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

Özetçe

Gözetimli öğrenme tekniği ile geliştirilen derin öğrenme modelleri, duygu tespitinde yüksek başarı elde etmektedir. 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. Sözlük temelli eğitim için toplanan veri kümesi içerdikleri ASCII emoji karakterlerine göre sınıflandırılmışlardır. Kullanılan derin öğrenme modeli, *Çift Yönlü Uzun Kısa Soluklu Bellekler* (ÇY-UKSB) içeren bir mimariye sahiptir. Önerilen yöntem, modelin performansını F1 metriğinde yaklaşık %4.0, eğri altı alan metriğinde ise (EAA) yaklaşık %7 artırmaktadır.

Abstract

Supervised deep leaning models show high performance on sentiment detection tasks. Deep learning models require high number of labeled data to show high performance. Insufficient data is the main target of the paper. For boost on performance, training set is merged with data set which is created based on lexicon-based features. Then deep learning model is trained on merged data set. Classification samples in this data set is based on textual emoji characters. Suggested method trains model architecture formed by bidirectional long-short term memory units (Bi-LSTMS). Suggested method boosts the performance of model by 4% on F1 score and 7% on area under curve (AUC).

Önerilen Yöntem

Sunulan yöntem dört aşamadan oluşmaktadır. İlk aşama veririnin sözlük temelli yöntem ile çoğaltılmasıdır. Bu aşamada dikkat edilmesi gereken nokta, etiketli verinin test kümesinin hazırlanmasıdır. Performans artışı bu küme üzerinde gözlemlenecektir. Bu sebeple model bu kümeden örnekler almamalıdır. Birleştirilen iki veri kümesi ikinci aşamada bir dizi ön işlemlerden geçmelidir. Üçüncü aşamada verinin matematiksel gösterimi yer almaktadır. Modele verilen her *tweet* cümlesi belirli bir dizi uzunlukta kelime vektörlerine dönüştürülmelidir. Önceden eğitilmiş gömme vektörleri *Fasttext* modelinin Türkçe *Wikipedia* kaynakları üzerinde eğitilmesi sonucu elde edilmiştir. Dördüncü ve son aşama ise verinin *İki Yönlü Kısa-Uzun Soluklu* modelinde eğitilmesidir. Şekilde önerilen yöntemin aşamaları gösterilmiştir.

A. Sözlük Temelli Veri Etiketlemesi

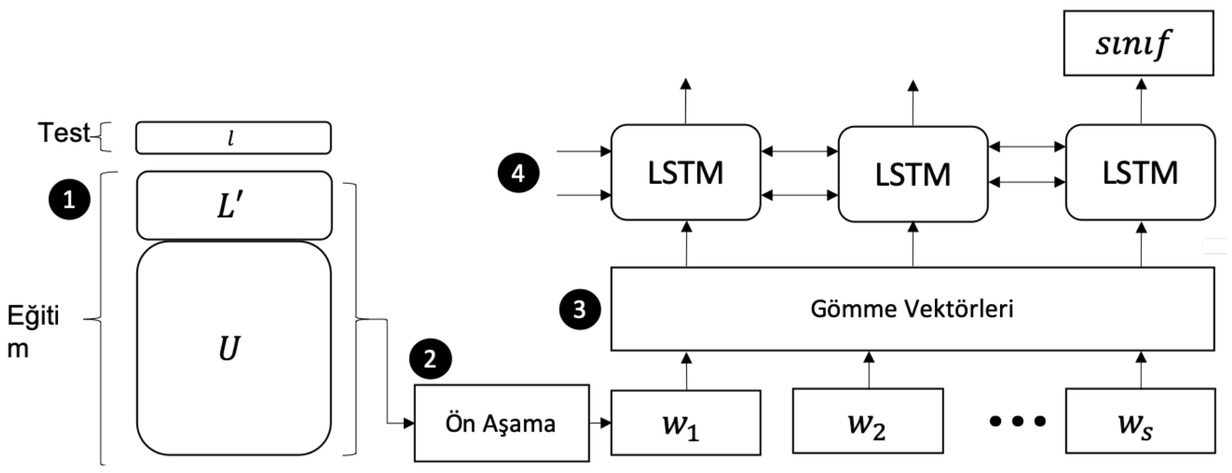
Sözlükte bulunan emoji karakterleri ASCII Emoji ve Etiketleri Tablosunda gösterilmiştir. Örneğin, pozitif karakterleri içeren *tweet* metinleri, pozitif sınıfına atanmıştır.

ASCII Emoji ve Etiketleri

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

B. Ön İşleme

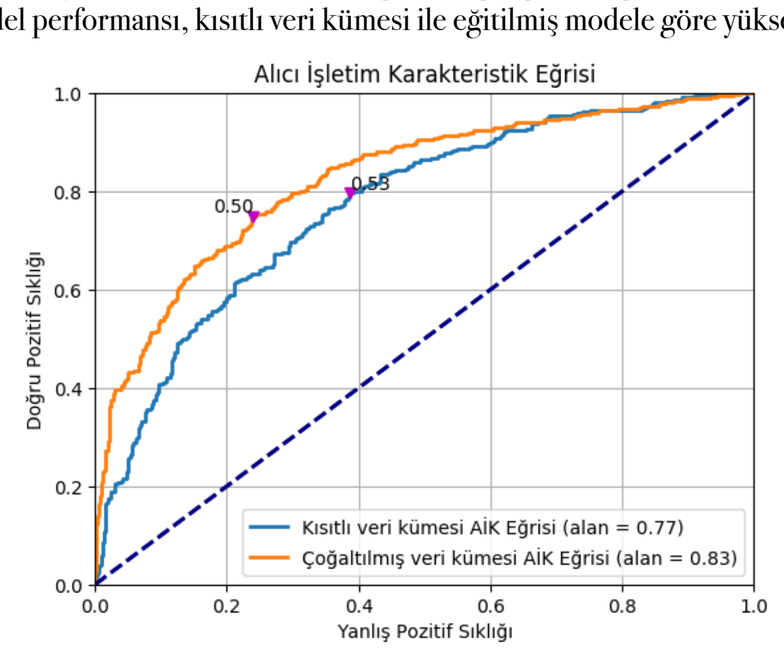
Kelimeler arası anlam bütünlüğünün yakalanabilmesi ve modelin metne duygu ataması yapabilmesi için modele verilen her girdinin anlamsal niteliği bulunmalıdır. Anlamsal niteliği bulunmadığını kabul ettiğimiz, URL, sosyal medya etiketi(#); kişi, lokasyon atıfları, yirmi karakterden fazla ve de üç karakterden az olan kelimeler, noktalama işaretleri ve de sayılar çıkartılmıştır. Bunlara ek olarak gereksiz kelime, *stopword*, olarak adlandırılan kelimeler çıkartılmıştır. İkinci verinin etiketlemesi ASCII emoji ifadelerine göre yapıldığı için, bu ifadeler metinlerden çıkartılmıştır.



Dört aşamalı önerilen yöntem. İlk aşamada, etiketli veriden L, test l ve eğitim L', verisinin çıkartılması; sonra eğitim verisinin sözlük temelli etiketlenmiş veri U ile birleştirilmesi. İkinci aşamada, birleştirilen U ve L' verisinin ön işlemi. Üçüncü aşamada, kelime dizilerinin $w_1, w_2, \ldots\ w_s$ gömme vektörlerine dönüştürülmesi. Son aşamada, kelime vektörlerinin *Çift Yönlü Uzun Kısa Soluklu Bellekler* modeline sırasıyla verilmesi. Model, dördüncü aşamanın sonunda, girdiye *sınıf* ataması yapmaktadır

Sonuçlar

Performans karşılaştırması için eğri altı alan (EAA) değerlendirme ölçütü kullanılmıştır. Aşağıda alıcı işletim karakteristik eğrisi gösterilmiştir. Sınıflandırma performansları bu grafiğe bakılarak yorumlanabilir. Grafikte görüldüğü gibi, çoğaltılmış veri kümesi ile eğitilmiş model performansı, kısıtlı veri kümesi ile eğitilmiş modele göre yüksektir.



Kullanılan yöntemin makine öğrenmesi yöntemlerindeki performansını gözlemlemek amacıyla iki farklı model üzerinde deneyler yapılmıştır. İlk model, Doğrusal Destek Vektör Makineleridir (DDVM). Bu model etiketli veri bildirisinde kullanılmıştır. İkinci model ise Rastgele Karar Ağaçlarıdır (RKA). Modellerin performansı aşağıdaki tabloda gösterilmiştir.

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**

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.