BİTİRME PROJESİ

Haftalık Rapor – 04.13.2022

4 MART 2022

KIRIKKALE ÜNİVERSİTESİ – BİLGİSAYAR MÜHENDİSLİĞİ İÖ

AHMET MUNGAN – 160255081

İÇİNDEKİLER

GİRİŞ

İnsanlar, dünya üzerinde milyarları aşan veriler ile sarmal bir yapı halindedir. Çevremizi kuşatmış olan veri yığınlarının kimi zaman farkında bile olmayız. Bu verilerin belirli kurallar yönergesinde kimi zaman hayat kurtarıcı olabileceği gerçeği son yüzyılda özümsenmiştir. Büyük boyutlardaki veriye sahip ortamların değeri, gün geçtikçe önemi artar hale gelmiştir. Bu veriler uluslararası korunma kanun ve denetim mekanizmalarına tabi olsa da, özünde bu verilerin belirli aşamalarda insanların yararı için kullanılması gerekir. Medya ortamlarında üretimini yapan sanatçı veya üreticilerin dijital ürünler üzerinde hak iddiası olabildiği gibi, yazarların da eserleri üzerinde hak iddia edebildiği bir gerçektir. Üretilen her türlü eserden çıkarımlarda bulunup, faydalı veriler haline getirmek mümkündür. Veri bilimi bu alanda uzman sayılabilen çalışmaları algoritmaya dayandırarak ortaya koymuştur. Teknoloji alt yapısı ile birlikte hızlı bir hal alan yaşam döngüsünde öz ve gerekli bilgi kıtlığını aşmak için veri bilimi üzerine çalışanlar yoğun bir çaba göstermektedir.

Tüm sebep ve gelişimlerden ötürü bu alanda çalışan kişilerin doğru amaca hizmet ettiği sürece her endüstriyel kesime doğrudan faydası olacağı kesindir. Bu alanda çalışmanın verdiği haz ve mutluluk, karşılığında kazanılan kazanımların yanında çok daha değerli olacağı da bir gerçektir.

Bitirme Projesi 2 dersi kapsamında veri bilimi alt alanlarından olan metin madenciliği konusunda çalışmalar yürütülecektir. Metin madenciliğinin de alt konularından olan metin özetleme (document summarization), metin konu çıkarımı (entitiy extraction) ve yerine göre metin sınıflandırması (classfication) konuları üzerinde durulacaktır. Bu çerçevede proje rapor ve dökümanlarının anlaşılması için bazı ön bilgilerin bilinmesi şartı sağlanmalıdır. Makine öğrenmesi, yapay zeka, veri madenciliği ve orta düzeyde python programlama alanlarında ön bilgi sahibi olmak gerekir.

Bitirme Projesi 2 kapsamında veri olarak kullanılan metinlerin, sahiplik belirten ve belki de izinsiz kullanımı durumlarında sorun yaratabilecek koşullarını tenzih ederek çalışmalar yapılacaktır. Projenin kaynakça ve referans kısımlarında bu kişilerin hakkının ihlal edildiği durumlarda, bunun eğitim amaçlı olduğunu unutmamaları en büyük ricadır. Maddi bir çıkar, kötüleyen veya karalayan herhangi

bir durum içerisinde kalmayacakları kesindir. İhlal durumlarında gerekli aksiyonların alınması halinde, projenin dökümanlarında bu verileri istedikleri gibi koruyabilecekleri gerçeğinin unutulmaması gerekmektedir. Makale, tez ve diğer tüm çalışmaların incelenip, kaynakça olarak gösterilmemesi (kasti olmayan, unutma durumlarından bahsedilmektedir) durumunda intihal olarak algılanmaması en büyük ricadır. Hem bir öğrencinin hem de bu alanın gelişmekte olduğu unutulmamalı ve istenmeyen durumlarda eser sahibinin anlayışlı olması beklenmektedir.

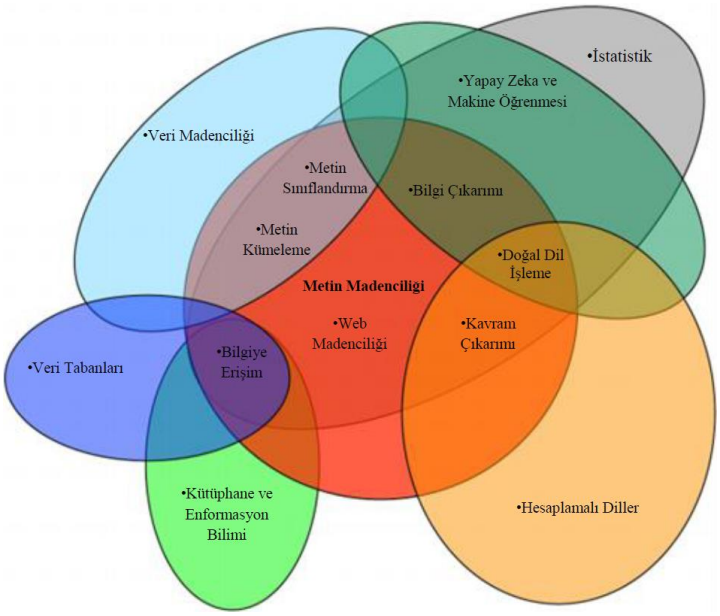
Bu alana gönül vermiş, meslektaş sayılabilecek tüm kişiler için motivasyonunun ve çalışma azminin en üst düzeyde olmasını canı gönülden diliyorum. Veri ile kal, verinin içinde ol, büyü, geliş, ama asla anlamsız olma!

ÖZET

Metin madenciliğine giriş yapılıp, temel kavramlar ve bu kavramların ne ifade ettiği öğrenilmiştir. Metin madenciliğine yardımcı, bilişim dünyasında sıkça kullanılan alanların ve yöntemlerin etkisi öğrenilip anlatılmıştır. Tübitak destekli Metin Madenciliği Yazılımları A.Ş.'nin "Metin Madencisi" projesi incelenmiştir.

METİN MADENCİLİĞİ ANLAMI VE UYGULAMALARI

Geçen Veri madenciliği alt alanı gibi gözüken fakat anlambilim ve çıkarımların bazı noktalarda semantik olarak incelendiği metin madenciliği özel bir noktaya değinmektedir. Diğer veri bilimi alanları ile ilişkisi küme olarak özet bir biçimde gösterilmiştir [1].



Şekil 1

Şekil 1'de görüldüğü gibi metin madenciliğini baz alındığı durumda yapay zeka, makine öğrenmesi, veri madenciliği, doğal dil işleme ve istatistik bilimini kapsadığı görülmektedir. Bu çerçevede Şekil 1'de görüldüğü gibi istatistik biliminin etkisi yüksektir.

Bu kapsamda Türkiye'de destek gören ve bu işle uğraşanların çıkardığı "Metin Madencisi" projesinde Türkçe diline ait kelime ve anlamların kapsandığı bir çalışma mevcuttur. Bu çalışma Türkçe dilinde yapılan kelime ve terim sayısı en yüksek projelerden biridir. Metin içeriği tespit etme, paragrafları cümleye, cümleleri kelimelere ayırabilme ve her türlü metnin incelenmesi gibi özellikleri vardır. Metnin başka metinlere benzerlikleri bulunabilmektedir. Bu benzerliklerden yola çıkarak telif hakları konusunda bu proje devletin farklı mekanizmalarında da kullanılmaktadır.

İstatistik Biliminin Metin Madenciliğine Etkisi

İstatistik bilimi metin madenciliğinde en sık kullanılan kısımları için söz edecek olursak; elde edilen sonuçların ne kadar etkili olduğunu anlamaya yardımcı olur. Metrik ve kriterlerin tamamı istatistik biliminden yararlanarak metin madenciliği için hesaplanır. Metin madenciliği özel uygulaması olmamakla birlikte en yaygın kullanımlarından Python programlamada bazı kütüphanelerin içerisinde metriklerin hazır fonksiyonlarının bulunduğunu söylemek mümkündür.

Code 1

```
>>> from sklearn.metrics import accuracy_score
>>> y_pred = [0, 2, 1, 3]
>>> y_true = [0, 1, 2, 3]
>>> accuracy_score(y_true, y_pred)
0.5
>>> accuracy_score(y_true, y_pred, normalize = False)
2
```

Code 1'de makine öğrenmesi uygulamaları için geliştirilmiş sklearn kütüphanesinde accuracy (doğruluk) metriğinin birkaç satır kodla nasıl hesaplanabildiği görülmektedir. Bu hazır kodun arkasında istatistik biliminin varlığından söz edilmelidir. Aynı şekilde istatistik bilimi kullanılarak F score, precision, recall gibi birçok farklı metrik çeşitleri bu tarz kütüphaneler ile hesaplanabilir. Dolayısıyla istatistiğin önemi metin madenciliği uygulamalarının çıktılarında başarının ne düzeyde olduğunu göstermekle birlikte, büyük veri üzerinde yapılan çalışmalarda elle hesaplama yeteneğinin azalacağı düşünüldüğünde hayat kurtarıcı olabilmektedir. Ayrıca istatistik biliminin programlama kabiliyetine büyük ölçüde ihtiyaç duymadığı da bir gerçektir. Örneğin hazır kütüphanelerde var olmayan farklı bir metrik ile ölçüm yapılacağı zaman, bu metriğin hesabının istatistiksel temellerini bilmek yeterli olacaktır. Bilindiği takdirde, Python veya R dili üzerinde kullanıma özel fonksiyonun yazılması ve hatta grafik üzerinde görselleştirilmesi söz konusudur.

İstatistik bilimi metriklerin dışında, metin madenciliğinde kullanılan, Zipfs kurallarına göre dijital ortamlara taşınan kelime frekans dağılımı (word frequecy distribution) yöntemlerinde de kullanılabilmektedir [2]. Bu yöntemde kelime

sayısından, kelimelerin uzunluklarına kadar istatistiksel hesaplamalar yapılabilmektedir.

Makine Öğrenmesi ve Yapay Zekanın Metin Madenciliğine Etkisi

Makine öğrenmesi algoritmalarının uygulamaları sektörün her alanına yayıldığı gibi, metin madenciliğinde de büyük bir rol üstlenir. Metin madenciliği ile ortaya koyulan algoritmaların, yöntemlerin, çıktıların denenmesi için gerekli olabilmektedir. Metin madenciliği ile ortaya çıkan matris ayrışımlı yapıların makine öğrenmesinde sistemin yeri geldiğinde nasıl karar verdiğini gözlemlemek gerekebilir. Bu sayede makine öğrenmesinin sunduğu regresyonel yapısallaşma, sınıflandırmalar, kümelemeler ve denetimsiz öğrenme algoritmalarını kullanmak avantaj olabilir.

Yapay zekalı uygulamalar ise, rastgeleliğin yönetilmesi kısımlarında işe yarar etki gösterebilir. Doğru yapay zeka modeli seçilerek evrimsel algoritmaların avantajları kullanılabilir [3]. Yapay zeka optimizasyon algoritmalarını barındıran ve veri konusunda uçtan uca karar verme mekanizması sunan bir alandır.

Veri Madenciliğinin Metin Madenciliğine Etkisi

Veriler kimi zaman çok anlamsız olabilmektedir. Bu verilerin süreçlerden geçirilerek uygun bir formata dönüştürmek gerekebilir. Eldeki veriye göre bir temizlik yapılması elzem olabilmektedir. Verilerin işlenmeden evvel ya da işlenirken belli başlı aşamaları takiben, daha anlaşılır duruma gelmesi için veri madenciliği yöntemleri kullanılabilir.

Örneğin CİMER'e kullanıcıların yazdığı yazıların özünü anlamak üzerine bir uygulama yapılacak olsun. Uygulamanın ne şekilde yapılacağından önce gelen verinin (kullanıcıların yazdığı yazıların) Türkçe dili kullanıldığı bilindiği takdirde anlaşılması gerekmektedir. Anlaşılması için kişilerin kendini metin içerisinde tanıttığı kısımların atılması (veriyi indirgeme aşamasında) gerekebilir. Hali hazırda CİMER sistemine yazan kullanıcıların kimlik bilgileri çıktığı için gereksiz bir öznitelikten kurtulma şansı olabilir. Daha sonra son dönemde sıkça kullanıcıların kullandığı bazı kelime dizgileri belirlenebilir. Bu kelimeler: alt yapı, elektrik kesintisi, su kesintisi, yol çalışması gibi

kelime dizgileri seçilip filtrelemeden geçip otomatik bir cevap mesajı hazırlanabilir. Fakat bu projede veri indirgeme ve veri seçme yapılmadığı takdirde, milyonlarca kullanıcının paragraflarca yazısını filtrelemeye çalışmak ciddi bir zaman kaybına yol açacaktır. Günlük 100.000 kelimeyi algoritmanın çalıştırmasında ciddi performans ve verim kayıpları yaşanırken, veri seçme yöntemlerinin sonucu 20.000 kelimeyi algoritmanın çalıştırması daha performanslı ve verimli olabilir. Ayrıca anahtar kelimelerin seçilmesinde de veri indirgeme yöntemlerinin doğru kullanımıyla sistemi yönetmenin başrol oyuncusu olabilir.

REFERANS VE KAYNAKÇA

[1] Metin Madenciliği üzerine Türkçe kaynak. Link için tıklayınız. – Ayrıca Şekil 1'in alıntılandığı yer.

[2] Li, Wentian. "Random texts exhibit Zipf's-law-like word frequency distribution." IEEE Transactions on information theory 38.6 (1992): 1842-1845.

[3] Chen, Hsinchun, et al. "Knowledge management, data mining, and text mining in medical informatics." Medical Informatics. Springer, Boston, MA, 2005. 3-33.

EKLER

Bitirme Projesi 2'ye ait doküman, haftalık rapor ve ek bilgilerin paylaşıldığı github linki için tıklayınız.