SETSCI Conference Indexing System,
Volume 2 (2018), 380-385

ISAS 2018
I[st] International Symposium on Innovative Approaches in Scientific Studies

DUYGUSAL TWITTER ANALİZİ

Tohid JAVADI
Necaattin BARIŞÇI
Bilgisayar mühendisliği, Gazi üniversitesi, Türkiye
[*]*Corresponding author: tohid.javadi@gmail.com*
[+]*Speaker: tohid.javadi@gmail.com*
Presentation/Paper Type: Oral / Full Paper

Özet **-** İnsanlar sanal âlemde her geçen gün sosyalleşmek için paylaşım yaparak, duygu ve düşüncelerini değişik platformlarda diğer insanlara duyurmakta ve onlarla paylaşım yapmaktadır. İnsanların duygu ve düşünceleri onların doğadaki olaylara ve bu olayların öznelerine karşı davranış ve söylemlerini şekillendirmektedir. Özellikle sanal platformlardan Twitter, her geçen gün popülerliği artmakta ve hızlı bir mikro blog tarzında kullanıcılarına anlık paylaşım yapma ve içerik görüntüleme imkânı vermektedir. Duygusal Twitter Analizi çalışması; insanların olaylar karşısındaki duygu ve düşüncelerinin Twitter üzerindeki yapmış oldukları paylaşımların içeriklerinden belirlenip belirlenemeyeceği üzerinedir. Makine öğrenimi, Yapay Zekâ ve önceden oluşturulmuş Corpus adı verilen büyük veriler sayesinde %98.8 oranında sınıflandırmamız mümkün olmuştur. Bu sınıflandırma sayesinde hedef kitlenin ya da kişinin eğilimleri önceden tahmin edilebilmekte ve bu da politika yapıcıları ile tanıtım-pazarlama sektörü için yapılacak yatırımlar ve söylemler hakkında bir fikir verebilmeyi mümkün kılmaktadır. Bu çalışmada önceki çalışmalardan farklı olarak tek bir algoritma kullanmak yerine üç farklı algoritma kullanılmıştır. Duygusal analizde doğal dil işleme teknikleriyle beraber hangi algoritma daha verimlidir sorusuna cevap vermiştir. Çalışmada 4500 öğe içeren bir veri seti kullanılmıştır. Bu veri setinin 4000 öğesi sistemi eğitmek ve geri kalan kısmı ise sistemin testi için kullanılmıştır. Kullanılan algoritmalar arasında en verimli çalışanı Naïve Bayes olmuştur. Bu algoritma %98.8 oranında doğru sonuç vermiştir. Sırayla Karar ağacı algoritması %93.1 ve N-Gram algoritması ile % 73.9 başarı elde edilmiştir. Bu çalışmada sistemin doğu çalışmasından emin olduktan sonra anlık olarak kullanıcı temelinde Twitler analiz edilmiştir.

Anahtar Kelimeler: Duygusal Zekâ Analizi, yapay zekâ, Makine öğrenmesi, Twitter, Naive Bayes, N-Gram, Karar ağacı

Abstract **-** Human shares lots of thing in cyberspace every day to socialize and shares their sense and ideas in social platforms to announce to others. One's thoughts and senses shape his behaviors and reaction to these events and subjects of these. Twitter, one of the social media platform has a popularity getting bigger in everyday and gives opportunity to its users to share their sense, thoughts, and view contents as a microblog. My work naming "Sentinel Twitter Analysis" is on detecting people's senses and thoughts by their shares in social media platform, Twitter. Thanks to Machine Learning (ML), Artificial Intelligence (AI) and big data naming PreSet Corpus, it could be possible to classify tweets with nearly 98.8% accuracy. In this study, tree different algorithms are used in spite of using only one regarding with prior ones. In Sentimental Analysis, it was answered to the question that "Which algorithm is the most efficient when it is used with Natural Language Processing. A data set containing 4500 tokens is used in study. 4000 of this data set were used to train the system and the rest was used to test the system. The most efficient algorithms among used tree was Naïve Bayes. This algorithm gave 98.8% percent accuracy while Decision Tree Algorithm gave 93.1% and N-grams algorithm gave 73.9% accuracy. After being ensured that system runs correctly, user-based Tweets were analyzed instantly.

Keywords: Sentimental Analysis, Twitter, Naive Bayes, Machine Learning, Artificial Intelligence, N-Gram, Decision Tree

1. GİRİŞ

İnsanlar günümüzde eskiden hiç olmadığı kadar sosyallikten uzak dijital dünyada bu yoksunluklarını ya da ihtiyaçlarını giderme peşindedir. Durum böyle olunca her gün sayıları ve içerikleri hızla artan sosyal medya mecraları ise daha da popüler hale gelmekte ve insanlar günlük yaşamdaki duygu, istek ve arzularını sosyal medyada siber dünyanın dili ile anlatmaya çalışmaktadır. Siber dünya her ne kadar sınırsız sayıda cihaz, altyapı ve kullanıcılardan oluşsa da genel anlamda insanlar çağlar boyunca edindiği bilgi, birikim ve kültürleri ile iletişim kurmaktadır. Bugün kullanılan yüzlerce sosyal medya araçlarında neredeyse kullanılan dil aynıdır. Bu dil kullanıcıların kendi ya da başka kullanıcıların paylaşmış oldukları içerikler hakkında yapmış olduğu yorumlar ya da beğeni-like ya da beğenmeme-unlike şeklinde karşımıza çıkmaktadır. Dahası son dönemde meşhur olan duygu anlatımı için oluşturulmuş özel karakterler-emojiler kullanılmaktadır. Sosyal medya bugün itibariyle neredeyse dünya nüfusunun yarısına ulaşmış ve her yaştan her gruptan insanın eylemde bulunduğu, paylaşım yaptığı dijital bir platform durumundadır. Bu kadar fazla kullanıcı web2.0 etkisi ile yapılan paylaşımlara katkı sağlamakta ya da bizzat paylaşımı kendisi oluşturmaktadır. Bu kadar çok kişinin kullandığı ve bu denli yoğun içeriğin olduğu dijital platformdan anlamlı sonuçlar çıkarmak mümkün müdür? Bu paylaşımlardan elde edilebilecek anlamlı veriler ne işe yarar?

Kısaca Duygu(Sentiment) Analizi(DA denilen sosyal medya ortamlarındaki her türlü içeriğin makine öğrenimi(machine learning, MÖ) ve yapay zekâ algoritmaları(artificial intelliegince-AI) ile işlenerek anlamlı sonuçlar elde etme sürecini ifade etmektedir. Günümüzün en çok tartışılan ve erişilen teknoloji çağının hızına ayak uydurmuş olan Twitter kitleleri en çok etkileyen bir alan haline gelmiştir. İşte "Duygusal(Sentiment) Twitter Analizi" isimli çalışmada bu sosyal medya ortamı hedef alınarak yukarıda sorulan sorulara cevap ararken multidisipliner bir çalışma ile yapılan paylaşımlardan elde edilen anlamlı sonuçların nasıl ve ne şekilde kullanılacağı yönünde Reklam ve Pazarlamadan Siyasete, Spora ve toplumun hemen her alanı ile alakalı ilgili kişilere fikir verme ve genel eğilimleri(trendleri) gösterme amacı taşımaktadır.

Bu çalışmaya benze daha önce Malaysiya'da Aliza Sarlan, Chayanit Nadam ve Shuib Basri tarafından yapılmış çalışmalar bulunmaktadır [1]. Bu çalışmada insanların olumlu veya olumsuz duyguları üzerinde çalışma yapılarak bu ülkede insanların duygu haritasını çıkarılmaya çalışılmıştır. Ayrıca Amerika birleşmiş devletlerinde 2012 seçimlerinde insanların düşüncelerini ve tuttukları siyasi partileri anlayabilmek için Hao Wang tarafından güney California üniversitesinde bir çalışma yapılmıştır [14]. Hindistan Kelkete üniversitesinde Nipun Bhatia ve Sarath Krishna tarafından insanların menfi ve müspet duygularını daha fazla anlayabilmek adına bir çalışma yapılmıştır [15].

2. DUYGU ANALİZİ

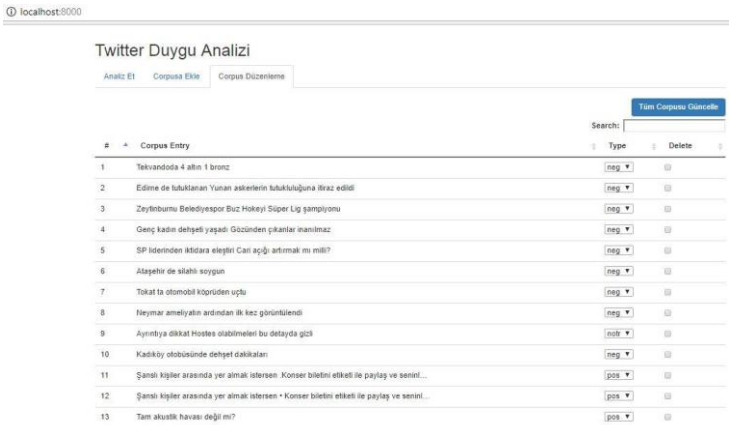
Günlük yaşamda insanlar karşılaştıkları olaylar karşısında karar vermek zorunda kalırlar. Tam da bu esnada yaşanılan sıkıntı kişiler için tam da katlanılmaz bir süreç haline gelmektedir. Hangi arabayı alsam? Hangi model olsa? Pişman olur muyum? Gibi sorular insanları sıkıntıya sokmakta ve yaşam standartlarını düşürmektedir. Duygu analizini bir fikir madenciliği olarak gören Illinois Universitesinden Bing Liu, DA'nin insanların tam da bu kararsızlık içerisindeki durumları için bir fikir verici olarak görülebileceğinden bahsetmektedir [1].

Bu çalışma esnasında kişilerin anlık olarak paylaşım yaptıkları Twitter isimli sosyal medya hedef olarak alınmıştır. İnsanlar duygu durumlarını genellikle kelimelere dökerek ifade ederler [2]. Kişilerin yapmış oldukları paylaşımları müspet, menfi ve tarafsız olarak üç gruba ayırabiliriz. Günlük hayatta fikirler ve paylaşımlar en genel olarak bu şekilde sınıflandırılabilir. Mutluluk, heyecan, neşe, ümit, sevinç gibi duygu barındıran ifadeler müspet iken; kin, nefret, korku, ümitsizlik, savaş, ölüm, hırs gibi duygu barındıran ifadeler ise menfi ve kişilerin objectif olarak durum bildirmeleri yani aktarım cümleleri gibi ifadeler ile diğer iki gruba da ait olamayacak ifadeler tarafsız duygular olarak ifade edilir.

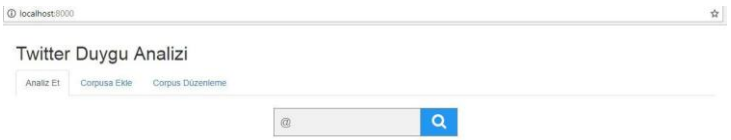
2006 yılında Twitter bugün itibariyle aylık 330 milyon aktif kullanıcısı ve günlük 500 milyon adet tweet paylaşımı ile dünyanın en büyük sosyal medya araçlarından biridir [3]. Saniyede yaklaşık 6 milyon adet tweetin atıldığı bu ortamdan anlık olarak doğru şekilde duygu analizi yapmak bir insan için mümkün değildir. O yüzden böylesine büyük verinin olduğu bir ortamda veriler iyi tanımlanmış bilgisayar destekli Yapay Zekâ Algoritmaları kullanan yazılımların işidir [4].

3. MATERYAL VE YÖNTEMLER

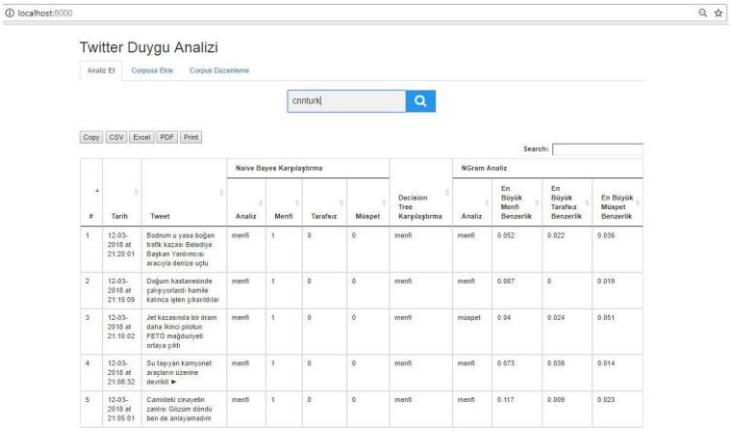
Bu çalışmada günümüzde dil işlemede en çok kullanılan programlama dili olan Python 3.6 [5] ve Django Web Framework [6] tercih edilmiştir. Çalışmanın anlamlı sonuçlar verebilmesi için ilk önce üç duygudan birisiyle etiketlenmiş olan geniş bir sözlük(corpus) oluşturulmuştur. Bu sözlüğün oluşturulması ve anlık olarak herhangi bir Twitter kullanıcısının tweetlerini analiz edebilmek için Twitter Geliştirici API kullanılmıştır [7], Twython [8] isimli Pyhton API ile Twitter verileri çekilmiştir. Programda üç sekme bulunmaktadır. İlk sekmede 1. Resimde görüldüğü üzere girilen twitter kullanıcısının son 200 twiti analiz edilmiş olarak gelmektedir. İkinci sekmede 2. Resimde görüldüğü gibi corpusa dinamik olarak ekleme yapabilmek için bir twitter adresi girilmektedir. Girilen kullanıcıya ait olan tweetler burada sınıflandırıldıktan sonra corpusa etiketlenmiş olarak kaydedilmektedir. Üçüncü sekmede ise 3. ve 4. Resimden anlaşılacağı üzere corpustaki tüm veriler görüntülenebilmekte ve gerektiğinde düzeltme ya da silme işleme gerçekleştirilmektedir. Twitter kullandığımız API sayesinde bize JSON formatında bir veri dönmektedir [9]. Bu veri içinde 140 dan fazla özellik olarak tanımlanan veri mevcut olup UTF-8 olarak kodlanmış veri JSON olarak verilmektedir. Twitterın bu çalışma mantığı ve Türkçe alfabenin UTF-8 içinde olmasından dolayı programımızda Python 3.6 tercih edilmiştir. ASCII formatı kullanan birçok programlama dili ve twitter API Türkçe karakterlerde muhtemel yaratacakları sıkıntılardan dolayı bu tercih yapılmıştır.



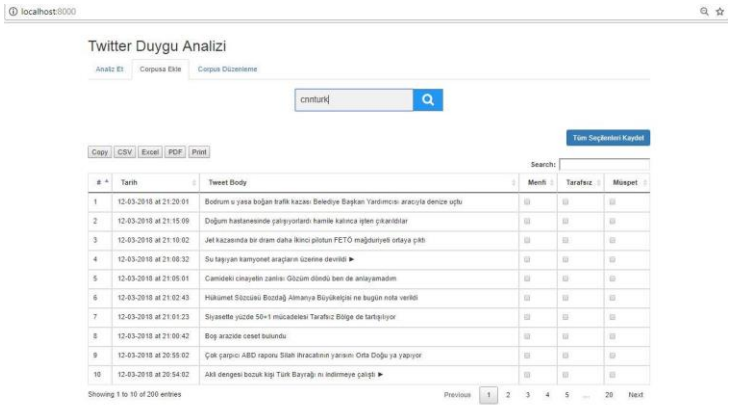
Resim 1. Son 200 tweet analizi



Resim 2. Corpus'a dinamik ekleme



Resim 3. Corpus'daki tüm verilerin görüntülenmesi



Resim 4. Corpus'daki verilerin düzenlenmesi

Programın etkili çalışabilmesi ve doğru sonucu verebilmesi için corpusa eklenecek olan veriler üzerinde temizleme işlemi yapılmıştır. Python programının etkili katar işlemi olan Regex Kütüphanesi kullanılmış ve özellikle noktalama işaretleri, @, #, RT, linkler gibi gereksiz karakterler temizlenmiştir.

Tüm bu işlemlerden sonra Ajax ve Jquery gibi dinamiklik sağlayan teknolojiler sayesinde kullanıcı kendisine gösterilen twitlerden hangi sınıfa uygunsa ona göre corpusa(sözlüğe) kaydeder. Sözlük sistem içinde bir JSON dosyasıdır ve veriler JSON objeleri olarak kaydedilir:

[{"text": "Tekvandoda 4 altın 1 bronz", "label": "pos"}]

3.1 Kullanılan Algoritmalar

Twitter Duygu Analizi işleminde Corpus oluşturulduktan sonraki aşama Makine Öğrenimi gerçekleştirmek ve Yapay Zekâ Algoritmaları ile anlamlı sonuçlar çıkarmaya çalışmaktır. Doğal Dil İşleme Tekniklerinin-Algoritmalarının başında gelen Naive Bayes, Karar Ağacı ve N-Gram Analiz yöntemleri yapılan çalışmalarda doğru sonuç verebilmesi için kullanılmıştır. Üç algoritmaya da aynı veri seti uygulanmış ve aynı girdilere verdikleri sonuçlar tablada sırasıyla karşılaştırılmıştır.

3.1.1 Naïve Bayes [10]

Bayes teorimi bir olayın kendisi ile ilişkili olayların gerçekleşme şartına bağlı olduğunu ortaya koyan istatistiksel bir teoridir. Dil işleme algoritmaları arasında verdiği sonuçlar sayesinde en popüler olan ve doğruluğa en yakın sonuç veren algoritmalardan biridir.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}. \quad (1)$$

Bu karşılaştırmada bir sonraki kelimenin mutlaka bir öncekiyle bağlı şartı göz önüne alınarak yapılan bir karşılaştırma söz konusudur. Bu karşılaştırmaları yapabilmek için TextBlob [11] isimli Python kütüphanesi kullanılmıştır. Bu kütüphane NLTK [12] isimli doğal dil işleme kütüphanesini kullanmaktadır. Bu karşılaştırma sayesinde en son kelime olarak etiket kabul edilmektedir. Etiketlerde ise cümlenin olumlu, olumsuz veya tarafsız olduğu ile ilgili 3 farklı etiket söz konusudur. Her bir tweet için ayrı ayrı sınıflandırma yapılıp tüm etiketlere benzerlik oranı ondalık sayı olarak virgülden sonra 3 hanesi yuvarlanarak kullanıcıya gösterilmektedir.

Naive Bayes karşılaştırma çok küçük kütüphanelerde dahi doğruluk olarak gayet güzel sonuçlar verebilmektedir.

3.1.2 Karar Ağacı (Decision Tree) [13]

Karar Ağacı algoritması oluşturmuş olduğumuz sözlükteki her bir cümleyi ele alarak bir karar mekanizması oluşturmaktadır. İkili (doğru-yanlış Binary) şekilde her bir durum dallanmaktadır. En nihayetinde verilen girdisi bize bu karar ağacı üzerinde yürüterek en doğru sonucu vermeyi amaçlamaktadır.

Bu algoritmanın kullanımı da TextBlob ve NLTK Kütüphanesi ile mümkün olmaktadır. TextBlob kütüphanesi NLTK sayesinde sahip olmuş olduğu DecisionTreeClassifier isimli sınıfı sayesinde kullanıcılara bu algoritma ile sınıflandırma yapabilme imkânı verebilmektedir.

Bu sınıfın bir özelliği de yaratılan DecisionTree yapısını preetyprint isimli metodu ile bizlere verebilmesidir.

Dikkat çeken en önemli özellik DecisionTree algoritmasının küçük kütüphanelerde bazen yanlış sonuçlar verse de kütüphanenin büyümesi halinde doğruluk değeri 1'e yani %100'e yaklaşmaktadır.

3.1.3 N-Gram Analizi

N-Gram analiz verilen texti n-gramlara ayırarak en çok benzer şekilde bize sözlükteki kelimeyi vermektedir. Bu haliyle N-Gram analizini kullanmak çok mümkün gözükmemektedir. Çünkü kütüphanede tüm veriler birlikte tutulmaktadır. Bu yüzden analiz aşamasında kütüphanedeki verileri sanal olarak 3 farklı sözlüğe çevirmekte ve her bir kütüphanedeki en çok benzer olanının oranları karşılaştırılmaktadır. Böylece analiz edilmek istenen tweetin en çok hangi label'e benzediği hakkında bize fikir vermektedir.

4. ALGORİTMALARIN KARŞILAŞTIRILMASI

Her üç algoritma da TextBlob ve NLTK kütüphaneleri sayesinde kullanılmakta ve analiz edilmek istenen kullanıcının tweetleri tek tek sınıflandırılmaktadır (Tablo 1).

Küçük kütüphanelerde en doğru sonuçlar Naive Bayes ile alınırken kütüphane büyüdükçe Karar ağacının daha doğru cevaplar verdiği ortaya çıkmaktadır. N_gram analizde ise sanal olarak

oluşturulmuş kütüphanelerden en çok benzeyen tweetlerin oranları karşılaştırılmakta ve netice de ençok benzeyenin label'i analiz edilen tweetin label'ı kabul edilmektedir.

Tablo 1. Kullanılan algoritmaların karşılaştırılması

Naïve Bayes	**Decision Tree**	**N-Gram**
%98.8	%93.1	%73.9

5. SONUÇ

Duygusal Twitter Analizi, gelişen teknoloji sayesinde insanların ilişkilerinin siber âleme kayması ve neredeyse tüm duygu ve düşüncelerinin bu platformlarda ifade etmelerinin bir sonucu olarak bu ortamda paylaşılan içeriklerden anlamlı sonuçlar çıkarma amacını taşımaktadır. Sosyal Medya araçlarından neredeyse en hızlı ve etkileme açısından en etkilisi olan Twitter çalışmasında sırasında model alınmıştır. Twitter içeriği olan Tweetlerden Makine Öğrenimi ve Yapay Zekâ Algoritmaları kullanılarak bir Corpus oluşturulmuştur. Bu Corpus Big Data olarak adlandırılan büyük veri tipleri içermektedir. Burada yapılması gereken algoritmalar ve bilgisayarların işlem gücü kullanılarak anlamlı sonuçlar çıkarmadır. Önceden oluşturduğumuz Corpus makinenin etiketlerden yola çıkarak kendisine sorulan içeriklerin ne olduğunu sınıflandırmaya yaramaktadır. Elde edilen sonuçlar bu şekilde bir analizin mümkün olacağını ortaya koymaktadır. Duygusal Twitter Analizi sayesinde amaç her ne kadar kullanıcıların paylaşımlarındaki duygu tiplerinin ortaya çıkarılması olsa da, ilerleyen teknik ve alt yapı ile sosyal medya ortamında işlenen suçlar, çocuk pornografisi, teröre destek ve eleman kazanma gibi birçok kriminal vakaların çözüme kavuşturulması için hem politika yapıcılara hem de bu işin uzmanlarına gerekli desteği verecektir.

KAYNAKLAR

[1] Liu B., Sentiment Analysis and Subjectivity, Handbook of Natural Language Processing, 2010
[2] PSYCHOLOGICAL ASPECTS OF NATURAL LANGUAGE USE: Our Words, Our Selves
[3] Aslam S., Twitter by the Numbers: Stats, Demographics & Fun Facts, 2018, https://www.omnicoreagency.com/twitter-statistics/ erişim tarihi:12.03.2018
[4] Akgül E. S. , Ertano C, Diri B.,Twitter verileri ile duygu analizi, Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 22(2), 106-110, 2016
[5] https://www.python.org/ , erişim tarihi:12.03.2018
[6] https://www.djangoproject.com/ , erişim tarihi:12.03.2018
[7] https://apps.twitter.com , erişim tarihi:12.03.2018
[8] https://twython.readthedocs.io/en/latest/ , erişim tarihi:12.03.2018
[9] https://www.json.org/ , erişim tarihi:19.03.2018
[10] https://en.wikipedia.org/wiki/Bayes%27_theorem , erişim tarihi 13.03.2018
[11] http://textblob.readthedocs.io/en/dev/index.html, erişim tarihi 13.03.2018
[12] https://www.nltk.org/ , erişim tarihi 13.03.2018
[13] https://en.wikipedia.org/wiki/Decision_tree , erişim tarihi 13.03.2018
[14] http://ieeexplore.ieee.org/abstract/document/7066632/?anchor=authors
[15] https://dl.acm.org/citation.cfm?id=2390490
[16]https://ccc37e68-a-62cb3a1a-s-sites.googlegroups.com/site/anthonykunneljose/home/MajorReport.pdf?attachauth=ANoY7cqekRyPXpc83IBr_aYDHqNqrfX-ZQPMZ3BuCnELO_-uaWBPJFmTUBoJgD_Jd0Iuvu1PVMSbix2CMhqcIl7nXWeymk_EYotUZy8r5dDJHVyZlsu3Nr7o93R7Nk4xS_nqvHNGCKQ9ASc6ac37OcZKXn1bKVOXaJ4h5BY7EHw3F5cLjscoBKvBgF__hV9vv20JnSpWL1oghz15YcMnGeWofXgA4OUbpNX6cUmOKOcuMwZAistQHY4%3D&attredirects=0