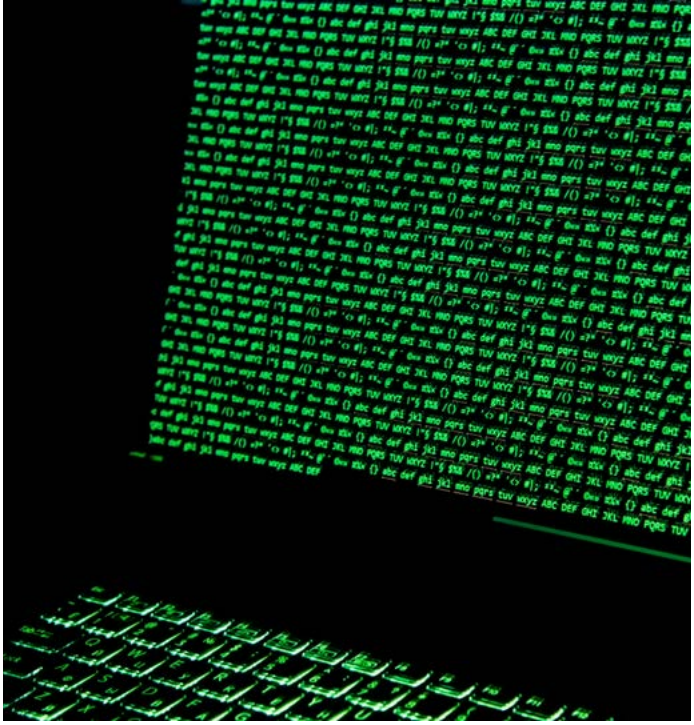
thinktech
STM Teknolojik Düşünce Merkezi

Metin Madenciliği: İmkânlar, Kısıtlar ve Geleceği



Yazı icat edildiğinde, bu harika aracı okuyacak olanların insanlar olacağı düşünüldü. Sümerler'den bugüne büyüsel amaçlarla yazılan ve kimsenin okuması beklenmeyen tılsımlar ve gizli tutulan günlükler dışında yazı her zaman insanlar için var olageldi. Bugün ise yakın geçmişimizde var olan Python, R, Julia, Java, C#, Go, Scala, SQL, Ruby ve Lua gibi bilgisayar dilleri insanlar için değil bilgisayarlar için tasarlandı ve kodları anlamaya çalışan yazılımcılar dışında hep bilgisayarlar tarafından okundu.

Bilgisayar dilleri insan dillerinin aksine bir düşünceyi iletmek gibi mütevazı bir amacın ötesinde insanların artık elle yapmak istemediği ve yapılması gereken işleri gerçekleştirmek için hep emir kipinde yazılan komutlardan oluşuyor. Uzun uzun yazılan bu komutlarda dünya görüşümüz yerine ağırlıklı ortalaması alınacak sayıların nasıl toplanacağını ifade ediyoruz. Neyse ki bilgisayar dilleri halen insanlar tarafından üretiliyor. Bilgisayarların kendi aralarındaki dilleri, bazı ileri deneyler dışında bilgisayarlar henüz oluşturabilmiş değiller.

İnsandan insana iletişimin aracı olarak dünyada günümüzde 6000'in üzerinde farklı dil var. Artık büyük veri paradigması dünyada işlenip de yorumlanmamış tek satır bilgi bırakmama idealiyle hareket ettiğine göre, milyarlarca satır sayısal veri analiz edilirken, yazılmış milyarlarca kelime analiz edilmeden kenarda duracak değildi.

Bilgisayarlara dayalı analitiğin imkânlarının tam olarak kullanılmadığı dönemlerde metinleri analiz etmek yine bu kompleks dilleri icat eden atalarımızın torunları olan bizlere kalmıştı. Terminolojik olarak dilsel komplekslik kavramının ötesinde, insan dilleri halen bilgisayar dillerine göre kuralları, kuralsızlıkları, istisnaları, lehçeleri ve şiveleriyle bilgisayar dillerinden daha kompleks kalmaya devam ediyor. Bu nedenle yazılan tweet'leri halen excel'e koyup olumlu ve olumsuz tweet'leri toplayamıyoruz.

Bilgisayar dillerinin esnekleşmesiyle insan dillerini de analiz edecek imkânları araştıran "metin madenciliği" alanı doğdu. Adından da tahmin edileceği üzere "veri madenciliği" alanın alt dalı olarak kabul ediliyor. Metin madenciliği yöntemleriyle örneğin yazılı herhangi bir doküman koleksiyonunda (buna metin madenciliği literatüründe korpus denir) kayıtlı tüm dokümanlardaki (bunlar haber, tweet, rapor, kitap veya yorum olabilir) kelimeleri, köklerine inerek sayabilir ve en basit haliyle kelimelerin sıklık ve dağılımını elde edebilirsiniz.

Daha ileri örneklerde ise kelimelerin sıklık ve dağılımına dayanarak dokümanların olumlu, olumsuz, nötr olduğunu anlayabilirsiniz. Tam olarak bunun canlı örneğini e-posta hesaplarımızdaki spam kutularının arkasındaki algoritmalar olan SpamClassifier'lar gerçekleştiriyor. Hangi e-postanın istek dışı reklam e-postası, hangisinin de kuzenimiz ya da patronumuzdan gelen gerçek e-posta olduğunu algılayabiliyor. Bunu yaparken

de sadece kelime istatistiği değil makine öğrenmesi kullanıyor. Naive Bayes ve kNN gibi algoritmalarla metin klasifikasyonu gerçekleştiriliyor.

Bilgisayarların bir ölçüde insan dilini anlamaya başlamalarının bize sağladığı en büyük kolaylık, artık okuyamayacağımız miktara ulaşan içeriklerle baş etmemizi sağlamaları oldu. Anlamı itibariyle, internet öncesi dönemde metin madenciliği bugünkü gücüne ulaşmış olsaydı (internet öncesinde de metin madenciliği çalışmalarının var olduğunu hatırlatalım) içerik sayısı zaten az olduğundan bu teknolojinin özel bir değeri olmayacaktı. Ancak artık her an farklı kaynaklarda ortaya çıkan milyonlarca farklı içeriği tek tek okumaya vaktimiz bulunmuyor ve bu nedenle bazı alanlarda (Spam örneği gibi) metin madenciliğine muhtacız.

Metin madenciliği arka planda işleyen ve hayatımızı kolaylaştıran yazılımsal çözümler dışında da kullanılıyor. Nicel yöntemlerle ekonomik ve finansal haberlerin sınıflandırılması ile piyasaların yönü hakkında bilinen parametrelere bir yenisi daha eklenmiş oluyor. Diğer taraftan, kullanıcıların sosyal medyada veya haber sitelerinde bir ürüne yönelik yazdığı mesajlardan da içerik analizi yapılmak suretiyle kullanıcıların tatminine dair belki de anketlerden daha güçlü analizler gerçekleştiriliyor.

Bu kadar olanak vadeden metin madenciliği alanının da kuşkusuz belirli kısıtları bulunmaktadır. Bu kısıtlar metin madenciliğini kullanırken ortaya çıkmaktan ziyade yorumlanırken ortaya çıkar. Bu kısıtların gözden kaçırılması nedeniyle metin madenciliğine dayalı istatistiklerden herhangi bir anketin veya sayısal analizin netliği beklenebilir. Ancak verilerin çokluğu ve "gürültü" olarak tabir edilen içeriğin fazlalığı nedeniyle bu analizleri ele alırken kesin hükümlerden ziyade içgörüler sağlamaya çalışmak daha sağlıklı olacaktır.

İnsan dilinin kompleks doğası düşünüldüğünde bu kısıtların temel kaynağı daha iyi anlaşılacaktır. Örneğin Türkçe gibi sondan eklemeli dillerde son ekler metin madenciliğinde "stemming" işlemi sonrası soruna dönüşmektedir. "Ümitsiz, ümitli, ümitvar" kelimelerinin kökü "ümit" kelimesi iken, bu kelimelerin her biri olumlu ve olumsuz farklı anlama gelmektedir. İroni ise insan dilini tümden muamma haline getirebilmektedir. "O sektörün geleceği de çok parlakmış..." şeklindeki ironik bir olumsuz yorumu makinelerin anlaması halen epey güçtür.

İnsanlar için icat edilen 6000'den fazla dilin artık başka aktörler -bilgisayarlar- tarafından da anlaşılabilmeye başlandığı günümüzde bildiğimiz veri kaynaklarının ötesinde, işletmelerimizin içinde ve dışında artık yeni bir kaynağa sahibiz. Bu kaynak anılan kısıtlara rağmen varsayımlara dayalı gerçekliği yeniden kurgulamak için önemli araçlar sunmaya devam ediyor. Bu alanda, makine öğrenmesi ve doğal dil işleme (NLP) çalışmaları ile de çok daha fazla imkânın ortaya çıkacağını tahmin etmek zor değildir.