Kısıtlı Türkçe Veri Üzerinde Derin Öğrenme
Deep Learning on Limited Turkish Data

Selim F. Tekin[1, 3], Selim F. Yılmaz[1] ve Ismail Balaban[2, 3]

[1]Elektrik ve Elektronik Mühendisliği Bölümü, Bilkent Üniversitesi, Ankara, Türkiye

{tekin, syilmaz}@ee.bilkent.edu.tr

[2]Çoklu Ortam Bilişimi Bölümü, Orta Doğu Teknik Üniversitesi, Ankara, Türkiye

ismail.balaban@metu.edu.tr

[3]DataBoss A.S., Ankara, Türkiye

{furkan.tekin, ismail.balaban}@data-boss.com.tr

1. Giriş

- Gözetimli öğrenme tekniği ile geliştirilen derin öğrenme modelleri, duygu tespitinde yüksek başarı elde edilmektedir.
- Ancak, derin öğrenme modellerinin yüksek başarı sağlayabilmesi için yüksek miktarda etiketli veriye ihtiyaç duymaktadır.
- Bu bildiri, etiketlenmiş veri eksikliği problemini hedef almaktadır.
- Performansı artırmak amacıyla, veri kümesi sözlük temelli eğitim tekniği ile çoğaltılmıştır. Sonra derin öğrenme modeli eğitilmiştir.

2. Yapılmış Çalışmalar

- Duygu tespiti üzerine yapılan çalışmalar:
  - Güdümlü öğrenmede yüksek miktarda etiketli veriye ihtiyaç duymaktadır.
  - Güdümsüz öğrenmede methodlar eğitim setindeki tanım kümesi dağılımını öğrenebilmekte fakat test setindeki farklılığı yakalayamamaktadır
- Kısıtlı veri problem üzerine yapılmış çalışmalar kısıtlı performans artışı sağlamaktadır:
  - Eğitilmiş kelime vektörleri
  - Eğitilmiş dönüştürücüler
  - Özyükleme, zayıf güdümleme tekniği

3. Problem Tanımı

- Etiketli veri kümesi $D_L: \{x_i, y_i\}_{i=1}^{N}$ kullanılarak öğrenilmek istenilen fonksiyon

$$f(x_i) = y_i$$

gösterilmektedir. x_i burada girdi *tweet* metnini ifade ederken $y_i \in \{0, 1\}$ girdiye ait etiket kümesini ifade etmektedir.

- Her girdi bir sözlüğe ait kelimelerden w_j oluşmaktadır $x_i: \{w_j\}_{j=1}^{S}$ $w_j \in K$.
- K sözlüğü ifade ederken, S girdi uzunluğunu temsil etmektedir.

3. Problem Tanımı

- Her bir kelimenin matematik karşılığı olarak kelime vektörleri oluşturulmuştur.
- Sözlük her bir kelimeye ait d boyutunda kelime vektörlerinden oluşmaktadır. $K \in \mathbb{R}^{d \times M}$
- Bildirideki amaç $f(.)$ fonksiyonuna yaklaşmaktır. Fonksiyon bir olasılık dağılımından geldiği kabul edilmektedir. Fonksiyona yaklaşımda elde edilecek tahmin olasılıkları, $\hat{y}$, kayıp fonksiyonunu hesaplamakta kullanılacaktır.

3. Problem Tanımı

- Kayıp fonksiyonu olarak ikili çapraz entropi kullanılacaktır,

$$\mathcal{L} = -\sum_{k=1}^{C=2} y_{i,k} \log(f(x_i))$$

- Bu fonksiyon, verinin mini balyaları kullanılarak Stokastik Gradyan İnişi (SG I) ile optimize edilecektir.
- sentetik veri kümesi D_U gösterilmektedir. $D_U : \{x_i, y_i\}_{i=1}^{N'}$
- Bildiride amaç, $D_L \cup D_U$ üzerinde eğitilen modelin, sadece D_L üzerinde eğitilen modelden daha iyi performans göstermesidir.

4. Yöntem

- Dört aşamadan oluşmaktadır:
  1. Test kümesinin oluşturulması. Verinin sözlük temelli yöntem ile çoğaltılması.
  2. Verinin ön işlemden geçmesi.
  3. Modele verilen her *tweet* cümlesi belirli bir dizi uzunlukta kelime vektörlerine dönüştürülmesi.
  4. Verinin *Çift Yönlü Kısa-Uzun Hafızalı Bellek* modelinde eğitilmesidir.

4. Yöntem

- Etiketli veri artırımı *tweet* içeriklerinde sıkça kullanılan metinsel emojilere göre gerçekleştirilmiştir.

ASCII Emoji ve Etiketleri

Emoji	Sınıf
:), :)), :-), =), :D, :d, :P	pozitif
:(, :((, :-(, :/, :S, :s	negatif

4. Yöntem

- Ön işleme
  - URL, sosyal medya etiketi(#); kişi, lokasyon atıfları, yirmi karakterden fazla ve de üç karakterden az olan kelimeler, noktalama işaretleri ve de sayılar çıkartılmıştır. Bunlara ek olarak gereksiz kelime, *stopword*, olarak adlandırılan kelimeler çıkartılmıştır.

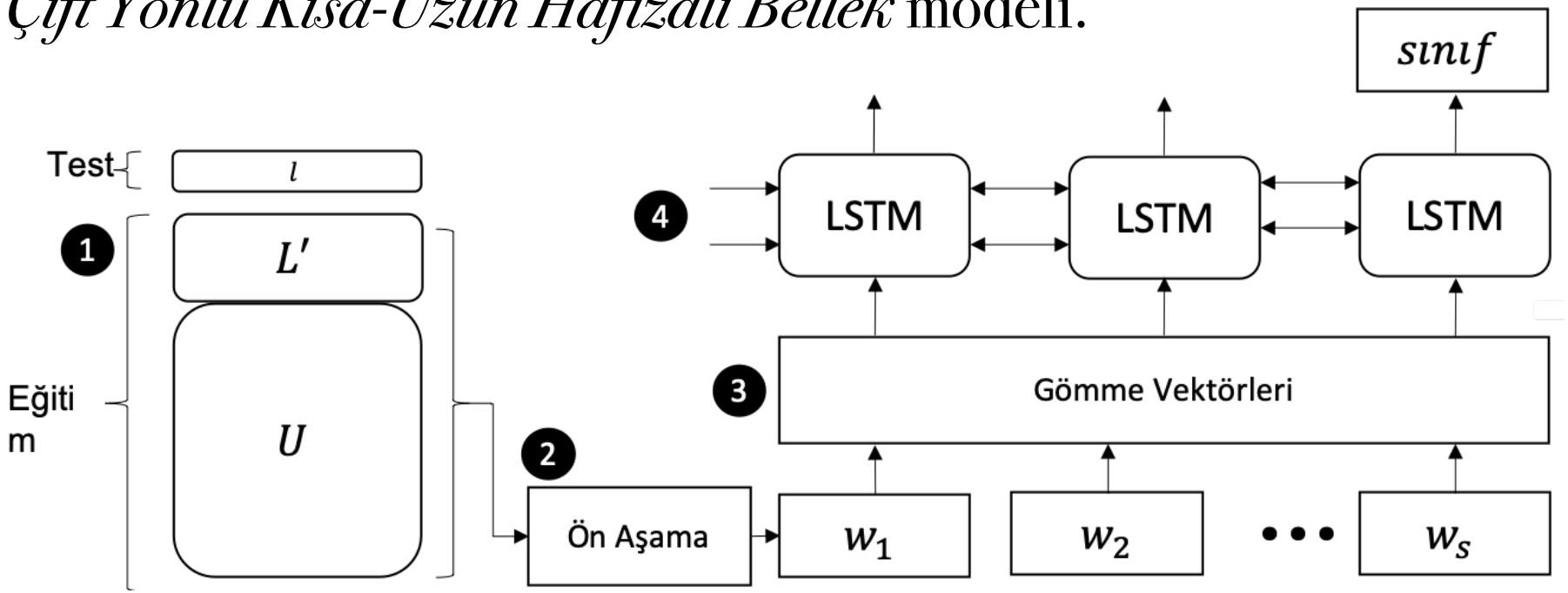
  - Ikinci verinin etiketlemesi metinsel emoji ifadelerine göre yapıldığı için, bu ifadeler metinlerden çıkartılmıştır.

4. Yöntem

- Gömme vektörleri
  - Kullanılan kelime gömme vektörleri *Fasttext* modelinin ürünüdür. Bu model *skipgram* modelinin geliştirilmiş ve genişletilmiş halidir.
  - *Fasttext* kullanılarak her bir kelimenin 300 boyutlu vektörel hali elde edilmiştir.
  - Her bir tweet 15 kelime uzunluğuna "<pad>" simgesi ile tamamlanmıştır.

4. Yöntem

- *Çift Yönlü Kısa-Uzun Hafızalı Bellek* modeli.



5. Deneyler

- Veri kümesi
  - Etiketli veri: 6000 tweet içermektedir, 3000 tanesi negatif, 1552 tanesi pozitif geri kalanı nötrdür.
  - Sentetik veri: 22000 tweet içermektedir. 12000 tanesi pozitif, 10000 tanesi negatifdir.
  - Sadece pozitif ve negatif veriler kullanılmıştır.
- Performans karşılaştırması için eğri altı alan (EAA) değerlendirme ölçütü kullanılmıştır.

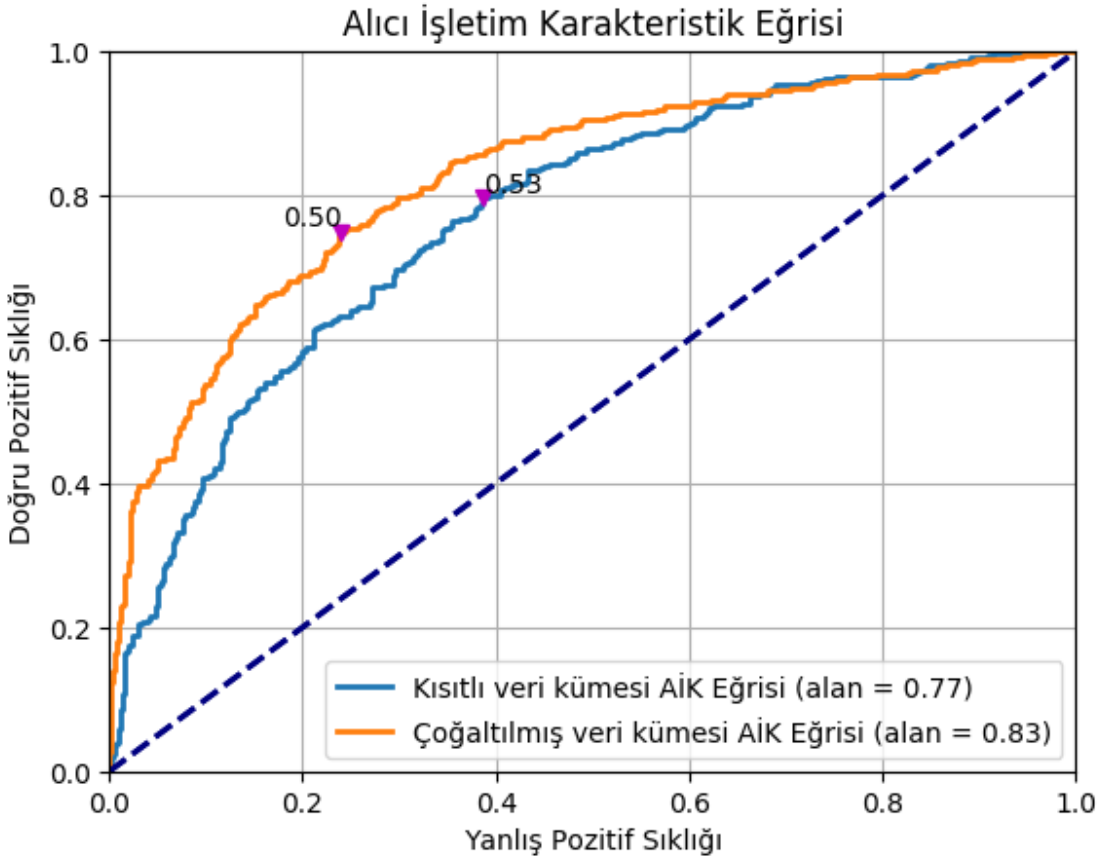
5. Deneyler

- Kullanılan yöntemin makine öğrenmesi metotlarındaki performansını gözlemlemek amacıyla iki farklı model üzerinde deneyler yapılmıştır.
  - Doğrusal Destek Vektör Makineleridir (DDVM).
  - Rastgele Karar Ağaçlarıdır(RKA).

Modeller ve F1, Kesinlik, Eğri Altı Alan Değerleri

Modeller	Kısıtlı Veri Kümesi			Çoğaltılmış Veri Kümesi		
	F1(%)	Kesinlik(%)	EAA(%)	F1(%)	Kesinlik(%)	EAA(%)
DDVM	65.0	79.0	87.0	69.0	75.0	82.0
RKA	56.0	76.0	82.0	61.0	73.0	80.0
ÇY-UKSB	**68.0**	**76.0**	**77.0**	**71.0**	**77.0**	**83.0**

5. Deneyler


Alıcı İşletim Karakteristik Eğrisi
Doğru Pozitif Sıklığı
Yanlış Pozitif Sıklığı
0.50
0.53
Kısıtlı veri kümesi AİK Eğrisi (alan = 0.77)
Çoğaltılmış veri kümesi AİK Eğrisi (alan = 0.83)
0.0
0.2
0.4
0.6
0.8
1.0

6. Sonuçlar

- Grafikte görüldüğü gibi, çoğaltılmış veri kümesi ile eğitilmiş model performansı, kısıtlı veri kümesi ile eğitilmiş modele göre yüksektir.
- ÇY-UKSB modelinde gözlemlenen başarı artışının sebebi, modelin daha çok veri gözlemleyerek kelimeler arasındaki örüntüyü daha iyi öğrenmesi ve bu örüntünün kelimenin sınıfı ile olan bağıntısını çözmesi olarak gösterilebilir.

6. Sonuçlar

- Temel modellerin başarı gösterememesinin sebebi olarakeğitim verisine katılan sözlük temelli etiketlenmiş verinin gürültü katması gösterilebilir.
- Bu metotların çizdikleri hiper düzlemin veri çokluğuna göre değişiklik göstermektedir. Bu hiper düzlem çoğunlukta olan veriyi ayrıştırmada başarılı olabilir fakat test veri setinin ortalama özelliklerini yansıtmada başarılı olamamaktadır.
- Kelime vektörlerinin ortalamasının alınması, kelimeler arası anlam ve örüntü kaybına sebep olmaktadır

Teşekkürler