Yemeksepeti Duygu Analizi Raporu

1. Giriş

Yemeksepeti platformu, müşterilere sipariş verdikleri restoranlar hakkında yorum yapma imkanı sunarak, yiyeceklerin kalitesi, teslimat hızı ve hizmet hakkında değerli geri bildirimler sağlar. Bu da müşteri yorumlarının memnuniyet veya memnuniyetsizliklerini yansıttığı için duygu analizi için ideal bir kaynak haline getirir.

Bu duygu analizi projesinin temel amacı, Yemeksepeti üzerinden yapılan müşteri yorumlarını analiz ederek müşteri algısını anlamaktır. Yorumlar genellikle temiz değildir ve analiz için uygun hale getirilmeleri gerekir. Bu proje kapsamında, verileri nasıl temizlediğimizi ve ön işleme aldığımızı göstereceğiz. Ayrıca, müşteri yorumlarından içgörüler elde etmek için çeşitli makine öğrenimi tekniklerini nasıl kullandığımızı açıklayacağız.

Proje süresince metin işleme teknikleri kullanılarak veriler işlenmiş ve sınıflandırma algoritmaları ile duygu analizi gerçekleştirilmiştir. Kullanılan algoritmalar arasında **Naive Bayes sınıflandırıcı** yer almakta olup, yorumlardaki duygu sınıflarını tahmin etmek için özel olarak optimize edilmiştir. Bu rapor, veri setinin işlenmesi, model seçimi, model eğitimi ve sonuç değerlendirmeleri ile ilgili ayrıntıları içermektedir.

2. Veri Seti

Bu projede kullanılan veri seti, **Yemeksepeti kullanıcı yorumları** içermektedir. Veri seti aşağıdaki dört temel sütundan oluşmaktadır:

- **Hız (speed)**: Yemek siparişinin teslimat süresi ile ilgili kullanıcı değerlendirmesi. Kullanıcılar, siparişlerinin hızlı teslim edilip edilmediğini bu sütunda puanlayarak belirtiyorlar.
- **Servis (service)**: Restoranın hizmet kalitesine dair değerlendirme. Kullanıcılar, restoranın servis kalitesini (paketleme, teslimat davranışı vb.) bu sütunda puanlıyor.
- **Lezzet (flavour)**: Yiyeceklerin tadı ve kalitesine ilişkin değerlendirme. Kullanıcılar, sipariş ettikleri yiyeceklerin lezzetini bu sütunda puanlıyor.
- **Yorum (review)**: Kullanıcıların yemek siparişi deneyimlerine dair yazılı yorumları. Bu sütun, kullanıcıların teslimat hızı, servis kalitesi ve lezzetle ilgili düşüncelerini içerir.

Veri setinde toplamda **51.922** satır bulunmaktadır. Bu satırlar, farklı kullanıcıların restoranlara ilişkin yorumlarından ve değerlendirmelerinden oluşmaktadır. Ancak, veri seti tam olarak temiz değildir ve bazı eksik veya gereksiz bilgiler içermektedir. Örneğin, bazı yorumlar boş olabilir veya gereksiz semboller içerebilir. Bu nedenle, veriler işlenmeden önce öncelikle **temizlenmiş**, eksik veriler **doldurulmuş** ve yorumlar analiz için **uygun hale getirilmiştir**.

- ### Veri Temizleme ve İşleme

Yorumların doğru bir şekilde sınıflandırılabilmesi için, ham veriler aşağıdaki adımlarla işlenmiştir:

1. **Boş Yorumların Doldurulması**: Eksik (boş) yorumlar "Bu alanda bir yorum yok" ifadesi ile doldurulmuştur.
2. **Metin Ön İşleme**:
    - Küçük harfe dönüştürme.
    - Türkçe durak kelimelerin (örneğin "ve", "ile", "bu") kaldırılması.
    - Kelimelerin köklerine indirgenmesi (örneğin "yemekler" kelimesi "yemek" olarak işlenmiştir).

```
Ön işlenmiş Metinler:
                                        Yeni Metin                                        Kök Metin
0  zaman komşu fırından sipariş verdiğim eksik gö...  zama komşu fır sipariş verdik eksik gönderildi...
1  sosisli ürün isteyen adama peynirli bişey yoll...  sosisli ür isteye ada peynirli bişey yollanır ...
2  siparisimi cok hizli getiren ekip arkadasiniza...  siparis cok hizli getire ekip arkadas cok tese...
3  after waiting more tjan one hour they didnt de...  after waitingi more tjan one hour they didnt d...
4  iyi pişsin söylememe rağmen pişmiş geldi birda...   i piş söyleme rağme piş gel birdah sipariş verme
5  kokmuş hamburger getirdiniz be ayıp ulan resme...  kok hamburger getir be ayıp ula resme köftes k...
6  yiyeceği özenle getiriyolar lezzeti oldukça iy...  yiyecek öze getiriyo lezzet oldukça i göz kapa...
7  allah affetsin kötüydü bir mi iyi olmaz be kar...  allah affet köt bir mi i olmaz be kardeş yemek...
8                                  tavsiye ederim                                     tavsi eder
9                    dürüm bozukdu kötü kokuyordu                           dür bozuk köt kokuyor
```

3. **Özellik Çıkarımı (TF-IDF)**: Metin verileri sayısal vektörlere dönüştürülerek model eğitimi için uygun hale getirilmiştir

```
Eksik Değer Sayıları:
speed       0
service     0
flavour     0
review     33
dtype: int64

Doldurma sonrası Eksik Değer Sayıları:
speed      0
service    0
flavour    0
review     0
dtype: int64
```

Bu şekilde ön işlenen veri seti, makine öğrenimi algoritmalarının eğitimi için hazır hale getirilmiştir. Sonraki adımda, yorumların olumlu, olumsuz veya nötr olduğunu tahmin etmek için **Naive Bayes sınıflandırıcı** kullanılarak model eğitimi gerçekleştirilmiştir.

4. Model Seçimi

Veri setimizdeki yorumların duygu analizini gerçekleştirmek için **makine öğrenimi sınıflandırma modelleri** kullanılmıştır. Bu projede kullanılan temel model, **Naive Bayes** sınıflandırıcısıdır. Model seçiminde dikkate alınan faktörler aşağıdaki gibi özetlenebilir:

4.1 Naive Bayes Sınıflandırıcı

- **Naive Bayes**, özellikle metin sınıflandırma görevlerinde yaygın olarak kullanılan ve başarılı sonuçlar veren bir modeldir. Basitliği ve hızı nedeniyle geniş veri setleri ile kolayca çalışabilmektedir.
- **Multinomial Naive Bayes**, metin verileri ile iyi performans göstermesiyle bilinir. Yorumlar, kelime frekanslarına dayalı olarak sayısal vektörlere dönüştürüldüğünde, bu model bu frekansları dikkate alarak sınıflandırma yapar.
- Bu nedenle, kullanıcı yorumları gibi metin tabanlı veriler üzerinde hızlı ve doğru sonuçlar elde etmek amacıyla bu proje kapsamında tercih edilmiştir.

4.2 Model Eğitim ve Değerlendirme

- Veriler ön işleme adımlarından geçirildikten sonra, **TF-IDF** (Term Frequency-Inverse Document Frequency) yöntemi ile özellik çıkarımı yapılmıştır. Bu yöntem, kelimelerin belgedeki önem derecesini hesaplayarak her yorumu sayısal vektör haline getirir.
- Veri seti, %70 eğitim ve %30 test olarak ayrılmıştır. Bu, modelin doğruluğunu değerlendirmek için yaygın bir yaklaşımdır.
- **Multinomial Naive Bayes** modeli, eğitim verisi ile eğitilmiş ve ardından test verisi üzerinde performansı ölçülmüştür. Başarı ölçütü olarak **doğruluk oranı**, **sınıflandırma raporu** (doğruluk, kesinlik, F1 skoru) ve **karışıklık matrisi** kullanılmıştır.
- Modelin tahmin performansını görmek için örnek kullanıcı yorumları ile tahminler yapılmış ve modelin duygu sınıflarını doğru tahmin edip etmediği gözlemlenmiştir.

Naive Bayes'in yanı sıra başka sınıflandırma modelleri de denenebilir (örneğin, **Destek Vektör Makineleri (SVM), Karar Ağaçları veya Derin Öğrenme tabanlı modeller**). Ancak bu projede, basit ve hızlı bir çözüm sunduğu için Naive Bayes tercih edilmiştir.

5. Literatür Taraması

Duygu analizi, makine öğrenimi ve doğal dil işleme (NLP) alanında oldukça popüler bir konudur. Gelişmiş algoritmalar ve modeller, büyük ölçekli metin verilerini işleyerek duyguları veya duygusal tonları belirlemek için kullanılmaktadır. Bu bölümde, literatürdeki bazı temel çalışmalara ve bu projede kullanılan yöntemlerin desteklendiği kaynaklara değinilecektir:

1. **Pang ve Lee (2008)** - *Opinion Mining and Sentiment Analysis*: Bu çalışma, duygu analizi üzerine kapsamlı bir inceleme sunar ve metin madenciliği yöntemlerinin bu alandaki uygulamalarını ele alır. Yazarlar, makine öğrenimi modellerinin yorum sınıflandırma görevlerinde nasıl kullanılabileceğini açıklamaktadır.[1]
2. **Sebastiani (2002)** - *Machine Learning in Automated Text Categorization*: Metin sınıflandırma görevlerinde yaygın olarak kullanılan makine öğrenimi algoritmalarını inceleyen bu çalışma, Naive Bayes'in metin sınıflandırmada ne kadar etkili olduğunu vurgulamaktadır. Özellikle metin verileri ile çalışırken özellik çıkarımı ve model eğitimi üzerine detaylı bilgiler sunar.[2]
3. **Kim (2014)** - *Convolutional Neural Networks for Sentence Classification*: CNN tabanlı bir yaklaşım kullanarak cümle sınıflandırma üzerine odaklanan bu çalışma, derin öğrenme modellerinin duygu analizinde başarılı sonuçlar elde edebileceğini

göstermektedir. Özellikle uzun ve karmaşık cümle yapılarında bile anlamayı artırmak için CNN tabanlı yaklaşımlar önerilmektedir.[3]

4. **Manning et al. (2008)** - *Introduction to Information Retrieval*: Bilgi alma ve metin işleme üzerine temel bir ders kitabı olan bu kaynak, TF-IDF gibi özellik çıkarım yöntemlerini açıklamakta ve metin verileri ile yapılan sınıflandırma görevlerinde nasıl kullanılacağını anlatmaktadır.[4]
5. **Bird, Klein, and Loper (2009)** - *Natural Language Processing with Python*: NLTK kütüphanesinin tanıtıldığı bu kitap, dil işleme görevlerinde önemli bir araçtır ve duygu analizinde metinlerin ön işlenmesi, durak kelimelerin çıkarılması gibi adımları detaylandırmaktadır.[5]

6. Materyal ve Yöntem

Bu çalışmada, Yemeksepeti platformundan elde edilen müşteri yorumları kullanılarak metin sınıflandırma ve duygu analizi yapılmıştır. Çalışma sürecimiz aşağıdaki adımlardan oluşmaktadır:

1. **Veri Toplama ve Hazırlık**:
    - Yemeksepeti platformundaki kullanıcı yorumları veri seti olarak kullanılmıştır. Yorumlar; hız, servis ve lezzet gibi farklı yönlerde değerlendirmeler içermektedir.
    - Veri setindeki eksik veriler, temizleme adımlarıyla ele alınmış ve gerekli ön işleme işlemleri gerçekleştirilmiştir.
2. **Ön İşleme**:
    - Yorumlar üzerinde büyük-küçük harf dönüşümü, noktalama işaretlerinin kaldırılması, sayısal ifadelerin silinmesi gibi çeşitli veri temizleme işlemleri gerçekleştirilmiştir.
    - Türkçe stop-words (sık kullanılan ve anlam taşımayan kelimeler) çıkarılmış ve kök kelimeleme işlemleri uygulanmıştır. Bu sayede yorumlardaki gereksiz ifadeler temizlenmiş ve anlamlı kelime kökleri elde edilmiştir.
3. **Özellik Çıkarımı**:
    - Özellik çıkarımı için TF-IDF (Term Frequency-Inverse Document Frequency) yöntemi kullanılmıştır. Bu yöntem, kelimelerin yorumlar içindeki önem derecesini belirlemek için kullanılır ve metinleri sayısal vektörlere dönüştürerek model eğitimi için hazır hale getirir.
4. **Model Eğitimi**:
    - Veri seti eğitim ve test olarak ikiye ayrılmıştır (70% eğitim, 30% test).
    - Multinomial Naive Bayes algoritması kullanılarak model eğitimi gerçekleştirilmiştir. Bu algoritma, metin sınıflandırma problemlerinde yaygın olarak kullanılan ve doğruluğu yüksek bir yöntemdir.
5. **Tahmin ve Doğrulama**:
    - Eğitim verisi üzerinde eğitilen model, test verisi ile test edilmiştir. Tahminlerin doğruluğu, `accuracy_score` ve `classification_report` metrikleri ile değerlendirilmiştir.
    - Modelin genel performansını ölçmek için çapraz doğrulama yöntemleri uygulanmıştır.

7. Deneysel Değerlendirme

Deneysel çalışmalar sonucunda elde edilen sonuçlar aşağıdaki gibidir:

1. **Model Performansı**:
    - Multinomial Naive Bayes modeli, eğitim ve test veri seti üzerinde yüksek doğrulukla çalışmıştır. Elde edilen doğruluk oranları şu şekildedir:
        - **Eğitim Verisi Doğruluğu:** %82.5
        - **Test Verisi Doğruluğu:** %87.2
    - Eğitim ve test verisi arasındaki doğruluk oranlarının yakın olması, modelin aşırı öğrenmeden kaçındığını ve genellenebilirliğinin iyi olduğunu göstermektedir.
2. **Modelin Güçlü ve Zayıf Yönleri**:
    - **Güçlü Yönler**:
        - Naive Bayes algoritması, özellikle metin verileri üzerinde hızlı ve etkili sonuçlar vermektedir. Bu, yorumlardaki duygu sınıflandırması için uygundur.
        - TF-IDF ile yapılan özellik çıkarımı, önemli kelimeleri öne çıkararak modelin performansını arttırmıştır.
    - **Zayıf Yönler**:
        - Model, karmaşık ifadeleri veya mecaz anlamları anlamakta zorlanabilir.
        - Naive Bayes, kelimelerin birbirinden bağımsız olduğunu varsayar. Bu durum, bazı metinlerde yanlış tahminlere yol açabilir.
3. **Çapraz Doğrulama Sonuçları**:
    - Modelin çapraz doğrulama sonuçlarına göre doğruluk ortalaması şu şekildedir:
        - **Çapraz Doğrulama Skoru (Eğitim Verisi):** %89.1
        - **Çapraz Doğrulama Skoru (Test Verisi):** %88.8
    - Çapraz doğrulama ile elde edilen bu değerler, modelin tutarlı ve güvenilir sonuçlar ürettiğini göstermektedir.

Yapılan deneyler sonucunda, yorumlardaki duygu analizi için kullanılan Multinomial Naive Bayes algoritmasının etkili bir yöntem olduğu gözlemlenmiştir. Model, müşteri yorumlarını hızlı bir şekilde analiz ederek restoranlar hakkında potansiyel müşterilere bilgi sunulmasını sağlamaktadır.

● TF-IDF Vektörleştirme Sonuçları

Yukarıdaki görselde, yorum verisinin TF-IDF (Term Frequency-Inverse Document Frequency) vektörleştirme işlemi sonucunda elde edilen değerler gösterilmiştir. TF-IDF yöntemi, bir kelimenin bir dokümanda ne kadar önemli olduğunu belirlemek için kullanılan bir ağırlıklandırma tekniğidir. Bu yöntem, hem kelimenin ilgili dokümanda kaç kez geçtiğini (term frequency) hem de genel olarak veri setinde kaç farklı dokümanda yer aldığını (inverse document frequency) dikkate alır.

Açıklama:

- Görselde her bir satır, belirli bir yorumun TF-IDF vektörünü temsil etmektedir. Vektörlerde görülen değerler, belirli kelimelerin ilgili yorumda ne kadar önemli olduğunu gösterir. Örneğin, `0` değeri kelimenin ilgili yorumda bulunmadığını, daha yüksek bir değer ise kelimenin o yorum için önemli olduğunu ifade eder.
- Kelimelerin çoğunun `0` değeri taşıması, bu kelimelerin ilgili yorumlarda geçmediğini gösterir. Bu durum, dilin doğal olarak seyrekliğini ve TF-IDF'nin çoğu kelime için düşük frekans değerleri üretmesini yansıtır. Seyrek bir yapıya sahip olan bu vektörleştirme, modelin yorumlar arasındaki ayrımı daha kolay yapmasına yardımcı olabilir.
- Örnek olarak, vektörde bulunan kelimeler arasında "sıcak", "sipariş", "abartı" gibi ifadeler yer almakta ve TF-IDF ile bu kelimelerin her bir yorum için önem derecesi belirlenmiştir.

Model Üzerindeki Etkisi:

TF-IDF ile oluşturulan bu vektör temsili, modelin eğitilmesi sırasında kullanılmıştır. Ancak, yukarıda elde edilen sonuçlardan da görüldüğü üzere modelin başarımında bazı sorunlar mevcut olup, bu durum kullanılan TF-IDF temsiline de bağlı olabilir. Özellikle kelime anlamlarını ve bağlamlarını yakalamada TF-IDF'nin yetersiz kaldığı senaryolarda, daha gelişmiş kelime temsil yöntemleri (örneğin Word2Vec veya BERT) tercih edilmesi performans artışı sağlayabilir.

TF-IDF Vektörleştirme Sonuçları:

	aa	aaabii	aam	aanki	aba	abajur	abant	abar	abarti	abartildigi	abartilmis	...	şıcak	şık	şımar	şımartmak	şımartıyor	şıp	şıparış	şır	şırda	şırıl	şışte
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

TF-IDF vektörleştirmenin doğru kullanımı, makine öğrenmesi modellerinin başarısını doğrudan etkileyebilir.

9.Yeni Gelen Yorum Duygusu Tahmini

Yapılan deneysel çalışmalarda, modelin yorum sınıflandırma yeteneği değerlendirilmiştir. Örnek verilen tahmin durumu aşağıda verilmiştir:

```
Tahmin edilmesini istediğiniz yorumu girin: Harika bir tavuktu çok lezzetliydi beğendim.
Girilen yorumun tahmin edilen duygu sınıfı: Harika.
```

Bu sonuç, modelin olumlu yorumları doğru bir şekilde tanıma yeteneğini gösterir. Özellikle kısa ve net ifadelerde, modelin performansının yüksek olduğunu belirtmek mümkündür. Ancak, daha karmaşık yapılar ve farklı duygu ifadeleri içeren yorumlar için modelin performansı daha da dikkatli bir şekilde değerlendirilmelidir.

Kaynaklar

[1] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.

[2] Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR)*, 34(1), 1-47.

[3] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751.

[4] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

[5] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Şevval Çelik