BİTİRME PROJESİ

SONUÇ RAPORU

9 HAZİRAN 2022

KIRIKKALE ÜNİVERSİTESİ – BİLGİSAYAR MÜHENDİSLİĞİ İÖ

AHMET MUNGAN – 160255081

İÇİNDEKİLER

GİRİŞ

Günümüz dünyasının gelişimi ve veri üreten insan toplulukları yeni iş kolları yaratmıştır. Veriyi doğru bir şekilde kullanmak, veriye yönelik aksiyonlar almak artık gereklilik düzeyindedir. Eğer bir toplum veri üretmiyorsa, ditijal ortamda sadece tüketici olarak nitelendirilebilir. Veri üreten toplumların en sık karşılaştığı problemlerden biri ise, verilerin depolandığı ortamlarda verilerin nasıl kullanıldığı sorunudur. Bu sorunun çözümü için veri depolayan birimlerin, ne gibi şartlarda verileri nasıl kullanacağı için kişisel verilerin kullanım kanunları oluşturulmuştur. Bu kanunlar uluslararası düzeyde tek düze gibi gözükse de, arka kapılarda neler yapıldığı biraz da olsa etik kurallara bağlıdır. Etik kurallar mühendisler birliklerinin belirledikleri kurallar olabileceği gibi, bu veriler üzerinde çalışanların da veri sahiplerine ne gibi zararlar verebileceğini düşündüğünde ortaya çıkan kurallardır.

Verilerin bu kadar önemli olduğu bir dönemde, kişiden ve etik değerlerden bağımsız kullanmak esastır. Aksi halde kalıcı depolama alanlarında sayıdan başka bir şeyi ifade etmeyen, anlamsız, gereksiz bir döküntü haline gelir. Veriler yerine göre zamana bağımlı, anlık ve geri dönülmez formatta olabilmektedir. Dolayısıyla elden kayıp gitmeden evvel verilerin bir an önce işe yarar örüntüsel ya da ilişkisel hale çevrilmesi hayati önem taşımaktadır.

Proje gereği dünyada en sık karşılaşılan, düzenlemesi en zor olan, depolaması en maliyetli olan ve dilden dile bağımlılık gösteren metin verisi kullanılmıştır. Proje kapsamında tüm bu zorlukların üstesinden gelinmiştir. Detaylı bir araştırma, teoride bilinenlerin pratikte uygulanması, Türkçe diline bağımlı bazı özel uygulamalar, makine öğrenmesi uygulamaları, gerçek hayat verileri ile proje uçtan uca sıfırdan sonuca kadar yapılmıştır. Projede başta danışman Doç. Dr. Fahrettin HORASAN hocamız olmak üzere, makalesinin okunup dolaylı yoldan destek olan tüm bilim insanlarına sonsuz teşekkürler!

ÖZET

Sonuç raporu kapsamında haftalık ilerleyen projenin genel hedefleri, amaçları ve teknik detayları paylaşılmıştır. Bu kapsamda proje konusunun etkilendiği ve çevresindeki ilgili bilim dalları, proje süreci boyunca projeye yansıtılmıştır. 14 haftalık süreçte elde edilen tüm bilgiler haftalık raporlar halinde paylaşılıp, haftalık raporların sonucu olarak bu rapor paylaşılmıştır. Projenin alanında tek olması ve spesifik bir konuya değinmesi gibi projeyi özel kılan bir umum unsurlar içerikte paylaşılmıştır.

PROJE

Proje adı: Metin Madenciliği ile Restoran Özetlerinin Çıkarılması

2021 – 2022 eğitim yılı bahar yarıyılı kapsamında Bitirme Projesi 2 dersi gereği danışman hocanın yönlendirmesi ile bu proje yapılmıştır. Proje; içerisinde farklı kollar barındıran, gerçek hayat verilerinin en aktif şekilde kullanıldığı, sık kullanılmayan makine öğrenmesi yöntemlerinin kullanıldığı, yüksek oranda uygulamalı bir projedir. Uygulama kısmında istatistik, modelleme, doğal dil işlemeye benzer bir metot, program kodu yazılmıştır. Bu sayede; öğrenilen teorik ve matematiksel yöntemlerin yazılımsal ortamlara aktarılıp aktarılamadığını ölçmek amaçlanmıştır.

Proje kapsamı: Bilgisayar Mühendisliği Bitirme Projesi, Online yemek sipariş platformlarının entegre bir sistem, metin özeti çıkarımında farklı bir yol.

Teknik Tanım

Metin verileri: GetirYemek ve Yemeksepeti ortak veya tekil restoranlarına yapılan yorumlardan faydalanılmıştır. Veri seti gibi bağıl ve bağlayıcı bir kavram mevcut değildir. Olası doküman yığını içeriği: Yemek, lezzet, teşekkür etmek, afiyet olsun, hız, sıcak, soğuk, kurye, restoran, tuz ve yağ gibi içeriklerin benzeşen kısımlarıdır. Veri seti yerine bir ekol düşünülmüştür, bu ekol dinamik ve maliyeti düşük bir sistem ile desteklenmiştir. Kişilerin özel verileri projeye dahil edilmemiştir. Veri gerçek hayata dair olduğu için, etiketsiz, yarı-yapılandırılmıştır, boyutu belirsiz ve özniteliği yoktur. İlgili web sitelerinden elde edilen veriler tek bir değişken altında toplanmıştır.

Koleksiyonun oluşturulması: Veri ilgili web sitelerinden Selenium ile çekilmiştir. Burada hangi restoranın seçileceği kullanıcının isteğine bırakılmış olup, restoranın her iki web sitesinde de bulunma zorunluluğu yoktur. Elde edilen verilerin yarı-yapılandırılmış halden (web etiketlerinin veride bulunduğu hali) yapılandırılmış hale (etiketlerin, script'lerin temizlenmesi ve zamana bağlı değişimin yok sayılması ile) getirilmiştir. Veri seti sadece yorumlardan oluşmaktadır.

Ön işleme: Gerçek hayat verisinde ön işleme detaylı bir şekilde yapılması gerektiğinden; Küçük harf dönüşümü, noktalama işaretlerinin temizlenmesi,

emojilerin temizlenmesi, sayıların temizlenmesi, cümlelerin kelimelere parçalanması, durak kelimelerin silinmesi, sözlük karşılaştırmalı doğal dilde örüntüler ile lemmatization, benzersiz kelimelerin bulunması. Bu işlemler ile ön işleme özel olarak gerçekleştirilmiştir.

Metin dönüşümü: Frekansların bulunması, sıfır matrisinde dengelemeler, olasılıkların hesaplanması, global ağırlıkların entropi ağırlık ile hesaplanması, local ağırlıkların logaritma ile ağırlıklandırma hesaplanması, ayırt edicilik matrisinin elde edilmesi uygulamaları gerçekleştirilmiştir.

Veri madenciliği: Denetimsiz öğrenme algoritmalarından K-Means ve metrikleri kullanılmıştır. K-Means için elbow yönteminde distortion score function yöntemleri kullanılmıştır.

Sonuç ve Değerlendirme: Kümeleme algoritmasından elde edilen en iyi sonuçlar özet olarak çıkarılmıştır ve algoritmanın performans ölçütleri hesaplanmıştır. Silhouette skoru, silhouette kesiti ve dunn indeksi hesaplanmıştır. Ngramlar kullanılmadığı ve etiketsiz bir veri olduğu için rouge metrikleri hesaplanamamıştır.

Hedef Kitle

Başta Kırıkkale Üniversitesi Mimarlık-Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü hocaları başta olmak üzere, online yemek siparişi verilen tüm siteler için kullanılabilir olduğu için, büyük e-yemek firmaları da bu projenin hedef kitlesidir.

Amaç

Projenin amacı aşağıda maddeler halinde verilmiştir.

1. Veri bilimi alanında ve topluluğu hakkında bilgi sahibi olmak. Bu alanda yapılan temel çalışma ve yöntemleri öğrenmek.
2. Veri düzenlemek, sabit olmayan veri seti üzerinde çalışmayı öğrenmek, doğal dilin zorluklarını anlamak ve Türkçe diline özel operasyonlar gerçekleştirmek.

3. Matris işlemlerini kompleks olarak çözmek, matrisleri indirgemek, matristen çıkarımlarda bulunmak, istatistik ve olasılık bilimini veri ile birleştirmek.
4. Makine öğrenmesi yöntemlerini uygulamak ve performans ölçütlerini hesaplamak.
5. Metin özeti metriklerini öğrenmek, uçtan uca sıfırdan sonuca bir metin madenciliği projesinin zorluklarıyla uğraşmak.

LİTERATÜR

Metin madenciliği üzerine literatürde birçok farklı uygulama mevcuttur. Bu uygulamaların en büyük özelliği hazır, sabit, İngilizce, kategorik ve yapılandırılmış veri seti üzerinde yapılan uygulamalardır. Ayrıca literatürde daha çok metin sınıflandırma, yazar tespiti, duygu analizi üzerine yapılmaktadır. Literatürde genellikle TF IDF matris dönüşümü yöntemi kullanılmaktadır.

Örnekler

Literatür geniş bir skalaya sahip olduğu için literatür örnekleri bire bir proje ile örtüşmeyebilir. Dolayısıyla sadece metin madenciliği adı altında örtüşen projelerden örnek verilmiştir.

1. *Cumhurbaşkanlığı Dijital Dönüşüm Ofisi* – Türkiye'de son dönemde, halkın iletişim başkanlıklarına yazı yazabildikleri bir ofistir. Bu ofis ile birlikte tüm vatandaşlar doğrudan ilgili birimlere yazı yazabilmektedir. Bu yazıların anahtar kelimeleri çıkarılarak ilgili kuruma veya kurumlara özet şeklinde iletilmesi söz konusudur. Aksi halde ilgisi olmayan kurumlara, uzun uzun yazıların gönderilmesi durumu ortaya çıkabilmektedir. Bu durum maliyet açısından dezavantaj yaratmaması için metin madenciliği süreci kullanılmaktadır.

2. *Metin Madencisi Projesi* – Tübitak tarafından desteklenen, Türkiye'nin en büyük metin madenciliği projesi Türkçe dilinde bir çalışmayı kapsamaktadır. Ücreti ile birlikte hazır kütüphaneler, kullanıma uygun framework'ler üretmektedir. Devletin bazı teşkilatlanma birimlerinde, istihbarat mekanizmalarında, dilek istek şikayet kutularının dijitalleşmesinde kullanılmaktadır. Projeye özgün Türkçe dilinde bir doğal dil işleme algoritmasına sahip olduğu söylenmektedir.

3. *Yemeksepeti ve GetirYemek* – Türk Yapılan projeye benzer, yine yorum ve restoran bilgilerinden metin madenciliği uygulamalarını firmalar kendi içlerinde yapmaktadır. Fakat bu firmalar özellikle online destek platformlarında bot yazmakla ünlüdürler. Botları, gerçek insan ayırt etmek kimi zaman zor olsa da,

bazı kullanıcı sorunları için çok yetersiz kaldığı durumlar da oluşmaktadır. Ayrıca geri bildirim yorumları için metin madenciliği uygulamalarının kullanıldığı söylenebilir.

ARAŞTIRMA VE ÇALIŞMALAR

Proje kapsamında haftalık raporlarda adım adım yaşanan sorunlar, yazılan kodlar ve algoritmanın gelişimi belirtilmiştir. Bu ilerlemeler ile sonuca yönelik bir bütünlük oluşturulduğunda algoritmanın BigO değeri n^3'e kadar düşürebilmiştir. Literatüre bakıldığında, genellikle makalesi bile yazılan projelerin hazır kütüphanelerin arka kapılarında nasıl bir maliyet olduğu tam bir muamma yaratmaktadır. Bir mühendislik projesinin olmazsa olmazı projenin maliyetidir. Çok iyi çalışmasından ziyade maliyete önem verilmesi de önemsenmesi gereken parametrelerdendir.

Öğrenilenler

Öğrenilenlerin listesi aşağıda özet olarak verilmiştir:

- Veri bilimi topluluğu ve çalışmaları,
- Veri bilimine yönelik programlama bilgileri,
- Makale türlerinin ve türe göre nasıl bir tutum içerisine girilmesi gerektiği,
- Makalelerin incelenmesi ve uygulanması,
- Forum, blog ve sözlük gibi kaynakçaları en doğru şekilde değerlendirmek,
- Anakonda ortamlarında bağımsız çalışmak,
- Metin madenciliği ile alakalı teorik bilgiler, bilgilerin kullanım yerleri, uygulama ve sonuçları,
- Matris işlemlerinin nasıl yönetilmesi gerektiği, matrisin boyutlarının önemi,
- Analize yönelik gerçek hayat verisini yönetmek ve yöntem öğrenmek, yöntemlerin matematiği ile kanıtlanmasını sağlamak,
- Grafik, matris ve metin yorumlamak,
- Ağırlıklandırmaları normaliza etmek ve makine öğrenmesine sıfırdan öznitelik çıkarmak,

- Dilbilim ve dil bilgisi ile disiplinler arası yöntemler geliştirmek, sözel bilimlerin sayısal ve analitik sonuçlara dönüştürülmesi,
- İstatistiğe yönelik yöntemleri incelemek, performansı ölçmek.

Hedeflerin Gerçekleşme Oranı

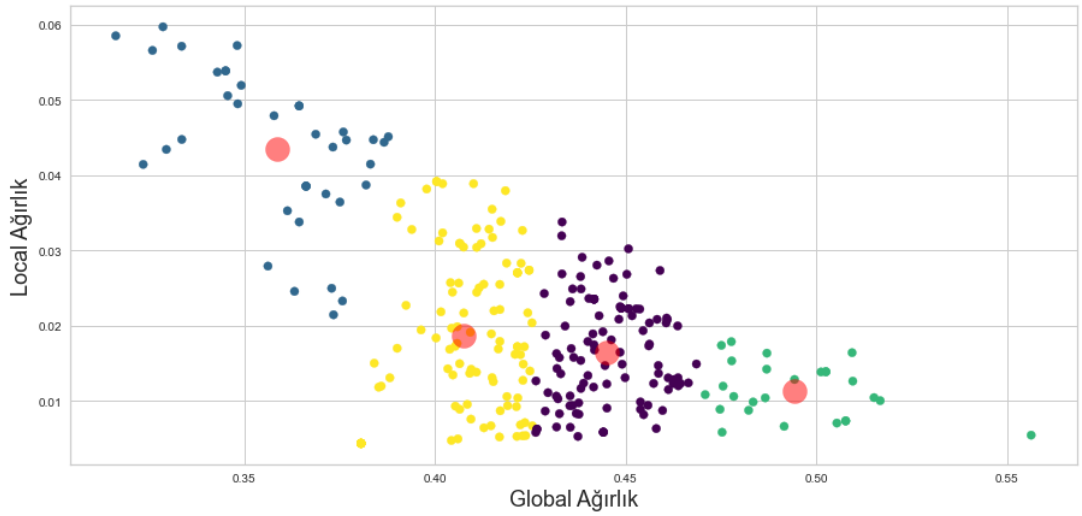
Proje konusu belirlenirken Yemeksepeti firmasının desteği ile farklı bir proje gerçekleştirilecekti. Fakat firma gerekli desteği sağlamayarak, gerçek hayata döndürülen bir proje haline gelmiştir. Dolayısıyla projenin başında bu projenin tavsiye sistemleri ile birleştirilip, firmanın sistemine doğrudan etki edecek bir proje olabileceği gündemdeydi. Projenin içeriği değişmesiyle hedef yorumların özetlerinin çıkarılması ile değiştirilmiştir. Değişen hedefe yönelik projenin gerçek hayata yönelik iyi sonuçlar verdiği söylenebilir.

Proje başlı başına sıfırdan sonuca yönelik olduğu için, performans ölçütleri yüksek başarılar göstermemektedir. Sonuçlar gerçekçi olmakla birlikte, restoranın özeti kişilere göre bile değişkenlik gösterdiği göz önünde bulundurulması gerektiği unutulmamalıdır. Dolayısıyla öznel bir durum olduğu için az da olsa makinenin yanılması bir avantajdır.

Elde edilen ayırt edicilik skorlarının makine öğrenmesine girdi olarak verilmesi hedeflenen bir durumdu. Bu hedefe yüksek oranda bağlı kalınarak, global ve local ağırlıklar normalize edilerek makine öğrenmesi algoritmalarının çıktılarının normal değerlerde çıkması hedeflenmiştir. Sonuç ve değerlendirme kısmında bu hedefin de gerçekleştiği grafik üzerinde görülmektedir.

Makine öğrenmesi algoritmasında, k değerinin seçimi çoğu projede büyük bir sorun teşkil etmektedir. Burada k değerinin minimum değeri ve maksimum değeri için en az 2 olmak şartıyla, en fazla verinin büyüklüğüne bağlı %3 olarak belirlenmiştir. Dolayısıyla her veri setinin farklı olacağı düşünülürse ve boyutunun da farklı olacağı düşünülürse dinamik bir doğruluktan söz etmek mümkündür. Ayrıca doküman sayısının az olması anlamsız sonuçlar çıkmasının da önüne geçecektir. Matematiksel ayrıntısı haftalık raporlarda bahsedilmiştir.

gerçektir. Metinsel verinin kullanıldığı her alanda, özellikle yemek için kullanılan verilerin analizinde etkili bir proje olduğu söylenebilir. Genel geçer özelliklerinin yanında, kendine has ve Türkçe diline duyarlı bir proje olduğu gözlemlenebilir. Bu projede sonucun en iyi gözlemlendiği makine öğrenmesi çıktısında kümelerin $f(x) = -kx$ düzlemine yerleşmiş hali aşağıda gözlemlenebilir.



Yukarıda dağılımların çok net olmadığı ve gerçekte ne kadar kararsız olduğumuzu da aslında simgeleyen bir çıktı görülmektedir. (4 küme örnek olarak bir restoran seçildiğinde elde edilmiştir.)

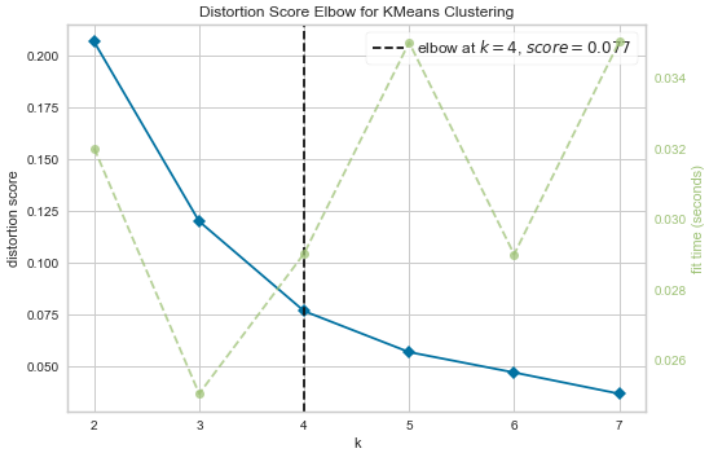
$$Silhouette_{score} = 0{,}461391$$

$$Silhouette_{samples} = \{value_{all} > 0 \quad if\ not \quad value_{[27,55,163]}\}$$

$$Dunn_{index} = \%66{,}242$$

Yukarıda K-Means algoritmasının performans çıktıları görülmektedir. Birkaç değer hariç dengeli bir dağılım olduğu ve sonuçların iyi olduğu gerçek hayat verisine göre söylenebilir. Dolayısıyla hedefler projenin başına göre yüksek oranda gerçekleşmiştir denebilir.

```
Elbow Yönteminden K Sayısı:  4
```




```
<AxesSubplot:title={'center':'Distortion Score Elbow for KMeans Clustering'}, xlabel='k', ylabel='distortion score'>
```

Yukarıda elbow yöntemiyle k sayısının nasıl elde edildiği ve en optimum sonucu makinenin nasıl bulduğuna dair bir grafik verilmiştir. Bu grafik tam sayı olarak k sayısının ne kadar doğru seçildiğini göstermektedir.

```
['çok geç teslimat sağlandı soğuk ürün para zaiyati sadece',
 'biraz soğuk geldi sipariş onun dışında güzeldi ',
 'eski tadı yok ',
 'sipariş sıcak ve hızlı geldi.lezzetide 10 üzerinden 10 ustamin ellerine sağlık.',
 ['soğuktu ve lezzetli degildi...'],
 ['çok soğuk ve kuru geldi hava soğuk anlıyorum ama bu kadarda soğuk gelmesin ustad yeme şansımız yoktu çok kurumustu teşekkür ederim....'],
 ['çok güzeldi hamurun kalınlığına güzeldi çok ince olunca güzel olmuyor']]
```

Yukarıda elde edilen özetler görülmektedir. Yorumlara bakılırsa örnek olarak seçilen restoranın gerçekten kurye sorunu yaşadığını restorandan edinilen bilgiler doğrultusunda doğrulanmıştır. Doğrulama sonucunda halen daha yakın bölgelere kurye ile değil de direkt restoran sahiplerinin paketleri hızlı bir şekilde götürmesiyle memnun olan bazı müşteri kitlelerinin de olduğunu söylemişlerdir. Restoranda yemekleri yapan ustanın değiştiği ve bazı müşterilerin eski lezzeti bulamadıklarını da belirtilmiştir. Bu edinilen bilgiler doğrultusunda, restoranın onaylamasıyla, özetlerin restorana gösterilmesiyle, doğruluk payının olduğu restoran tarafından onaylanmıştır. Bu sayede etiketsiz veriyi ancak bu şekilde doğrulamak mümkündür ve restoran bu özet için teşekkürlerini iletmiştir.

Eğer bu proje değerlendirilirse; elde hiçbir veri seti ya da bağlı bulunan bir birim olmadan, neler yapılabileceğini göstermek için bu proje idealdir. Sonuçların çok iyi olmaması elbette eksiklerinin olduğunu göstermektedir. Fakat günü kurtaran, anlık olarak kendini düzeltmek isteyen restoranlar için biçilmiş bir kaftandır. Restoranlar burada özetlerden yola çıkarak iyi yaptıklarını ve kötü yaptıklarını tartabilir, bu sayede kendilerini düzeltebilir ya da iyi yönlerini perçinleyebilir. Günümüzde her şeyin dijital

olduğu düşünülürse, en çok satışı online sipariş üzerinden yapan restoranlar buradaki yorumları göz ardı etmemelidir. Ayrıca yorum yapan kullanıcıların büyük heves taşıyarak yorum yaptığı, gerçekten bir şeyleri değişmesi için uğraşan bu kullanıcıların emekleri göz ardı edilmemelidir. Herhangi bir müşteri sadece puan verip hiçbir yorum yapmazken, önemseyen ve gerçekten kızgın ya da mutlu müşteri emek verip yorum yapıyorsa, bu yorumların özeti çıkarılarak değerlendirilmelidir. Burada atlanmaması gereken kısım ise, ipe sapa gelmez, eleştiriden ziyade saldırı niteliği barındıran yorumların bu proje ile devre dışı kalmasıdır. Dolayısıyla restoran bu projenin çıktılarını direkt bir şekilde kendi faydasına kullanabilir. Bu da, projeyi kullanan restoran için artı bir puan olarak düşünülebilir.

EKLER

Projeye ait tüm içeriklerin bulunduğu kişisel github linki için tıklayınız. (Github, güvenli bağlantı. Tüm rapor ve içeriklere erişebilirsiniz.)

REFERANS VE KAYNAKÇA

Literatürden örnek verilen firma ve kuruluşların web siteleri:

1. https://cbddo.gov.tr/
2. https://www.yemeksepeti.com/
3. https://getir.com/yemek/
4. http://www.metinmadencisi.com/