Şiir Türlerinde Sınıflandırma

Hilal Yüce
Bilişim Sistemleri Mühendisliği
Kocaeli Üniversitesi
Kocaeli, Türkiye
221307070

***Öz*-Bu raporda Yazılım Geliştirme Laboratuvarı-I dersinin projesinin veri toplama ve veri ön işleme aşamaları ele alınmıştır .Çalışmada Antoloji şiir web sitesinden şiir türlerine uygun olacak şekilde, farklı birçok kategori arasından seçim yapılarak veri kazınmıştır. Veri kazıma işlemi Python dilinde Selenium, BeautifulSoup, Requests ve Pandas kütüphaneleri kullanılarak gerçekleştirilmiştir. Şiir türlerindeki sınıflandırmamız Epik Şiir, Lirik Şiir Pastorel Şiir, Satirik Şiir ve Felsefi Şiir olmak üzere 5 ana başlık altında toplanmıştır. Veri ön işleme kısmında tokenizasyon, lemmatizasyon, kök bulma ve stopwords temizleme aşamaları gerçekleştirildi Bu işlemler sonucunda verilerimiz istenilen duruma getirilmiştir.**

Anahtar Kelimeler—veri ön işleme, şiir, python, veri kazıma, kütüphane, şiir türleri.

I. GİRİŞ

Metin Verileri, genellikle bir dizi metnin sıralanması veya belirli bir dilsel, mantıksal ya da sayısal kurala göre düzenlenmesi anlamında kullanılır. Metin verisi ifadesi belirli bir dilde yazılmış her türlü içerik için kullanılır. Bu projede ise şiir metinleri ve bu Metinlerin sınıflandırılması amacıyla kullanılmıştır. [1]

II.VERİ TOPLAMA

A.Veri Kaynağı Seçimi

Veri kaynağı seçimi projede en önem verdiğimiz noktalardan biridir. Kaynağın güvenilir, anlaşılır, net olması web kazıma işlemi açısından gereklidir. Türkiye'nin en büyük ve kapsamlı şiir sitelerinden biri olan Antoloji.com web sitesini kullanılarak veri çekme işlemi gerçekleştirilmiştir. Bu sitede şiirlerin kategorilerine göre sınıflandırılmış olması sebebiyle burdan yararlanılmıştır.

B. Kategoriler Ve Sınıflandırma

Antoloji şiir web sitesinde şiir türleri sınıflandırmasına uygun olacak şekilde kategoriler seçilmiştir.

1.Lirik Şiir: Acı, Aşk, Ayrılık, Hasret, Keder, Sevinç, Üzüntü
2.Epik Şiir: Türkiye, Savaş, Vatan, Bayrak, Kahraman, Asker,Şehit
3.Pastorel Şiir: Ağaç, Doğa, Gökyüzü, Bahçe, Çiçek, Dağ
4.Satirik Şiir: Yoksulluk, Düşman, Adalet, Sitem,Akıl Kötülük, İhanet
5.Felsefi Şiir: Dünya, Hayat, İnanç, Zaman,Tanrı,Din

Şiir kategorileri seçimi , şiir türlerinin gerekliliklerine göre yapılmıştır. Lirik şiir de şairin bireysel hisleri ve duygusal durumları ön plandadır. Epik şiirde ise vatan için yapılmış fedakarlıklar, kahramanlıklar ile ilgili konular ön plandadır. Pastorel şiir de daha çok doğa, kırsal yaşam anlatılmaktadır. Satirik şiir ise toplumsal ***a***ynı zamanda bireysel yanlışlıkları, aksayan yönleri eleştirel bir biçimde ele alır. Son olarak Felsefi şiir de felsefe, din, ahlak gibi konular üzerinde durur [2]Tüm bunlar göz önünde bulundurulup yukarıdaki kategori seçimleri yapılmıştır.

C. Web Kazıma Aşaması

Antoloji şiir web sitesinden python ve kütüphaneleri kullanılarak veri çekme işlemleri gerçekleştirilmiştir. Sayfa içeriğini çözümlemek ve analiz etmek için en yaygın olarak kullanılan BeautifulShoup kütüphanesi kullanılmıştır.Htttp isteği göndermek ve yanıtlar Requests kütüphanesi kullanılmıştır.
Çekilen verilerin csv dosyalarına kayıt edilmesi kullandığımız Pandas kütüphanesi ile sağlandı. Web sayfalarını otomatikleştirmek için kullanılan selenıum şiir sayfalarını tarayıcıda açmak ve HTML elemanları erişimi için kullanıldı.

III. VERİ ÖN İŞLEME

Veri ön işleme aşamasında; bir metni daha küçük parçalara ayıran **tokenizasyon işlemi**, kelimelerin kök veya temel şekillerine indirgenmesi işlemini sağlayan **lemmatizasyon**, dilde anlam taşımayan veya çok küçük anlam taşıyan kelimeleri veri setinden çıkaran **stopwords işlemi**, kelimelerin son eklerini keserek anlam bütünlüğü dışında kök haline getiren **kök bulma(stemming)** işlemi, kelime türü belirleyen **POS Tagging** işlemi ile veri ön işleme adımları tamamlanmıştır.

A. Tokenizasyon

Öncelikle python ile Java arasında köprü kuran **Jtype ve Türkçe** doğal dil işleme kütüphanesi olan **Zemberek** kullanılmıştır. Şiir mısrasındaki her kelimeyi morfolojik olarak analiz eden Zemberek fonksiyonu uygulanmıştır. Bu işlem metni kelime kelime böler yani tokenize edilmiş hale getirir. Bu olay Zemberek in içsel olarak kelimeleri tanıyıp her birine morfolojik analiz yapması ile gerçekleşir.[3]

B. Lemmatizasyon

Lemmatizasyondaki temel amaç, metindeki kelimeleri daha anlamlı ve doğru bir biçimde analiz etmek, böylece daha etkili dil işleme sonuçları elde etmektir. **Lemmatizasyon**, kelimenin dilbilgisel özelliklerini göz önünde bulundurarak, kelimelerin ek ve türevlerinden arınmış temel sözlük formuna indirir.[4]

```
Yıllar var hasretim, yârana dosta,,"[{'input': 'Yıllar', 'root': 'yıllamak', 'type': 'Verb'}
Isıcak çorban görmedim tasta,,"[{'input': 'çorban', 'root': 'çorba', 'type': 'Noun'}, {'
Sana sağlam geldim, eyledin hasta,,"[{'input': 'Sana', 'root': 'sanmak', 'type': 'Verb'
Zıkkım olsun paran, pulun Almanya.,"[{'input': 'Zıkkım', 'root': 'zıkkım', 'type': 'Nou
El sıraya katışamadım,,"[{'input': 'El', 'root': 'el', 'type': 'Noun'}, {'input': 'sıraya', 'ro
Evlat oldum elden tutuşamadım,,"[{'input': 'Evlat', 'root': 'evlât', 'type': 'Noun'}, {'i
Anam, babam öldü yetişemedim,,"[{'input': 'Anam', 'root': 'ana', 'type': 'Noun'}, {'i
Köyüme uzak yolun Almanya,"[{'input': 'Köyüme', 'root': 'köy', 'type': 'Noun'}, {'inp
```

C. **Kök Bulma (Stemming)**

Stemming kelimenin anlamını korumadan ekler ve türevlerinden arındırarak kelimenin en temel biçimine indirgenmesi işlemidir. Stemming dildeki kelimelerin anlamları kaybolmasına rağmen bir kelimenin kökünü bulma sürecidir. Lemmatizasyon ise kelimenin anlamını koruyarak dilbilgisel kurallara uygun şekilde doğru sözlük formuna indirger.

```
Haramdan eksik tartıp helalı,
{'input': 'Haramdan', 'root': 'haram', 'type': 'Noun'}
{'input': 'Haramdan', 'root': 'hara', 'type': 'Noun'}
```

D. **POS Tagging**

Metindeki her bir kelimenin dilbilgisel kategorisini belirlemek için uygulanmıştır. Kelimenin ismini, fiil, zarf sıfat şeklinde etiketlenmesi sağlanmıştır. Bu işlem veri kullanım sırasında kolaylık sağlar. Çünkü bu etiketler, kelimelerin şiirdeki rolünü ve ilişkisini anlatmaktadır.[5]

E. **Stopwords Temizleme**

Stopwords temizleme, genellikle dikkate alınmayan dilde sıkça yer alan ancak metnin anlamını belirlemede çok fazla katkı sağlamayan kelimelerin metinden kaldırılması işlemidir. Öncelikle Türkçe dilinde yaygın olarak kullanılan Stopwords kelimelerinin bir listesi tanımlanmıştır.

```
stopwords = set([
    "ve", "bu", "şu", "o", "bir", "ancak", "fakat", "çünkü", "ki", "de", "da",
    "ile", "mi", "mı", "mu", "mü", "sen", "ben", "biz", "siz", "onlar", "ya",
    "ne", "her", "birçok", "çok", "az", "biraz", "bile", "hem", "gibi", "ama",
    "eğer", "ise", "hala", "sadece", "ya", "yine", "için", "kadar", "ile",
    "dolayı", "dolayısıyla", "aslında", "başka", "herhangi", "bazı", "bazısı"
])
```

Remove_stopwords() fonksiyonu ile cümle alınıp, o cümledeki kelimeler ayrıştırıldı. Stop words listesinde yer almayan kelimeler geri döndürüldü.[6]

REFERANSLAR

[1] https://en.wikipedia.org/wiki/Text_mining

[2] https://www.turkedebiyati.org/siir-bilgisi/#google_vignette

[3] https://tahtaciburak.medium.com/zemberek-k%C3%BCt%C3%BCphanesi-pythonda-nas%C4%B1l-kullan%C4%B1l%C4%B1r-8993ec1c3f0e

[4] https://en.wikipedia.org/wiki/Natural_language_processing#Morphological_analysis

[5]https://en.wikipedia.org/wiki/Natural_language_processing#Higher-level_NLP_applications

[6] https://turhancankargin.com/2020/10/18/dogal-dil-isleme-nedir/