

Sosyal ağlarda makine öğrenmesi yöntemi ile gönderi tabanlı cinsiyet tespiti ve kullanım alanları

Özgür Öney¹

¹ Gazi Üniversitesi, Ankara, Türkiye

ozgur.oney@gazi.edu.tr

Öz Özellikle son 15 yılda kayda değer biçimde insan hayatında alan kazanan sosyal medya olgusu, beraberinde milyarlarca veriyi içinde tutan dev bir kümeyi de getirir. Bu veri kümesinden anlamlı veriler çıkarabilmek, özellikle kişilerin ayırt edilmesi ve bu bağlamda güvenlik, pazarlama, erişilebilirlik gibi alt dalların kişilere özel oluşturulması noktasında değer arz etmektedir. Sosyal medya üzerinde cinsiyet tespitinin yapılması, kişiye özel uygulamaların en başında gelir. Sosyal medyada paylaşılan bilgilerin ışığında cinsiyet analizi yapılabilmesi için, günümüzde veri analizi araçları ve makine öğrenmesi yöntemleri birlikte ilerlemektedir. Makalede, sosyal ağların eriştiği tarihçeden başlayarak sosyal ağdan cinsiyet analizi yapılmasının artı ve eksileri değerlendirilecek, ardından daha önce denenmiş yöntemlere eleştirel bir bakış getirilecek ve son kısımda da tüm bu işlemlerin uygulanabilirliği tartışılacaktır.

Anahtar kelimeler Sosyal medya analizi, makine öğrenmesi, veri madenciliği, cinsiyet tahmini, çıkarım, tahmin yöntemleri.

Abstract The phenomenon of social media, which has gained considerable impact on human life especially in the last 15 years, the, brings with it a huge set of billions of data. To be able to extract meaningful data from such data set is especially worthwhile to distinguish between individuals and to create sub-branches such as security, marketing and accessibility. Gender determination on social media is one of the most important applications. Today, data analysis tools and machine learning methods are progressing together in order to perform gender analysis in the light of the information shared on social media. In the article, the pros and cons of social analysis from the social network will be evaluated starting from the history of social networks, then a critical look will be given to the previously tried methods and the applicability of all these processes will be discussed in the last part.

Keywords: Social media analysis, machine learning, data mining, gender recognition, inference, estimation methods.

1. Sosyal Medya

1.1 Sosyal medya kavramı

Sosyal medya siteleri, kendi sınırlı çerçeveleri ve içerikleri doğrultusunda kullanıcılarına halka açık ya da kısmi şekilde halka açık bir profil oluşturma imkanı sağlayan, kullanıcıların bağlantı içinde bulunduğu kişiler arasında deneyim, bilgi, anı ve benzer bir çok ögeyi paylaşma imkanı sağlayan ve diğerlerinin paylaştığı öğeleri görmemize imkanı veren, üzerine ek olarak arkadaşların çevresinden kişileri de kolayca görüp kendi ağına ekleyebildiğiniz ağ servisleri olarak tanımlanabilir.[1] Tüm bu özelliklere ek olarak, sosyal medyanın kişilere, “normal hayatta kuramayacakları arkadaşlık ya da paylaşım ortamını sağlayabilen bir platform” olduğundan söz edilebilir, ki bu durum “örtülü bağ” kurma durumu olarak anılır.[2]

Piyasaya sürüldükten sonra geçen yıllar içinde, başta Facebook, MySpace, Youtube ve Twitter gibi websiteleri olmak üzere, sosyal medyada yer alan internet siteleri dünya üzerinde farklı bölgelerden milyonlarca kullanıcıya gündelik hayatları ile bütünleşik bir ortam sağlamıştır. Sosyal medya uygulamalarının, sıradan hayatların içine bu denli bütünleşik kalabilmesinin nedeni ise, insanların temel olarak kendilerini istedikleri gibi anlatabildikleri[3] ve sosyal ortamlarındaki bağları dijital ortama yansıtabildikleri bir mecra arayışında bulunmalarıdır. Dolayısıyla, bu kadar insanın kendi ile özdeşleştirebildiği mecralarda yer alması, akademik ve ticari araştırmaların doğması ile doğrudan ilişkilidir. Nitekim, makalenin ilerleyen kısmında anlatılacağı şekilde, özellikle 2000li yılların sonlarında, kişilerin sosyal medya paylaşımlarının analizi ve bu analizden doğan makul öneri mekanizmalarının üretilmesi konusunda ciddi çalışmalar yapılmıştır.

1.2 Sosyal medyanın tarihçesi

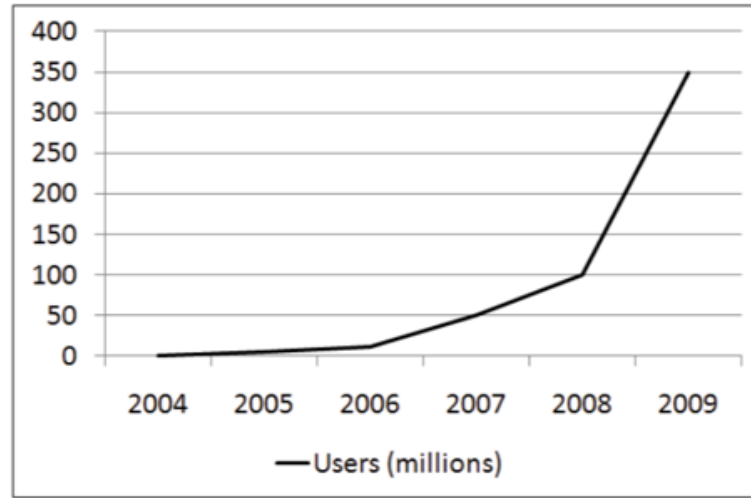
Sosyal medya olgusu, günümüzde popüler olan sosyal medya platformlarının tarihçesinden bildiğimizin aksine, daha eskilere dayanmaktadır. Bilinen ilk sosyal medya sitesi olan SixDegrees.com, kullanıcılarına profil oluşturmak, paylaşımlar yapmak ve arkadaşları ile irtibata geçmek gibi zamanının ötesinde[4] olanaklar sunsa da, o günlerdeki internet altyapısının tüm bu olanakları güçlü şekilde destekleyici biçimde olmaması, sosyal medyanın kuruluş aşamasında gecikmeye sebep olmuştur.

1997 yılından 2001 yılına değin, özellikle kadın-erkek ilişkilerini hedef alan sosyal medya platformlarının internet ağında yükselişi göz ardı edilmemelidir. AsianAvenue, MiGente ve Friends gibi siteler bu dört yıllık zaman aralığında, kadın-erkek ilişkilerinin sosyokültürel olarak zahmetli görüldüğü Uzakdoğu pazarında tutunmaya çalışmıştır. Ancak, sitelerin sınırlı bir yaş kitlesine hitap etmesi, internet ağının erişilebilirlik konusunda yaşadığı problemler ve tüm bunların üstüne, özellikle 21. Yüzyılın başında Amerika Birleşik Devletleri’nde yaşanan internet yatırımlarının fazlalığına kadar olan internete yatırım yapmaya şüpheli bakan fonların neticesinde, bu siteler dar kapsamlı ve niş bir kitleye hitap eden oluşumlar olarak kalmıştır.

2001 yılında, daha sonraları LinkedIn isimli iş ağının kurucuları da olacak bir ekip, Ryze.com isimli iş ağını kurarak günümüz sosyal medya sitelerinin kabaca temelini atar. Her ne kadar Ryze, çok fazla kişinin ilgisini çekememiş olsa da, özellikle iş konusunda sağlayabileceği avantajların, Silikon Vadisi yatırımcıları tarafından keşfedilmesi

neticesinde, sonraki dönemde emsal sitelerin arkasına aldığı rüzgarın en önemli sebeplerinden biri olarak görülmektedir.

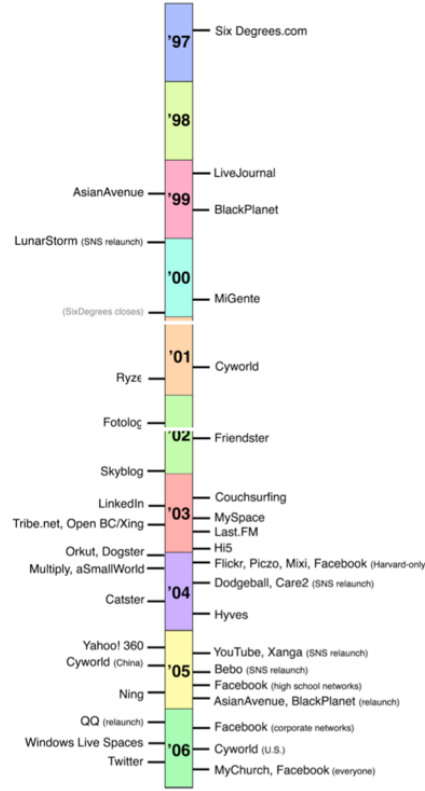
2004 yılına gelindiğinde ise, Harvard Üniversitesi'nin eski bir öğrencisinin, üniversite mailine sahip olan kişileri müdahil edecek şekilde kurduğu arkadaşlık sitesi Facebook (önceki ismi ile thefacebook.com), sırası ile önce diğer üniversitelere, sonra liselere, şirket mailine sahip olan kişilere ve son olarak herkese, daha önce sunulan profil oluşturma ve paylaşım yapma seçeneklerini daha cazip hale getirerek, sosyal medyanın karşı konulamaz büyümesinin fitilini ateşlemiştir. Nitekim, Facebook'un bu agresif büyümesi ile aynı dönemlerde ortaya çıkan akıllı telefon olgusu, sosyal medya sitelerinin daha geniş kitlelere ulaşmasını sağladığı gibi, her an her yerde kullanılarak günlük rutinin bir parçası olmasına olanak vermiştir.



Şekil 1: Senelere göre Facebook kullanıcı sayısı

Öte yandan, çalışmanın yürütüldüğü sosyal medya sitesi Twitter, 2006 yılında yine Amerika Birleşik Devletleri sınırları içerisinde kapılarını kullanıcılarına açmıştır. Son kullanıcıya sınırlı sayıda karakterle paylaşımlar yapma imkanı sunan site, özellikle mikro-blogging olarak isimlendirilen paylaşım tarzına öncülük etmesi ve daha önemlisi, ünlü kişiler tarafından kullanılması neticesinde bilinirlik kazanarak[5], sosyal medya platformları ağının büyük oyuncularından biri haline gelmiştir.

Sonuç olarak, tarihsel gelişim evresi içinde sosyal medya değerlendirilecek olursa, kullanıcı sayılarında yaşanan artışla birlikte sosyal medya platformlarının çeşitliliklerindeki artışla, günümüzde bir gerçek haline gelmiş sosyal medya kullanımının ilerleyen dönemde benzer trendin seyriyle birlikte artacağını ve bu durumun sosyal medyayı baz alan akademik ve endüstriyel araştırmalara hatırı sayılır genişlikte bir alan yaratacağını söylemek mümkündür. Özellikle, gelişim evresi içinde kullanıcıya doymuş platformların yerini, yeni ve dinamik platformların alması ve buna ek olarak kullanıcıların da kitlesel ve üstel şekilde bu platformları tercih etmesi, bu durumun temel sebebi olarak gösterilebilir. Nitekim, şimdiden sosyal medya platformunda oligapol yapının devam ettirilebilmesi adına, kökleri daha eskiye dayanan sosyal medya platformlarının yenilerine satın alma ya da fonlama yöntemleri ile teşvikler verdiği gerçeği göz ardı edilmemelidir.



Şekil 2: 1997 yılından itibaren kurulan sosyal ağ sitelerinin kronolojik sıralaması[6]

1.3 Sosyal medyanın günümüzde durumu

Sosyal medya, kurulduğu ilk günden günümüze, hayatımızda her geçen gün daha fazla yer etmektedir. Bunun ilk sebebi olarak, sosyal medyaya erişimin kolaylaşması gösterilebilir. 2005 yılında, sosyoekonomik çizelgede sadece yukarıda kalan kişilerin kullanabildiği ve düşük gelirden oluşan kesimde kullanım oranı yalnızca %4 olan internet, 2015 yılı itibarı ile geçen 10 yıllık süre zarfında kullanım oranını neredeyse 14 kat artırarak, gelir seviyesi düşük kesimde neredeyse her 10 kişiden 6'sının kullandığı bir olgu haline gelmiştir. [7] Ek olarak, altyapı çalışmalarının AR-GE faaliyetleri neticesinde hızlandırılması ve sosyal medya kullanıcılarını tatmin edecek düzeyde hızlı internet altyapısının fiyat olarak görece ucuzlaması da, internetin yalnızca şehirde değil kırsalda da kullanılabilmesi anlamında gelir ki nitekim bu önerme, 2005 yılında %5 olarak gözlenen kırsal kesimde internet kullanma oranının 2015 yılına gelindiğinde %58 olarak gözlenmesi sonuçları ile doğrulanabilmektedir. [8] İkincil bir sebep olarak ise, sosyal medyadaki sağlayıcıların çeşitliliğinin artması gösterilebilir. Nitekim, 2000li yılların başında sadece kadın-erkek ilişkilerine odaklanmış bir sosyal medya olgusu gözlenmekteyken, günümüzde sosyal medya iş hayatından fotoğrafçılığa, küçük günlükler tutmaktan ilişkilere, anlık hikayeler paylaşmaktan müziğe ve videoya kadar uzanan geniş bir yelpazede hizmet vermekte olan sitelerin bütününi temsilen kullanır. Bu bağlamda, sosyal medyanın kullanımının çeşitliliğe bağlı olarak arttığı ve günlük hayatımızdaki her bir eyleme karşılık neredeyse bir sosyal medya sitesi ile karşı karşıya olduğumuz gerçeği ortadadır. Son olarak, giderek artan teknoloji kullanım yüzdesi

neticesinde, sosyal medyanın da günümüzde son kullanıcı sayısında bir artış yaşanması neticesi görülebilmektedir.

1.4 Sosyal medyadan yapılan çıkarımlar

Bahsedilen şekilde, sosyal medyanın kullanıcıların kendi isteği ile kendine ait verileri yüklemesi ya da olmayı istediği kişiye dair ipuçlarını paylaşıyor olması, sosyal medyanın bir cazibe merkezine dönüşmesinde büyük rol oynamıştır. Bu bağlamda, sosyal medya sitelerinden elde edilen verilerin anlamlı hale getirilmesi neticesinde, başta pazarlama olmak üzere birçok alanda kullanılabilecek şekilde çıkarımlar elde edilebilmektedir.

Genel yaklaşım, müşteriye sadece satın alabileceği ve kar getirebileceği şekilde ele almak olsa da, sosyal medya sayesinde müşterinin etkileyebileceği kişilerin de olduğu varsayımı son dönemde popülerlik kazanmıştır.[9] Dolayısıyla, “müşterinin değerine” ek olarak bir de “müşterinin ağ değeri” pazarlama stratejilerinde hedef haline gelmiştir. Günümüz şartlarında, ilerleyen paragraflarda nasıl yapılacağı açıklanacak veri çıkarım yöntemlerinin sosyal medya üzerinden gerçekleştirilmesi, pazarda büyük önem teşkil etmektedir.

2. Büyük Veri

2.1 Büyük veri tanımı ve sosyal medya ilintilemesi

Milyon ve hatta milyar düzeyindeki kullanıcı sayısına erişen sosyal medya uygulamalarında, birikimsel olarak ilerleyen, gün geçtikçe büyüyen ve kayda değer miktarda verinin kullanılan her gün oluşturulduğunu söylemek mümkündür. Düzlemsel olarak, sosyal medya uygulamalarının büyük bir veri havuzunun bir parçası olduğunu iddia etmek yanlış olmayacaktır.

Bu noktada, sosyal medyadan gelen akışın bu denli yoğun olduğu günümüzde, büyük verinin ne olduğuna dair bir tanım koymak gereklidir. Büyük veri, her ne kadar farklı firmalar ve kişiler farklı tanımlara erişerek tanımlıyor olsa da, temel olarak değere dönüşümü için belirli teknoloji ve analitik yöntemler gerektiren, yüksek hacim, hız ve çeşitlilik ile karakterize edilen bilgi varlıklarını temsil eder. [10]. Hacim verisinin büyüklüğünü, hız verinin hacmindeki değişim oranını ve çeşitlilik de verinin alt dallarındaki farklı girdi sayısını temsilen aktarılırken, tüm bunların sebebi olarak da sosyal medya kullanıcılarının sayısının milyar düzeyinde olması neticesinde yaşanan hızlı dolanım olarak görmek mümkündür.

Kısacası; Facebook, Twitter ve Weibo gibi sosyal ağların artan kullanımı, çok miktarda veri üretmiştir ve üretmektedir. Ticari firmalar ve diğer organizasyonlar, iş performansını arttırmak için yeni iş anlayışlarını keşfetmekle ilgilendiğinden; şirketler gelişmiş analitik yöntemler kullanarak, bireylerin ve grupların sosyal davranışlarını karakterize eden sosyal ağların altında yatan ilişkiler hakkında bilgi edinmek için büyük verileri analiz edebilir. İlişkileri tanımlayan büyük verileri kullanarak, ağdaki diğer kişilerin davranışlarını etkileyen sosyal liderleri ve diğer insanların hangi ağ katılımcılarından en çok etkilendiklerini belirlenebilmektedir.[11]

2.2 Büyük veri araştırmalarının önemi ve karşılaşılan zorluklar

Büyük veri, tüm şüphenin ötesinde, büyük bir değerdir. Bilgi endüstrisinin bakış açısından, büyük veri, esas olarak büyük veri, bulut bilişim, mobil internet ve sosyal sektöre atıfta bulunan üçüncü platform üzerine inşa edilen yeni nesil BT endüstrisine güçlü bir ivme kazandırmaktadır. IDC, 2020 yılına kadar üçüncü BT platformunun pazar büyüklüğünün 5.3 trilyon ABD dolarına ulaşacağını ve 2013'ten 2020'ye kadar, BT sektöründeki büyümenin %90'ı üçüncü BT platformu tarafından yönlendirileceğini tahmin etmiştir.[12] Sosyoekonomik açıdan bakıldığında büyük veri, Amerikan ekonomisti W.B. Arthur'un ileri sürdüğü şekilde[13], işlemci, konektörler, sensörler ve yürütücüler üzerinde çalışan ekonomik faaliyetleri üzerinde barındırarak ikinci ekonominin çekirdek çağrışımı ve eleştirel desteğidir. 2030 yılına kadar bu bahsi geçen ikinci ekonominin büyüklüğünün ilk ekonomiye (yani geleneksel fiziksel ekonomiye) yaklaşacağı tahmin edilmektedir. İkinci ekonominin temel desteği, tükenmez ve sürekli zenginleştirici bir kaynak olduğu için büyük veridir. Gelecekte, büyük verilerden dolayı, ikinci ekonominin altındaki rekabet artık emek üretkenliği değil, bilgi üretkenliği olacaktır.

Büyük verinin, ulusal, bilimsel, endüstriyel ve beşeri bakımdan büyük önemi bulunmaktadır. Öncelikle, ulusal bazda düşünüldüğünde, bilhassa İkinci Dünya Savaşı ile başlayan süreç itibari ile, ulusların ölçek bazında gücü sahip olunan jeopolitik, sosyolojik ya da askeri kaynaklar ile değil, bilgi ile ölçülmektedir. Dolayısıyla, veri üzerinde egemen olan ülkelerin, gelecekte kara, hava, deniz ve dış uzay ile alakalı üstünlük kuracakları aşıkardır. Bu doğrultuda, başka Amerika Birleşik Devletleri'nin başı çektiği Batı ülkeleri ve Çin'in başını çektiği gelişmekte olan ülkeler bloğu, hızla büyük veri işleme merkezleri kurmakta ve alana yatırım yapmaktadır. Büyük verinin bir ikinci önemi ise, hiç şüphesiz sanayi alanındaki gelişmelere sağladığı ufuk açıcı roldür. Son yıllarda büyük perakende şirketlerinin de önemin farkına varmaları neticesinde, kullanıcılardan gelen verilerin doğru ve bütüncül şekilde işlenmesi neticesinde pazarlama ve tedarik sektörlerinde hızla gelişen sonuç odaklı yaklaşımlar görülmüştür. Büyük veri, akademik camianın bilimsel araştırma yöntemlerini yeniden gözden geçirmesine sebep olarak[14], bilimsel düşüncede de değişimlere sebep olmuştur. Her ne kadar günümüze değin bilimsel araştırma sırasıyla deneysel, teorik ve simülasyon tabanlı şekilde ilerlese de, günümüzde yalnızca ihtiyaç duyulan verilerin tüm veri havuzundan çekilerek kullanılabileceği algısının getirdiği yaklaşımlar, büyük verinin bir devrim yarattığı söylenebilir. Son olarak, beşeri boyutta büyük veri, günü daha iyi anlamak ve günümüzdeki verilerin ışığında geleceği daha iyi tahmin etmek konusunda da yardımcı olmuştur.

Büyük veri araştırmalarında, üç farklı zorluktan söz edilebilir. Bunlardan ilki, verinin karmaşıklığının yarattığı zorluk hissiyatıdır. Temel problem, büyük verilerin karmaşıklığının temel özelliklerini nasıl formüle edileceği veya nicel olarak tanımlanacağıdır. Büyük verilerin karmaşıklık teorisi üzerine yapılan çalışma, büyük verilerde temel özellikleri ve karmaşık kalıpların oluşumunu anlama, temsiliyi basitleştirme, daha iyi bilgi birikimi elde etme ve büyük veri üzerinde hesaplama modelleri ve algoritmalarının tasarımı yönlendirmeye yardımcı olacaktır. Ancak, bunu gerçekleştirebilmek için çok modellenmiş ilişkilerde veri dağıtım teorisini ve algoritmalarını kurmak gerekecektir. Ayrıca, veri karmaşıklığı ile uzaysal hesaplamalı karmaşıklık arasındaki içsel bağlantılarının ayrılması gerekecektir. Dahası, veri karmaşıklığının içsel mekanizmalarını modellemek ve analiz ederek, büyük veriyi işlemek için büyük veriyi sağlam bir temel haline getirme ilkelerini ve mekanizmalarını açıklanabilecektir. Değişen veri kümesi üzerinde yapılmasını güçleştiren bir diğer konu

olarak hesapsal karmaşıklık gösterilebilir. Öte yandan, karmaşık hesapları çalıştırmak için gereken sistemlerin, enerji tüketimi ve operasyonel verimlilik konusundaki karmaşıklıkları da, büyük verinin işlenmesi hususunda karşımıza çıkan zorluklardan biri olarak gösterilmektedir.

3. Sosyal medyada veri madenciliği

3.1 Kullanım alanları

Sosyal medyada veri madenciliğinin kullanımı, her ne kadar yeni bir alan olsa da, gün geçtikçe popülerlik kazanmaktadır. Tıpkı gerçek madencilerin ellerinde gerekli aletler ile toprağı kazması ve sonucunda değer yaratan bir elementi gün yüzüne çıkarması gibi, veri madenciliği de temel olarak değerli ve değersiz verinin bir arada bulunduğu bir veri yığınının, gerekli aletlerin kullanımı neticesinde anlamlı ve değer yaratan verinin açığa çıkarılması amacını gütmektedir. Veri madenciliğinin sosyal medyaya uyarlanması, her gün inanılmaz miktarlarda verinin üretildiği internet ortamında insanların davranışlarını ve etkileşimlerinin analiz edilmesi adına önem arz etmektedir. [15]

Bu bağlamda, sosyal medyanın veri madenciliğinde sunduğu tabana en güzel örneklerden biri, 2008 yılında Amerika Birleşik Devletleri'nde yapılan başkanlık seçimleridir. Seçimler esnasında, YouTube ve Facebook dahil olmak üzere sosyal medya siteleri, fonları artırma ve adayların mesajlarını seçmenlere alma konusunda önemli bir rol oynamıştır[16]. Massachusetts Teknoloji Enstitüsü (MIT), Kolektif Zeka Merkezi'ndeki araştırmacılar, adayların kullandıkları sosyal medya miktarı ile 2008 başkanlık kampanyasının kazananı arasındaki ilişkiyi göstermek için blog verilerini çıkardılar[17]. Bu güçlü örnek, veri madenciliği sosyal medya verilerinin, ulusal düzeyde elde edilen sonuçları öngörme potansiyelini vurgulamaktadır.

Yukarıdaki ulusal düzeydeki örneğe ek olarak, sosyal medyanın toplumsal, ekonomik ve sosyolojik etkilerine direkt bağlı şekilde, veri madenciliği araçlarının bahsi geçen mecraları hedef aldığı da söylenebilir. Veri madenciliği olgusu, buradan sağladığı anlamlı çıkarımlar neticesinde, toplumsal güvenliğe de katkıda bulunmaktadır.

3.2 Cinsiyet tespitinde veri madenciliği kullanımı

Bahsedildiği şekilde, sosyal medyanın sağladığı veriler temel alınarak yapılan araştırmalarda, sosyal medyadaki gönderiyi yayınlayan kişinin cinsiyetini tahmin etmek de bir araştırma konusudur. Her ne kadar internet ve sosyal medya, ilk çıktığı zamanlarda erkek egemen bir mecra olarak görünse de, bu fark günden güne kapanmakta ve dolayısıyla yüzdeler eşitlenmeye yaklaşmaktadır. Bu durumda, cinsiyet tespitinin yapılmasına dair çalışmaların sayısının ilerleyen günlerde artacağından da bahsetmek yanlış olmayacaktır.

Geçmişten günümüze gelen erkek egemen internet kavramının sebebi olarak, üç farklı görüş öne sürülebilir. Bunlardan ilki, kadınların temel olarak dışa dönük değil içe dönük, yani okul ya da benzeri kurumlardan değil aile kurumundan daha yoğunluklu eğitim alması neticesinde, kariyer aşamasında erkeklere göre daha az aktif rol almalarıdır. Böylece, kadınların günlük hayattan ziyade iş hayatında daha aktif olarak kullanılan internetten daha az faydalanıyor olmaları makul görünebilir.[18] Bir ikinci sebep, kadınların “çifte gün” denilen olguyla iç içe olmasıdır. Çifte gün kavramı, çalışan

kadınların işlerinden evlerine döndükleri andan itibaren eve dair sorumlulukları yerine getirmede bir kez daha çalışıyor olmalarına yapılan naziredir. [19] Erkeklerin ev işlerinde aldıkları sorumlulukların sayısındaki azlık, kadınların internet kullanımına dair bir asimetri yaratmıştır. Son olarak da, Georgia Tech GVU Merkezi'nde yapılan araştırmaya göre, erkeklerin internet kullanımında kadınlara oranla daha rahat hissetmeleri gerçeği de[20], kadınların internet kullanımındaki yüzdesinin şu zamana kadar erkeklere oranlar daha düşük oluşunu açıklar. Ancak, tüm bu önermelerin dışında, gittikçe artan kullanım ve kitlesel teşvikler neticesinde internette cinsiyetler arasındaki kullanım yüzdelerinin farklılığının gittikçe azaldığı sonucu da göz ardı edilmemelidir.

Bu farkın kapanması ile, internet üzerinde cinsiyet farklılıklarının tespiti çalışmaları arasında korelasyon vardır. Cinsiyet tespit edilmesinin temel amaçlarından ilki, elbette ki güvenlidir. Sosyal medya üzerinde, sahte isim, yaş, cinsiyet ve yaşanılan yer yaratma imkanı, başta terör, dolandırıcılık ve pedofili olmak üzere ağır suçlara zemin hazırlayabilmektedir. [21]

Bu suçlardan cinsiyet tespiti gerektireni, kuşkusuz ki pedofilidir. Suçlular, sosyal medya siteleri üzerinden karşı cinse ait hesaplar açarak kurbanlarında sempati ve yakınlık yaratma ve bu durumun yardımıyla da kötü emellerine ulaşmayı hedeflemektedirler. Bu bağlamda, kullanıcıların sosyal medyada yaptıkları paylaşımlardan ve metinde kullandıkları ifadelerden gerekli çıkarımların yapılması ile suçlu kişiye ulaşılması, sosyal medyada veri madenciliği ile cinsiyet analizinin nihai hedefini açıklar. Belçika'da yapılan araştırmalar neticesinde, yerel bir sosyal medya ve mesajlaşma uygulaması olan Netlog üzerinde yapılan deneysel çalışmalarda, elde edilen mesajlaşma kümesi içerisinde %90'a varan kesinlikte cinsiyet tespiti sağlanmış ve bu çalışmalar yine Belçika polisi ile paylaşılarak ileride yaşanması olası güvenlik problemlerinin önüne geçilmiştir.[22]

Cinsiyet tespitine dair çalışmaların faydalı olduğu bir diğer alan ise, pazarlamadır. Reklamcılık sektörünün kitle bazlı hedeflemeden ziyade birey bazlı hedeflemeye geçtiği bu dönemde, sosyal medya paylaşımlarından ya da sadece oluşturulan profilin kendisinden kullanıcının cinsiyetini belirlemek ve kişiye bu doğrultuda reklamlar öneriyor olmak, daha fazla getiriyi daha az harcama ile sağladığından ötürü şirketlerin ilgisini cezbetmektedir.

3.3 Cinsiyet tespitinde veri madenciliği yöntemleri

Veri madencileri, sosyal medya üzerinde bireylerin cinsiyetlerine dair çıkarımlar yaparken, farklı yöntemler kullanırlar. Çalışmanın hedefi, istenen kesinlik aralığı ve ciddiyetine göre, kullanılan yöntemler tekil ya da hibrid yöntemler olabilmektedir.

Bu yöntemlerden ilki, en basit şekilde yapılan paylaşımın gönderildiği yere bakarak kabaca tahmin yürütmektir. Her ne kadar bu yöntem daha çok destekleyici pozisyonda yer alıyor olsa da, özellikle erkeklerin yoğun olarak bulunduğu sanal oyun salonları, stadyumlar ve kıışlalar ile kadınların daha yoğun olarak bulunduğu rahat mekanlar, alışveriş merkezleri ve bakım merkezleri gibi belirleyici noktaların varlığı, cinsiyet tahminini kolaylaştırmaktadır. [23]Bölge bazlı cinsiyet tahmininin yardımı ile, en basitinden hangi reklam içeriğinin nereye konumlandırılacağına dair basit bir çıkarım elde edilebilmektedir.

Bu yöntemlerden ikincisi, kişinin profiline bakarak yapılacak cinsiyet tahminidir. Bu tahmin yöntemi ile, kişilerin kullanıcı adı, **hashtag** olarak adlandırılan konu başlıklarının kullanım sıklığı ve bu hashtagların hangilerinin kullanıldığı, profil fotoğrafları, profilde yer alan arka plan resmi ve bu resmin baskın rengi, yapılan paylaşımların uzunluğu ve içeriği, paylaşımlar arasındaki süre gibi kontrol adımları ele alınarak, kişilerin cinsiyetlerine dair çıkarımlarda bulunulmaktadır.

Bir diğer anlamlı veri elde etme yöntemi ise, Belçika’da yapılan araştırmada gösterildiği şekilde, kullanıcıların attığı metinlerde kullandıkları kelimelerin ve figürlerin analiz edilmesidir. Bu doğrultuda, kadın ile erkeklerin sıkça kullandıkları kelimeler ve bu kelimelerin türevleri işaretlenerek, dillerdeki doğal kadın erkek ayrımından faydalanılmaya çalışılmaktadır. [22] Örneğin, kadınlar daha çok dili görselleştirme amacı güdüp hareketli karakterler ya da gülen suratlar gibi dil dışı ifadeler kullanmaya eğilimliler iken, erkekler daha sadece, daha kısa ve daha az karakter kullanarak paylaşımlar yapma konusunda eğilimlidirler. Cinsiyet tespitinde veri madenciliği potansiyel görüşler

Sosyal medyadan alınan verilerin işlenmesi neticesinde anlamlı ve kesine yakın cinsiyet tahminlerinde bulunmak, veri madenciliğinin bir alt dalı olarak ilerleyen dönemde daha fazla önem kazanacaktır. Dünyada son 25 yılda artan güvenlik ihtiyacı ve kişilerin olası saldırılara karşı yüzey alanı internet kullanımı ile birlikte kayda değer şekilde arttığından, güvenlik amaçlı veri madenciliği araçlarının kullanımı da önem arz etmektedir ve zamanın ruhuna uygun şekilde artış gösterecektir. Eş zamanlı olarak, pazarlama sektörünün neredeyse yüz yıldır sabit yöntemleri kullanıyor olması neticesinde yaşanan durgunluklar ve sektörde sürekli canlı tutulmak istenen yaratıcılık ihtiyacı ile, veri madenciliği ile cinsiyet analizi yapılması desteklenecek ve günden güne sektörde hakim pozisyonundaki yöntemlerden biri haline gelecektir.

4. Sonuçlar

Sonuç olarak, 1997 yılından günümüze gelişmekte olan ve teknolojinin akıllı telefon olgusunu yaratması ile birlikte herkesin ulaşabileceği hale gelen sosyal medya, yapılan paylaşımlar neticesinde her gün devasa bir verinin üretildiği mecra haline gelmiştir. Bu karmaşık veri havuzundan anlamlı örüntülerin çıkarılması ve bilgiler elde edilmesi arayışı, sosyal medyada veri madenciliği kavramının çıkış noktasıdır. Cinsiyetin belirlenmesi de, özellikle güvenlik ve pazarlama sektörüne katkıları sayesinde bu veri madenciliği trendinin önemli bir ayağını oluşturmaktadır. İlerleyen dönemde, artan sosyal medya mecralarının sayısı ve kullanıcı sayısı ile altyapıdaki gelişmelerin de destekleyeceği şekilde, sosyal medyada veri madenciliği daha da güçlenecek ve önem kazanacaktır.

5. Referanslar

- [1] Boyd, Danah, and Nicole Ellison. "Social Network Sites: Definition, History, and Scholarship." *IEEE Engineering Management Review*, vol. 38, no. 3, 2010, pp. 16–31., doi:10.1109/emr.2010.5559139.
- [2] Haythornthwaite, Caroline. "Social Networks and Internet Connectivity Effects." *Information, Communication & Society*, vol. 8, no. 2, 2005, pp. 125–147., doi:10.1080/13691180500146185.
- [3] Boyd, Danah. "Book Review: Material Virtualities: Approaching Online Textual Embodiment." *New Media & Society*, vol. 7, no. 1, 2005, pp. 139–141., doi:10.1177/146144480500700110.
- [4] A. Weinreich, pers communication, July 11, 2007
- [5] Edosomwan, Simeon O., "The history of social media and its impact on business" , *The Journal of Applied Management & Entrepreneurship*, vol. 16, no. 3, pp. 42-48.
- [6] Boyd, Danah, and Nicole Ellison. "Social Network Sites: Definition, History, and Scholarship." *IEEE Engineering Management Review*, vol. 38, no. 3, 2010, pp. 16–31., doi:10.1109/emr.2010.5559139.
- [7] Andrew Perrin. "Social Networking Usage: 2005-2015." Pew Research Center. October 2015. Available at: <http://www.pewinternet.org/2015/10/08/2015/Social-Networking-Usage-2005-2015/>
- [8] Andrew Perrin. "Social Networking Usage: 2005-2015." Pew Research Center. October 2015. Available at: <http://www.pewinternet.org/2015/10/08/2015/Social-Networking-Usage-2005-2015/>
- [9] Zhang, Jiawei, et al. "Intertwined Viral Marketing in Social Networks." *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, doi:10.1109/asonam.2016.7752241.
- [10] De Mauro, Andrea & Greco, Marco & Grimaldi, Michele. (2014). What is Big Data? A Consensual Definition and a Review of Key Research Topics. 10.13140/2.1.2341.5048.
- [11] O. Liu, K.L. Man, W. Chong, "Social network analysis using big data", *IMECS*, vol. 2, 2016

- [12] Jin, Xiaolong, et al. "Significance and Challenges of Big Data Research." *Big Data Research*, vol. 2, no. 2, 2015, pp. 59–64., doi:10.1016/j.bdr.2015.01.006.
- [13] .B. Arthur, The second economy, available at: <http://www.images-et-reseaux.com/sites/default/files/medias/blog/2011/12/the-2nd-economy.pdf>, 2011.
- [14] T. Hey, S. Tansley, K. Tolle (Eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Corporation, 2009.
- [15] Barbier, Geoffrey & Liu, Huan. (2011). Data Mining in Social Media. *Social Network Data Analytics*. 327-352. 10.1007/978-1-4419-8462-3_12.
- [16] E. Qualman. *Socialnomics*. Knopf Books for Young Readers, New York, 2009.
- [17] P. Gloor, J. Krauss, S. Nann, K. Fischbach, and D. Schoder. Web science 2.0: Identifying trends through semantic social network analysis. volume 4, pages 215 –222, aug. 2009.
- [18] Weiser, Eric B. "Gender Differences in Internet Use Patterns and Internet Application Preferences: A Two-Sample Comparison." *CyberPsychology & Behavior*, vol. 3, no. 2, 2000, pp. 167–178., doi:10.1089/109493100316012.
- [19] Rosen, E.I. (1987). *Bitter choices: Blue-collar women in and out of work*. Chicago, IL: University of Chicago Press.
- [20] Graphics, Visualization, and Usability Center (GVU). (1999). *Results of Gvu's tenth world wide user survey* Atlanta: Georgia Tech Research Corporation. Online document: http://www.gatech.edu/gvu/user_surveys/survey-1998-10/tenthreport.html
- [21] Peersman, Claudia, et al. "Predicting Age and Gender in Online Social Networks." *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents - SMUC 11*, 2011, doi:10.1145/2065023.2065035.
- [22] Peersman, Claudia, et al. "Predicting Age and Gender in Online Social Networks." *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents - SMUC 11*, 2011, doi:10.1145/2065023.2065035.
- [23] Fayoumi, Amjad & Jackson, Christopher & Lewis, Conor & Straw, Jordon & Sharpe, Joshua & Nicol, Duane. (2017). What They Are Tweeting About Me? Social Media Data Analytics with Geographical Visualisation.