılandırılmış olarak nitelendirilebilir (Eggers ve Spirling, 2018). Diğer yandan, meclis tutanakları gibi siyasi konuşmaları içeren kısımlar genellikle yapılandırılmamıştır ve anlam çıkarmak için doğal dil işleme tekniklerinin kullanımını gerektirebilmektedir.
- *Resmî belgeler:* Belgenin türüne bağlı olarak biçim ve içerik açısından farklılıklar gösterebilmektedirler. Aynı gruptan olan mevzuat veya anlaşma gibi belgeler genellikle belirli bir formatı takip ettiği için tarih, yazar ve kurum gibi meta verileri içerebilmektedir (Bouchart, 2020, s.522). Bu sebeple bu tür metinsel veriler kısmen yapılandırılmış olarak nitelendirilmektedir.
- *Basın açıklamaları, röportajlar ve gazete makaleleri:* Gazete makaleleri ve basın açıklamaları, biçim ve içerik açısından büyük farklılıklar gösterebileceği için bu tür metinsel veriler genellikle yapılandırılmamış olarak kabul edilmektedir. Bununla birlikte, yayın tarihi, yazar ve kaynak gibi bilgiler, verilere bir miktar yapı sağlayabilmektedir.
- *Sosyal medya paylaşımları:* Sosyal medya platformları serbest biçimli ifadeye izin verdiğinden ve günlük dil, argo ve kısaltmalar içerebileceği için bu tür metinsel veriler genellikle yapılandırılmamış olarak nitelendirilmektedir (Hatipoğlu ve diğerleri, 2022). Ancak, kullanıcı adı, tarih ve konum gibi meta veriler, verilere bir miktar yapı sağlayabilmektedir. Sosyal

medyadaki ses ve video kayıtlarının transkriptleri de metinsel veriye dönüştürülerek kullanılabilmektedir.

Veri boyutundaki artış, bu verilerin analizi için etkili yöntemlere olan ihtiyacı doğurmuştur. Büyük veri olarak da nitelenen büyük miktarlardaki metin verisinin sistematik ve nesnel analizini mümkün kılan otomatik metinsel veri analizi, metinlerin içindeki eğilimlerin ve ilişkilerin keşfedilmesini sağlayarak siyasi olgulara yeni fikirler sağlamaktadır (Slapin ve Proksch, 2008; Grimmer, 2010). Siyaset biliminde otomatik metin analizi yöntemleri önemli avantajlar taşımaktadır. Kelime sıklığı, benzerlik/uzaklık ve okunabilirlik ölçüleri gibi istatistiksel analizler, büyük miktarda metin verisini işlemek için basit ve verimli bir yol sağlamaktadır. Dil kullanımındaki kalıpları ve eğilimleri belirlemek amacıyla yararlı bir ilk adımdır. Metinleri önceden belirlenmiş kategoriler halinde sınıflandırmak için kullanılabilen denetimli makine öğrenmesi, metni daha derinlikli bir şekilde analiz etmek için güçlü bir araç olarak karşımıza çıkmaktadır. Bu yöntem, insanlar tarafından hemen fark edilemeyecek metinler arasındaki kalıpları ve bağlantıları ortaya çıkarabilecek potansiyele sahiptir. Kümeleme ve konu modelleme de dahil olmak üzere denetimsiz makine öğrenmesi, metin verilerinde daha önce bilinmeyen veya varsayılmayan yeni kalıpları ve temaları keşfetmek için kullanışlı bir yöntem olmaktadır. Metinlerden basitçe belirlenemeyecek temel kavramları veya konuları belirlemeye yardımcı olma kapasitesine sahiptir. Son olarak, dağılımsal semantik modeller ve kelime gömmeleri, kelimeler arasındaki karmaşık anlamsal ilişkilerin keşfedilmesini sağlayarak, metinlerin anlamı ve bağlamı hakkında daha zengin ve daha ayrıntılı bir anlayış sağlamaktadır. Sağladıkları bu avantajlar sebebiyle, metnin anlamı hakkında tahminlerin doğruluğunu artırmaya yardımcı olabilecek makine öğrenimi modellerinde özellik çıkarma amacıyla kullanılabilmektedir.

Bununla birlikte, bu yöntemlerin bazı kısıtlılıkları da vardır. Basit istatistiksel analizler, karmaşık dil kullanımını aşırı basitleştirilerek metindeki önemli noktaları bu sebeple gözden kaçırabilmektedir (Roberts, 2020, s.57). Denetimli makine öğrenmesi, metinlerin yanlı veya yanlış sınıflandırılmasına yol açabilecek şekilde önceden var olan kategorilere çok fazla güvenebilme riskini taşımaktadır. Denetimsiz makine öğrenmesi, yorumlanması veya genelleştirilmesi zor sonuçlar üretebilmekte ve sonuçları her zaman güvenilir veya tekrarlanabilir olmayabilmektedir (Athey, 2018, s.520). Dağılımsal semantik modeller ve kelime gömmeleri, eğitim verisi ve algoritma seçimine aşırı duyarlı olduğu için kelimelerin bağlama özgü anlamlarını yakalayamayabilir. Ayrıca, bu alandaki ileri teknikler, büyük bilgi işlem kaynaklarına veya özel yazılımlara erişim gerektirebilmektedir.

Sonuç olarak, otomatik metin analizi yöntemleri, siyaset biliminde büyük miktarda metin verisini analiz etmek için birçok avantaj sunarken, araştırmacıların hesaba katması gereken bazı sınırlamaları da beraberinde getirmektedir. Yöntem seçimi; araştırma sorusuna, metin verilerinin boyutuna ve analizde gereken ayrıntı düzeyine bağlı olmaktadır. Doğru ve geçerli sonuçlar elde etmek için her yöntemin güçlü ve zayıf yönlerinin dikkatli bir şekilde değerlendirilmesi gerekmektedir. Bu bölümde, basit istatistiksel analizler ve denetimli ve denetimsiz yöntemler, dağılımsal semantik modeller ve kelime gömmeleri gibi siyaset biliminde kullanılan çeşitli otomatik metin analizi yöntemlerine genel bir bakış sağlanacaktır. Bölüm boyunca, her yöntemin tanımı ve amacı belirtilerek, avantajları ve sınırlılıkları tartışılacaktır. Ek olarak, bu yöntemler siyaset bilimi alanındaki uygulamalarıyla örneklendirilecektir.

2.1. Basit İstatistiksel Analizler

Otomatik içerik analizi yöntemleri, mevcut metinlerin miktar ve özelliklerini kullanarak bilinmeyen sosyal gerçeklerin bazı yönlerini ve detaylarını ortaya çıkarmayı amaçlamaktadır

(Gökçe, 2006, ss.20–21). Siyaset bilimi alanında bilgisayarlaşma öncesinde uzun yıllar boyunca Osgood (1959) tarafından geliştirilen ve metinde tekrar eden temaların sayıldığı değerlendirici beyan analizi *(evaluative assertion analysis)* kullanılmıştır. Elle kodlamanın çok zaman alması sebebiyle Harvard Sosyal İlişkiler Bölümünden Philip Stone tarafından, otomatik içerik analizi için ilk bilgisayar programı olan "General Inquirer" geliştirilmiştir (Neuendorf, 2017, s.48). İçerik analizi tarihinde bir dönüm noktası olarak kabul edilen bu gelişme, siyaset bilimi alanına 1993'teki ARPA projesi ile uygulanmaya başlanmıştır. Aynı yıllarda internetin yaygınlaşmaya başlamasıyla birlikte resmi açıklamalar, parlamento tutanakları ve basın açıklamaları gibi pek çok siyasi metin ücretsiz olarak kullanılabilir hale gelmiştir (Monroe ve Schrodt, 2008). Geleneksel elle kodlama tekniklerinin yanı sıra otomatik yöntemlerin benimsenmesiyle içerik analizinin kullanımı zaman içinde önemli ölçüde artmış ve bu tür yöntemleri içeren projelerin sayısı 1980'den 2002'ye en az altı kat arttığı görülmüştür (Neuendorf, 2004).

Bu metinleri otomatik olarak analiz edilmesini mümkün kılan siyaset bilimine özel ilk örneklerden biri Benoit ve Laver (2003) tarafından geliştirilen *Wordscores* yöntemidir. Yazarlar, İngiltere, İrlanda ve Almanya'daki siyasi partilerin konumlarını incelemek için yasama konuşmalarını kelimeler biçiminde verilere dönüştürerek "dil körü" kelime puanlama tekniğini geliştirmişlerdir. *Wordscores* ile kelimeler puanlandıktan sonra metin verileri siyasi içeriği özetleyen sayısal bir temsile dönüştürülebilmektedir. Bu temsil daha sonra, siyasi söylemdeki zaman içindeki değişiklikleri izlemek, farklı politikacıların konuşmalarının içeriklerini karşılaştırmak veya konuşmaların içeriği ile siyasi sonuçlar arasındaki ilişkiyi analiz etmek gibi çeşitli amaçlar için kullanılabilmektedir. *Wordscores* yöntemi, siyaset biliminde yaygın olarak kullanılmaktadır ve metin verilerinin siyasi içeriğini yakalamakta etkili olduğu gösterilmiştir. Klemmensen ve diğerleri (2007) bu yöntemin ayrıntılı bir değerlendirmesini sağlamak için, 1945'ten 2005'e kadar Danimarka siyasi metinlerinden *Wordscores* kullanarak çıkarılan pozisyonları, elle kodlama yapan Karşılaştırmalı Manifesto Projesi'ndeki pozisyonlarla ve uzman anketlerinin sonuçlarıyla karşılaştırmıştır. Sonuçlar, *Wordscores*'un diğer kaynaklara benzer sonuçlar vererek siyasi pozisyonlar için makul zaman serileri ürettiğini göstermektedir.

Siyaset bilimi araştırmalarında basit istatistiksel analizler, frekans dağılımları ve benzerlik/uzaklık ölçüleri gibi temel veri ölçütlerinin hesaplanmasını içermektedir. Bu yöntemler, genellikle bir metindeki belli kelimelerin ne sıklıkta görüldüğüne ait dağılımı, metnin okunabilirliğini (*readability*) ve sözcüksel çeşitliliğini (*lexical diversity*) ölçmek için kullanılmaktadır. Frekans dağılımları, bir derlemdeki kelimelerin sıklığını listeleyen bir tablodur. Bu dağılımlar, en yaygın ya da en az kullanılan sözcüklerin hangileri olduğu hakkında fikir verebilir ve farklı metinleri karşılaştırmak için kullanılabilirler (Zanini ve Dhawan, 2015). Bu yöntemler; siyasi aktör, kurum veya gruplar gibi belli bir kaynağa ait siyasi konuşmalarda en sık hangi kelimelerin kullanıldığının belirlenmesi veya birden farklı kaynağa ait konuşmalardaki sıklık dağılımlarının karşılaştırılmasıyla dil kullanımlarındaki farklılıkları ortaya çıkarmak için kullanılabilmektedir.

Benzerlik ve mesafe ölçüleri ise, iki veya daha fazla metnin içeriğini karşılaştırmak ve aralarındaki benzerlik veya farklılık derecesinin sayısal ifadesi için kullanılan istatistiksel yöntemlerdir (Grimmer ve diğerleri, 2022, s.71). Siyaset biliminde farklı siyasi metinlerdeki ortak konuların belirlenmesi için benzerlik ölçüsü veya farklı siyasi partilerin ifade biçimlerini karşılaştırmak için mesafe ölçüsü kullanılabilmektedir. Tumasjan ve diğerleri (2010) 2009'daki Almanya federal seçimlerinden önce politikacıların ve partilerin Twitter'daki duygu profilleri ile siyasi programları, aday profilleri ve miting konuşmalarının medyada yer alması arasındaki benzerliği analiz etmiştir. Çalışmada ayrıca, Twitter'ın siyasi duyarlılığın geçerli bir göstergesi olup olamayacağını görmek

için siyasi partilerden bahseden tweet'lerin frekans dağılımı geleneksel seçim anketleriyle karşılaştırılırken benzerlik ölçüsü kullanılmıştır.

Okunabilirlik ölçüsü; metindeki cümlelerin ve kelimelerin uzunluğu, çok heceli kelimelerin sıklığı ve teknik terimlerin kullanımı gibi faktörlerin dikkate alınarak o metnin karmaşıklığının sayısal ifadesini sağlamaktadır. Siyaset biliminde bu ölçü bir siyasi partinin mitinglerdeki konuşmalarının ne ölçüde halkın geneli tarafından anlaşılır olduğunun tespiti için kullanılabilmektedir. Gyasi (2023) çalışmasında, Gana'da parlamentoda temsil edilen üç siyasi partinin manifestolarını okunabilirlik ölçüsü açısından değerlendirmiştir. Sonuç olarak, her üç partinin de seçim manifestolarının benzer derecede okunabilir ve Ganalı ortalama seçmen için okuması hayli zor metinler olduğu ortaya çıkmıştır. Sözcüksel çeşitlilik ölçüsü ise bir metinde kullanılan kelimelerin çeşitliliğinin sayısal ifadesidir. Shrestha ve Spezzano (2021), üç farklı siyasi ve güncel haber veri setini kullanarak, sahte haberleri gerçek haberlerden ayırmaya yardımcı olabilecek özellikleri belirlemeyi amaçlamıştır. Haber uzunluğu, noktalama işaretlerinin kullanımı ve sözcüksel çeşitliliği gibi çeşitli özellikler yönünden karşılaştırarak sahte haberlerin gerçek haberlere göre daha kısa, daha az noktalama işareti kullanılmış ve daha az sözcüksel çeşitliliğe sahip olduğu bulunmuştur.

Siyaset biliminde bu ölçüler farklı siyasetçilerin veya siyasi partilerin dil kullanımlarındaki zenginliği ölçmek için kullanılabilir[3]. Siyaset biliminde özellikle siyasi iletişim ve siyasi söylem alanlarında otomatik metin içerik analizi yöntemleri sıkça kullanılmasına rağmen bazı sınırlamalara sahiptir. Örneğin, frekans dağılımının sonuçları, kullanılan derlemin boyutundan ve analize dahil edilen sözcüklerin seçiminden etkilenmeye açıktır. Ek olarak, benzerlik ve mesafe ölçüleri de bu faktörlerden etkilenir. Siyaset bilimi araştırmalarında bu sınırlamaları göz önünde bulundurmak ve istatistiksel sonuçları dikkatlice yorumlamak önemlidir.

2.2. Makine Öğrenmesi

Makine öğrenmesi; algoritmalar ve istatistiksel modeller aracılığıyla büyük boyuttaki verileri analiz etmek ve verilerden öğrenmek için kullanılan yöntemlerdir (Di Cocco ve Monechi, 2022). Son yıllarda makine öğrenmesi yöntemleri, siyaset biliminde büyük veriyi analiz edip anlamanın ve siyasi olgular hakkında tahminlerde bulunmanın bir yolu olarak yaygın kullanılmaya başlanmıştır. Siyaset biliminde makine öğrenmesinin; metin sınıflandırma, duygu analizi, konu modelleme ve tahmine dayalı modelleme gibi çok çeşitli uygulama alanları vardır[4]. Örneğin, metin sınıflandırma algoritmaları, siyasi demeçlerin veya haber metinlerinin ana konularını belirlemek için kullanılabilecek önemli bir araçtır. Duygu analizi, siyasi mesajların tonunu (olumlu ya da olumsuz) ölçmek için etkili bir yöntemdir. Konu modelleme ise, geniş bir metin derlemindeki baskın konuları belirlemek için faydalıdır. Tahmine dayalı modelleme ise, seçim sonuçları gibi belirli olaylar hakkında tahminler yapmak için etkili bir araçtır.

Makine öğrenmesi, siyasi sonuçları anlamak ve tahmin etmek için geniş bir uygulama yelpazesiyle siyaset biliminde büyüyen bir araştırma yöntemi alanıdır. Makine öğrenmesi, siyaset bilimi araştırmaları için güçlü bir yöntemsel araç olsa da sınırlamalarının ve potansiyel hata kaynaklarının farkında olmak önemlidir. Örneğin, makine öğrenmesi algoritmaları yalnızca üzerinde eğitildikleri veriler kadar iyidir ve verilerdeki sapmalar, yanlı sonuçlara yol açabilir.

[3] Bu analizler için kullanılabilecek bazı örnek programları ve kütüphaneleri şunlardır: Python programlama dilinde *spaCy* (Vasiliev, 2020), R programlama dilinde *tidytext* (Silge ve Robinson, 2016) ve Google Cloud platformu üzerinden çalışan *Google AutoML* (Bisong, 2019).

[4] Denetimli makine öğrenmesiyle sınıflandırma için Evans ve diğerleri (2007), Yu, Kaufmann ve Diermeier (2008) ve Peterson ve Spirling (2018); denetimsiz makine öğrenmesiyle sınıflandırma için Sanders, Lisi ve Schonhardt-Bailey (2017), Nguyen ve diğerleri (2015) ve Godel (2022) başlıca temel kaynaklar ve örnek uygulamalar arasında sayılabilir.

Sınırlılıklarına rağmen makine öğrenmesi, siyasi olgulara ilişkin elle analizin mümkün olamayacağı boyutta veriden siyasi olgulara ilişkin yeni bilgilerin üretilmesi ve buna dayanarak gelecek hakkında daha iyi tahminlerin yapılabilmesine fırsat sunmaktadır. Makine öğrenmesi, denetimli ve denetimsiz makine öğrenmesi olarak genellikle iki ana kategoriye ayrılmaktadır (Grimmer ve Stewart, 2013). Denetimli makine öğrenmesi, verilerin analiz öncesinde bir etiket veya hedef değişkeniyle eşleştirilmesini gerektirmektedir. Denetimsiz makine öğrenmesi ise, etiketleme olmadan verileri gruplandırmaya veya özelliklerini öğrenmeye odaklanır.

2.2.1. Denetimli Makine Öğrenmesi

Denetimli makine öğrenmesi (*supervised machine learning*), önceden var olan bir etiket grubuna dayalı olarak bir veri kümesinin sonuçlarını tahmin etmek için bir makine öğrenmesi algoritmasının eğitilmesini içermektedir (Di Cocco ve Monechi, 2022). Denetimli makine öğrenmesinde, belirlenen etiketlere dağıtılmış metinlerin oluşturduğu veri seti kullanılarak makine öğrenmesi algoritması eğitilmektedir. Denetimli makine öğrenmesinde metin sınıflandırması için yaygın olarak kullanılan algoritmalardan bazıları; Naive Bayes, Destek Vektör Makineleri (SVM), Rastgele Ormanlar ve Sinir Ağlarıdır (Osisanwo ve diğerleri, 2017). Bu algoritmaların her birinin güçlü ve zayıf kaldığı yönleri vardır ve kullanılacak algoritma araştırma sorusu, veri setinin türü ve boyutuna göre belirlenmelidir.

Naive Bayes, metin verilerinin belirli bir özelliğinden dolayı hangi sınıfa ait olma ihtimalini tahmin edebilmek için Bayes teoremini kullanan basit ve verimli bir algoritmadır. Bu algoritma, metin verilerinin özelliklerinin koşullu olarak bağımsız olduğunu varsayar ve olasılık temelli olarak sınıflandırmak için basit hesaplamalar yapar (Şahinaslan ve diğerleri, 2022). Kapočiūtė-Dzikienė ve Krupavičius (2014) Litvanya parlamentosunda bulunan siyasi parti gruplarının sıklıkla değişkenlik göstermesine rağmen Naive Bayes algoritmasını kullanarak parlamento konuşmalarından parti grubunu tahmin etme konusunda diğer algoritmalara kıyasla daha fazla verim elde etmiştir. Uslu ve Özmen-Akyol (2021) Türkçe haber makalelerini sınıflandırmak için Destek vektör makineleri, Rastgele ormanlar ve Naive Bayes algoritmalarının sonuçlarını karşılaştırmıştır. Analiz sonuçları, Türkçe için Naive Bayes sınıflandırıcısının %91 doğruluk oranı ile en başarılı performansa sahip olduğunu göstermiştir.

- *Destek Vektör Makineleri (Support Vector Machines - SVM)*, kümeler arasındaki maksimum mesafeyi bularak verileri sınıflandırmak için kullanılabilen bir tür makine öğrenmesi algoritmasıdır (Kaynar ve diğerleri, 2016). Siyaset biliminde, SVM'ler genellikle siyasi metinleri içerdikleri duygu veya ideolojiye göre sınıflandırmak için kullanılmaktadır. Diermeier ve diğerleri (2012) destek vektör makinelerini kullanarak ABD'li muhafazakâr ve liberal senatörlerin kullandığı kelimeleri ve deyimleri belirleyerek bir senatörün siyasi eğilimini %92 doğrulukla tahmin edebilmiştir. Ekonomik referanslara kıyasla kültürel referansların muhafazakâr ve liberal tutumları tahmin etmede daha doğru sonuçlar ürettiğini göstermişlerdir.
- *Rastgele Ormanlar (Random Forests)*, tahmin yapmak için eğitim verileriyle öğrenme aşamasında anlamlı rastgele öğe olarak ormanlar olarak adlandırılan birkaç karar ağacı algoritması kullanan bir öğrenme yöntemidir. Rastgelelik sebebiyle, düşük yanlılık ürettiği için daha yüksek doğruluk elde etmek için birçok karar ağacının sonuçlarını yumuşatma eğilimindedir (Nayak ve Natarajan 2016). Bu yöntem, metindeki kelime ve kelime gruplarını ve bunlar arasındaki ilişkileri dikkate alarak metin verilerini analiz etmek için kullanılabilmektedir. Regresyon ve sınıflandırma ağacı modelleri, birçok değişkeni olan karmaşık veri kümelerini analiz etmek için iyi bir seçenek olmaktadır. Önceki çalışmalar, değişkenler arasındaki etkileşimler, karmaşık ilişkiler ve ani değişiklikler içerdiğinde ağaç

tabanlı yöntemlerin doğru tahminler yapmak için etkili olduğunu göstermiştir (Montgomery ve Olivella, 2018).

- *Sinir Ağları*, insan beyninin yapısına benzer modellenen bir tür makine öğrenmesi algoritmasıdır. Özellikle siyaset bilimi bağlamında, küçük metin veri kümelerindeki kelimelerin değişen anlamlarını analiz etmek için sinir ağlarının kullanımı tavsiye edilmektedir (Osisanwo ve diğerleri, 2017). Rodman (2020) *word2vec* adlı bir tür sinir ağını kullanarak 161 yıllık gazete haber metinlerinden oluşan bir veri kümesindeki kelimelerin anlamındaki değişiklikleri izlemiş ve diğer yaygın yöntemlere kıyasla değişen sözcük anlamlarının daha ayrıntılı bir analizine izin verdiği sonucuna varmıştır. Sonuç olarak denetimli makine öğrenmesi, siyaset bilimciler için güçlü ve esnek bir araçtır ve siyasi analizler için yeni olanaklar sunmaktadır.

2.2.2. Denetimsiz Makine Öğrenmesi

Denetimsiz makine öğrenmesinin amacı, bilgisayarın herhangi bir ön bilgi veya denetim olmaksızın verilerdeki kalıpları ve yapıları öğrenmesini sağlamaktır (Wilkerson ve Casas, 2017). Bu yöntem, araştırmacıların analiz öncesinde farkında olmayabilecekleri konuları ve ilişkileri ortaya çıkarmalarına olanak tanımaktadır. Siyaset biliminde denetimsiz makine öğrenmesi çeşitli şekillerde kullanılabilmektedir:

- Kümeleme algoritmaları; benzer öğeleri kümeler halinde gruplandıran popüler bir denetimsiz makine öğrenmesi tekniğidir. Grimmer ve Stewart'a (2013) göre, metinlerin kümelendirilmesi siyaset biliminde en sık kullanılan metin analizi yöntemlerindendir. Siyaset Biliminde bu yöntem, içeriklerine göre benzer belge, konuşma veya sosyal medya paylaşımlarını gruplamak için kullanılmaktadır. Sagarzazu ve Klüver (2017), seçimlerin koalisyon hükümetlerini oluşturan partiler üzerindeki etkisini ölçmeyi amaçlamıştır. Bu partiler, bir yandan birlikte çalışmaya uygun hareket etmeleri gerekirken seçimler öncesinde ise seçmenlere kendilerine özgü fikirler ve politikalara sahip olduklarını göstermeleri gerekmektedir. Bu çelişkili durumunda varlığının tespiti için 2000-2010 yılları arasında koalisyon partileri tarafından yayınlanan 21.000'den fazla basın bültenini denetimsiz makine öğrenmesi teknikleriyle konularına sınıflandırmıştır. Bu yöntem, çevrenin korunmasını tartışırken "çevre", "doğa" veya "koruma" gibi belirli kelimelerin ve "işsizlik politikasını tartışırken "işler", "işsizler" veya "iş" gibi belirli kelimelerin birlikte anılmasının daha muhtemel olması fikrine dayanmaktadır. Yazarlar, bu ayırt edici kelime gruplarını tanımlayarak, bu alanların ne olduğuna dair önceden bilgi sahibi olmadan, basın bültenlerini farklı konu alanlarına göre sınıflandırabilmiştir.
- Gizli Dirichlet Ayrımı (*Latent Dirichlet Allocation - LDA*), metin analizinde yaygın olarak kullanılan başka bir denetimsiz makine öğrenmesi tekniğidir. LDA, her belgenin konuların bir karışımı olduğunu varsaymaktadır. Bu yöntem, bir grup konuşmada ele alınan konuları anlayabilir ve kısıtlayıcı varsayımlar yapmadan bu konuşmaların konusunu belirleyebilmektedir (Benoit, 2020). Quinn ve diğerleri (2010), ABD Senatosu'nda 1997-2004 yılları arasında yapılan 118 bin konuşmayı (70 milyon kelime) içeren bir veri tabanını kullanarak, Senato gündemini incelemiştir Bu analiz sayesinde, belirli konulara ait konuşmaların yoğunluğu ve sıklığı gibi özellikleri ortaya çıkararak demokratik gündemin dinamiklerini daha ayrıntılı şekilde gösterebilmişlerdir. Spirling (2012) ABD hükümeti ile Yerli Amerikalılar arasında imzalanan yaklaşık 600 anlaşmayı analiz ederken, anlaşmaların şartlarının katılığının, anlaşma yapma mekanizmalarındaki değişikliklerden ziyade ABD'nin göreli pazarlık gücüyle ilgili olduğunu göstermek için ölçekleme tekniklerini kullanmıştır. Bu yöntem, daha önce belirlenmiş bir konu hakkında bilgi sahibi olmaya gerek olmaması ve

analiz yapmak için verilerin içeriğine yönelik kısıtlayıcı varsayımlar yapılmadığı için verileri daha tarafsız bir şekilde analiz etmek için kullanılabilmektedir.

- *Wordfish*, büyük miktarlarda yapılandırılmamış metin verileri içinde anlamsal boyutları tanımlamayı ve farklı sözcük ve belgelerin anlamlarını karşılaştırmayı amaçlayan bir ölçekleme modelidir (Hjorth ve diğerleri, 2015). Sonuçları, metin verileri içindeki eğilimleri, konuları ve ilişkileri belirlemeye yardımcı olmak ve mevcut ilişkilere dayanarak yeni metin verileri hakkında tahminlerde bulunmak için de kullanılabilmektedir. Siyaset biliminde bu yöntem, siyasi görüşleri şekillendirebilecek isimlendirilmemiş faktörleri ortaya çıkarmak veya siyasi aktörlerin ideolojik konumlarında zaman içinde meydana gelen değişimleri sözlerine ve ifadelerine dayalı olarak incelemek için kullanılabilmektedir. Frid-Nielsen (2018) Avrupa Parlamentosu'nda 2004'ten 2014'e kadar iltica hakkında yapılan 876 konuşmayı incelemiş ve ulusal partiler ile Avrupa Parlamentosu Üyelerinin tercihlerini ölçeklendirmek için Wordfish analizi yöntemi kullanmıştır. Genel olarak çalışma, Wordfish'in AB'yi incelemek için etkili ve geçerli bir araç olduğunu göstermiş ve siyaset bilimi araştırmalarında otomatik içerik analizinin önemini doğrulamıştır.

Siyaset bilimi de dâhil olmak üzere çeşitli alanlarda uygulanan makine öğrenmesi yöntemleri, sağladıkları avantajların yanında, dikkatle ele alınması gereken sınırlamalar ve güvenilirlikle ilgili riskler taşımaktadır. Analizler sonucunda elde edilen bulguların geçerliliğini ve genellenebilirliğini etkileyebileceğinden bu sınırlamaların belirtilmesi çok önemlidir. Makine öğrenmesinde eğitim veri seti, temsil kabiliyeti düşük veya yanlı şekilde oluşturulmuşsa üreteceği sonuç da yetersiz veya yanlı olacaktır. Belirli kişilerin, grupların veya konuların aşırı temsiline ve diğerlerinin yetersiz temsili yanlı sonuçlara yol açabilmektedir. Makine öğrenmesi modellerinin güvenilirliği, eğitim ve test için kullanılan verilerin kalitesine ve miktarına bağlıdır. Gürültülü veya değişkenlikten yoksun veriler, model performansının düşük olmasına ve güvenirliği düşük sonuçların üretilmesine sebep olmaktadır (Grimmer ve Stewart, 2013). Bu nedenle, makine öğrenmesi modelleri için kullanılan verilerin yüksek kalitede olmasını ve araştırılan olguyu yeterli düzeyde temsil etmesini sağlamak önemlidir. Sonuç olarak, bu noktalar dikkatlice göz önünde bulundurularak siyaset biliminde makine öğrenmesi uygulamalarının, karmaşık siyasi olaylara ilişkin değerli sonuçlar sunma potansiyeli bu konulardaki anlayışımızı ilerletmek için etkili bir şekilde kullanılabilecektir.

2.3. Dağılımsal Semantik Modeller ve Kelime Gömmeleri

Dağılımsal Semantik Modeller (*Distributional Semantic Models - DSM*) ve Kelime Gömmeleri (*Word Embeddings*), bir metin derlemindeki sözcükler arasındaki anlamsal ilişkileri ortaya çıkarmak ve metin sınıflandırma, duygu analizi ve belge özetleme gibi uygulamalar için kullanılmaktadır (Grimmer ve diğerleri, 2022, s.78). Kelimeleri sayılara dönüştürerek büyük bir metin derleminde nasıl kullanıldıklarına bağlı olarak kelimelerin birbirleriyle nasıl ilişkili olduğunu göstermektedirler (Benoit, 2020). Gram Atla (*Skip-gram*) ve Sözcük Torbası (*bag-of-words*) modelleri gibi farklı DSM modelleri, siyasi söylem ve duyguya ilişkin sonuçlar elde etmeyi sağladığı için siyaset bilimi alanında da kullanılmaktadır. Rodriguez ve Spirling (2022) çalışmalarında, siyaset bilimi araştırmalarında sözcük yerleştirmelerin performansını incelemektedir. Önceden eğitilmiş yerleştirmelerin, insan kodlayıcılara kıyasla iyi çalıştığı bulmuştur. Ayrıca, analizlerinin sonuçları farklı diller için de geçerliliğini korumaktadır.

Aydoğan ve Karcı (2019) Python'daki *Beautiful Soup* kütüphanesini kullanarak geniş bir Türkçe metin derlemi oluşturmaya odaklanmıştır. Çalışmalarında, Word2Vec modelini kullanarak derlemdeki Türkçe sözcükler arasındaki anlamsal ilişkileri belirlemeyi amaçlamaktadır. Ayrıca modelin performansını artırmak için Türkçe'deki dolu kelimelerinin bir listesi de

oluşturulmuştur. Geliştirilen model, Türkçe metinleri sınıflandırmak için test edilmiş ve olumlu sonuçlar vermiştir. Onan (2020) Türkçe bir metin parçasının olumlu, olumsuz veya tarafsız bir duyguyu ifade edip etmediğini belirleme süreci olan duygu analizi için Word2vec, fastText, GloVe ve LDA2vec olmak üzere dört temel kelime gömme yöntemi incelemiştir. Sonuçlar, evrişimli sinir ağlarının, diğer makine öğrenme yöntemlerinden veya diğer derin öğrenme mimarilerinden daha etkili olduğunu ve Word2vec tabanlı kelime gömme şeması ile %92,53'lük bir sınıflandırma doğruluğu elde edildiğini göstermiştir.

DSM'lerin ve kelime gömmelerinin bu alanda ana kullanımı, siyasi söylemin analizi üzerinedir. Rheault ve Cochrane (2020) kelime gömme yöntemini kullanarak İngiltere, Kanada ve ABD'deki parlamento konuşmalarını incelemiştir. Sonuçlar, bu yeni modellerin siyasi ideoloji gibi kavramları anlamada başarılı olduğunu ve siyasi dili incelemek için yararlı bir araç olabileceğini göstermiştir. Siyasi demeçler, haber metinleri ve sosyal medya paylaşımlarındaki kelimelerin anlamlarının incelenmesine olanak tanımaktadır. Kelimeler ve kelime grupları vektörlere göre eşleştirilerek, farklı kelimelerin farklı bağlamlardaki anlamları karşılaştırılabilmektedir. Polat ve Körpe (2018) çalışmalarında, veri seti olarak Türkiye Büyük Millet Meclisi (TBMM) tutanaklarını kullanarak, derlemdeki semantik olarak benzer kavramları ve bağlamlarını çıkarsamak için Word2vec ve GloVe modellerini kullanmıştır. Sonuçlar, Word2vec modelinin daha iyi sonuçlar ürettiğini göstermiştir. Çalışma ayrıca her iki modelin de kavramlar arasında analojiler bulma ve metin sınıflandırma gibi doğal dil işleme uygulamalarında kullanılabileceğini göstermiştir.

Bunların yanında, DSM'ler ve kelime gömmeleri, metnin duygusunu tahmin etmek için kullanılabilmektedir. Genel olarak, DSM'ler ve kelime gömmeleri, kelimeler arasındaki anlamsal ilişkilerden siyasi tutumlar ve davranışlar hakkında çıkarım yapılabilmesi açısında siyaset biliminde yeni yollar açabilmektedir.

3. TARTIŞMA VE DEĞERLENDİRME

Otomatik metin analizi yöntemleri; metin verilerini analiz etmek için basit istatistiksel analizlerden denetimli ve denetimsiz makine öğrenmesi, dağılımsal semantik modeller ve kelime gömmeleri gibi daha gelişmiş tekniklere kadar bir dizi yöntemi içermektedir. Bu alanda kullanılan tüm yöntem ve modellerin bu makalede yer alması mümkün olmadığı için siyaset biliminde en sık kullanılan dört farklı yöntem ele alınmıştır. Tablo 1 bu yöntemlerin karşılaştırmasını içermektedir. Tablodaki her satır, yöntemin bir özelliğini ve her yöntem için karşılık gelen değeri listelemektedir.

Tablo 1. Otomatik Metin Analizi Yöntemlerinin Karşılaştırması

	Analizin Amacı	**Veri gereksinimleri**	**Varsayımları**	**Analizin Çıktısı**
Basit İstatistiksel Analizler	Büyük miktarda metin verisini özetlemek ve açıklamak	Önceden temizlenmiş, yapılandırılmış metin verileri	Verilerin normalliğini varsayar	Ortalama ve standart sapma gibi sayısal özetler
Denetimli Makine Öğrenmesi	Eğitim verilerine dayalı olarak belirli bir sonucu veya etiketi tahmin etmek	Tahmin görevleri için etiketli eğitim verileri	Eğitim verilerine bağlıdır, yanlılıkları yansıtabilir	Yeni metin verileri için tahmin edilen küme etiketleri
Denetimsiz Makine Öğrenmesi	Önceden tanımlanmış sonuçlar olmadan verilerdeki kalıpları ve ilişkileri keşfetmek	Yapılandırılmış veya yapılandırılmamış metin verileri	Başlangıç koşullarına bağlıdır, sapmaları yansıtabilir	Küme atamaları, konu modelleri vb.
Dağılımsal Semantik Modeller ve Kelime Gömmeleri	Belirli bir bağlamda kelimelerin anlamlarını modellemek ve ilişkilerini keşfetmek	Büyük miktarda metin verisi	Birlikte oluşun anlamın göstergesi olduğunu varsayar	Benzerlik puanları gibi kelimelerin sayısal temsilleri

Kaynak: Araştırmacı tarafından Benoit'dan (2020) yararlanılarak oluşturulmuştur.

Basit istatistiksel analizler, otomatik metin analizinin en basit ve en temel biçimidir. Metin verilerinde en sık kullanılan kelimelere ve kelime öbeklerine hızlı bir genel bakışın yanı sıra kelime sayısı, ortalama cümle uzunluğu ve sözcük çeşitliliği gibi diğer tanımlayıcı istatistikleri de sağlamaktadır. Basit istatistiksel analizlerin ana avantajı, hızlı ve kolay üretilmeleri ve büyük bir metin verisi derlemin içeriğine dair ilk bakışı sağlayabilmeleridir (Benoit, 2020). Bununla birlikte, metnin anlamı veya bağlamı hakkında bilgi sağlama yönünden kısıtlı ve verilerdeki eğilimleri veya ilişkileri belirleme yetenekleri sınırlıdır (Grimmer ve Stewart, 2013).

Denetimli makine öğrenmesi, metin verilerini önceden belirlenmiş kategorilere veya etiketlere göre sınıflandırmak için kullanılabilmektedir. Bu yöntem, tümdengelimci yaklaşımla (Kroon, van der Meer ve Vliegenthart, 2022) metindeki kalıpları öğrenmek için eğitim verilerini kullanır ve ardından tahminler yapmak için bu kalıpları yeni metin verilerine uygulamaktadır. Bu yöntemin temel avantajı, metin verilerini yüksek doğruluk oranlarıyla kategoriler halinde sınıflandırabilmesi ve ayrıca metin verilerindeki önemli özellikleri veya konuları tanımlayabilmesidir. Bununla birlikte, bu yöntemler hesaplama açısından basit istatistiksel analizlere kıyasla karmaşıktır, büyük miktarda eğitim verisi gerektirir ve eğitim verilerindeki sapmalardan etkilenirler (Grimmer ve diğerleri, 2021).

Denetimsiz makine öğrenmesi, denetimli makine öğrenmesine benzer, ancak önceden belirlenmiş kategorileri veya etiketleri kullanmamaktadır. Bu bağlamda, tümevarımcı yaklaşımla (Nelson, 2020) metnin içeriği veya anlamı hakkında önceden herhangi bir bilgi sahibi olmadan, metin verilerindeki eğilimleri ve ilişkileri eğitim verisi olmadan aramaktadır. Denetimsiz makine öğrenmesinin en büyük avantajı, önceden var olduğu bilinmeyen eğilimleri ve ilişkileri tanımlayabilmesidir (Wilkerson ve Casas, 2017). Ancak, bu yöntem aynı zamanda hesaplama açısından yoğun olma ve yorumlanmasının zor olması gibi kısıtlılıklara sahiptir.

Dağılımsal semantik modeller ve kelime gömmeleri, nispeten yeni sayılabilecek otomatik metin analizi yöntemleridir. Bu yöntemler, metin verilerindeki sözcüklerin anlamını ve bağlamını vektörler veya yüksek boyutlu boşluklar olarak temsil etmek için matematiksel algoritmalar kullanmaktadır. Bu yöntemlerin temel avantajı, kelimelerin anlamını ve bağlamını diğer yöntemlerle mümkün olmayan bir şekilde yakalayabilmeleri ve metin verilerindeki kalıpları ve ilişkileri basit istatistiksel analizlerden çıkarsanamayacak şekilde tanımlamak için kullanılabilmeleridir (Benoit, 2020). Bununla birlikte, bu yöntemler de hesaplama açısından yoğundur ve ürettiği sonuçların bilgisayar bilimlerinde güçlü bir altyapıya sahip olmayanlarca yorumlanması ve anlaşılması zor olabilmektedir.

Her yöntemin doğasında varsayımlar ve yanlılıklar vardır. Basit istatistiksel analizler, metin verilerinin normal bir dağılım izlediği varsayımına dayanır ve bu sebeple verileri açıklamak için kullanılan özet istatistiklerin seçiminde yanlı olabilir. Makine öğrenmesi yöntemleri, kullanılan özelliklerin ve algoritmaların seçiminin yanı sıra eğitim verilerinin kalitesi ve temsil edilebilirliğine göre yanlı olabilmektedir. Dağılımsal semantik modeller ve kelime gömmeleri, bir kelimenin anlamının göründüğü bağlamdan çıkarılabileceğini varsayan ve modelleri eğitmek için kullanılan derlem seçimiyle yanlı olabilen dağılımsal hipoteze dayanmaktadır (Young ve Soroka, 2012). Yöntemlerin sunduğu birtakım artılar ve bazı kısıtlılıklar vardır. Basit istatistiksel analizler nispeten basit ve hesaplanması hızlıdır, ancak karmaşık yöntemler kadar fazla bilgi sağlamayabilir. Makine öğrenmesi yöntemleri, özellikle büyük ve karmaşık veri kümeleri için önemli hesaplama kaynağı ve zaman gerektirebilir, ancak daha doğru ve bilgilendirici sonuçlar sağlayabilmektedir. Dağılımsal semantik modeller ve kelime gömmeleri, eğitmek için daha da fazla hesaplama kaynağı ve zaman gerektirir, ancak kelimelerin ve kelime öbeklerinin anlamlarının zengin ve incelikli temsillerini sağlayabilir.

Bu yöntemler, analizlerin sonuçlarının kullanıcılar tarafından anlaşılma ve yorumlanma konusunda da farklılıklara sahiptir. Basit istatistiksel analizler nispeten basit ve yorumlanabilir. Makine öğrenmesi yöntemlerinin yorumlanması, özellikle karmaşık modeller için zor olabilmektedir. Dağılımsal semantik modelleri ve kelime gömmelerini yorumlamak genellikle zordur. Sonuç olarak, dört yöntemin her birinin ayrı ayrı güçlü ve zayıf kaldığı yönler vardır. Yöntem seçimi, analizin özel amaçlarına ve gereksinimlerine bağlı olmalıdır. Örneğin, yorumlanabilirlik ve basitlik öncelikli kaygılar ise, basit istatistiksel analizler iyi bir seçim olabilir. Doğruluk ve ayrıntı birincil endişelerse, makine öğrenmesi veya dağılımsal semantik modeller daha iyi bir seçim olabilir.

4. SONUÇ

Siyaset biliminde otomatik metin analizi yöntemleri, büyük veri olarak tanımlanan boyuttaki metin tabanlı verinin analizi için giderek daha sık kullanılır hale gelmiştir. Bu alana önemli katkı sunma potansiyeli taşıyan bu yöntemler üzerine Türkçe yayın konusunda önemli bir eksiklik mevcuttur. Bu makale, otomatik metin analizi yöntemlerine genel bir bakış sunmakta ve bu yöntemlerin siyaset bilimi alanı için potansiyellerine ve kullanımlarına ilişkin hususları ele almaktadır.

Otomatik metin analizi, siyaset bilimciler için metin verilerinin içeriğini ve anlamını incelemek için önemli bir araç olarak karşımıza çıkmaktadır. Bu alanda kullanılabilecek her biri kendi güçlü ve zayıf yönleri olan çeşitli yöntemler vardır. Bu makalede, basit istatistiksel analizler, denetimli makine öğrenmesi, denetimsiz makine öğrenmesi, dağılımsal anlamsal modeller ve kelime gömmeleri olmak üzere otomatik metin analizinin farklı yöntemleri karşılaştırmalı olarak ele alınmıştır. Basit istatistiksel analizler için analizin amacı; genellikle kelime frekanslarının ortalamasını ve standart sapmasını hesaplayarak metin verilerini açıklamak ve özetlemektir. Makine öğrenmesi yöntemlerinde amaç, genellikle metin verilerini önceden tanımlanmış kategoriler halinde sınıflandırmak veya kümelemek ya da metin verilerine dayalı olarak bir hedef değişkeni tahmin etmektir. Dağılımsal semantik modeller ve kelime gömmeleri için amaç, sözcüklerin ve sözcük öbeklerinin anlamını, anlamsal olarak benzer sözcüklerin boşlukta yakın olduğu yüksek boyutlu bir vektör uzayında temsil etmektir.

Sonuç olarak, bu çalışmada ele alınan tüm otomatik metin analizi yöntemlerinin güçlü ve zayıf yönleri vardır. Kullanılacak en iyi yöntem; analizin amaçlarına, metinsel verilerinin türü ve boyutuna ve hesaplama yöntemine bağlı olacaktır. Siyaset bilimciler, bir otomatik metin analizi yöntemi seçmeden önce hedeflerini ve kaynaklarını dikkatlice değerlendirmeli ve her yöntemin doğasında bulunan varsayımların ve yanlılıkların farkında olmalıdır.

Geliş Tarihi | Kabul Tarihi | Yayım Tarihi	6 Mart 2023 | 27 Haziran 2023 | 30 Haziran 2023
Yazar Katkısı	Betül Aydoğan Ünal (%100)
Hakem Değerlendirmesi	Dış bağımsız
Etik Onay	Bu makale, insan veya hayvanlar ile ilgili etik onay gerektiren herhangi bir araştırma içermemektedir.
Çıkar Çatışması	Yazar çıkar çatışması bildirmemiştir.
Finansal Destek	Yazar bu çalışma için finansal destek almadığını beyan etmiştir.
Telif Hakkı & Lisans	Yazar dergide yayınlanan çalışmalarının telif hakkına sahiptirler ve çalışmaları CC BY-NC 4.0 lisansı altında yayımlanır. https://creativecommons.org/licenses/by-nc/4.0/deed.tr
Submission | Acceptance | Publication	6 March 2023 | 27 June 2023 | 30 June 2023
Author Contribution	Betül Aydoğan Ünal (100%)
Peer-review	Externally peer-reviewed.
Ethical Approval	This article does not contain any studies with human participants or animals performed by the authors.
Conflicts of Interest	The author declares that there is no conflict of interest.
Grant Support	The author received no financial support for the research, authorship and/or publication of this article.
Copyright & License	Author publishing with the journal retain(s) the copyright to their work licensed under the CC BY-NC 4.0. https://creativecommons.org/licenses/by-nc/4.0/

KAYNAKÇA | REFERENCES

Atalay, M. ve Çelik, E. (2017). Büyük veri analizinde yapay zekâ ve makine öğrenmesi uygulamalari-artificial intelligence and machine learning applications in big data analysis. *Mehmet* Akif *Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *9*(22), 155-172. doi:10.20875/makusobed.309727

Athey, S. (2018). *The impact of machine learning on economics*. A. Agrawal, J. Gans ve A. Goldfarb (Ed.), *The economics of artificial intelligence: An agenda* (s.507-547) içinde. Chicago: University of Chicago Press.

Aydoğan, M. ve Karcı, A. (2019). Kelime temsil yöntemleri ile kelime benzerliklerinin incelenmesi. *Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, *34(2),* 181-196. doi:10.21605/cukurovaummfd.609119

Benoit, K. (2020). *Text as data: An overview*. L. Curini and R. Franzese (Ed.), *The handbook of research methods in political science and international relations* (ss. 461-497) içinde. Tthousand Oaks: Sage.

Benoit, K. ve Laver, M. (2003). Estimating Irish party policy positions using computer *wordscoring*: The 2002 election–a research note. *Irish political studies*, *18(1),* 97-107. doi:10.1080/07907180312331293249

Bisong, E. (2019). Google AutoML: cloud natural language processing. *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners,* 599-612. doi: 10.1007/978-1-4842-4470-8_43

Bouchart, S. (2020). *Classification and clustering*. SAGE Publications Ltd. doi:10.4135/9781526486387

Budge, I. ve Pennings, P. (2007). Do they work? Validating computerised word frequency estimates against policy series. *Electoral Studies*, *26(1),* 121-129. doi:10.1016/j.electstud.2006.04.002

Di Cocco, J. ve Monechi, B. (2022). How populist are parties? Measuring degrees of populism in party *manifestos* using supervised machine learning. *Political Analysis*, *30(3),* 311-327. doi:10.1017/pan.2021.29

Diermeier, D., Godbout, J. F., Yu, B. ve Kaufmann, S. (2012). Language and ideology in Congress. *British Journal of Political Science*, *42(1),* 31-55. doi: 10.1017/S0007123411000160

Eggers, A. C., ve Spirling, A. (2018). The shadow cabinet in Westminster systems: modeling opposition *agenda* setting in the House of Commons, 1832–1915. *British Journal of Political Science, 48(2)*, 343-367. doi:10.1017/S0007123416000016

Evans, M., McIntosh, W., Lin, J. ve Cates, C. (2007). Recounting the courts? Applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies, 4(4),* 1007-*1039*. doi: 10.1111/j.1740-1461.2007.00113.x

Frid-Nielsen, S. S. (2018). Human rights or security? Positions on asylum in European Parliament speeches. *European union politics, 19(2),* 344-362. doi: 10.1613/jair.1.13112

Gee, J. P. (2018). Reading as situated language: A sociocognitive perspective. In *Theoretical models and processes of literacy* (s.105-117). New York: Routledge.

Godel, W. (2022). Ideology, Social Media and Fake News: New Machine Learning Methods for Political Science (Yayımlanmamış doktora tezi). Wilf Family Department of Politics, New York University.

Gökçe, O. (2006). *İçerik analizi-kuramsal ve pratik bilgiler*. Ankara: Siyasal Kitabevi

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: measuring expressed agendas in Senate press releases. *Political Analysis*, *18(1),* 1-35. doi: 10.1093/pan/mpp034

Grimmer, J., Roberts, M.E. ve Stewart, B.M. (2021). Machine learning for social science: an agnostic approach. *Annual Review of Political Science*, *24*, 395-419. doi: 10.1146/annurev-polisci-053119-015921

Grimmer, J., Roberts, M.E. ve Stewart, B.M. (2022). *Text as data: a new framework for machine learning and the social sciences*. New Jersey: Princeton University Press.

Grimmer, J. ve 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. doi:10.1093/pan/mps028

Gül, S.S. ve Nizam, Ö.K. (2021). Sosyal bilimlerde içerik ve söylem analizi. *Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *42,* 181-198. doi: 10.30794/pausbed.803182

Gyasi, W.K. (2023). The readability of political party manifestos of the 2016 general elections in Ghana. *Athens Journal of Mass Media and Communications*, *9(1),* 57-70. doi:10.30958/ajmmc

Hatipoğlu, E., Gökçe, O.Z., Arın, İ. ve Saygın, Y. (2022). *Otomatik metin analizi ve uluslararası ilişkiler.* E. Aydınlı (Der.). *Uluslararası İlişkiler Metodolojisi* içinde (s.135-166). İstanbul: Koç Üniversitesi Yayınları.

Hjorth, F., Klemmensen, R., Hobolt, S., Hansen, M.E. ve Kurrild-Klitgaard, P. (2015). Computers, coders, and voters: comparing automated methods for estimating party positions. *Research & Politics*, *2(2),* 1-9. doi: 10.1177/2053168015580476

Kapočiūtė-Dzikienė, J. ve Krupavičius, A. (2014). Predicting party group from the Lithuanian parliamentary speeches. *Information Technology and Control*, *43(3),* 321-332. doi:10.5755/j01.itc.43.3.5871

Kaynar, O., Görmez, Y., Yıldız, M. ve Albayrak, A. (2016). Makine öğrenmesi yöntemleri ile duygu analizi. *International Artificial Intelligence and Data Processing Symposium (IDAP'16)*, 234-241.

Kılıç, H., Atalay, E. ve Yurtsever, A.E. (2019). Büyük veri (Bigdata) ve müşteri ilişkileri yönetimi (CRM) işbirliğinin pazarlama iletişimi stratejilerindeki rolü: büyük ölçekli özel bir banka *örneği. Stratejik ve Sosyal Araştırmalar Dergisi*, *3(2),* 289-310. doi: 10.30692/sisad.574133

Klemmensen, R., Hobolt, S.B. ve Hansen, M.E. (2007). Estimating policy positions using political texts: an evaluation of the wordscores approach. *Electoral Studies*, *26(4),* 746-755. doi:10.1016/j.electstud.2007.07.006

Konşuk Ünlü, H. (2022). Başlığında "data science" ifadesi geçen uluslararası kongrelerde sunulan bildiri *özetlerinin* metin madenciliği yöntemleri ile incelenmesi. *Nicel Bilimler Dergisi*, *4(1),* 1-21. doi:10.51541/nicel.1075225

Kroon, A.C., van der Meer, T. ve Vliegenthart, R. (2022). Beyond counting words: assessing performance of dictionaries, supervised machine learning, and embeddings in topic and frame *classification. Computational Communication Research, 4(2),* 528-570. doi:10.5117/CCR2022.2.006.KROO

Monroe, B.L. ve Schrodt, P.A. (2008). Introduction to the special issue: the statistical analysis of political text. *Political Analysis, 16(4),* 351-355. doi: 10.1093/pan/mpn017

Montgomery, J.M. ve Olivella, S. (2018). Tree-Based Models for Political Science Data. *American Journal of Political Science*, *62*(3), 729-744. doi: 10.1111/ajps.12361

Nayak, A. ve Natarajan, D. (2016). Comparative study of naive Bayes, support vector machine and random forest *classifiers* in sentiment analysis of twitter feeds. *International Journal of Advance Studies in Computer Science and Engineering (IJASCSE)*, *5*(1), 16. Erişim adresi: https://rb.gy/964f1h

Nelson, L.K. (2020). Computational grounded theory: a methodological framework. *Sociological Methods & Research, 49(1),* 3-42. doi: 10.1177/0049124117729703

Neuendorf, K.A. (2004). Content analysis: a contrast and complement to discourse analysis. *Qualitative methods, 2(1),* 33-36. Erişim adresi: https://zenodo.org/record/998700

Neuendorf, K.A. (2017). *The content analysis guidebook.* New Delhi: SAGE.

Nguyen, V.A., Boyd-Graber, J., Resnik, P. ve Miler, K. (2015). Tea party in the house: a hierarchical ideal point topic model and its application to republican legislators in the 112th congress. In *Proceedings* of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 1438-1448.

Onan, A. (2020). Evrişimli sinir ağı mimarilerine dayalı türkçe duygu analizi. *Avrupa Bilim ve Teknoloji Dergi*si, 374-*380*. doi: 10.31590/ejosat.780609

Osgood, C.E. (1959). Representational model ve relevant research methods. In I. Pool (Ed.), Trends in content analysis (ss. 33-38). Urbana, IL : Illinois Press.

Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O. ve Akinjobi, J. (2017). Supervised *machine* learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48(3),* 128-138. doi:10.14445/22312803/IJCTT-V48P126

Özyiğit, H. (2022). Muhasebe alanına güncel yaklaşımlar: metin madenciliği. *Muhasebe ve Vergi Uygulamaları Dergisi*, *15*(3), 637-663. doi: 10.29067/muvu.1104525

Özoran, B.A. (2022). Bir halkla ilişkiler aracı olarak twitter: dünya sağlık örgütü paylaşımlarının içerik analizi ve metin madenciliği ile incelenmesi. *Celal Bayar Üniversitesi Sosyal Bilimler Dergisi, 20(04),* 125-146. doi: 10.18026/cbayarsos.1083191

Quinn, K.M., Monroe, B.L., Colaresi, M., Crespin, M.H. ve Radev, D.R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54(1),* 209-228. doi: 10.1111/j.1540-5907.2009.00427.x.

Peterson, A. ve Spirling, A. (2018). Classification accuracy as a substantive quantity of interest: measuring polarization in westminster systems. *Political Analysis, 26(1),* 120-128. doi:10.1017/pan.2017.39

Polat, H. ve Körpe, M. (2018). TBMM genel kurul tutanaklarından yakın anlamlı kavramların çıkarılması. *Bilişim Teknolojileri Dergisi, 11(3),* 235-244. doi: 10.17671/gazibtd.402468

Rheault, L. ve Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis, 28(1),* 112-133. doi: 10.1017/pan.2019.26.

Roberts, C.W. (Ed.). (2020). *Text analysis for the social sciences: methods for drawing statistical inferences from texts and transcripts*. New York: Routledge.

Rodman, E. (2020). A timely intervention: tracking the changing meanings of political concepts with word vectors. *Political Analysis, 28(1),* 87-111. doi: 10.1017/pan.2019.23.

Rodriguez, P. L. ve Spirling, A. (2022). Word embeddings: what works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics, 84(1),* 101-115. doi:10.1086/715162.

Sagarzazu, I. ve Klüver, H. (2017). Coalition governments and party competition: political communication strategies of coalition parties. *Political Science Research and Methods*, 5(2), 333-349. doi: 10.1017/psrm.2015.56

Sanders, J., Lisi, G. ve Schonhardt-Bailey, C. (2017). Themes and topics in parliamentary oversight hearings: a new direction in textual data analysis. *Statistics, Politics and Policy, 8(2),* 153-194. doi: 10.1515/spp-2017-0012

Schoonvelde, M., Schumacher, G. ve Bakker, B.N. (2019). Friends with text as data benefits: assessing and extending the use of automated text analysis in political science and political psychology. *Journal of Social and Political Psychology, 7(1),* 124-143. doi:10.5964/jspp.v7i1.964

Shrestha, A. ve Spezzano, F. (2021). Textual characteristics of news title and body to detect fake news: a reproducibility study. *Advances in Information Retrieval: 43rd European* Conference *on IR Research, 43*, 120-133. doi: 10.1007/978-3-030-72240-1_9

Silge, J. ve Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. Journal *of Open Source Software, 1(3),* 37. doi: 10.21105/joss.00037

Slapin, J.B. ve Proksch, S.O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science, 52(3),* 705-722. doi: 10.1111/j.1540-5907.2008.00338.x

Spirling, A. (2012). US treaty making with American Indians: Institutional change and relative power, 1784–1911. *American Journal of Political Science, 56(1),* 84-97. doi: 10.1111/j.1540-5907.2011.00558.x

Şahinaslan, Ö., Dalyan, H. ve Şahinaslan, E. (2022). Naive bayes sınıflandırıcısı kullanılarak youtube verileri üzerinden çok dilli duygu analizi. *Bilişim Teknolojileri Dergisi, 15(2),* 221-229. doi: 10.17671/gazibtd.999960

Tumasjan, A., Sprenger, T., Sandner, P. ve Welpe, I. (2010). Predicting elections with twitter: what 140 characters reveal about political sentiment. *Proceedings of the international AAAI conference on web and social media, 4(1),* 178-185. doi: 10.1609/icwsm.v4i1.14009

Uslu, O. ve Özmen-Akyol, S. (2021). Türkçe haber metinlerinin makine öğrenmesi yöntemleri kullanılarak sınıflandırılması. *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, *2(1),* 15-20. Erişim adresi: https://dergipark.org.tr/en/download/article-file/1483397

Van Loon, A. (2022). Three families of automated text analysis. *Social Science Research*, *108*, 102798. doi: 10.1016/j.ssresearch.2022.102798

Vasiliev, Y. (2020). *Natural language processing with Python and spaCy: A practical introduction*. San Francisco: No Starch Press.

Wesley, J.J. (2014). *The qualitative analysis of political documents*. Bertie Kaal, Isa Maks ve Annemarie van Elfrinkhof (Ed.), *From text to political positions: text analysis across disciplines* (ss.135-160) içinde. Amsterdam: John Benjamins

Wilkerson, J. ve Casas, A. (2017). Large-scale computerized text analysis in political science: opportunities and challenges. *Annual Review of Political Science*, *20*, 529-544. doi: 10.1146/annurev-polisci-052615-025542

Young, L. ve Soroka, S. (2012). Affective news: the automated coding of sentiment in political texts. *Political Communication*, *29(2),* 205-231. doi: 10.1080/10584609.2012.671234

Yu, B., Kaufmann, S. ve Diermeier, D. (2008). Classifying party affiliation from political speech. Journal *of Information Technology & Politics, 5(1),* 33-48. doi:10.1080/19331680802149608

Zanini, N. ve Dhawan, V. (2015). Text Mining: an introduction to theory and some applications. *Research Matters*, *19*, 38-45. Erişim adresi: https://rb.gy/q4rwu5

EXTENDED SUMMARY

This article provides an overview of the most commonly used automatic text analysis methods in the field of political science. These methods range from simple statistical analysis to more advanced techniques such as supervised and unsupervised machine learning, distributional semantic models, and word embeddings. The main objective of this study is to explain how these methods can be employed to extract meaning from texts, illustrated with examples from the literature. Thus, it can serve as a starting point for political science researchers and students, equipping them with the necessary information to make informed decisions when designing their own research projects. In order to promote informed decision-making, each method is presented along with its main assumptions, biases, and limitations.

Simple statistical analysis represents the most basic and straightforward form of automated text analysis. It offers a rapid overview of frequently used words and phrases in textual data, along with descriptive statistics such as word count, average sentence length, and word variety. The primary advantage of simple statistical analysis lies in its quick and easy production, providing an initial glimpse into the content of extensive text corpora. However, its limitations become apparent when attempting to extract information regarding the meaning or context of the text, as well as identifying trends or relationships within the data. Simple statistical analysis assumes that text data follow a normal distribution, which may introduce bias in the selection of summary statistics used to describe the data.

Supervised machine learning can classify text data based on pre-established categories or tags. This method employs training data to deduce patterns within the text and applies these learned patterns to new text data for predictions. The significant advantage of this approach is its ability to categorize text data with high accuracy while identifying crucial features or topics within the text. However, these methods are computationally complex compared to simple statistical analysis, necessitate large amounts of training data, and are influenced by biases present in the training data. Unsupervised machine learning resembles supervised machine learning but does not rely on predefined categories or tags. With an inductive approach, this method explores trends and relationships within the text data without prior knowledge of its content or meaning, thus operating without training data. The primary benefit of unsupervised machine learning lies in its capacity to discover previously unknown trends and relationships. However, this method also has limitations, including its computational intensity and difficulty in interpretation. Machine learning methods can be influenced by the quality and representativeness of the training data, as well as the choice of features and algorithms employed, leading to potential biases.

Distributive semantic models and word embeddings represent relatively new approaches to automated text analysis. These methods employ mathematical algorithms to represent the meaning and context of words within the text data as vectors or high-dimensional spaces. The main advantage of these techniques is their ability to capture meaning and context in a manner that other methods cannot, enabling the identification of patterns and relationships not deducible through simple statistical analysis. However, these methods also require significant computational resources, and their results can be challenging to interpret and understand for individuals without a solid background in computer science. Distributional semantic models and word embeddings are based on the distributional hypothesis, which posits that the meaning of a word can be inferred from its context. However, they can be affected by the corpus selection used to train the models, introducing bias.

These methods offer advantages as well as limitations. Simple statistical analyses are relatively easy and quick to perform but may not provide as much information as more complex methods. Machine learning methods may require significant computational resources and time, particularly for large and intricate datasets, but they can yield more accurate and informative results. Distributional semantic models and word embeddings demand even more computational resources and training time but can provide rich and nuanced representations of word and phrase meanings.

Furthermore, these methods differ in terms of users' understanding and interpretation of the analysis results. Simple statistical analyses are relatively straightforward and interpretable. On the other hand, interpreting machine learning methods can be challenging, especially for complex models. Distributional semantic models and word embeddings are often difficult to interpret. As a result, each method has its own strengths and weaknesses. The choice of method should depend on the specific objectives and requirements of the analysis. For instance, if interpretability and simplicity are primary concerns, simple statistical analysis may be a suitable choice. On the other hand, if accuracy and detail are of utmost importance, machine learning or distributional semantic models may be more appropriate.