Makine öğrenmesi yöntemleri ile türk

lehçelerinin tahminlenmesi

Yasin ŞAHİN1

*1Kocaeli Üniversitesi  
Bilişim Sistemleri Mühendisliği*

1181307026@kocaeli.edu.tr

***Abstract*—**

***Keywords***— **lehce tahmini, ses bölümlemei**

1. Introduction

Makine öğrenmesi modelleri sonrası rapor bitiminde yazılacaktır.

1. Literatür taraması

Mustafa, türk lehçeleri arasındaki benzer kelimelerin eş değerlik durumu hakkında bir çalışma yapmıştır [1]. Çalışmasında türk lehçelerinin dini, coğrafi, kültürel değişimlerinin kelimeler üzerindeki değişimlerini incelemiştir. Bire bir veya bire çok eş anlamlısı bulunan kelimelerin köklerine inerek tam benzerlik veya kabul edilebilir benzerliklerini göstermiştir. Hasan, yaptığı çalışmada parkinson hastalarının ses öznitelikleri üzerinde çeşitli makine öğrenmesi yöntemleri uygulamıştır [2]. PDC veri seti üzerinde TBA ve DDA boyut indirgeme yöntemi uygulayarak doğruluk değerlerini karşılaştırmıştır. Ahmet rulman arızalarının tespit edilebilmesi için arızalı ve sağlam rulman seslerini toplamış ve fourier dönüşümü kullanarak özniteliklerini çıkarmıştır [3]. Ses sinyallerini on farklı sınıflandırma yöntemi ile gerçekleşmiş ve doğruluk değerlerini karşılaştırmıştır.

1. Ses dosyalarının toplanması

Ses dosyalarının elde edilmesi için öncelikle arama kriterleri belirlenmiştir. Arama kriterleri için sınıflandırma işlemi yapılacak lehçelerin kendi dillerinde ‘haber’ içeriği taşıyan videolar aranmıştır. Ses ve video için sonsuz kaynağa sahip olan youtube aracı kullanılmıştır. Youtube üzerinde Azerice, Kazakca, Kırgızca, Özbekce, Tatarca, Türkmence, Uygurca dillerinde haber videoları araştırılmış ve spiker sunumu bulunan 20-30 dakikalık video linkleri toplanmıştır. Toplanan linklerin örnek görüntüsü Figür 1’de verilmiştir. Link bağlantıları kaynakların altında url olarak verilmiştir. Ses dosyaları toplama aşamasında sınıflara eşit ses dosyası düşmesi için ses süreleri dikkate alınmıştır. Ses dosyalarının toplanması için toplanan linkler bir dosya ile python dilinde yazılmış olan scripte girdi olarak verilmiştir. Hazırlanan script aracılığı ile linkler “wav” formatına dönüştürülerek yerel dosya hiyerarşisine kayıt edilmiştir. Ham verilerin kayıt aşamasında her bir sınıfa ait dosyalar oluşturulmuştur.

| **Link** | **Lehce** |
| --- | --- |
| https://youtu.be/SNxlwSeu2Ng | Azerice |
| https://youtu.be/I5ijLszpAGY | Tatarca |
| https://youtu.be/spB2wkwISXE | Turkmence |
| https://youtu.be/Sfn9CnpvYcs | Uygurca |
| https://youtu.be/-3A28X5CKw0 | Kazakca |
| https://youtu.be/0yJsBMgV1hM | Kırgızca |
| https://youtu.be/7q1qGWCW7sk | Özbekce |

Fig. 1 Link ve yöreye ait linklerin örnek gösterimi

1. Ses dosyalarının analiz edilmesi ve temizlenmesi

Toplanan ses dosyalarından her bir lehçeye ait 10 saniyelik 20’şer adet veri alınmıştır. Alınan ses kesitleri Audacity yazılımı aracılığı ile incelenmiştir. Her bir ses dosyası aynı proje içerisinde açılarak spektrogramları incelenmiştir. Veriler üzerinde bilgi sahibi olduktan sonra seslerin bölümlendirilmesi adımına geçilmiştir. Spektrogram analizinin ekran kesiti Figür .2’ de verilmiştir.

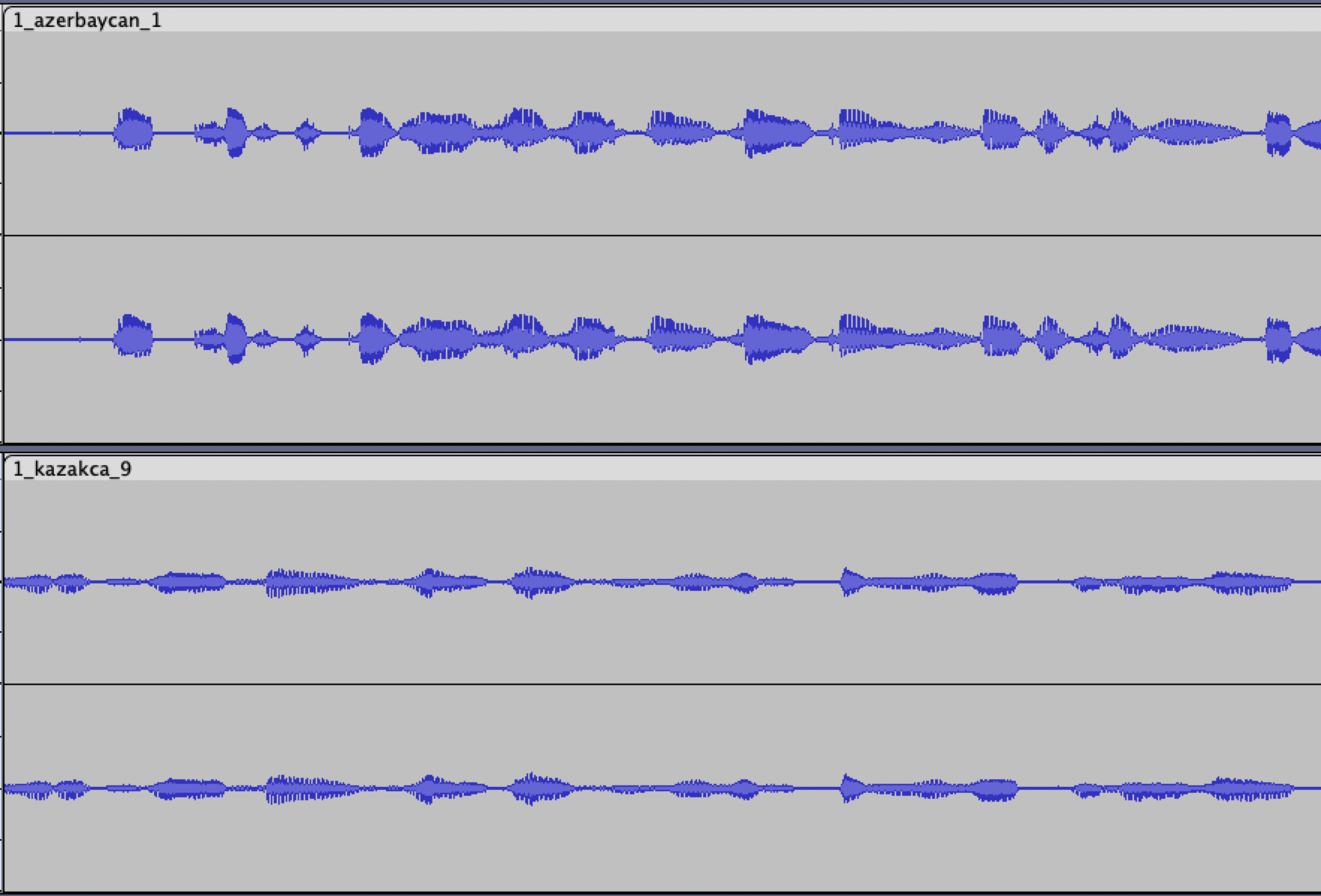


Fig. 2 Audacity yazılımı üzerinde spektrogram analizi

Örnek olarak alınan ses dosyalarını MFCC ve Spectral centroid niteliklerini çıkararak grafik ortalama değerlerini grafik üzerinde gösterdim. Ses dosyaları arasındaki korelasyonu gözlemleyerek veri seti hakkında daha detaylı bilgi edinmiş oldum. Bu çalışma sonucunda grafiğe dökülmüş görseller Figür 3 ve 4 de gösterilmiştir.

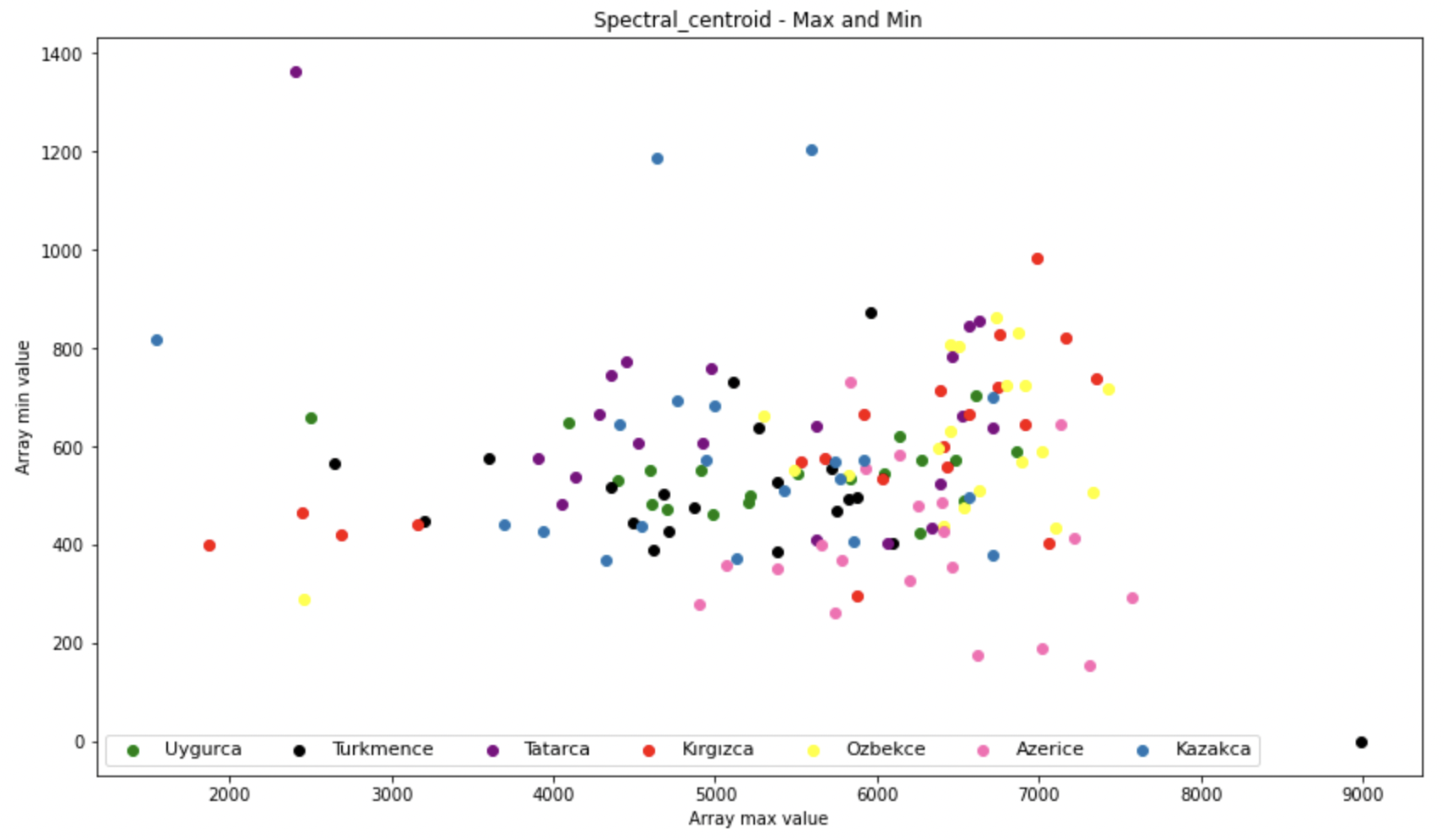


Fig. 3 Spectral centroid ortalama değerleri üzerinde korelasyon

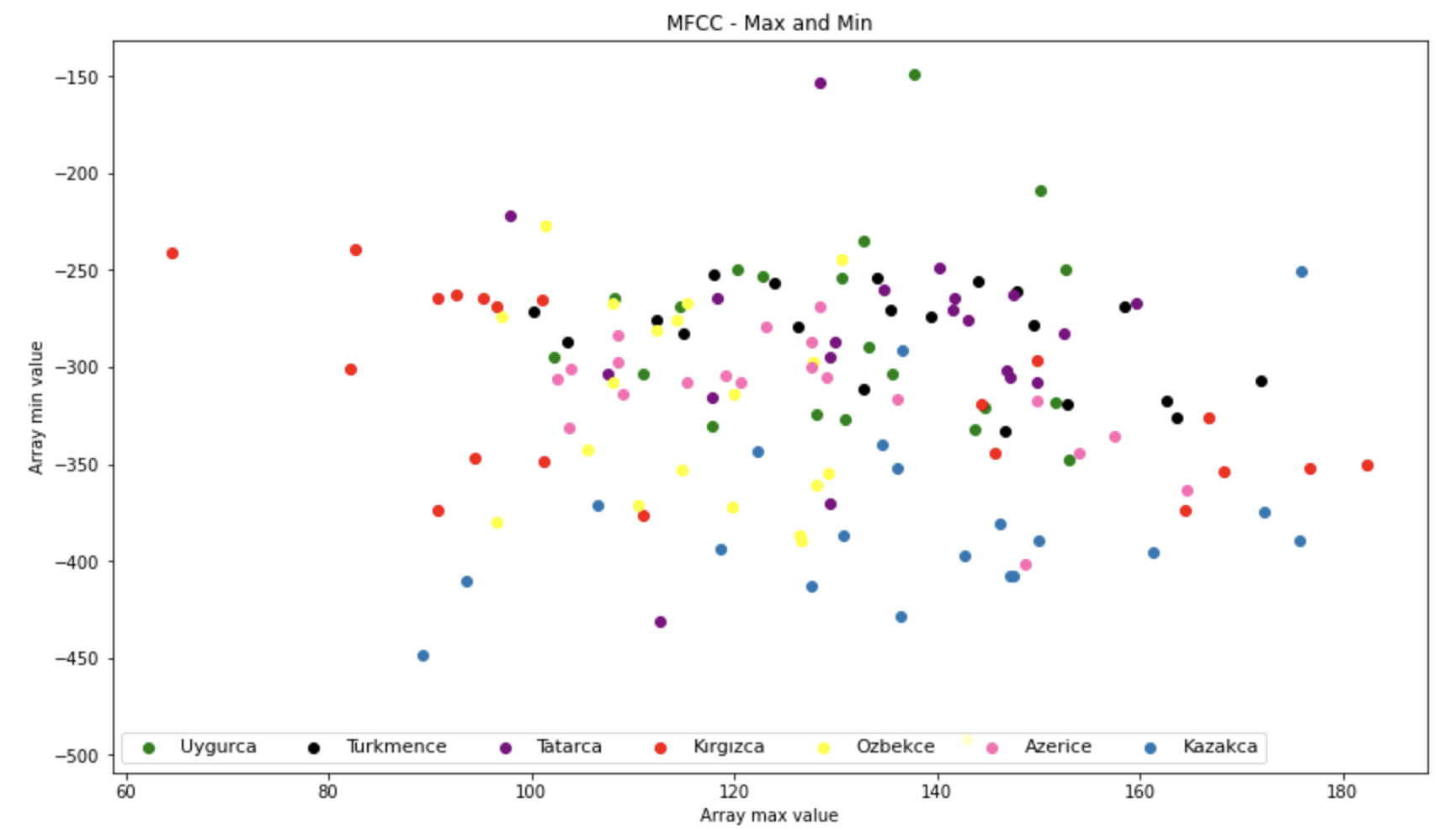


Fig. 4 MFCC ortalama değerleri üzerinde korelasyon

1. Ses dosyalarının bölümlenmesi

Etiketlenmiş ve indirilmiş olan ham ses verileri üzerinde bölümlendirme yapmak gereklidir. Bu işlem için python dilinde script yazılmış ve librosa kütüphanesi kullanılarak 4 saniyelik parçalara ayrılmıştır. Parçalara ayrılan ses dosyaları ve etiketleri csv dosyasına kayıt edilmiştir. Kayıt edilirken kullanılan isim ve sınıf özellikleri korunmuştur. Sınıflara ait dosyalar işaretlenmiş ve tek bir klasör altında toplanmıştır. Kayıt edilirken kullanılan isimlendirme ve dosya yolu Figür 5’de tabloda gösterilmiştir.

| **file\_name** | **class\_name** |
| --- | --- |
| 4\_azerbaycan\_12.wav | Azerice |
| 2\_turkmence\_42.wav | Turkmence |
| 1\_kirgizca\_3.wav | Kırgızca |
| 3\_ozbekce\_44.wav | Özbekce |

Fig. 5 Bölümlenmiş ses dosya örnekleri

Bütün parçalanan ses dosyaları yeniden isimlendirilerek bir klasör altında toplanmıştır. Herbir lehçe için toplanan ses dosyaları sayısı Figür.6’de verilmiştir.

| **Lehce** | **Veri adedi** | **Ses uzunluğu (saniye)** |
| --- | --- | --- |
| Azerice | 369 | 4 |
| Kazakca | 685 | 4 |
| Kırgızca | 524 | 4 |
| Özbekce | 372 | 4 |
| Tatarca | 610 | 4 |
| Uygurca | 1074 | 4 |
| Türkmence | 998 | 4 |
| Toplam | 4632 | 4 |

Fig. 6 Ses dosyaları sayısı

Sonuç

Kaynaklar

1. Uğurlu, Mustafa. "TÜRK LEHÇELERİ ARASINDA BENZER KELİMELERİN EŞ DEĞERLİK DURUMU." *Electronic Turkish Studies* 7.4 (2012).
2. BADEM, Hasan. "Parkinson Hastaliğinin Ses Sinyalleri Üzerinden Makine Öğrenmesi Teknikleri ile Tanimlanmasi." *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi* 8.2 (2019): 630-637.
3. TEKTAŞ, Ahmet Burak, and İsmail KIRBAŞ. "RULMAN ARIZALARININ MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE SES ANALİZİ YAPILARAK SINIFLANDIRILMASI."

Linkler

[link][https://www.youtube.com/playlist?list=PLyY6UmEUDM\_lwIt1uC0txRo5QHH-Qqqa](https://www.youtube.com/playlist?list=PLyY6UmEUDM_lwIt1uC0txRo5QHH-QqqaO)