Distributed City Sentiment

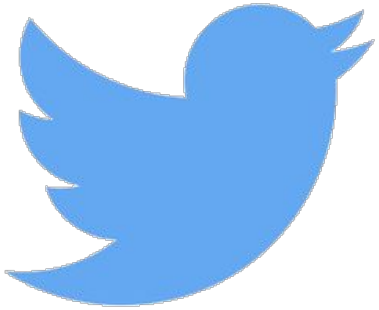
Rabia Arslan
Muhammed Said Zengin

Amaç

Günümüzde yaygın olarak kullanılan yapay zeka algoritmalarının dağıtık veri teknolojileri kullanılarak hızlandırılmasıdır.

Çözmeye çalıştığımız problem, gerçek zamanlı stream veri üzerinde tahminleme hızının artırılmasıdır. (bu çalışmada hızlandırma yöntemi stream veri üzerinde değil, 16 milyonluk sabit bir veri üzerinde yapılacaktır.)

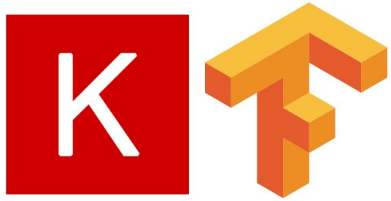
Veri



Toplam 40.500 kullanıcı ve 16.116.035 tweet elde edilmiştir.

En az tweet 88.390 adet Ardahan'da, en çok tweet 1.054.767 adet Ankara'da bulunmaktadır. Toplam 16.116.035 tweet vardır ve bir şehirdeki ortalama tweet sayısı 198.964 olarak hesaplanmıştır.

Duygu Analiz Modeli



Duygu analizi modeli oluşturmak için TensorFlow Keras kullanımını kolaylaştıran ktrain kütüphanesi kullanılmıştır. Bu kütüphane ile Huggingface üzerinde açık kaynak olarak paylaşılan BERTurk modeli ince ayar yapılmıştır. BERTurk modeli eğitilirken 35 GB ve 44 milyon token boyutundaki bir veri kullanılmıştır. BERTurk Türkçe için özel olarak hazırlanmış bir BERT modelidir. BERTurk modeli ince ayar yapılırken, 2000 adet duygu analizi için etiketlenmiş veri kullanılmıştır. Learning rate olarak 5e-5, batch size olarak 6 kullanılmıştır. 3 epoch eğitilmiştir. Tüm random seed'ler 42 olarak ayarlanmıştır. Train verisi %90, test verisi %10 olarak bölünmüştür. Sonucunda ise test verisi üzerinde %93.7 F1 skoru elde edilmiştir.

Dağıtık Tahminleme

Dağıtık hesaplama yapmak için PySpark kullanılmıştır. Baseline için kullanılan aynı sunucu, aynı veri ve aynı duygu analizi modeli kullanılmıştır. Tek fark olarak tahminleme dağıtık olarak yapılmıştır. Burada yapılan ilk deney, tüm veri üzerindeki maksimum hızlanma yüzdesini bulma olmuştur. İkinci deneyde ise N adet veri üzerinde, farklı dağıtma (partition) sayılarındaki hızlanma oranları incelenmiştir.

Sonuçlar

Duygu analizi için veri çekilmiş, veri etiketlenmiş ve bir model eğitilmiştir. Ardından Twitter üzerinden 16 milyon veri çekilmiştir. Bu veriler duygu analizi modeli kullanılarak hem normal bir şekilde hem de dağıtık bir şekilde tahmin edilmiştir. Bu sonuçlara göre dağıtık veri tahminlemenin normal tahminlemeye göre oldukça hızlı çalıştığı gözlemlenmiştir. Ayrıca duygu analizi sonuçlarına göre şehir bazlı mutluluklar ölçülmüştür.