



TWİTTER VERİSİ YARDIMI İLE SOSYAL MEDYADA CİNSİYET ANALİZİ

ÖZGÜR ÖNEY

188333203

SOSYAL AĞ ANALİZİ DERSİ DÖNEM SONU PROJESİ

Temel olarak...

- ▶ Çevrimiçi sosyal ağların (OSN) kullanımı yaygın bir şekilde popüler hale gelmiştir ve bu, halka açık olan büyük ölçekli kullanıcı verilerinin üretilmesine yol açmaktadır.
- ▶ Bu veriler kullanılarak; kişiselleştirme, reklam ve kanun yürütütmeleri için kullanılabilcek kullanıcılar hakkında tahminler yapılabilir.
- ▶ Projede, Twitter kullanıcıları Twitter'daki farklı özellikler kullanılarak cinsiyetlerine göre sınıflandırılacaktır.
- ▶ Projede yüksek doğrulukla bir tahminde bulunmak için en güvenilir olduğuna inanılan özellikler el ile seçilmiştir.
- ▶ Çerçeveümüz sonuçları, yeni çıkarılan özellikler ekleyerek daha da genişletilebilir ve tavsiye sistemlerinde, pazarlamada kullanılabilir.

Yazılım ve donanım araçları

- ▶ Projenin uygulanmasında MatLab programlama dili kullanılmıştır,
- ▶ MatLab'ın İstatistik ve Makine Öğrenimi Araç Kutusu (MatLab's Statistics and Machine Learning Toolbox) ve Eğri Uydurma Araç Kutusu (Curve Fitting Toolbox) tercih edilmiştir
- ▶ Kullanıcı adına göre cinsiyet çıkarımı yapılabilmesi için ise Gender API kullanılır
- ▶ Kullanıcının tweetlerinin konularını çıkarmak için LDA algoritması kullanılacaktır
- ▶ Görüntü işlemede, cinsiyetten profil çıkarımı elde etmek için bir yüz tanıma aracı olan MatLab Toolbox kullanılacaktır
- ▶ Programlama dillerini derleme yeteneğine sahip derleyicileri ve gerekli diğer yazılımsal altyapıyı barındıran bir bilgisayar dışında farklı araçlara ihtiyaç duyulmamıştır.

Veri seti

- ▶ Projede kullanılacak olan veri seti, kaggle.com'da bulunmuştur.
 - ▶ Kaggle, Google, Inc.'in sahip olduğu çevrimiçi bir veri bilimci ve makine öğrencisi topluluğudur
- ▶ Veri seti 20.000 satırdan oluşmaktadır ve eksik değerler içermemektedir.
- ▶ Veri setinin boyutu, çalışmanın yürütülebilmesi için yeterlidir ve projenin ana motivasyonu sınıflandırma olduğundan, eksik değerlerin bulunmadığı bir veri seti çalışmanın amacına uygundur.

unit_id	name
unit_state	profile_yn_gold
trusted_judgments	profileimage
last_judgement_at	retweet_count
gender	sidebar_color
gender:confidence	tweet_created
profile_yn	user_timezone_text
profile_yn:confidence	tweet_coord
created, description	tweet_id
fav_number	tweet_count
gender_gold	tweet_location
link_color	

Veri seti(devamı)

- ▶ Veri kümesi, kullanıcılar hakkındaki gerçek cinsiyet bilgilerini içерdiğinde, sonuçta algoritma ve gerçek değerler tarafından bulunan tahminler karşılaştırılarak değerlendirilebilmektedir.
- ▶ Karmaşık ve ham verilerle çalışırken, bunların seçimi makine öğrenme algoritmasının performansını önemli ölçüde etkileyebilmektedir
- ▶ Cinsiyet tahmininde kullanılacak özellikler el ile seçilmektedir

Karşılaşılan zorluklar

- Kaggle'dan indirilen veri kümesi, tablo verilerini depolamak için kullanılan ve farklı araçlarda kullanılmak üzere başka bir dosya uzantısı şeklinde içe veya dışa aktarılabilen basit bir dosya formatı olan .csv dosya uzantısına sahiptir.
- .csv uzantısı kolay işlemek için .xlsx'e çevrilmiştir.
- Veri kümesi dikkatlice incelendiğinde, bazı satırların alt satırda kaydığı görülmüştür, çünkü kullanıcının tweet'ini gösteren sütunlar gibi çok uzun bazı sütunlar, kendi doğal hücre boyu limitlerini aşarak alt satırda sunulan veri kümesinde kalan sütunların üzerine yazılmıştır.
- Indirilen veri kümesindeki verinin kirliliği, yani farklı karakterleri kullanımı gibi problemler de görülmüştür, nitekim ASCII karakterleri MatLab'daki veri kümesinin temsilinde görünmemektedir

Karşılaşılan zorluklar(devamı)

- ▶ Karşılaşılan bir başka zorluk, üç etiketin mümkün olduğu veri kümesi cinsiyet kategorisindedir: kadın, erkek ve marka.
- ▶ Kullanıcının profil resminden cinsiyet çıkarımı kısmında ise bazı internet adreslerinde resim bulunmadığı fark edilmiş, bu durum da bazı hesapların kapalı olduğunu düşündürmüştür.

Tüm bu işlemlerin sonunda yaklaşık olarak **yüzde 35lik bir sadeleşme** elde edilerek 13000 satırdan oluşan bir veri setine indirgenme sağlanmıştır.

Kullanılan yöntemler

- Yapılan araştırma ve çalışma süresince, sağlıklı ve kullanılabilir çıkarımlar oluşturmak için ham verilere birçok yöntem uygulanmıştır. Her şeyden önce, bir kullanıcının erkek veya kadın olduğuna dair öngörüde bulunurken altı niteliği takip etmek gerekmektedir:
 - ❖ Kullanıcı adı
 - ❖ Kullanıcının profilinin kenar çubuğu rengi
 - ❖ Kullanıcının profilinin rengi
 - ❖ Kullanıcının yazdığı tweetlerin sayısı
 - ❖ Kullanıcının yazdığı tweetlerin konusu
 - ❖ Kullanıcının profil resmi

Yürütmeye planı

- Veri kümesi, biri eğitim için diğer test için eşit büyüklükte iki alt gruba bölünecektir.
- Amaç mevcut cinsiyet tahmini yöntemlerinin doğruluk seviyesine ulaşmak veya en azından mevcut yöntem doğruluğunu yüzde 20'sinden daha az olmayan bir doğruluk seviyesine ulaşmak olacaktır.
- Profil görüntülerini kullanarak cinsiyeti tahmin etmek için "COSFIRE_Gender_recognition_1_0" isminde bir araç kutusu kullanılmaktadır.
- Hesap adları veri kümesinde bulunduğuundan ve bu adları kullanarak faydalı bir bilgi alınamadığından, kullanıcı profili sayfasından HTML kodunu alan ve bu HTML kodunu işleyen Java programı kullanılmış olup kişilerin isimleri HTML kodundan tarama yöntemi ile elde edilmiş ve .txt uzantılı bir dosyaya yazılmıştır.

Kaggle Dataset

.csv format dataset is taken from www.kaggle.com

That dataset is raw, which simply includes mixed rows and non-ascii inprocessable entries.

Cleaning Data

A Java application that arranges row order, fixes mixed and separated rows and loops for removing non-ascii characters has been created.

Java is **optional**, made it for the sake of easiness but Python possible way faster and easy-to-implement.

Purifying Data

Columns, indicating features that will not be used in the training and testing, are removed. Just assign all cells to **empty**.

Put this version of data to matrix called **raw**

Profile Pic Usage

Use a toolbox, that is trained by some part of (I used half) dataset and some images.

Now got 3 output, the one named result keeps the success of the algorithm.

Get the error distribution from that struct. Put this to the regression matrix

Link and side color

Put the numbers in the preferences matrix

Get Hex color info selected half of users, convert it the data. to CIE-Lab color space to find meaningful distances of preferences

Get the error distribution from that struct. Put this to the regression matrix

Regression

Perform multinomial logistic regression by using randomly

Get Hex color info selected half of users, convert it the data.

Remaining half is tested.

Sonuçlar

- ▶ 90%'a yaklaşan isabet oranı ile,
 - ❖ Sadece profil resmi kullanımı ile yapılan çalışmalardan 12%
 - ❖ Sadece kenar çubuğu kullanımı ile yapılan çalışmalardan 54%
 - ❖ Sadece isim kullanımı ile yapılan çalışmalardan 35%
- Daha **başarılı, etkin ve isabetli** sonuçlara erişilmiştir.
- ▶ Elde edilen grafiklere göre verinin renk faktörü değerlendirilmeye katılmadan önceki ve sonraki dağılımı da çeşitlilik göstermektedir. Sonuçlara göre, kullanıcıların seçtikleri renklerin değişkenlik göstermesi, verinin homojenliğine katkıda bulunmaktadır.