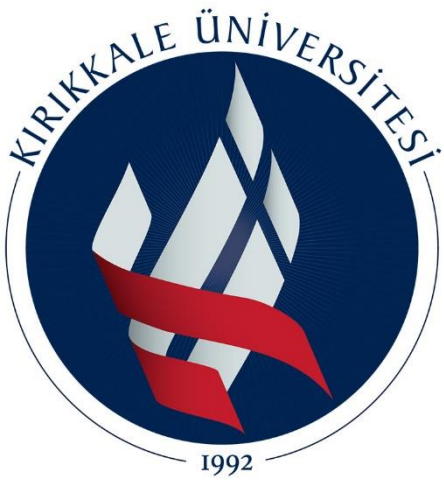

BİTİRME PROJESİ

Haftalık Rapor – 11.03.2022

11 MART 2022

KIRIKKALE ÜNİVERSİTESİ – BİLGİSAYAR MÜHENDİSLİĞİ İÖ

AHMET MUNGAN – 160255081

İÇİNDEKİLER

ÖZET

Metin madenciliğine yönelik piyasada bulunan veri setlerinin genel itibariyle ne biçimde olduğu, nasıl düzenlenebileceği, düzenlenme süreçlerinin nasıl yönetilmesi gerektiği ve düzenlendikten sonra yapılan analiz ve çalışmalara etkisi araştırılmıştır. Bu araştırmalar çerçevesinde bir örnek gerçekleştirilip, bu örneğin gerçeğe yakın bir veri seti ile uygulamalı olarak gösterilmiştir. Elde edilen ufak çaplı sonuçlar rapor kapsamında paylaşılmıştır.

METİN MADENCİLİĞİNDE VERİ SETLERİ

Veri madenciliği sürecine benzer fakat metin olmasından kaynaklı birtakım değişikliklere sebebiyet veren süreçler metin madenciliğinde bulunmaktadır. Sürecin en önemli kısımları elbette veri madenciliği ile bildiğimiz tekniklerin ve verinin işlemeye hazırlanmasından değerlendirilmesine kadar geçen süreçtir.

Bu süreçler başta verinin elde edilmesi ile ilgilidir. Metin verisi halihazırda elde edildiğinde piyasada ve hazır veri setlerinde görüleceği üzere yapılandırılmamış bir biçimdedir. Örnek olması açısından; genellikle makine öğrenmesi için veri setini ücretsiz sağlayan kaggle, metin verilerinin büyük bir kısmında gerçek dünyada verinin görüldüğü şekliyle kullanıcılar tarafından paylaşılmış veri setleri mevcuttur [1]. Bu veri seti yapılandırılmamış olarak kaggle platformunda yerini almıştır.

	train_id	name	item_condition_id	category_name	brand_name	price	shipping	item_description
0	0	MLB Cincinnati Reds T Shirt Size XL	3	Men/Tops/T-shirts	NaN	10.0	1	No description yet
1	1	Razer BlackWidow Chroma Keyboard	3	Electronics/Computers & Tablets/Components & P...	Razer	52.0	0	This keyboard is in great condition and works ...
2	2	AVA-VIV Blouse	1	Women/Tops & Blouses/Blouse	Target	10.0	1	Adorable top with a hint of lace and a key hol...
3	3	Leather Horse Statues	1	Home/Home Décor/Home Décor Accents	NaN	35.0	1	New with tags. Leather horses. Retail for [rm]...
4	4	24K GOLD plated rose	1	Women/Jewelry/Necklaces	NaN	44.0	0	Complete with certificate of authenticity

Tablo 1

Tablo 1'de görüldüğü üzere konu modelleme (topic modelling) üzerine olan örnek verilerin bazı kısımları NaN olarak, bazı kısımlarında ise açıklamaların bulunmadığına dair ibareler yer almaktadır. Tablo 1'de verinin etiketlenmiş olduğunu price özniteliğinin 1 ve 0'lardan oluştuğu görülse de, çok fazla hatalı veri olması sebebiyle yanıltıcıdır. Bu sebeple anlamlandırılması gereken bir veri seti olduğu açık bir şekilde Tablo 1'deki anlamsız kısımlardan anlaşılmaktadır.

Yapılandırılmamış veriden günümüzde geliştiriciler kaçınsa da, üzerinde çalışmalar yapanların sayısı da fazladır. Yapılandırılmamış ya da yarı-yapılandırılmış verinin kesinlikle ön işleme süreçlerinden geçirilmesi gerekmektedir. Tablo 1'deki veri setinde anlam bütünlüğünü bozan örneğin train_id=0 ile referanslandırılmış vektör veri setinden çıkarılabilir veya anlam taşıyorsa vektörün bazı öznitelikleri

yeniden değerlendirilebilir. Bu ve bunun gibi işlemleri arttırarak daha anlamlı ve yapılandırılmış veri elde etmek gerekir. Bu süreç veri madenciliği açısından değerlendirilmeli, yöntem ve teknikleri bu çerçeveden uygulamak çözümleyici olabilir. Ayrıca veri setinin paylaşıldığı platformda bu veri setinin görselleştirilmesi yapılmış, doğru ve tutarlı olmasa da sonuç alınmaya çalışılmıştır. Bu çabanın, madencilik geçirmeyen yapılandırılmamış veri setinde aykırı değer analizleriyle sınıfta kalacağı ortadadır.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	UserName,ScreenName,Location,TweetAt,OriginalTweet,Sentiment															
2	1,44953,NYC,02-03-2020,"TRENDING: New Yorkers encounter empty supermarket shelves (pictured, Wegmans in Brooklyn), sold-out online grocers (FoodKick, Max															
3	2,44954,"Seattle, WA",02-03-2020,"When I couldn't find hand sanitizer at Fred Meyer, I turned to #Amazon. But $114.97 for a 2 pack of Purell??!!Check out how #coron															
4	3,44955,,02-03-2020,Find out how you can protect yourself and loved ones from #coronavirus. ?,Extremely Positive															
5	4,44956,Cl medical supplies after #healthcare worker in her 30s becomes #BigApple 1st confirmed #coronavirus patient OR a #Bloomberg staged event?															
6																

Tablo 2

Tablo 2'de ise covid-19 üzerine atılmış tweet'lerin yapılandırılmamış veri setinden 4 kullanıcılı bir örneği mevcuttur [2]. Daha evvel bahsedilen ön işlemenin yapıldığı takdirde, tek satırdan oluşan bu veri yığınının daha düzenli hale getirilmesi gerekir. Ayrıca indirgemeler yapılarak "ScreenName" ve "UserName" olarak geçen özniteliklerin bu veri setinde faydasız olacağından indirgeme işlemlerinde atılması durumları söz konusu olabilir.

	A	B	C	D
1	Location	TweetAt	OriginalTweet	Sentiment
2	NYC	2.03.2020	"TRENDING: New Yorkers encounter en	Extremely Negative
3	Seattle, WA	2.03.2020	"When I couldn't find hand sanitizer at	Positive
4				

Tablo 3

Tablo 2'de görülen yapılandırılmamış ve anlamsız veri setini; basit bir ön işleme ve ufak çaplı bir indirgeme sürecinden geçirerek Tablo 3'de ilk 2 kayıt elde edilebilmiştir. Burada öncelikle öznitelikleri tek bir hücreye yazmadan, ayırarak ve daha anlaşılabilir bir hale getirilmesi sağlanmıştır. Daha sonra asıl sınıflandırma niteliği taşıyan, bu veri setinden çıkarımda bulunmada başrol oynayan sentiment (duygusallık) özniteliği kesin çizgiler ile veri içeriğinden ayrılmıştır. Ayrıca Tablo 2'de görüldüğü üzere kişilere özgü spesifik veriler (id gibi kişileri ayırt edici öznitelikler) veri setinden atılmıştır. Özniteliklerin belirlenmesi için, her bir öznitelik hücresi sarı arka plan renginde boyanmıştır ve bu sayede veri setinin okunabilirliği de artmıştır. "Seattle'da yaşayan insanların attığı tweet dizgilerine bakılarak, covid-19 ile ilgili mutsuz oldukları ortaya koyulabilir." şeklinde veri setinden duygu analizi ve

metin madenciliği ile örnek çıkarımlarda bulunulabilir. Burada veri setinin iyi olmasından kaynaklı olarak herhangi bir kişi hedef gösterilmeden ya da bir kişinin açık adresi belirtilmeden, hukuki açıdan yasaları çiğnemeden bir analiz ve bilgi keşfi süreci gerçekleştirilebilir. Veri setinin düzgün, anlaşılır, işlenebilir, okunabilir ve sade olmasından yola çıkarak hiçbir kişisel hakkı ihlaline gerek kalmadan analiz yapılabilmesi mümkündür. Bu da veri setinin sadece makine öğrenmesi veya derin öğrenme algoritmalarının iyi çalışması için değil, aynı zamanda yapılan analizin anlamlı ve yasal çerçevede olmasını sağladığı yadsınamaz bir gerçektir.

REFERANS VE KAYNAKÇA

[1] Kaggle'da yapılandırılmamış metin veri seti örneği. Link için tıklayınız.

[2] Kaggle'da covid-19 üzerine atılan tweet'lerin veri seti örneği. Link için tıklayınız.

EKLER

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