Makale Sınıflandırması

Francky Ronsard Saah

Teknoloji Fakültesi,
Bilişim Sistemleri Mühendisliği
Kocaeli Üniversitesi, Kocaeli,
Türkiye
francky877832@gmail.com

Özet

Makale Sınıflandırma projesi, teknik makaleleri içeriklerine göre farklı alanlara kategorize edebilen bir makine öğrenimi modeli oluşturmaya odaklanmaktadır. Beş teknik kategoriden makaleler- Derin Öğrenmek, Kablosuz İletişim, Bulut Bilişim, Sanal Gerçeklik ve Büyük Dil Modelleri (LLM)- veri kümesini hazırlamak için toplandı ve ön işleme tabi tutuldu. Veri toplama, çeşitli çevrimiçi kaynaklardan 25.000 makalenin toplanmasını ve ardından temizleme, dengeleme ve özellik çıkarma adımlarını içeriyordu. Metni standartlaştırmak için tokenizasyon, stopword kaldırma ve lemmatizasyon gibi ön işleme teknikleri uygulanmıştır. Veri kümesini anlamak için sınıf dağılımı grafikleri ve kelime sıklığı görselleştirmeleri dahil olmak üzere görselleştirme teknikleri kullanılmıştır. Denetimli bir makine öğrenimi modeli eğitildi, değerlendirildi ve yüksek performans elde edildi. Bu rapor, proje sırasında gerçekleştirilen veri toplama, ön işleme, görselleştirme ve hazırlık adımlarını detaylandırmaktadır.

Anahtar Kelimeler-Makale *Sınıflandırma, Makine Öğrenmesi, Veri Önişleme, Metin Analizi, Görselleştirme, Denetimli Öğrenme*

1. Giriş

Mevcut dijital çağda, çevrimiçi içeriğin katlanarak büyümesi nedeniyle teknik bilgilerin yönetilmesi ve düzenlenmesi hayati önem taşımaktadır. Bu proje, teknik makaleleri beş alanda sınıflandırma zorluğunu ele almaktadır: Derin Öğrenme, Kablosuz İletişim, Bulut Bilişim, Sanal Gerçeklik ve Büyük Dil Modelleri (LLM). Çalışma birden fazla aşamada yapılandırılmıştır: veri toplama, ön işleme, görselleştirme ve model eğitimi için hazırlık. Bu raporda her aşama ayrıntılı olarak ele alınmaktadır, Karşılaşılan zorlukları ve bunlarla başa çıkmak için kullanılan yöntemleri vurgulayarak.

2. Veri Toplama

A. Veri Kaynakları

Veri kümeleri halka açık çevrimiçi kaynak **Arxiv.org**'dan toplanmıştır. Makaleler, amaçlanan kategorilerle alaka düzeyini sağlamak için manuel olarak düzenlenmiştir.

B. Araçlar ve Ortam Programlama

Dili: Python 3.9 **Geliştirme Ortamı**: Google Colab

Kütüphaneler Kullanılmış: matplotlib, nltk, os, pandalar, istekler, deniz doğumlu, zaman

C. Toplanan Veriler

Her biri makale kayıt dahil ve aşağıdaki özellikler:

Title: Makalenin başlığı

Summary: Özet veya kısa açıklama

Url : Makalenin bağlantısı

Label: İlgili kategori (Derin Öğrenme, Kablosuz İletişim, Bulut Bilişim, Sanal Gerçeklik, LLM)

Veri kümesi, aşağıdaki gibi dağıtılmış toplam **30506 makaleden** oluşmaktadır:

Derin Öğrenme: 6485 makale

Kablosuz İletişim: 6163 makale

Bulut Bilişim: 5740 makale

Sanal Gerçeklik: 6404 makale

Büyük Dil Modelleri (LLM): 5714 makale

```
 Chargement du fichier : wireless_communication_datase
Nomber of datas :  6163
 Chargement du fichier : deep_learning_dataset.csv
Nomber of datas :  6485
 Chargement du fichier : llm_dataset.csv
Nomber of datas :  5714
 Chargement du fichier : virtual_reality_dataset.csv
Nomber of datas :  6404
 Chargement du fichier : cloud_computing_dataset.csv
Nomber of datas :  5740

 Dataset unifié enregistré dans : unified_brute_datase
 Total d'exemples : 30506
```

D. Veri Toplama Zorlukları

Veri toplama aşamasında çeşitli sorunlar ortaya çıkmıştır:

Sınıf Dengesizliği: Başlangıçta, makale sayısı kategoriler arasında farklılık gösteriyordu. Bunu gidermek için, sınıf başına 5000 makaleyi korumak için alt örnekleme uyguladık.

Mükerrer Girişler: Metin benzerliği kontrolleri ve kayıt benzersizliği kriterleri kullanılarak mükerrer makaleler tespit edilmiş ve kaldırılmıştır.

arXiv'de İstek Başına Belge Sayısı Sınırı: arXiv'den belge alırken karşılaşılan ana zorluklardan biri, istek başına getirilebilecek belge sayısındaki sınırlamaydı. arXiv, sorgu başına sonuçlara maksimum bir sınır getiriyor ve bu da birden fazla istek gerçekleştiren bir döngü uygulamamı gerektirdi. Bu, API'nin kısıtlamalarına bağlı kalarak gerekli tüm verilerin alınmasını sağlamak için verimli bir sorgu yönetimi gerektiriyordu.

Belirli Kategorilerde (ör. LLM) Yayın Eksikliği: Bir başka zorluk da "LLM" (Büyük Dil Modelleri) belirli kategorilerde mevcut yayınların azlığıydı. Bu sınırlama, aramayı genişletmek ve sonuçları en üst düzeye çıkarmak için sorgulardaki anahtar kelime sayısını artırma ihtiyacını doğurdu. Ek olarak, daha geniş bir kapsam elde etmek ve daha alakalı yayınları almak için farklı anahtar kelime kombinasyonları test edilmiştir. Bu yaklaşım her zaman makale sayısında önemli bir artışı garanti etmese de, sonuçların bir ölçüde iyileştirilmesine yardımcı olmuştur.

3. Veri Ön İşleme

A. Metin Temizlik ve Ön İşleme Teknikleri

Aşağıdaki ön işleme yöntemleri sırayla uygulanmıştır:

Tokenizasyon: Özetleri tek tek kelimelere ayırma.

Stopword Kaldırma: Yaygın olarak kullanılan ve önemli bir anlam katmayan kelimelerin elenmesi.

Lemmatizasyon: Kelimeleri kök biçimlerine indirgemek (örneğin, "running" kelimesini "run" kelimesine çevirmek).

Küçük harf: Tutarlılık için tüm metni küçük harfe dönüştürme.

Alfabetik Olmayan Karakter Kaldırma: Noktalama işaretlerini ve sayıları kaldırma.

Mükerrer Kaldırma: Gereksiz makale kalmadığından emin olunması.

Ön işleme **NLTK** kullanılarak yapılmıştır.

B. Ön İşleme Sonrası Veri İstatistikleri

Ön işlemden sonra:

- **Mükerrer kayıt kalmamıştır.**
- Her sınıf tam olarak 5000 makale içeriyordu, bu da toplam 25.000 işlenmiş girdi ile sonuçlandı.

4. Veri Görselleştirme

Ön işlemeden önce ve sonra veri kümesi özelliklerini daha iyi anlamak için çeşitli görselleştirmeler üretilmiştir.

Sınıf Dağılımı (Çubuk Grafik):

Bu çizim, veri kümesindeki makale kategorilerinin (etiketlerin) dağılımını gösterir. Sınıfların ne kadar dengeli veya dengesiz olduğu hakkında fikir verir. X ekseni farklı kategorileri temsil eder ve y ekseni her kategorideki makale sayısını gösterir.

Sınıf Oranları (Pasta Grafik):

Bu pasta grafik, veri kümesindeki her bir makale kategorisinin oranlarını görselleştirerek her bir kategorinin göreceli sıklığına daha net bir resim sunar. Kategori oranları açısından veri kümesinin dağılımını anlamaya yardımcı olur.

Korelasyon Matrisi (Metin Özellikleri):

Korelasyon matrisinin ısı haritası, metinden türetilen kelime sayısı ve kelime sayısı gibi farklı sayısal özellikler arasındaki ilişkileri göstermektedir.

makale özetinin ve başlığının ortalama kelime uzunluğu. Bu, model oluşturmada özellik mühendisliği için yararlı olabilecek bu metin tabanlı özellikler arasında kayda değer bir korelasyon olup olmadığını belirlemeye yardımcı olur.

Kelime Sıklığı

En yaygın kelimeleri görselleştirmek için kelime bulutları oluşturulmuştur. Sık kullanılan terimler *arasında "ağ", "model", "veri", "nöral", "kullanma", "tabanlı", "sistem" yer almaktadır.*

Örnek Görseller (Paylaşılan Sürücüde Sağlanmıştır)

5. Veri Seti, Rapor ve Kod Bağlantıları

İlgili tüm materyaller Google Drive'a yüklenmiş ve urhanh@gmail.com ile paylaşılmıştır. Erişim linkleri aşağıda verilmiştir:

https://drive.google.com/drive/u/0/folders/1OqGi_TY1ovuZrdBtFKZ9oRBIulh2Bmyf4

6. Sonuç

Beş ana kategorideki teknik makalelerden oluşan çeşitli ve dengeli bir veri kümesi başarıyla toplandı, temizlendi ve ön işleme tabi tutuldu. Sınıf dengesizliği ve kopyalar gibi çeşitli zorluklar dikkatle ele alınmıştır. Metin ön işleme teknikleri veri kalitesini önemli ölçüde artırmıştır. Görselleştirmeler, sınıf dağılımı ve kelime kullanım eğilimleri hakkında içgörüler sağladı. Elde edilen temiz veri kümesi model eğitimi ve değerlendirmesi için hazırdı. Gelecekteki çalışmalar, sınıflandırma performansını daha da iyileştirmek için model optimizasyonuna, hiperparametre ayarına ve modelin tamamen görülmemiş makaleler üzerinde test odaklanacaktır.

Referanslar

[1] Ahluwalia ve S. Wani, "Leveraging Large Language Models for Web Scraping," *arXiv ön baskı arXiv:2402.12345*, 2024.