<KATEGORİ>

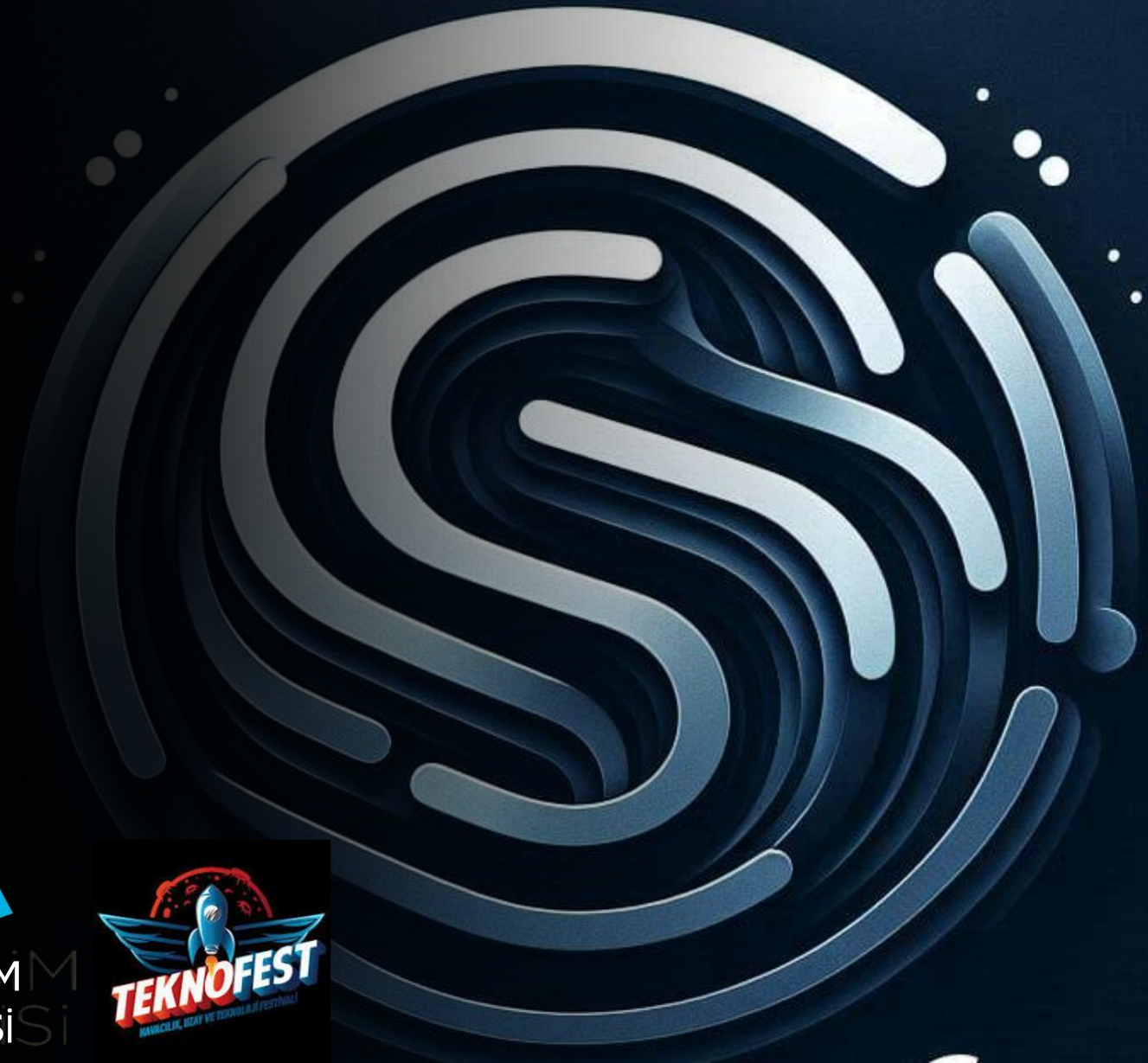
TÜRKÇE DOĞAL DİL İŞLEME

8 - 9 AĞUSTOS 2024

LISOS
TAKIMI

LİSOS TEAM

LİSANI DEVLETİ OSMANİ



#MİLLİ TEKNOLOJİ HAMLESİ

BİLİŞİM VADİSİ

TEKNOFEST
HAVACILIK, UZAY VE TEKNOLOJİ FESTİVALİ

EKİBİMİZ

Takım üyelerinin tanıtılması, rolleri ve katkıları hakkında bilgi verilmesi.

KAPTAN

FAZIL AHMED AZİZİ

Problem Tespiti- Çözüm

Bilgisayar Mühendisliği okuyorum. Anadilim Özbekçe olup Farsça-Darice, ve Türkçe bilmekteyim. Ayrıca liseden beri merak edip Osmanlıca eserler okuyorum. Türkiye Teknoloji Vakfı Gönüllüsüyüm.

ÜYE

ABDUL MUDASER GHAFURY

Proje-Süreç Planlaması

Bilgisayar Mühendisliği Okuyorum. Java dilini biliyor ve NLP ile ilerleme hedefi taşımaktayım. Türkiye Teknoloji Vakfı Gönüllüsüyüm.

ÜYE

SÜLEYMAN SÜLEYMANOV

Araştırma-Analiz

Ana dilim Azerbaycan Türkçesi olup Bilgisayar Mühendisliği okuyorum. Osmanlıcayı, Doğal Dil İşleme teknolojilerini merak ediyor ve bu alanda güzel işler başarabileceğimize inanıyorum

Türkiye Teknoloji Vakfı Gönüllüsüyüm.

ÜYE

AHMAD ABİD YAMİN

Yöntem-Teknik Önerileri

3.Sınıf Bilgisayar Mühendisliği okuyorum. Java, Python, C++, PHP, HTML ve temel düzeyde Assembly bilgisine sahibimTürkiye Teknoloji Vakfı Gönüllüsüyüm.

.

PROJENİN TANIMI

Çalışmanın Ana Teması:

Osmanlıca El Yazma ve Matbaa metinlerini, Arapça ve Farsça dillerinin özelliklerini de dikkate alarak, metin madenciliği teknikleriyle analiz edilmesi. Bu analizler sonucunda tarihî olaylar, kişiler ve yerler hakkında değerli bilgilerin ortaya çıkarılması ve bu bilgilerin anlamlı veri kümeleri, analiz raporları ve görselleştirme araçları ile sunulması.

Ele alınan problemin tanıtılması:

Günümüzde, Osmanlıca el yazma ve diğer metinlerin tam anlamıyla doğal dil işleme yöntemiyle analiz edilmesi ve pek çok değerli bilgilerin ortaya çıkarılması konusunda her hangi kapsamlı bir çalışma bulunmamaktadır. Ayrıca Osmanlı Türkçesini Doğal Dil İşleme(NLP) yöntemleriyle işlenip dijital ortama aktarılabilmesi ve çevrilebilmesi için de; Arapça ve Farsça dillerinin bazı gramer kurallarına ve özellikle Osmanlıcaya uyarlanan tamlamalar için bu iki dile de hakim olmak gerekiyor. Mevcut doğal dil işleme araçları, Osmanlı Türkçesi ve bu iki dilin özelliklerini tam olarak kapsayamadığı için Osmanlıca metinlerdeki anlam, duygu ve tarihi bilgilerin doğru bir şekilde çıkarılması yeterli düzeyde olmayıp nakıs kalmış, daha ayrıntı çalışmalara ihtiyaç duyulmuştur.

PROJENİN SAĞLADIĞI ÇÖZÜM

Çalışmanın nasıl bir çözüm sunar:

1.En başta NLP aşamalarından önce mevcut el yazma ve diğer yazılı metinleri dijital ortama aktarmak gerekiyor. Bunun için hem zaman tasarrufu hem yüksek doğruluk için geliştirilen «Transkribus Yazılımı»na benzer bir çalıştırma oluşturup ve geliştirmek.

2. Hedefimize ulaşmada gerekli dillerin ekibimizce biliniyor olması ve Dilbilimciler, Osmanlı Dili, Tarihi, Kültürü Araştıran Uzmanlar ile de işbirliği yapmak, sonraki süreçte ise NLP aşamalarını tatbik etmek olacaktır.

Çözümden faydalanacak hedef kitle:

Çözüm en başta değerli bilgilerin ortaya çıkarılmasına yani edebi, tarihi, bilimsel ve günümüze kadar ortaya çıkarılmayı bekleyen diğer birçok bilgi mahzeninin keşfedilmesine vesile olacaktır. Böylece Türkiye'nin ulusal ve kültürel mirasına katkıda bulunarak, Osmanlıca'ya olan ilgiyi ve anlayışı artırmak Osmanlı tarihi ve kültürüyle ilgilenen araştırmacılar, öğrenciler ve genel halk için değerli bir kaynak oluşturarak bu dili dijital ortamda kullanma-deneyimleme imkanı yaratmak yani bir OstGPT oluşturmaktır.

PROJE İŞ AKIŞI

Projenin başarıyla tamamlanması için gereken görevlerin ve süreçler:

- Proje Planlama ve Ekip görev dağılımı
- Veri Toplama
  - Verilerin Dijital Ortama Aktarımı (Transkribus benzeri)
    - NLP süreci
      - ❖ Doğal Dil İşleme Modellerinin Geliştirilmesi
      - ❖ Dil Modelleme: Osmanlıca ile ilişkili dil unsurlarının bağlantılarını inceleme ve duygusal tonların belirlenmesi.
      - ❖ İsim Tanıma (NER): Tarihî olaylar, kişiler ve yerlerin tespiti.
- Veri Analizi ve Görselleştirme
- Raporlama ve Paylaşım

VERİ SETİ

Kullanılan veri setinin kaynağı, özellikleri ve hazırlanma süreci.

✓ **Veri Toplama ve Dijitalleştirme:**

- Tarama: El yazması belgeler, yüksek çözünürlüklü tarayıcılar ile dijital formata aktarılarak tarama sonrası elde edilen görüntüler, yapay zeka algoritmalarının daha iyi çalışabilmesi için; Tokenizasyon,, Lemmatizasyon, Normalizasyon gibi aşamaları barındıran "Ön İşleme" tabi tutulur.
- Tanıma: (Hem el yazısı hem de basılı metinleri tanıma teknolojisi) Transkribus ve (karakter tanıma algoritmalarıyla çalışan) OCR yazılımları, ön işlenmiş görüntüler üzerinde el yazısı tanıma işlemi gerçekleştirir. Bu işlem, metinlerin otomatik olarak okunmasını sağlar.
- Düzeltme ve Doğrulama: Otomatik tanıma sonrasında, elde edilen metinler insan uzmanlar tarafından kontrol edilir ve gerekli düzeltmeler yapılır.
- Dijital Arşivleme: Son aşamada, dijitalleştirilen metinler uygun formatlarda arşivlenir ve erişime açılır.

✓ **. Arapça ve Farsça Unsurları İnceleme:**

- Osmanlıca metinlerde sıkça kullanılan Arapça ve Farsça kelimeleri ve terkipleri inceleyerek bu unsurları oluşturulacak modelimize entegre etmek.

✓ **Dil Modeli Geliştirme**: Dil modeli eğitiminde Osmanlıca metinlerini kullanarak, bu dilin özgün yapısını ve özelliklerini yansıtan bir model oluşturmak.

YÖNTEM VE TEKNİKLER

Kullanılan algoritmalar, modeller ve doğal dil işleme tekniklerinin açıklanması.

- ❑ Algoritmalar:
  - ✓ Sinirsel Makine Çevirisi (NMT) ; Derin Öğrenme Algoritmaları, Türkçe için geliştirilmiş derin öğrenme tabanlı modeller (GPT2, GPT3, BERT, ELMo vd.)Stemming ve Morfolojik Analiz.
- ❑ Modeller:
  - ✓ Transkribus Modelleri, TurkishDelightNLP ve Özelleştirilmiş Modeller arasında yer almasını hedeflediğimiz OstGPT.
- ❑ NLP Teknikleri:
  - ✓ Sinirsel Makine Çevirisi (NMT), Metin Benzerliği ve Anlamsal Analiz, Osmanlı Türkçesi metinlerinde anahtar kelimelerin ve konuların belirlenmesine yardımcı olması için TF-IDF ve Konu Modelleme yöntemleri.

MODEL EĞİTİMİ VE DEĞERLENDİRME

Modelin nasıl eğitildiği, hangi metriklerle değerlendirildiği:

Modelin Eğitilmesi

Veri Hazırlığı:

Veri Toplama: Osmanlıca metinlerin mevcut dijital arşivlerden ve kaynaklardan toplanması ve Veri Temizleme yoluyla metinlerin dilsel ve anlamsal analiz için temizlenmesi ve ön işleme tabi tutulması,

Dil Kaynaklarının Hazırlanması: Osmanlıca ile ilişkili Arapça ve Farsça yapıların entegre edilmesi ve etiketlenmesi.

Model Seçimi ve Hazırlık:

- ✓ Dil Modeli: Dil modeli eğitiminde Osmanlıca metinlerini kullanarak, bu dilin özgün yapısını ve özelliklerini yansıtan bir model oluşturmak.
- ✓ Duygu Analizi: RNN (Recurrent Neural Network) veya LSTM (Long Short-Term Memory) gibi derin öğrenme yöntemlerinin uygulanması.
- ✓ İsim Tanıma (NER): CRF (Conditional Random Fields) veya BERT tabanlı NER modellerinin kullanılması.

Model Eğitimi:

Eğitim Verisi: Eğitim için hazırlanan etiketli verilerin kullanılması ve «Model Eğitimi» de seçilen modellerin eğitim verisi üzerinde eğitilmesi.

SONUÇLAR

Proje kapsamında elde edilen bulgular ve sonuçlar:

Proje sonucunda, anlamlı veri kümeleri oluşturulacak, analiz raporları hazırlanacak ve elde edilen bilgiler görselleştirme araçları ile sunulacaktır. Bu sayede, Osmanlıca metinlerden elde edilen bilgiler tarihçiler, araştırmacılar ve meraklılar için erişilebilir hale gelecektir. Aynı zaman da geliştirilecek Dil Modeli(OstGPT) hizmete sunulacaktır.

PROJE YOL HARİTASI

Projenin gelecekte nasıl geliştirilebileceği ve olası araştırma konuları:

- Dil Modelleri ve Eğitim Verileri:

  Osmanlıca için özel olarak eğitilmiş dil modelleri ve bu modelleri eğitmek için gerekli büyük veri setlerine ihtiyaç vardır bu alanda.

- Osmanlıca Dilbilgisi ve Sözlük Geliştirme:

  Osmanlıca dilbilgisi kurallarını ve kelime hazinesini içeren dijital sözlük ve dilbilgisi kaynakları oluşturulup zenginleştirme.

- Benchmarking: OCR modellerinin performansı diğer modellerle karşılaştırılması ve sürekli iyileştirmesi diğer bir araştırma konusudur.

DEMO VİDEO

Projenin demo videosunun ve linkinin eklenmesi.

TEŞEKKÜRLER