Haber Makalelerinden Protesto Bilgilerinin Çıkartılması

Ali SAFAYA
Bilgisayar Mühendisliği Bölümü
Sakarya Üniversitesi
alisafaya@gmail.com

DANIŞMAN
Dr. Öğr. Üyesi Mustafa AKPINAR

Giriş

Büyük Veri, son zamanların en yaygın teknoloji kavramlarından birisidir. Verilerin doğrudan insanlar tarafından işlenmesi neredeyse imkansız bir hale gelmiştir. Kaldı ki verinin büyüme hızı üstel bir şekilde artmaktadır. Veriler genel olarak iki ana türden oluşmaktadır, Görüntü ve Metin. Bu da Metinlerin evrensel olmamasından kaynaklanmaktadır. Metinler doğal dil (Arapça, İngilizce, Türkçe gibi) veya Sembolik diller (Programlama ve İşaretleme dilleri gibi) veya birden fazla dilin birleşiminden oluşabilir. Doğal Dil İşleme ise hesaplamalı dilbilimi ve yapay zekanın ortak çalışma alanlarını kapsayan bilgisayar bilim dalıdır.

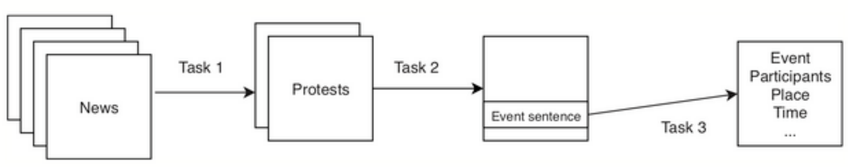
Bu çalışmada Koç Üniversitesi ve Avrupa Araştırma Konseyi (European Research Council) iş birliğinde düzenlenen Emerging Markets Welfare sosyal araştırma projesinin bir parçasıdır.

EMW araştırma projesi, gelişmekte olan piyasa ekonomilerinde yeni bir refah rejimi belirlemeyi ve neden ortaya çıktığını açıklamayı amaçlamaktadır. Proje Brezilya, Çin, Hindistan, Endonezya, Meksika, Güney Afrika ve Türkiye'yi iki hipotezi test etmek için karşılaştıracak: I. Gelişmekte olan piyasa ekonomileri, geniş kapsamlı ve deforme edici sosyal yardım programlarına dayanan, küresel kuzeyin liberal, korporatist ve sosyal demokratik refah rejimlerinden farklı yeni bir refah rejimi oluşturmaktadır. II. Yeni refah rejimi, temel olarak, hükümetler için ikili bir tehdit ve destek kaynağı oluşturan, yoksulların artan politik gücüne bir cevap olarak ortaya çıkmaktadır.

Bu projede CLEF (Conference ands Labs of the Evaluation Forum) kapsamında oluşturulan ve Koç Üniversitesi koordinatörlüğünde yürütülen ‘Lab on ProtestNews’ çalışmaları yapılmıştır.

ProtestNews laboratuvarının temel amacı, birçok ülkedeki haber makalelerinden protesto bilgilerinin çıkarılmasıdır. Özellikle isyanlar ve toplumsal hareketlerle nitelendirilen ve çekişmeli politikalar çerçevesinde olan olaylara odaklanılmıştır. Amaç bir ülkeden alınan verilerle metin sınıflandırma ve bilgi çıkarma araçlarını geliştirmek ve bunları farklı ülkelerden gelen veriler üzerinde test etmektir. Metin verileri İngilizcedir ve Hindistan, Çin ve Güney Afrika'dan toplanmıştır.

Yapılan Çalışmalar



Şekil 1: Lab on ProtestNews kapsamında planlanan araştırma akışı

ProtestNews laboratuvarı çalışmaları üç alt göreve ayrılmıştır:

- **Task 1**: Protesto olaylarıyla ilgili haber makaleleri ve diğer herhangi bir haber makalesi arasında ayrım yapmayı amaçlayan makale sınıflandırma görevi.
- **Task 2**: Olay cümle tespiti, bir olay tetikleyicisini veya sözünü içeren olay cümlelerini belirleme görevi.
- **Task 3**: Olay bilgisi çıkarma, olayla ilgili bilgileri çıkarma görevidir.

Bu projede kapsamında ise ilk ikisine değinilmiştir.

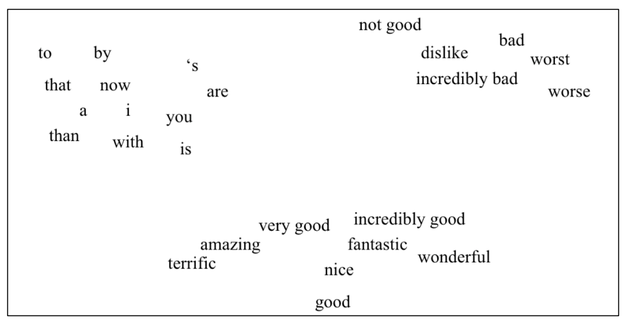
Kullanılan Yöntem

Task 1 ve Task 2 için izlenen yol hemen hemen aynıdır. İlk adımda crawler yardımıyla gazete sitelerinden makalelerin metinleri elde eldilmiştir. Ardından bu metin verilerinin temizlenme işlemi gerçekleştirilmiştir. Sonra da her Task için özel yapay zeka modelleri tasarlanmıştır ve değerlendirmeler yapılmıştır;

Şekil 2: Gerçekleştirme planı

Derin Öğrenme ve End-2-End Modeller

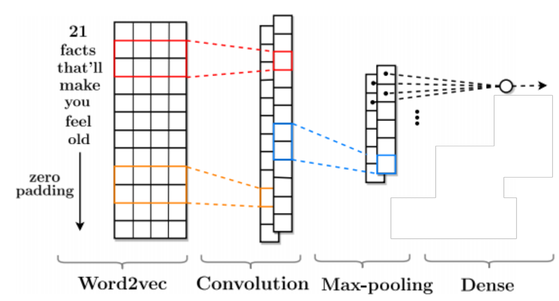
Yapısal makine öğrenmesi mimarilerindeki tekniklerin hepsini tek bir modele gömen Derin Öğrenme tabanlı sistemler, doğal dil işleme kullanılarak yapılacak bir çok işlemi (Özellik Çıkarma, Etiketleme ..vs gibi işlemler) tek bir yapay sinir ağı üzerinden öğrenmeye çalışır. Bu Derin Öğrenme modellerinin girişlerinde ham metinlerin yerini Kelime Vektörleri (Word Embeddings) almaktadır. Vektör anlambilimi fikri, bir kelimeyi çok boyutlu anlamsal bir uzayda bir nokta olarak temsil etmektir. Kelimeleri temsil etmek için vektörler genellikle gömme (embeddings) olarak adlandırılır, çünkü kelime belirli bir vektör uzayına gömülüdür.



Şekil 3: Bazı kelimeler ve cümleler için vektör gömme işlemlerinin (t-SNE) ile iki boyutlu izdüşümü

Task 1 İçin Geliştirilen Model

Task 1 görevi doküman sınıflandırma işlemi sınıfına girmektedir. Burada modelden istenilen işlev dokümandaki (Haber makalesindeki) metni okuyup bir bütün olarak (Protest veya Non-Protest) etiketlemesidir. Bu Task için en iyi performans Kıvrımlı Sinir Ağları (Convolutional Neural Network) kullanılarak elde edilmiştir. İlk olarak metinlerdeki kelimelerin vektör değerleri hazır eğitilmiş bir modelden alınmıştır. Bu vektör değerleri fasttext algoritması ile ingilizce wikipedia haber verilerini kullanılarak eğitilmiştir.



Şekil 4: Task 1 için geliştirilen CNN Derin Öğrenme modeli

Bu model sıra ile Convolution 1D, Max Pooling 1D olarak 6 adet katmandan oluşmaktadır. Hindistanlı kaynaklardan alınan 3430 eğitim ve 457 geçerleme veri örnekleri kullanılarak eğitilen bu model yine Hindistanlı kaynaklardan alınan 687 makale üzerinde test edilmiştir. Ayrıca modelin genellenebilirliğini ölçmek için Çinli kaynaklardan alınan 1801 makale üzerinde test edilmiştir. Sonuçlar aşağıdaki tabloda gösterilmiştir.

Task 1	Training-India	Validation-India	Test-India	Test-China
F1	0.9282	0.8364	0.7928	0.5755
Precision	0.8791	0.7797	*	*
Recall	0.9831	0.9020	*	*

Tablo 1: Task 1 Modelinin performans analizi. (*) Saklı tutuldu

Task 2 İçin Geliştirilen Model

Task 2 görevi daha çok dizin sınıflandırma işlemi sınıfına girmektedir. Buradaki modelde yine Kelime Vektörleri kullanılmıştır. Cümle (Kelime dizini) sınıflandırma işleminde girişlerin sırasının önemli olduğu için bu sıraya önem veren bir Derin Öğrenme modeli tercih edilmiştir. Dizin sınıflandırma işlemlerinde Yinelemeli Sinir Ağları (Recurrent Neural Network) ve türevleri özellikle de LSTM ve BiLSTM en iyi performansı vermektedir. Geliştirilen modelde BiLSTM kullanılmıştır;

Şekil 5: Task 2 için geliştirilen BiLSTM Derin Öğrenme modeli

Bu model her yönde 64 LSTM hücresinden oluşmaktadır. Hindistanlı kaynaklardan alınan 5885 eğitim ve 663 geçerleme veri örnekleri kullanılarak eğitilen bu model yine Hindistanlı kaynaklardan alınan 1107 makale cümlesi üzerinde test edilmiştir. Ayrıca modelin genellenebilirliğini ölçmek için Çinli kaynaklardan alınan 1235 makale cümlesi üzerinde test edilmiştir. Sonuçlar aşağıdaki tabloda gösterilmiştir.

Task 2	Training-India	Validation-India	Test-India	Test-China
F1	0.6127	0.6116	0.6274	0.4820
Precision	0.6667	0.7115	*	*
Recall	0.5668	0.5362	*	*

Tablo 2: Task 2 Modelinin performans analizi. (*) Saklı tutuldu