Mağaza Yorumlarından Makine Öğrenmesi ve Regresyon Modellerinin Oluşturulması ve Doğruluk Oranlarının Karşılaştırılması

İbrahim AYDIN1

1Yazılım Mühendisliği

İzmir Kâtip Çelebi Üniversitesi

ibraydin@gmail.com

Özet

*Bu çalışmada MediaMarkt mağazasının ürünlerine dair 11429 yorumun bulunduğu bir veri seti kullanılarak, metin madenciliğinin en çok kullanılan yöntemlerinden olan metin sınıflandırma işlemlerinin ön işleme adımları sırasıyla gerçekleştirilmiştir. Çalışmanın ikinci bölümünde makine* *öğrenmesi modelleri (Logistic Regression, Naive Bayes, Rondom Forest, XGBoost) oluşturulmuştur**. Oluşturulan her bir modelin Count Vectors, Word Level TF-IDF, N-Gram TF-IDF ve Charlevel doğruluk oranları hesaplanmış ve karşılaştırılmıştır.*

*Anahtar kelimeler: Metin Madenciliği; Yapay Zeka, Metin Sınıflandırma, Makine Öğrenmesi*

Abstract

*In this study, the preprocessing steps of text classification, which is one of the most used methods of text mining, were carried out sequentially by using a data set with more than 11000 comments on the products of the MediaMarkt store. In the second part of the study, machine learning models (Logistic Regression, Naive Bayes, Rondom Forest, XGBoost) were created. Count Vectors, Word Level TF-IDF, N-Gram TF-IDF and Charlevel accuracy rates of each created model were calculated and compared.*

*Keywords: Text Mining; Artificial Intelligence, Text Classification, Machine Learning*

# Giriş

Sosyal medya, forumlar, müşteri yorumlarını içeren sitelerin ortaya çıkıp yaygın olarak kullanılmasıyla birlikte, bu platformlarda araştırma yapan işletmeler müşterileri davranışları ile ilgili daha büyük bir orandaki veriye ulaşabilmektedirler. Bu tür işletmelerin sahip oldukları metin verisi sayısal veriye oranla daha fazladır. Karmaşık ve düzensiz bir halde elde edilmiş olan bu tür veri çeşidi, işletmelerin iş potansiyellerini artırmalarında etkin bir bilgi kaynağıdır. Bu nedenle doğal dil işleme ya da diğer bir ifadeyle metin madenciliği veri bilimciler tarafından gittikçe artan oranda kullanılmaktadır. Dolayısıyla, bir veri bilimcisinin sahip olması gereken yeteneklerden birisi etkin bir metin verisi analizi gerçekleştirebilmektir.

Metin madenciliği, dijital ortamda yapılandırılmamış halde kayıtlı olan doğal dil metinlerini kaynak veri olarak ele alan bir veri madenciliği yöntemidir. Metin madenciliğinde her kayıt bir belgeye karşılık gelmektedir. Yapılandırılmış veritabanları yerine, doğal dil metinlerinden faydalı bilgiyi ortaya çıkarması nedeniyle veri madenciliğinden farklılık gösterir (Hearst, 1999). Bu özelliği nedeniyle, doğal dil işleme yöntemlerinden biri olarak adlandırılan metin madenciliği bilgisayarların insan dilini anlamasına yardımcı olmaktadır. Yapılandırılmamış ham veriden, faydalı bilgi sağlanmasına yarayan çeşitli yöntemler içerir. Bu yöntemler karmaşık bir metin veri setinin makine öğrenmesinde kullanılabilen yapılandırılmış bir forma dönüştürülmesini sağlar.

Her ne kadar veri madenciliğinden farklı olsa da metin madenciliği sürecinde yapılandırılmış bir metin veri setinin değerlendirilmesi aşamasında, sözcük frekans dağılımı, duygu analizi (sentiment analysis), sözcük bulutu (word cloud) gibi çeşitli metin madenciliği yöntemlerinin yanısıra veri madenciliği modellerine de başvurulur. Bu doğrultuda metin verisi üzerinde, başlıca veri madenciliği yöntemlerinden biri olan sınıflandırma yöntemiyle ilgili çeşitli metin sınıflandırma algoritmaları ile tahminler gerçekleştirilir ve ayrıca metin verisi belgeleri arasındaki benzerlikleri ve ilişkileri ortaya koymak amacıyla kümeleme algoritmalarına başvurulur (Çalış, Gazdağı ve Yıldız, 2013).

Yapılandırılmamış bir metin verisi ön işlemler olarak da isimlendirilen çeşitli işlemler uygulanarak, sütunlarda metnin içerdiği sözcükler ve satırlarda metin verisinin elde edildiği kaynak (belge) isimleri yer alan bir veritabanı biçimine dönüştürülür. Elde edilen bu veritabanı üzerinde sınıflandırma yapmak amacıyla çeşitli veri madenciliği yöntemleri uygulanarak hedeflenen bilgi elde edilir.

Metin bir şekilde veri evrenini genişlettiği için yapılandırılmış veri setleri çok fazla sayıda satır ve sütuna sahip olan büyük boyuta sahiptirler. Bu nedenle, metin verisi analizinde eğitim sürecinin artmasından kaçınmak için makine öğrenme algoritmalarının seçiminde dikkatli olmak gerekir. Metin verisi üzerinde en iyi sonuç alınan algoritmalar arasında genellikle Logistic Regression, Naive Bayes, Rondom Forest, XGBoost ve Derin Öğrenme algoritmaları yer almaktadır.

Bu süreç, yapılandırılmamış bir metin verisine erişerek, bu metin verisini sözcüklerden oluşan niteliklere dönüştürmek amacıyla yapılan işlemleri kapsar. Bu nitelikler daha sonra veri madenciliği yöntemleri ile tahmin yapmak ya da tanımlayıcı analiz gerçekleştirmek amacıyla kullanılacaklardır. Söz konusu işlemler aşağıdaki gibidir (Şekil-1):

Metin Verisi Üzerinde Ön İşlemler: Analiz öncesi metin verisi üzerinde aşağıdaki gibi çeşitli işlemler gerçekleştirilir:

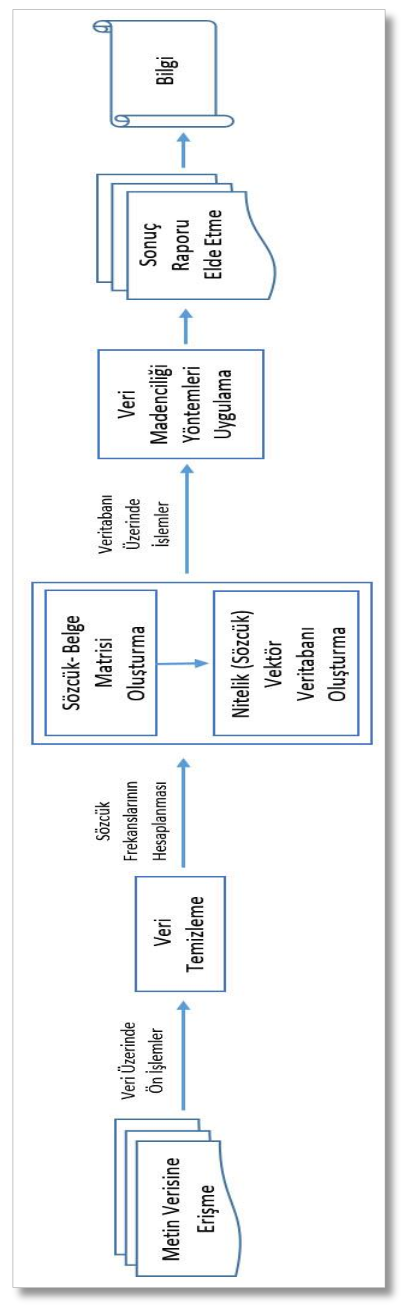
-Metin verisinin temizlenmesi: Başlıca veri temizleme işlemleri arasında; analize dâhil edilmeyen sözcüklerin, sayısal değerlerin, noktalama işaretlerinin, web adreslerinin ve gereksiz boşlukların silinmesi yer alır.

-Veri içinde yer alan sözcüklerin kök hâlinin elde edilmesi (Stemming): Metin madenciliğinde terim, Corpus ile elde edilmiş bir sözcük ya da sözcük dizisidir (Feldman ve Sanger,2007). Sözcüklerin ortaya çıkma sıklıklarının belirlenmesi aşamasında başvurulan bir işlemdir. Sözcüklerin ortaya çıkma sıklıklarının belirlenmesi aşamasında, veri setinde aynı kökten gelen ama içinde bulundukları cümle gereği farklı ek almış olan sözcüklerin aynı sözcük olarak ele alınması için köklerine çevrilmesidir. Bu amaçla, çoğul ek almış isimler ve fiil çekim eki almış fiiller, almış oldukları ekler kaldırılarak kök hâllerine çevrilirler.

-Büyük harf/küçük harf dönüşümü: Metin verisi içerisinde yer alan bütün sözcükler üzerinde büyük harften küçük harfe dönüşüm işlemleri gerçekleştirilir.

-İşaretleme (Tokenization): Metin verisinin “token” olarak isimlendirilen sözcük, sözcük grubu ya da simge halinde parçalara ayrılması işlemidir (Welbersa, Atteveldtb ve Benoit, 2017). Metin verisinin tekli sözcüklere ayrılması işlemi “unigram” olarak isimlendirilir. Benzer şekilde metin verisinin iki, üç ve daha fazla sayıda ardışık sözcük gruplarına ayrılma işlemleri sırasıyla; “bigram”, “trigram” ve “n-gram” olarak isimlendirilir.

-Sözcük-Belge Matrisi Oluşturma: Ön işlemlerin ardından “Sözcük- Belge Matrisi” oluşturulur. Bu matristen sözcüklerin sıklık değeri ve ilişkileri elde edilir (Zhao, 2015). Söz konusu matrisin her bir sütunundaki sözcük sayısı tercih edilen n-gram yöntemi doğrultusunda belirlenir. Elde edilen Sözcük-Belge Matrisi matrisinin satır sayısı, işlenmemiş metin verisinin yer aldığı kaynak veri setindeki belge sayısına eşit olacaktır.



Şekil 1: Metin Madenciliği Aşamaları

Metin veri seti üzerinde gerçekleştirilen ön işlemler ve Sözcük-Belge matrisinin oluşturulması aşamasından sonra, yapılandırılmış hâle gelen metin verisi üzerinde veri madenciliği yöntemleri uygulamak suretiyle, bir analiz modeli elde edilebilir. Bu bağlamda, metin madenciliğinde başvurulan çeşitli yöntemler aşağıda görülmektedir

* Duygu analizi (sentiment analysis)
* Makine öğrenmesi (machine learning)
  + Danışmanlı öğrenme (supervised learning)
    - Metin sınıflandırma (classification)
    - Yapay sinir ağları (neural network)
    - Derin öğrenme (deep learning)
  + Danışmansız öğrenme (unsupervised learning)
    - Metin kümeleme (clustering)
* Metinlerden konu çıkarma (concept/entity extraction)
* Metin özetleme (document summarization)
* Varlık ilişki modellemesi (entity relationship modelling)
* Sınıf taneciklerinin üretilmesi (production of granular taxonomy)

Sınıflandırma algoritmaları ile metin madenciliği yapmak üzere, aşağıdaki işlemler aşamalar halinde gerçekleştirilir:

1. Metin verisinin okunması.

2. Metin verisinin sözcüklerden oluşan niteliklere dönüştürülmesi.

3. Nitelik (sözcük) veritabanının oluşturulması.

4. Veritabanının eğitim ve test verisi olarak ikiye bölünmesi.

5. Eğitim veri seti üzerinde sınıflandırma algoritmaları ile model kurulması.

* Test veri seti ele alınarak, sınıflandırma modelleri ile elde edilen kurallar doğrultusunda tahminler yapılması.
* Doğru tahmin oranları göz önüne alınarak, algoritmaların performans değerlendirmesinin gerçekleştirilmesi.
* En iyi performansa sahip olan sınıflandırma algoritmasının seçilmesi.

6. En iyi performansa sahip olan sınıflandırma algoritmasının tahmin sonuçlarının değerlendirilmesi suretiyle, müşteri yorumları ile ilgili anlamlı bilgi elde edilmesi.

# Materyal ve Metotlar

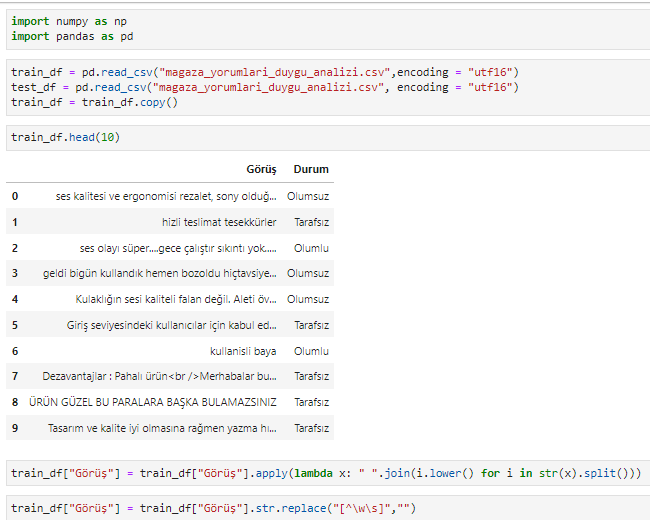
Bu çalışmada MediaMarkt mağazasının ürünlerine dair 11429 yorumun bulunduğu bir veri seti kullanılmıştır. Veri setindeki her yorum, “Olumlu”, “Olumsuz” ve “Tarafsız” olarak sınıflandırılmıştır. Veri setine aşağıdaki linkten ulaşılabilir:

<https://github.com/Gulfem93/MagazaYorumlariSementicAnalysis/blob/main/magaza_yorumlari_duygu_analizi.csv>

Bu çalışmada, Python programlama dili, Anaconda ve Jupyter Notebook web çatıları kullanılmıştır. Projede öncelikle metin ön işlemleri ve sonrasında öğrenmesi modelleri (Logistic Regression, Naive Bayes, Rondom Forest, XGBoost) oluşturulmuştur. Oluşturulan her bir modelin Count Vectors, Word Level TF-IDF, N-Gram TF-IDF ve Charlevel doğruluk oranları hesaplanmış ve karşılaştırılmıştır.

# Bulgular

Öncelikle numpy ve pandas kütüphaneleri ile veri setine ait işlemler gerçekleştirildi. Veri seti içindeki yorumlar sütunu başlığı “Görüş”, olumlu, olumsuz ve tarafsız olan metin sınıflandırma sütunu başlığı “Durum” olarak atandı. Tüm yorumlar küçük harfe çevrildi. Noktalama işaretleri temizlendi. uygulaması Jüpiter Notebook kullanılarak gerçekleştirilmiştir. Kütüphane yüklemeleri ve veri seti ile ilgili Python kodlamaları açıklamaları ile birlikte aşağıda gösterilmiştir:



Aşağıda veri setinin noktalama işaretlerinden temizlenmiş ve küçük harflere çevrilmiş durumu görülmektedir:



Natural Language Toolkit veya daha yaygın olarak NLTK, Python programlama dilinde yazılmış İngilizce için sembolik ve istatistiksel doğal dil işlemeye yönelik bir kitaplıklar ve programlar paketidir. Sınıflandırma, tokenizasyon, kök çıkarma, etiketleme, ayrıştırma ve anlamsal akıl yürütme işlevlerini destekler.

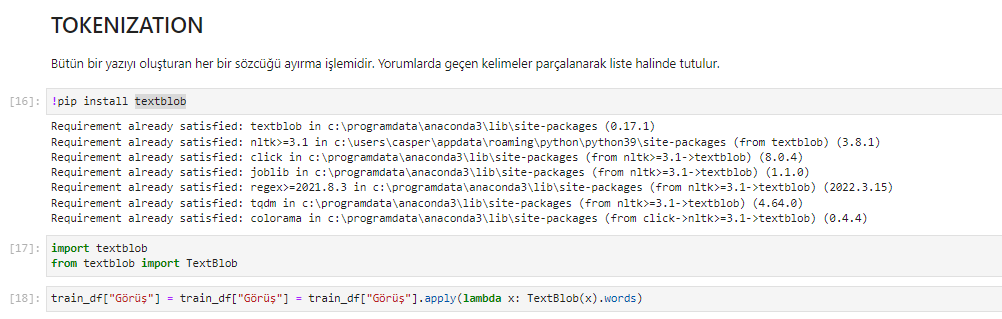
NTLK kurulumu yapıldı. Sonrasında gereksiz kelimeler görüşlerden temizlendi. Temizlenen gereksiz kelime sayısının 35941 olduğu görüldü:



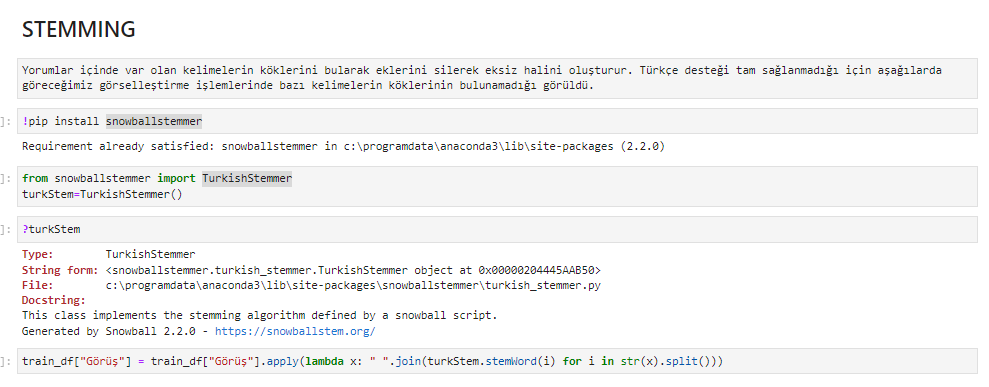
Tokenization (işaretleme) işlemi, bütün bir yazıyı oluşturan her bir sözcüğü ayırma işlemidir. Yorumlarda geçen kelimeler parçalanarak liste halinde tutulur.

TextBlob, metin verilerini işlemek için kullanılan bir Python kütüphanesidir. Duygu analizi,sınıflandırma ve çeviri gibi doğal dil işleme'de (NLP-Natural Language Processing) görevlerine dağıtmak için kullanılan basit bir API sağlar.

Tokenization işlemi için TextBlob yüklenmiştir.



Kök Bulma işlemi için snowballstemmer yüklendi ve TurkishStemmer kütüphanesi kullanıldı. Sonrasında kök alma işlemi yapıldı. Yani yorumlar içinde var olan kelimelerin köklerini bularak eklerini silerek eksiz halini oluşturuldu. Türkçe desteği tam sağlanmadığı için aşağılarda göreceğimiz görselleştirme işlemlerinde bazı kelimelerin köklerinin bulunamadığı görüldü.



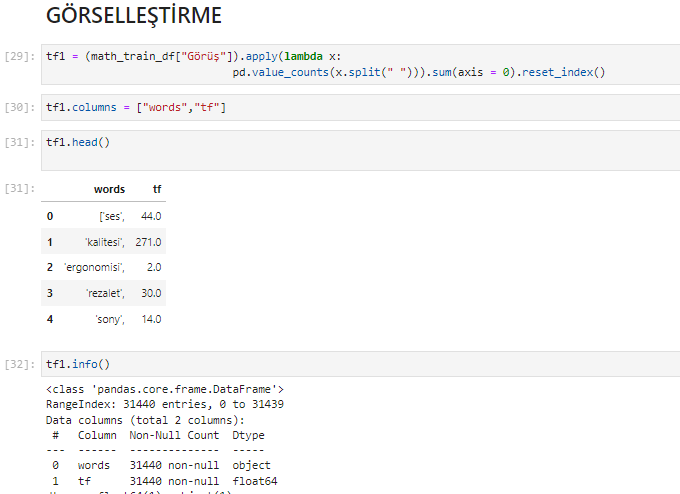
Sonra NLP uygulamasına geçildi. (ngrams=3 alındı)

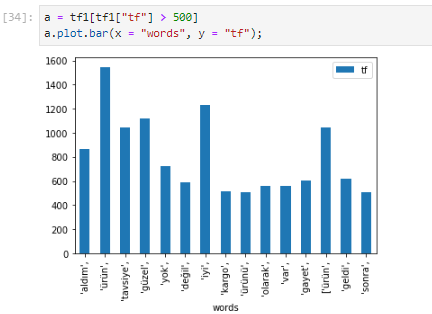


Harf ve kelime sayısını görmek için yapılan bazı matematiksel işlemlerden sonra tarafsız olan yorumlar silindi. Bu metin ön işlerimden sonra veri setinde 8491 kaydın kaldığı görüldü:



Hangi kelimeden kaç tane olduğu görselleştirildi:





Wordcloud yüklemesi yaparak kelime görselleştirme işlemi başka boyutta görüldü:



Projenin ilerleyen kısımlarında “Model Oluşturma” işlemlerine geçildi. Textblob, sklearn, pandas, keras ve warnings kütüphaneleri kullanarak Logistic Regression, Naive Bayes, Rondom Forest, XGBoost modelleri oluşturuldu.

Oluşturulan her bir modelin Count Vectors, Word Level TF-IDF, N-Gram TF-IDF ve Charlevel doğruluk oranları hesaplandı ve değerler karşılaştırıldı.



# Sonuç

Veri setine yapılan metin ön işlemleri sonrasında oluşturulan makine öğrenmesi modellerinin doğruluk karşılaştırma değer tablosu aşağıda görülmektedir:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Count Vectors** | **Word-Level TF-IDF** | **N-GRAM TF-IDF** | **CHARLEVEL** |
| **Logistic Regresion** | 0.8813 | 0.8765 | 0.7781 | 0.8605 |
| **Naive Bayes** | **0.8921** | 0.8911 | 0.7654 | 0.8431 |
| **Random Forest** | 0.8600 | 0.8586 | 0.7526 | 0.8328 |
| **XGBoost** | 0.8445 | 0.8421 | 0.6980 | 0.8469 |

Sonuç olarak **% 89,21** ile en yüksek doğruluk oranı **Naive Bayes** algoritmasında **Count Vectors** metodundan alınmıştır.

# Kaynaklar

Çalış, K., Gazdağı, O. ve Yıldız, O. (2013). Reklam İçerikli Epostaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti, Bilişim Teknolojileri Dergisi, 6(1), 1-7.

Feldman, R. ve Sanger. J. (2007). The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data, New York: Cambridge University Press.

Hearst, M. A. (1999). Untangling Text Data Mining. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics içinde, ACL ’99 (ss. 3–10). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1034678.1034679.

Welbersa, K., Atteveldtb W. V. ve Benoit, K. (2017). Text Analysis in R, Communication Methods and Measures, 11(4): 245-265.

Zhao, Y. (2015). R and Data Mining: Examples and Case Studies,https://www.webpages.uidaho.edu/~stevel/517/RDataMining-book.pdf (Son Erişim: 21.05.2023)