

İSTANBUL TEKNİK ÜNİVERSİTESİ -FEN EDEBİYAT FAKÜLTESİ
MATEMATİK MÜHENDİSLİĞİ PROGRAMI



NLTK İle Twitter Üzerinde Duygu Analizi

BİTİRME ÖDEVİ

Betül Durkaya 090120410

Teslim Tarihi: 29.05.2017

Tez Danışmanı: Doç.Dr.Atabey Kaygun

MAYIS 2017

ÖNSÖZ

Bu çalışmada Twitter üzerinde duygu analizi ele alınmıştır. Twitter üzerinden veri toplanıp Twitter kullanıcılarının görüşleri analiz edilmiştir.

Bu proje için bana yol gösteren hocam Doç.Dr.Atabey Kaygun'a teşekkürlerimi sunarım.

Mayıs 2017

Betül Durkaya
İTÜ- Matematik Mühendisliği

İÇİNDEKİLER

ÖNSÖZ.....	ii
İÇİNDEKİLER.....	iii
ÖZET.....	1
ŞEKİL LİSTESİ.....	2
1. GİRİŞ.....	3
2. PROJENİN AÇIKLANMASI.....	4
2.1. Duygu Analizi.....	4
2.1.1. Duygu Analizi Uygulamaları.....	5
2.1.2. Duygu Analizi Araştırması.....	6
3. ANALİZİN FARKLI SEVİYELERİ.....	6
3.1 Doküman Seviyesi.....	6
3.2 Cümle Seviyesi.....	7
3.3 Varlık ve Yön Düzeyi.....	7
4. DUYARLILIK SINIFLANDIRMA TEKNİKLERİ.....	7
4.1. Sözlüğe Dayalı Yaklaşım.....	8
4.2. Makine Öğrenme Yaklaşımı.....	9
4.2.1. Denetimsiz Öğrenme.....	9

4.2.2. Denetimli Öğrenme.....	10
4.2.2.1 Naive Bayes Sınıflayıcı.....	10
4.2.2.2 Maksimum Entropi Sınıflaması.....	10
4.2.2.3. Destek Vektör Makinesi.....	11
4.2.3. Zayıf denetlenen Öğrenme.....	11
5. TWITTER ÜZERİNDE DUYGU ANALİZİ.....	11
5.1. Twitter API.....	12
6. ANALİZ İÇİN YARARLANILAN KÜTÜPHANELER.....	14
6.1. NLTK.....	14
6.2. TextBlob.....	15
7. VERİ.....	15
7.1. JSON Data.....	15
7.1.1. JSON Data Formatı Parçaları.....	18
8. KOD.....	21
8.1.Listening.py.....	21
8.2.Authentication.py.....	22
9. ANALİZ SONUCU.....	24

9.1.Dođru Analiz Örnekleri.....	24
9.2.Yanlıř Analiz Örnekleri.....	25
9.2.1.Pozitif Olması Gereken Örnekler.....	25
9.2.2. Negatif Olması Gereken Örnekler.....	26
9.2.3. Nötr Olması Gereken Örnekler.....	26
10. TÜRKÇE DUYGU ANALİZİ.....	26
11. KAYNAKÇA.....	28

ÖZET

Duygu analizi; insanların bir konu hakkındaki görüşlerini analiz etmektedir. Twitter da birçok ülkeden kullanıcıların olduğu ve çeşitli konularda fikirler belirttiği bir sosyal medya platformudur. Bu yüzden duygu analizi için önemli bir yere sahiptir. Bundan dolayı duygu analizi için Twitter uygulaması tercih edilmiştir.

Projenin giriş bölümünde çalışmanın genel amacından bahsedilmiştir. Sonrasında proje ayrıntılı bir şekilde açıklanmıştır. Daha sonra projenin kod kısmı anlatılıp sonuçlar incelenmiştir.

ŞEKİL LİSTESİ

Şekil 3.1: Ürün incelemelerinde duygu analizi süreci 6

Şekil 4.1: Duyarlılık sınıflandırma teknikleri 8

Şekil 4.2: Bayes teoremi 10

Şekil 4.3: Naive Bayes’e göre formülün yeniden yazımı 10

Şekil 4.4: Maksimum entropide sınıflamasında olasılık hesabı 11

Şekil 4.5: Bir sınıflandırma probleminde destek vektör makinesinin kullanılması 11

Şekil 7.1: Nesne gösterimi 16

Şekil 7.2: Dizi gösterimi 16

Şekil 7.3: Değer gösterimi 16

Şekil 7.4: Bir tweetin JSON formatlı hali 17

Şekil 7.5: JSON formatındaki tweetin pretty print gösterimi 17

Şekil 7.6: Şekil 7.5’in devamı 18

Şekil 7.7: Bir tweetin JSON’daki ‘user’ bölümü 21

Şekil 8.1: Listening.py dosyası 22

Şekil 8.2: Authentication.py dosyası 23

Şekil 8.3: Toplanan tweetlerin ekrandaki çıktısı 23

Şekil 8.4: Toplanan tweetlerin tweets.txt dosyasındaki hali 24

1.GİRİŞ

Bu projenin amacı Twitter kullanıcılarının belirli bir konu hakkındaki görüşlerini analiz etmektir. Proje için Python programlama dili kullanılmıştır. Fedora işletim sisteminde PyCharm ve Jupyter Notebook kullanılarak işlem yapılmıştır.

Bundan sonraki kısımlarda duygu analizi, analizin farklı seviyeleri ve duyarlılık sınıflandırma teknikleri anlatılmıştır. Daha sonra Twitter üzerinde duygu analizinden ve analiz için yararlanılan kütüphanelerden bahsedilmiştir. Sonrasında veri ve kod kısmı anlatılmış, son olarak analiz sonuçları ve Türkçe duygu analizi incelenmiştir.

2. PROJENİN AÇIKLANMASI

2.1. Duygu Analizi (Sentiment Analysis)

Duygu analizi (sentiment analysis), fikir madenciliği (opinion mining) olarak da bilinir; insanların hizmetler, kuruluşlar, bireyler, konular, etkinlikler varlıklar ve ürünlere yönelik duygularını, fikirlerini, değerlendirmelerini ve tutumlarını analiz eden bir çalışma alanıdır. Bu büyük bir sorun alanı temsil eder. Ayrıca duygu analizi (sentiment analysis), görüş incelemesi (opinion mining), fikir çıkarımı (opinion extraction), duygu madenciliği (sentiment mining), öznel analiz (subjectivity analysis), etki analizi (affect analysis), inceleme madenciliği (review mining) gibi birçok isim ve biraz farklı görevler de vardır. Ancak, artık hepsi duyarlılık analizi (sentiment analysis) ya da fikir madenciliği (opinion mining) çatısı altında toplanmaktadır. Endüstride iken, duygu analizi terimi daha yaygın olarak kullanılmaktadır; akademik çevrelerde hem duyarlılık analizi hem de görüş incelemesi terimi sıklıkla kullanılmaktadır. Ne olursa olsun, temelde aynı çalışma alanını temsil ediyorlar. Ancak, bazı araştırmacılar duygu analizi ve görüş incelemesinin (opinion mining) biraz farklı kavramlar olduğunu belirtiyorlar. Görüş incelemesi; insanın bir varlık hakkındaki fikirlerini alıp analiz eder. Duygu analizi ise bir metinde ifade edilen duyguyu tanımlarken daha sonra bunu analiz eder. Dolayısıyla, duygu analizinin hedefi, görüşleri bulmak, ifade ettikleri duyguları belirlemek ve daha sonra kutuplarını sınıflandırmaktır. Duyarlılık analizi, çoğunlukla olumlu veya olumsuz düşünceleri ifade eden veya ima eden görüşlere odaklanır. Bir fikrin duygusal olarak hangi sınıfa girdiğini belirlemeyi amaçlar. Bu sınıflar olumlu (positive), olumsuz (negative) ve nötr (neutral) olarak adlandırılabilir.

Dilbilimin ve doğal dil işlemenin (NLP-Natural Language Process) uzun bir geçmişi olmasına rağmen, 2000 yılından önce insanların görüş ve düşünceleri konusunda çok az araştırma yapılmıştır. O günden bu yana, alan çok aktif bir araştırma alanı haline gelmiştir. Bunun birkaç nedeni vardır: Birincisi, geniş bir uygulama alanı mevcuttur. Duygu analizi de ticari uygulamaların çoğalmasıyla gelişmiştir. Uygulama alanının genişliği sayesinde alan aktif bir araştırma alanı haline gelmiştir. İkincisi, daha önce üzerinde hiç çalışılmamış olan birçok zorlu araştırma problemi sunuyor. Bu da araştırma alanının genişlemesi için teşvik edici bir sebeptir. Üçüncüsü, insanlık tarihinde ilk kez, sosyal medyada çok sayıda fikre sahip veri mevcuttur. Bu veriler olmadan, bir sürü araştırma mümkün olmazdı. Şaşırtıcı olmayan bir şekilde, duygu analizinin başlangıcı ve hızlı bir şekilde büyümesi, sosyal medyanın gelişimi ile çakışmaktadır. Aslında, duygu analizi şu anda sosyal medya araştırmasının merkezindedir. Bu nedenle, duyarlılık analizi araştırması yalnızca

NLP üzerinde önemli bir etkiye sahip olmakla kalmaksızın aynı zamanda yönetim bilimleri, siyaset bilimi, ekonomi ve sosyal bilimler üzerinde de etkiler yapabilir, çünkü bütün hepsi insanların görüşlerinden etkilenir. Bu etkenlerden ötürü duygu analizi hakkındaki araştırmalar artmıştır. Duygu analizi araştırması ağırlıklı olarak 2000 yılı başından itibaren başlamış olmasına rağmen, duygu sıfatları, öznellik, bakış açısı ve etkileri üzerine daha önce yapılmış bazı çalışmalar vardır.

2.1.1. Duygu Analizi Uygulamaları (Sentiment Analysis Applications)

Fikirler neredeyse tüm insan faaliyetlerinin merkezinde yer alırlar çünkü davranışlarımızın önemli etkeni bunlardır. Ne zaman bir karar vermemiz gerekse başkalarının görüşlerini bilmek isteriz. Gerçek dünyada, işletmeler ve kuruluşlar her zaman ürün ve hizmetleri hakkında tüketici veya halkın fikirlerini bulmak istemektedir. Tüketiciler, bir ürünün satın almadan önce mevcut kullanıcılarının görüşlerini veya siyasi seçimler için bir oy vermeden önce başkalarının siyasi adaylarla ilgili görüşlerini bilmek istemektedir. Geçmişte, bir kişinin görüşe ihtiyacı olduğunda arkadaşları ve ailesine danışırdı. Bir kuruluşun veya bir işletmenin halk veya tüketici görüşlerine ihtiyacı olduğunda incelemeler ve anketler gerçekleştirirdi. Halk ve tüketici görüşlerini edinmek uzun zamandır pazarlama, halkla ilişkiler ve siyasi kampanya şirketleri için büyük bir ticari iştir.

Web'de sosyal medyanın (örneğin incelemeler, forum tartışmaları, bloglar, mikrobloglar, Twitter, yorumlar ve sosyal sitelerdeki yayınlar) büyümesiyle, bireyler ve kuruluşlar karar vermede bu medyadaki içeriği daha fazla kullanıyor. Günümüzde bir kişi bir ürün satın almak istiyorsa, bu kişi sadece arkadaşları ve ailesinden fikir istemekle sınırlı değildir, çünkü ürünle ilgili Web'de kamuya açık forumlarda çok sayıda kullanıcı görüşleri bulunmaktadır. Bir organizasyon açısından; halkın fikirlerini toplamak için anketler yapmak artık gerekli değildir. Çünkü böyle bir bilgi kamuya açık bir şekilde mevcuttur. Ancak, çeşitli sitelerin çoğalmasından dolayı Web'deki görüş sitelerini bulma ve bunlarda yer alan bilgileri dağıtma zorlu bir görev olmayı sürdürüyor. Her site genellikle uzun bloglarda ve forum ilanlarında kolaylıkla çözilemeyen büyük bir fikir metni içeriyor. Ortalama bir okuyucu, ilgili siteleri belirlemekte ve bunlardaki fikirleri özetlemekte zorlanacaktır. Bu nedenle otomatik duyarlılık analiz sistemlerine ihtiyaç duyulmaktadır.

Son yıllarda, toplumsal ve siyasal sistemler üzerinde derin etkilere sahip olan sosyal medyada görüşlü mesajların işletmelerin ve halk duygularının yeniden şekillenmesine yardımcı olduğunu gördük. Bu tür ilanlar, 2011'de bazı Arap ülkelerinde yaşananlar gibi siyasi değişim için kitleleri harekete geçirdi. Dolayısıyla, Web üzerinde fikir toplamak ve incelemek bir zorunluluk haline

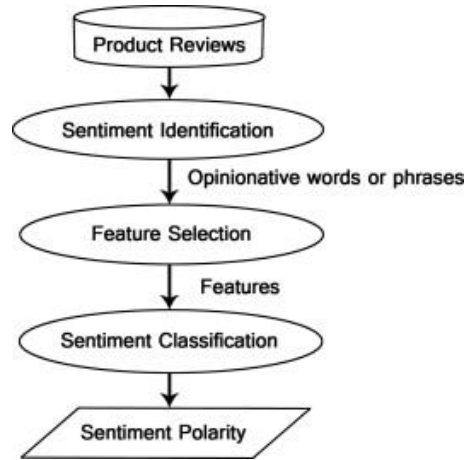
geldi. Elbette, fikir verilen belgeler yalnızca Web'de var olmakla kalmaz, aynı zamanda birçok kurumda dâhili veriler bulunur. Örneğin; e-postalardan ve çağrı merkezlerinden toplanan müşteri geri bildirimleri veya kuruluşlar tarafından yapılan anketlerden alınan sonuçlar.

Duyarlılık analizi uygulamaları; tüketici ürünleri, servisler, sağlık hizmetleri ve sosyal etkinliklere finansal hizmetler ve siyasi seçimler arasında neredeyse her alana yayılmıştır.

2.1.2. Duygu Analizi Araştırması (Sentiment Analysis Research)

Gerçek hayattaki yaygın uygulamalar, duyarlılık analizinin popüler bir araştırma problemi haline gelmesinin yalnızca bir parçası. Bir NLP araştırma konusu olarak da son derece zorlayıcıdır. Ek olarak, 2000 yılından önce NLP'de veya dilbilimde çok az araştırma yapıldı. Bunun nedenlerinden biri, o zamandan beri sayısal formlarda az sayıda görüş metninin mevcut olmasıdır. 2000 yılından beri bu alan, NLP'nin en aktif araştırma alanlarından biri haline geldi. Ayrıca, veri madenciliği, Web madenciliği ve bilgi edinme konularında geniş çapta araştırılmaktadır.

3. ANALİZİN FARKLI SEVİYELERİ (DIFFERENT LEVELS OF ANALYSIS)



Şekil 3.1: Ürün incelemelerinde duygu analizi süreci

Duygu analizi bir sınıflandırma işlemi olarak düşünülebilir. Temel olarak üç düzeyde incelenmiştir:

3.1. Doküman Seviyesi (Document Level)

Bu seviyedeki görev, bir düşünce belgesinin olumlu veya olumsuz bir düşüncüyü ifade edip etmediğini sınıflandırmaktır. Örneğin, bir ürün incelemesi göz önüne alındığında; sistem,

incelemenin ürünle ilgili genel bir olumlu veya olumsuz görüş bildirip bildirmediğini belirler. Bu görev yaygın olarak belge seviyesinde (document level) duyarlılık sınıflandırması (sentiment classification) olarak bilinir. Bu analiz düzeyi, her belgenin tek bir varlık üstünde görüş bildirdiğini varsayar. Bu nedenle, birden çok varlığı değerlendiren veya karşılaştıran belgeler için geçerli değildir.

3.2. Cümle Seviyesi (Sentence Level)

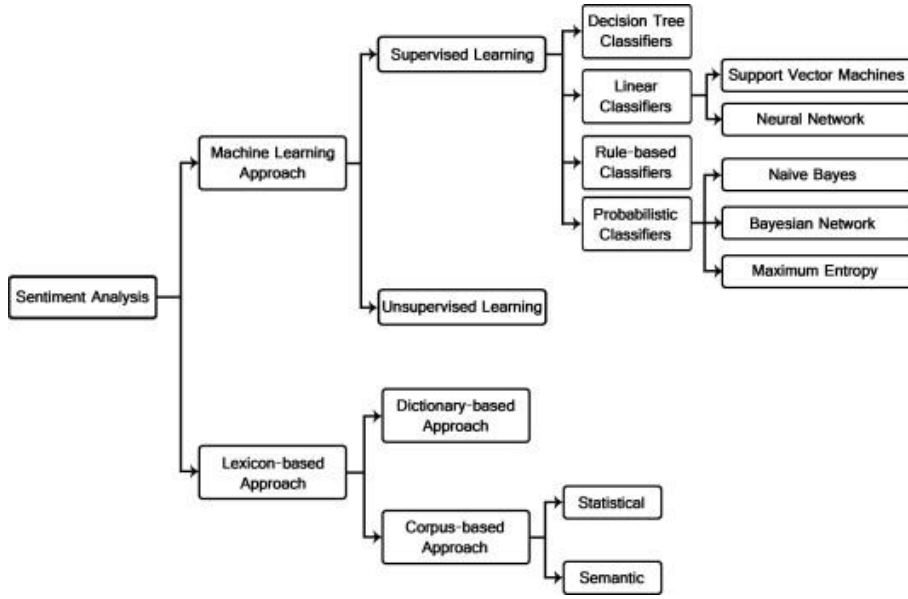
Bu seviyedeki görev, her cümlenin olumlu (positive), olumsuz (negative) veya nötr (neutral) görüşlerini ifade edip etmediğini belirler. Nötr genellikle bir görüş içermez. Bu analiz seviyesi, öznel düşünceleri ifade eden cümlelerle nesnel bilgi ifade eden cümleleri ayıran öznellik sınıflandırmasıyla(subjectivity classification) yakından ilişkilidir. Doküman seviyesinde analizden büyük bir farkı yoktur, çünkü cümleler de kısa dokümanlardır.

3.3. Varlık ve Yön Düzeyi (Entity and Aspect Level)

Hem belge düzeyi hem de cümle düzeyinde yapılan analizler tam olarak insanların ne beğendikleri ve beğenmediğini bulmazlar. Yön düzeyi daha ayrıntılı analiz yapar. Yön düzeyi daha önce özellik seviyesi olarak adlandırılmıştır. Dil yapılarına bakmak yerine, yön düzeyi doğrudan görüşe bakar. Görüş sahipleri, aynı varlıkta farklı yönlerden farklı fikirler verebilirler. Örneğin; "Although the service is not that great, I still love this restaurant." Cümle restaurant hakkında pozitif ancak servis hakkında negatiftir.

4. DUYARLILIK SINIFLANDIRMA TEKNİKLERİ (SENTIMENT CLASSIFICATION TECHNIQUES)

Duyarlılık Sınıflandırma teknikleri; makine öğrenme yaklaşımı (the machine learning approach), sözlüğe dayalı yaklaşım (lexicon based approach) ve hibrid yaklaşım (hybrid approach) olarak ayrılır. Makine Öğrenme Yaklaşımı, ünlü machine learning algoritmalarını uygular ve dilsel özellikleri kullanır. Sözlük tabanlı yaklaşım, bilinen ve önceden derlenmiş duygu terimlerinin bir koleksiyonu olan bir duygu sözlüğüne dayanır. Melez yaklaşım, her iki yaklaşımı birleştirir.



Şekil 4.1: Duyarlılık sınıflandırma teknikleri

Machine learning ile sınıflandırma, denetimli ve denetimsiz öğrenme yöntemlerine bölünebilir. Denetimli öğrenme, çok sayıda etiketli eğitim belgelerinden yararlanmaktadır. Etiketleme işlemi; bir girdiden (x) bir algoritmadan geçirilerek bir çıktı değeri (y) elde edilmesidir. Etiketli eğitim belgelerinin bulunması güç olduğunda denetimsiz öğrenme kullanılır.

Lexicon yaklaşımı, metnin analizinde kullanılan görüş sözlüğünün bulunmasına bağlıdır. Bu yaklaşımda iki yöntem vardır: Sözlük tabanlı yaklaşım (dictionary-based approach); kök sözcüklerini bulmaya dayanır, eşanlamlılarının ve zıt anlamlılarının sözlüğünü araştırır. Korpus tabanlı yaklaşım (corpus-based approach); görüş kelimelerinin bir kök listesi ile başlar ve diğer görüş kelimelerini büyük bir korpus içerisinde bulur. Bu, istatistiksel veya semantik yöntemler kullanılarak yapılabilir.

4.1. Sözlüğe Dayalı Yaklaşım (Lexicon Based Approach)

Şaşırtıcı olmayan bir şekilde, duyguların en önemli göstergeleri, fikir sözcükleri (opinion words) olarak da bilinen duygu sözcükleridir (sentiment words). Bunlar, pozitif veya negatif duyguları ifade etmek için sıklıkla kullanılan kelimelerdir. Örneğin; "iyi", "harika" ve "şaşırtıcı" olumlu duygu kelimeleridir ve "kötü", "yoksul" ve "korkunç" olumsuz duygu sözcükleridir. Bu kelimeler duygu analizine yardımcı olur. Ayrıca kelimelerin yanı sıra, deyimler de kullanılabilir. Deyimler söz gruplarından oluşur, "pahalya patlamak" (cost someone an arm and a leg) deymi gibi. Deyimler de kelimelerde olduğu gibi duygu analizine yardımcı olurlar. Bu sözcük ve deyimlerin

bir listesine duygu sözlüğü (sentiment lexicon) denir. Araştırmacılar, bu sözlükleri derlemek için çok sayıda algoritma tasarlamışlardır.

Duygu kelimeleri ve cümleleri duygu analizi için önemli olmasına rağmen, bunları kullanmak yeterli değildir. Sorun çok daha karmaşıktır. Olumlu veya olumsuz bir duygu kelimesi, farklı uygulama alanlarında zıt anlam ifade edebilir. Örneğin; "suck" kelimesi genellikle olumsuz düşünceyi gösterir. "This camera sucks." örneğindeki gibi. Fakat pozitif görüşte bildirebilir. Örneğin "This vacuum cleaner really sucks." Bu yüzden duygu sözlüğü, duygu analizi için yeterli değildir.

4.2. Makine Öğrenme Yaklaşımı (Machine Learning Approach)

Duyarlılık analizinin altında yatan temel ilke, çeşitli kategorideki kelimeleri alt kategorilere ayırma veya her bir anahtar kelimedede önceden tanımlanmış bir kriter kümesi kullanarak puan atayan matematiksel beceridir. Bu sınıflandırma yöntemi, daha çok makine öğrenme tekniği (machine learning technique) olarak bilinir.

Daha önce Naive Bayes Sınıflayıcı (Naive Bayes Classifier - NBC), Maksimum Entropi Sınıflaması (Maximum Entropy Classification - MEC) ve Destek Vektör Makinesi (Support Vector Machine - SVM) kullanılarak makine öğrenme tekniği uygulanmıştır. Bir makine öğrenme yaklaşımında; bir belge veya cümle, tipik olarak bir kelime dizisinden oluşur. Bu yaklaşımdaki altta yatan varsayım, tüm cümle ya da belgenin duygusal yöneliminin, tek tek kelimelerin duygu kutupsallık puanlarının toplamına bağlı olmasıdır.

Makine öğrenme teknikleri denetimsiz (unsupervised), zayıf denetlenen (weakly-supervised) ve denetimli (supervised) öğrenmeyi kullanır. Metin sınıflandırmasının temel amacı, belgeleri belirli bir sayıda önceden tanımlanmış kategorilere sınıflamaktır. Bunu başarmak için, çok sayıda etiketli eğitim dokümanı denetimli öğrenme için kullanılır. Metin sınıflandırmasında, bu etiketli eğitim belgelerini oluşturmak bazen zordur, ancak etiketlenmemiş belgeleri toplamak kolaydır. Denetimsiz öğrenme yöntemleri bu zorlukların üstesinden gelir.

4.2.1. Denetimsiz Öğrenme (Unsupervised Learning)

Denetimsiz yöntemler yalnızca etiketlenmemiş bir set alır. Etiketlenmemiş veriler, dünyadan kolayca elde edilebilir, doğal olarak oluşur ya da insanlar tarafından üretilir. Etiketsiz verilere örnek olarak fotoğraflar, ses kayıtları, videolar, tweetler vb. verilebilir. Yani, denetimsiz öğrenmede sadece bir input verisi (x) vardır. Bunun sonucunda bir output değeri oluşmaz.

4.2.2. Denetimli Öğrenme (Supervised Learning)

Denetimli makine öğrenme teknikleri, belirli bir sınıflandırma fonksiyonunu öğrenmek için etiketli bir eğitim kitaplığını kullanır. Etiketlenmiş veriler genellikle etiketlenmemiş bir veri kümesi alır ve etiketlenmemiş verilerin her bir parçasını anlamlı bir şekilde etiketler. Yukarıdaki etiketsiz veri için etiket örnekleri şunlar olabilir: "Bu fotoğraf bir kedi içeriyor mu?", "Bu ses kaydında hangi kelimeler söyleniyor?", "Bu videoda hangi eylem gerçekleştiriliyor?", "Bu tweetin genel düşüncesi nedir?" vb. Yani, denetimli öğrenmede bir input değerinden bir output değeri ($y=f(x)$) elde edilir.

4.2.2.1. Naive Bayes Sınıflayıcı (NBC)

En basit ve en çok kullanılan sınıflandırıcıdır. Değişken bir değer verilecek sınıfa etkisinin diğer değişken değerlerden bağımsız olduğunu varsayar. Verilen bir özellik kümesinin belirli bir etikete ait olma ihtimalini tahmin etmek için Bayes teoremini kullanır.

$$P(\text{label}|\text{features}) = \frac{P(\text{label}) * P(\text{features}|\text{label})}{P(\text{features})}$$

Şekil 4.2: Bayes teoremi

$P(\text{label})$, bir etiketin önceden oluşma olasılığı ya da rastgele bir özelliğin etiketi ayarlama olasılığıdır. $P(\text{features}|\text{label})$, belirli bir özellik kümesinin bir etiket olarak sınıflandırılmasının öncelikli olasılığıdır. $P(\text{features})$, belirli bir özellik kümesinin oluşma olasılığıdır. Tüm özelliklerin bağımsız olduğunu belirten Naive varsayımına göre denklem aşağıdaki şekilde yeniden yazılabilir:

$$P(\text{label}|\text{features}) = \frac{P(\text{label}) * P(f_1|\text{label}) * \dots * P(f_n|\text{label})}{P(\text{features})}$$

Şekil 4.3: Naive Bayes'e göre formülün yeniden yazımı

4.2.2.2. Maksimum Entropi Sınıflaması (MEC)

Maxent Sınıflayıcı koşullu üssel sınıflandırıcı olarak bilinir. Kodlamayı(encoding) kullanarak etiketli özellik kümelerini vektörlere dönüştürür. Bu vektör daha sonra her özellik için ağırlıkları(weights) hesaplamak için kullanılır. Bu sınıflandırıcı $X\{\text{weights}\}$ kümesi tarafından parametrelendirilir. $X\{\text{weights}\}$, bir özellik kümesinden $X\{\text{encoding}\}$ ile üretilen ortak özellikleri

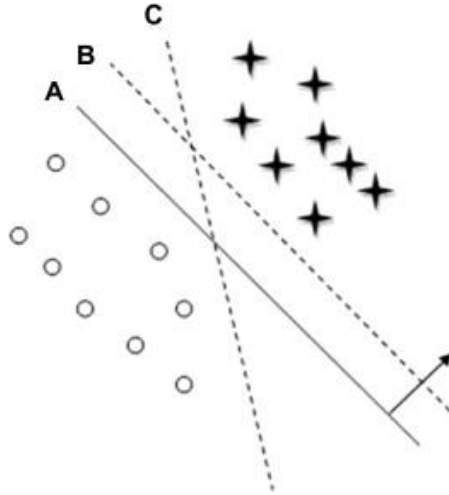
birleştirmek için kullanılır. Kodlama, her $C \{(\text{featureset}, \text{label})\}$ çiftini bir vektöre eşler. Daha sonra her etiket olasılığı aşağıdaki denklem kullanılarak hesaplanır:

$$P(fs|label) = \frac{\text{dotprod}(\text{weights}, \text{encode}(fs, \text{label}))}{\sum(\text{dotprod}(\text{weights}, \text{encode}(fs, l)) \text{ for } l \text{ in labels})}$$

Şekil 4.4: Maksimum entropide sınıflamada olasılık hesabı

4.2.2.3. Destek Vektör Makinesi (SVM)

SVM'lerin ana ilkesi, arama alanındaki doğrusal ayırıcıları belirlemektir. Aşağıdaki şekilde x ve o olmak üzere iki sınıf, ayrıca A, B ve C olmak üzere üç hiperdüzlem vardır. Hiperdüzlem A sınıflar arası en iyi ayrımı sağlar çünkü veri noktalarının herhangi birinin normal mesafesi en büyüktür. Bu yüzden maksimum ayırım marjını temsil eder.



Şekil 4.5: Bir sınıflandırma probleminde destek vektör makinesinin kullanılması

4.2.3. Zayıf Denetlenen Öğrenme (Weakly-Supervised Learning)

Zayıf denetlenen öğrenme veya yarı denetimli öğrenme (semi supervised learning), küçük bir etiketlenmiş veri kümesi ve büyük bir etiketsiz veri havuzu kullanılır.

5. TWITTER ÜZERİNDE DUYGU ANALİZİ (SENTIMENT ANALYSIS ON TWITTER)

Sosyal medya platformları, farklı konularda düşüncelerini ifade etmek için farklı kişiler tarafından kullanılır, bu nedenle de insanların görüşlerinin değerli bir kaynağıdır. Twitter, kullanıcıların oluşturduğu çok sayıda metin mesajı içerir ve her geçen gün büyür. Twitter'ın izleyici kitleleri

sıradan kullanıcılardan ünlülere, şirket temsilcilerine, siyasetçilere ve hatta ülke başkanlarına kadar değişir. Bu nedenle, farklı sosyal gruplardan kullanıcıların metin mesajlarını toplamak mümkündür. Mesajların içeriği kişisel düşüncelerden kamuya açıklamalara kadar değişir.

Dünyanın herhangi bir yerinden veri edinebilirsiniz. Twitter'ın kitlesi birçok ülkeden kullanıcılar tarafından temsil edilmektedir. ABD'li kullanıcılar hâkim olsa da, farklı dillerde veri toplamak mümkündür.

Twitter API'si iyi tasarlanmış ve kolay erişilebilirdir. Twitter verilerini analiz için uygun bir biçimdedir. Twitter'ın kullanım şartları, diğer API'lere kıyasla nispeten serbesttir. Bu özelliklerden ötürü duygu analizi için Twitter'ı kullandık.

5.1. Twitter API (Application Programming Interface)

Uygulama Programlama Arayüzü(API), bir yazılım uygulaması oluşturmak için kullanılan komutlar, fonksiyonlar, objeler ve protokoller kümesidir. İyi bir API temel parçaları sağlayarak program geliştirmeyi kolaylaştırır. Programcı da parçaları bir araya getirir. API'lerin çeşitli türleri vardır. Örneğin; işletim sistemleri, web siteleri, uygulamalar vb. için API'ler mevcuttur. Google Maps, Twitter, YouTube, Flickr and Amazon Product Advertising en popüler API'lerden bazılarıdır.

Twitter, temelde iki tür API'si mevcuttur: Streaming API ve REST API (Representational State Transfer). REST mimarisi, kaynakları tanımlayan ve veriye erişim yollarını belirleyen ağ tasarım ilkelerinin bir koleksiyonunu ifade eder. Popüler REST API'lerden biri Search API'dir. Search API, tarihsel olarak tweet verir. Örneğin; bir hafta için birkaç yüz tweet verebilir. Ancak Streaming API, anlık olarak tweet verir.

Twitter'a erişim için, bir twitter uygulaması oluşturulması ve anahtar edinilmesi gereklidir. Uygulamayı oluşturabilmek için de geçerli bir telefon numarasına sahip bir twitter hesabının olması gereklidir. Tarayıcıdan <https://apps.twitter.com/> adresine twitter hesabıyla giriş yapılır ve "Create New App" butonuna tıklanır. Uygulama oluşturma sayfasına yönlendirilince gerekli alanlar doldurulmalıdır. Ardından "Twitter Developer Agreement" okunup onaylanıyorsa "Yes, I have read and agree to the Twitter Developer Agreement." yazısının yanındaki kutucuk işaretlenir ve "Create your Twitter application" butonuna basılır. Uygulama başarılı bir şekilde oluşturulduktan sonra, uygulamanın "Details" sayfasına yönlendirme gerçekleşir. Bu sayfada uygulama hakkında genel bilgi verilmektedir.

Uygulama anahtarlarını oluşturmak için uygun erişim düzeyine sahip olunmalıdır. Bunu kontrol etmek için "Permissions" sayfasına geçilir. Varsayılan olarak, Twitter uygulaması Okuma ve Yazma (Read and Write) erişimine sahip olmalıdır. Durum böyle değilse izinler bu yönde değiştirilmelidir. Daha sonra "Keys and Access Tokens" sekmesine girilir. Buradaki sayfa Consumer Key ve Consumer Secret'ı listeler. Consumer key, hangi uygulamanın isteği oluşturduğunu tanımlar. Ayrıca Access Token ve Access Token Secret oluşturulmasına izin verir. Token, genellikle uygulamanın izinlerini temsil eder. Çoğu istekte access token olarak adlandırılır. Bunlar uygulamayı Twitter'la doğrulamak için gereklidir. Access Token ve Access Token Secret'ı oluşturmak için Access Token başlığı altındaki "Create my access token" butonu tıklanır. Bu şekilde access token keyleri elde edilmiş olur.

Twitter, API'ye yetkili erişim sağlanması için OAuth'u kullanmaktadır. Web, mobil ve masaüstü uygulamalarından basit ve standart bir yöntemle güvenli yetkilendirmeye izin veren açık bir protokoldür. Kullanıcıların hesap bilgilerini üçüncü parti uygulamalarıyla paylaşmaları gerekmez, bu yüzden güvenlidir. OAuth 2.0 yetkilendirme yapısı, bir üçüncü parti uygulamasının bir HTTP servisine sınırlı erişim elde etmesini sağlar. OAuth, korunan veriyi yayınlamak ve etkileşimde bulunmak için basit bir yoldur.

Twitter API'nin iki tür kimlik doğrulama modeli vardır: Kullanıcı doğrulama (user authentication) ve yalnızca uygulama doğrulama (application-only authentication). Kullanıcı doğrulama; Twitter'ın OAuth 1.0a uygulamasında kaynak kimlik doğrulamasının en yaygın biçimidir. İstek, son kullanıcının izinleriyle verilen kimliğe ek olarak bir uygulamanın kimliğini tanımlar. Yalnızca uygulama doğrulama; bir uygulamanın kendi adına bir kullanıcı bağlamı olmaksızın API istekleri yaptığı kimlik doğrulama biçimidir.

Twitter API'si OAuth 1.0a protokolünü kullanır. Twitter'ın tüm API'leri HTTP protokolüne dayanır. Bu, Twitter API'lerini kullanan herhangi bir yazılımın Twitter sunucularına bir dizi yapılandırılmış mesaj gönderdiği anlamı taşır. Bu mesajın (GET) başlığında (header) yukarıda oluşturulan şifrelerin (key) birleştirilmiş hali hashlenerek tutulur ve kimlik doğrulama cevabı beklenir.

Streaming API'ye bağlantı kurmak, çok uzun süre yaşayan bir HTTP isteği yapmak ve cevabı adım adım ayırtırmak anlamına gelir. Twitter bazı nedenlerle bu bağlantıyı kapatabilir. Bu nedenlerden bazıları aşağıdadır:

- Bir istemci aynı kimlik bilgilerini kullanıp çok fazla bağlantı kurduğunda en eski bağlantı sonlandırılır.
- Akıştan okunan tweetlerin oranı birdenbire düşerse bağlantı kapatılır.
- İstemciye gönderilecek mesaj kuyruğu çok fazla büyürse bağlantı kapanır.

HTTP hata kodları: Çoğu hata kodu ek ayrıntılarla birlikte bir dizi ile döndürülür. 200'ün üzerindeki tüm hata kodları için istemciler başka bir bağlantı denemeden önce beklemelidir.

- *401 (unauthorized)*: Kimlik doğrulama hatası; geçersiz kimlik doğrulama bilgileri veya geçersiz bir OAuth isteği nedeniyle oluşabilir.
- *403 (not acceptable)*: En az bir istek parametresi geçersizdir. İzlenecek anahtar sözcük (track) belirtilmemiş, çok uzun ya da çok kısa olması nedeniyle oluşabilir.
- *420 (rate limited)*: İstemcinin kısa bir süre içinde çok fazla giriş denemesi yapması durumunda oluşur.

6. ANALİZ İÇİN YARARLANILAN KÜTÜPHANELER

6.1. NLTK (Natural Language Toolkit)

NLTK(natural language toolkit), Python'da doğal dil işleme (natural language processing) için kütüphaneler ve programlar içeren bir araçtır. Başlangıçta Steven Bird, Edward Loper ve Ewan Klein tarafından geliştirilmiştir. Tokenization, ayrıştırma (parsing), sınıflandırma (classification), köklendirme (stemming), etiketleme (tagging), anlamsal akıl yürütme (semantic reasoning) için metin işleme (text processing) kütüphaneleri içerir.

NLTK; Hesaba dayalı dilbilim konularının yanı sıra programlama temellerini tanıtan kılavuz ve kapsamlı bir API dokümantasyonuna da sahip olması sayesinde dilbilimciler, mühendisler, öğrenciler, eğitimciler ve araştırmacılar için kullanıma uygundur. NLTK; Windows, Mac OS X ve Linux işletim sistemleri için kullanılabilir. Ücretsiz ve açık kaynak kodlu bir projedir.

NLTK, "hesaba dayalı dilbilimde Python kullanılarak çalışma ve öğretim için harika bir araç" ve "doğal dille oynamak için muhteşem bir kütüphane" olarak adlandırılmıştır.

NLTK'in yaptığı bazı işlemler:

- Tokenization: Stringlerin birer parçası olan kelimeler, cümlecikler ve semboller "token" olarak isimlendirilir ve bu işleme tokenization ya da tokenizing denir.

- Stemming: Sözcüğün kökünü bulmak için kelime son eklerini değiştirme ve kaldırma yöntemidir.
- Lemmatization: Farklılaşmış sözcük biçimlerini bir araya getirme işlemidir.

6.2. TextBlob

TextBlob, metinsel verileri işlemek için kullanılan bir Python kütüphanesidir. Duygu analizi, sınıflandırma ve çeviri gibi NLP görevlerini gerçekleştirmek için bir API sağlar. TextBlob metin işleme işlemlerine bir arayüz aracılığıyla erişim sağlamayı amaçlar. TextBlob objeleri, NLP işlemini nasıl yapacağını öğrenen Python stringleri gibi işlenebilir.

textblob.sentiments modülü iki duygu analizi uygulamasını içerir; bunlar NaiveBayesAnalyzer ve PatternAnalyzer'dır. NaiveBayesAnalyzer, bir NLTK sınıflandırıcısıdır. PatternAnalyzer ise pattern kütüphanesine dayalıdır. Pattern; Python programlama dili için bir web araştırma modülüdür. Veri madenciliği, NLP, makine öğrenmesi için araçlara(tools) sahiptir. Duygu analizi için pattern.en modülü kullanılır. Varsayılan uygulama PatternAnalyzer'dır. Ancak TextBlob kurucusuna başka bir uygulama geçirilerek bu çözümleyici(analyzer) etkisiz kılınabilir.

Duygu analizinde; NaiveBayesAnalyzer sonucu Sentiment(classification, p_pos, p_neg) şeklinde bir değişken grubu olarak döndürür. PatternAnalyzer ise sonucu Sentiment(polarity, subjectivity) şeklinde bir grup döndürür.

Kutupsallık; cümlelerin pozitif veya negatif olup olmadığını anlamamıza bağlı olarak sayısal bir değer veren ölçümdür. Kutupsallık değeri [-1.0, 1.0] aralığında bir kesirli sayıdır (float). Burada -1.0 değeri çok negatif, 1.0 değeri çok pozitif, 0.0 değeri ise nötrdür. Öznellik değeri de [0.0, 1.0] aralığında bir kesirli sayıdır (float). Burada 1.0 değeri çok öznel, 0.0 değeri ise çok nesneldir.

7. VERİ

Veri olarak Twitter üzerinden tweetler kullanılmış ve daha rahat işlenebilmesi için JSON formatına çevrilmiştir.

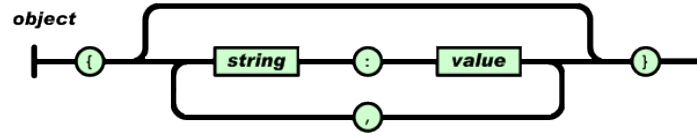
7.1. JSON Data

Twitter üzerinden tweetler data olarak kullanılacak ve tweetler JSON formatında toplanacak. JSON (JavaScript Object Notation - JavaScript Nesne Gösterimi) hafif bir veri alışverişi

biçimidir. İnsanlar için verilerin okunmasını ve yazılmasını, makineler için verilerin üretilmesini ve ayrıştırılmasını kolaylaştırır. JSON, tamamen programlama dilinden bağımsız bir yazı formatı olmasına rağmen JavaScript Programlama dili kavramına dayanmaktadır ve C dil ailesi programcılarının aşina olduğu gelenekleri kullanır. Bu özellikler, JSON'u ideal bir veri değiş tokuş dili yapar.

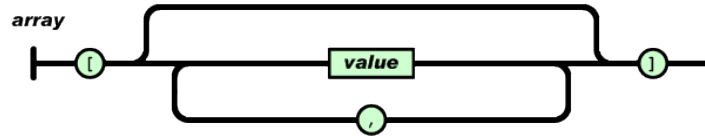
JSON iki yapı üzerine kurulmuştur: İsim-değer çifti koleksiyonu ve sıralı değer listesi. Jsonda bu yapılar aşağıdaki şekillerdeki gibi gösterilir:

Bir nesne, isim-değer çiftlerinin sırasız birleşiminden oluşur. Nesne { (sol süslü parantez) ile başlar } (sağ süslü parantez) ile biter. Her isimden sonra : (iki nokta üst üste) gelir ve isim-değer çiftleri , (virgül) ile ayrılır.



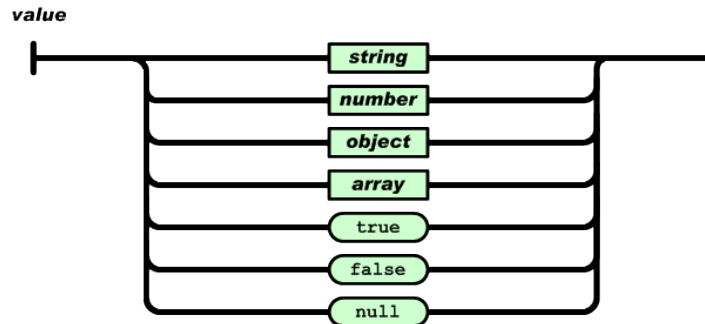
Şekil 7.1: Nesne gösterimi

Diziler, sıralı değer listesidir. Bir dizi [(sol köşeli parantez) ile başlar ve] (sağ köşeli parantez) ile biter. Değerler , (virgül) ile ayrılır.



Şekil 7.2: Dizi gösterimi

Bir değer, " (çift tırnak) içinde bir string, sayı, doğru (true), yanlış (false), boş değer (null), nesne (object) veya dizi (array) olabilir. Bu yapılar birbirlerinin içinde tekrar edebilirler.



Şekil 7.3: Değer gösterimi

Twitter üzerinden elde ettiğimiz her bir tweet için JSON data formatı aşağıda gösterilmiştir:

```
import json
obj = {'filter_level': 'low', 'retweeted': False, 'in_reply_to_user_id_str': None, 'id': 862303446493523968, 'id_str': '862303446493523968', 'coordinates': None, 'timestamp_ms': '1494424137124', 'in_reply_to_screen_name': None, 'in_reply_to_status_id_str': None, 'text': 'Featured: Swappa listing for Apple iPhone 6S Plus (Unlocked): $395 https://t.co/SaqVt8hVab', 'favorited': False, 'user': {'geo_enabled': False, 'id': 3002701812, 'id_str': '3002701812', 'profile_sidebar_fill_color': '000000', 'following': None, 'listed_count': 24, 'verified': False, 'location': None, 'contributors_enabled': False, 'name': 'Swappa Mucho', 'default_profile': False, 'description': 'A lot of Swappa, including featured listings and other promotions.', 'followers_count': 780, 'statuses_count': 68691, 'profile_sidebar_border_color': '000000', 'is_translator': False, 'lang': 'en', 'follow_request_sent': None, 'utc_offset': None, 'profile_image_url_https': 'https://pbs.twimg.com/profile_images/572504904745246720/UZQcFWhj_normal.png', 'profile_use_background_image': False, 'time_zone': None, 'protected': False, 'profile_image_url': 'http://pbs.twimg.com/profile_images/572504904745246720/UZQcFWhj_normal.png', 'profile_text_color': '000000', 'profile_background_image_url_https': 'https://pbs.twimg.com/profile_background_images/573522939548594176/7vv7IU8B.png', 'profile_banner_url': 'https://pbs.twimg.com/profile_banners/3002701812/1425573613', 'profile_background_image_url': 'http://pbs.twimg.com/profile_background_images/573522939548594176/7vv7IU8B.png', 'default_profile_image': False, 'url': 'http://swappa.com/', 'profile_background_color': '000000', 'created_at': 'Fri Jan 30 02:56:24 +0000 2015', 'notifications': None, 'favourites_count': 11, 'screen_name': 'SwappaMucho', 'friends_count': 1700, 'profile_background_tile': False, 'profile_link_color': '94D487'}, 'truncated': False, 'in_reply_to_user_id': None, 'source': '<a href="http://swappa.com" rel="nofollow">Swappa Mucho</a>', 'possibly_sensitive': False, 'place': None, 'contributors': None, 'favorite_count': 0, 'geo': None, 'entities': {'user_mentions': [], 'symbols': [], 'urls': [{'indices': [67, 90], 'expanded_url': 'https://swappa.com/listing/FYZ093/view', 'url': 'https://t.co/SaqVt8hVab', 'display_url': 'swappa.com/listing/FYZ093...'}]}, 'hashtags': []}, 'in_reply_to_status_id': None, 'is_quote_status': False, 'created_at': 'Wed May 10 13:48:57 +0000 2017', 'lang': 'en', 'retweet_count': 0}
print(json.dumps(obj, indent=2))
```

Şekil 7.4: Bir tweetin JSON formatlı hali

```
{
  "filter_level": "low",
  "retweeted": false,
  "in_reply_to_user_id_str": null,
  "id": 862303446493523968,
  "id_str": "862303446493523968",
  "coordinates": null,
  "timestamp_ms": "1494424137124",
  "in_reply_to_screen_name": null,
  "in_reply_to_status_id_str": null,
  "text": "Featured: Swappa listing for Apple iPhone 6S Plus (Unlocked): $395 https://t.co/SaqVt8hVab",
  "favorited": false,
  "user": {
    "geo_enabled": false,
    "id": 3002701812,
    "id_str": "3002701812",
    "profile_sidebar_fill_color": "000000",
    "following": null,
    "listed_count": 24,
    "verified": false,
    "location": null,
    "contributors_enabled": false,
    "name": "Swappa Mucho",
    "default_profile": false,
    "description": "A lot of Swappa, including featured listings and other promotions.",
    "followers_count": 780,
    "statuses_count": 68691,
    "profile_sidebar_border_color": "000000",
    "is_translator": false,
    "lang": "en",
    "follow_request_sent": null,
    "utc_offset": null,
    "profile_image_url_https": "https://pbs.twimg.com/profile_images/572504904745246720/UZQcFWhj_normal.png",
    "profile_use_background_image": false,
    "time_zone": null,
    "protected": false,
    "profile_image_url": "http://pbs.twimg.com/profile_images/572504904745246720/UZQcFWhj_normal.png",
    "profile_text_color": "000000",
    "profile_background_image_url_https": "https://pbs.twimg.com/profile_background_images/573522939548594176/7vv7IU8B.png",
    "profile_banner_url": "https://pbs.twimg.com/profile_banners/3002701812/1425573613"
  },
  "truncated": false,
  "in_reply_to_user_id": null,
  "source": "<a href='\"http://swappa.com\"' rel='\"nofollow\"'>Swappa Mucho</a>",
  "possibly_sensitive": false,
  "place": null,
  "contributors": null,
  "favorite_count": 0,
  "geo": null,
  "entities": {
    "user_mentions": [],
    "symbols": [],
    "urls": [
      {
        "indices": [67, 90],
        "expanded_url": "https://swappa.com/listing/FYZ093/view",
        "url": "https://t.co/SaqVt8hVab",
        "display_url": "swappa.com/listing/FYZ093..."
      }
    ],
    "hashtags": []
  },
  "in_reply_to_status_id": null,
  "is_quote_status": false,
  "created_at": "Wed May 10 13:48:57 +0000 2017",
  "lang": "en",
  "retweet_count": 0
}
```

Şekil 7.5: JSON formatındaki tweetin pretty print gösterimi

```

    "default_profile_image": false,
    "url": "http://swappa.com/",
    "profile_background_color": "000000",
    "created_at": "Fri Jan 30 02:56:24 +0000 2015",
    "notifications": null,
    "favourites_count": 11,
    "screen_name": "SwappaMucho",
    "friends_count": 1700,
    "profile_background_tile": false,
    "profile_link_color": "940487"
  },
  "truncated": false,
  "in_reply_to_user_id": null,
  "source": "<a href='\"http://swappa.com/\"' rel='\"nofollow\"'>Swappa Mucho</a>",
  "possibly_sensitive": false,
  "place": null,
  "contributors": null,
  "favorite_count": 0,
  "geo": null,
  "entities": {
    "user_mentions": [],
    "symbols": [],
    "urls": [
      {
        "indices": [
          67,
          90
        ],
        "expanded_url": "https://swappa.com/listing/FYZ093/view",
        "url": "https://t.co/SaqVt8hVab",
        "display_url": "swappa.com/listing/FYZ093\u2026"
      }
    ],
    "hashtags": []
  },
  "in_reply_to_status_id": null,
  "is_quote_status": false,
  "created_at": "Wed May 10 13:48:57 +0000 2017",
  "lang": "en",
  "retweet_count": 0
}

```

Şekil 7.6: Şekil 7.5'in devamı

7.1.1. JSON Data Formatı Parçaları

contributors: Tweet yazarı adına, tweetin yazarlığına katkıda bulunan kullanıcıları belirtir. Artık aktif olarak kullanılmayan bir değerdir.

Örneğin; 'contributors': None

coordinates: Tweetin coğrafi olarak yerini temsil eder.

Örneğin; 'coordinates': None

created_at: Tweetin oluşturulduğu zamanı gösterir.

Örneğin; 'created_at': 'Wed May 10 13:48:57 +0000 2017'

entities: Tweet metninden ayrıştırılan varlıklardır.

Örneğin; 'entities': {'user_mentions': [], 'symbols': [], 'urls': [{'indices': [67, 90], 'expanded_url': 'https://swappa.com/listing/FYZ093/view', 'url': 'https://t.co/SaqVt8hVab', 'display_url': 'swappa.com/listin/FYZ093\u2026'}], 'hashtags': []}

favorite_count: Tweetin twitter kullanıcıları tarafından kaç kere beğenildiğini gösterir.

Örneğin; 'favorite_count': 0

favorited: Tweetin kimliği doğrulanmış kullanıcı tarafından beğenilip beğenilmediğini gösterir.

Örneğin; 'favorited': False

filter_level: Kullanılabilen filter_level parametresinin maksimum değerini gösterir.

Örneğin; 'filter_level': 'low'

id: Tweeti tanımlayan tamsayı gösterimidir. Bu sayı 53 bitten büyüktür.

Örneğin; 'id': 3002701812

id_str: Tweeti tanımlar, id'den farklı olarak bu değer bir stringtir.

Örneğin; 'id_str': '3002701812'

in_reply_to_screen_name: Tweet bir cevapsa bu alanda tweet yazarının kullanıcı adı bulunur.

Örneğin; 'in_reply_to_screen_name': None

in_reply_to_status_id: Tweet bir cevapsa bu alanda tweetin id si bulunur.

Örneğin; 'in_reply_to_status_id': None

in_reply_to_status_id_str: Tweet bir cevapsa bu alanda tweetin id_str değeri bulunur.

Örneğin; 'in_reply_to_status_id_str': None

in_reply_to_user_id: Tweet bir cevapsa bu alanda tweet yazarının id si bulunur.

Örneğin; 'in_reply_to_user_id': None

in_reply_to_user_id_str: Tweet bir cevapsa bu alanda tweet yazarının id si string olarak bulunur.

Örneğin; 'in_reply_to_user_id_str': None

lang: Tweetin dilini belirtir.

Örneğin; 'lang': 'en'

place: Varsa tweetin ilişkili olduğu yeri belirtir.

Örneğin; 'place': None

retweet_count: Tweetin retweetlenme sayısını gösterir.

Örneğin; 'retweet_count': 0

retweeted: Tweetin kimliği doğrulanmış kullanıcı tarafından retweetlenip retweetlenmediğini gösterir.

Örneğin; 'retweeted': False

source: Tweetin html formatında string olarak gönderilirken kullanılan halidir.

Örneğin; 'source': 'Swappa Mucho'

text: Tweetin gerçek halidir.

Örneğin; 'text': 'Featured: Swappa listing for Apple iPhone 6S Plus (Unlocked): \$395
<https://t.co/SaqVt8hVab>'

truncated: Text parametresinin değerinin kesilip kesilmediğini gösterir. Örneğin; tweetin 140 karakteri aşması durumu.

Örneğin; 'truncated': False

user: Tweeti gönderen kullanıcıdır. Kullanıcıyla ilgili bilgileri içerir. Buradaki screen_name parametresi yazarın Twitter'daki kullanıcı adıdır.

Örneğin;

```
'user': {'geo_enabled': False, 'id': 3002701812, 'id_str': '3002701812', 'profile_sidebar_fill_color': '000000', 'following': None, 'listed_count': 24, 'verified': False, 'location': None, 'contributors_enabled': False, 'name': 'Swappa Mucho', 'default_profile': False, 'description': 'A lot of Swap pa, including featured listings and other promotions.', 'followers_count': 780, 'statuses_count': 68691, 'profile_sidebar_border_color': '000000', 'is_translator': False, 'lang': 'en', 'follow_request_sent': None, 'utc_offset': None, 'profile_image_url_https': 'https://pbs.twimg.com/profile_images/572504904745246720/UZQcFWhj_normal.png', 'profile_use_background_image': False, 'time_zone': None, 'protected': False, 'profile_image_url': 'http://pbs.twimg.com/profile_images/572504904745246720/UZQcFWhj_normal.png', 'profile_text_color': '000000', 'profile_background_image_url_https': 'https://pbs.twimg.com/profile_background_images/573522939548594176/7vv7IU8B.png', 'profile_banner_url': 'https://pbs.twimg.com/profile_banners/3002701812/1425573613', 'profile_background_image_url': 'http://pbs.twimg.com/profile_background_images/573522939548594176/7vv7IU8B.png', 'default_profile_image': False, 'url': 'http://swappa.com/', 'profile_background_color': '000000', 'created_at': 'Fri Jan 30 02:56:24 +0000 2015', 'notifications': None, 'favourites_count': 11, 'screen_name': 'SwappaMucho', 'friends_count': 1700, 'profile_background_tile': False, 'profile_link_color': '94D487'}
```

Şekil 7.7: Bir tweetin JSON'daki 'user' bölümü

Bunlardan id_str, created_at, user(yalnızca screen name) ve text parçaları kullanılıp, text parametresi TextBlob kütüphanesi ile analiz edildi ve bunlar .txt formatında bir dosyaya aktarıldı.

8. KOD

Programlama dili olarak Python kullanılmıştır. Kodlar PyCharm'da yazılmış ayrıca daha sonra Jupyter Notebook üzerinde de denenmiştir. Pycharm üzerinde listening.py ve authentication.py adında iki tane Python dosyası oluşturulmuştur.

8.1. Listening.py

StreamListener'den türetilen bir sınıf oluşturulur, bu basit akış dinleyicisi on_connect metoduyla bağlantı sağlanırsa akışın başladığını belirtir, on_error metoduyla hata oluşması durumunda bu hata kodunu ekrana yazar.

Tweepy'nin tekrar düzenlenen StreamListener kümesinin on_data metodu ile veriler uygun bir şekilde alınır. Veriler json kütüphanesinin loads metoduyla daha rahat işlenmek amacıyla json formatına dönüştürülür. Bu yöntem sonrasında retweetlenmemiş tweetler alınır ve textblob kütüphanesinin TextBlob isimli metoduyla bu tweetlerin öznellik ve kutupsallık değerleri oluşturularak hem bir dosyaya hem de ekrana yazılır. Kutupsallık değeri negatif, pozitif ve nötr olan veriler tweets.txt dosyasında tutulmuştur.

```

import tweepy
import json
from textblob import TextBlob

# organizing tweet information
class StreamListener(tweepy.StreamListener):
    def on_connect(self):
        print("Tweet streaming begin.")

    def on_error(self, status):
        print('Error Type: ' + status)
        return False

    def on_data(self, raw):
        data = json.loads(raw)
        tweet_id = data['id_str'] # The ID of tweet from Twitter in string format
        time = data['created_at'] # The time of creation of the tweet
        username = data['user']['screen_name'] # The Tweet author's username
        text = data['text'] # The entire body of the Tweet
        fs = open("tweets.txt", "a")

        try:
            # insert tweet data to tweet.txt file if RT is not exist
            if data['text'].find('RT @') is -1:
                sample = TextBlob(text)
                polarity = sample.sentiment.polarity
                subjectivity = sample.sentiment.subjectivity
                print(tweet_id + '\t' + time + '\t' + username + '\n' + text + '\n' +
                    'Sentiment Result: polarity=' + str(polarity) +
                    ', subjectivity' + str(subjectivity) + '\n\n')
                fs.write(tweet_id + '\t' + time + '\t' + username + '\n' + text + '\n' +
                    'Sentiment Result: polarity=' + str(polarity) +
                    ', subjectivity' + str(subjectivity) + '\n\n')
            fs.close()
        except Exception as e:
            print(e)
        return True

```

Şekil 8.1: Listening.py dosyası

8.2. Authentication.py

Tweet akışını almak için tweepy ve listening kütüphaneleri eklenir. Tweepy kütüphanesindeki OAuthHandler metodu çağırılarak consumer_key, consumer_secret, access_token ve access_token_secret şifreleri sayesinde Twitter API ile kimlik denetimi sağlanır.

Tweepy, kimliği doğrulama, bağlantı, oturum oluşturma ve yok etme, gelen iletileri okuma ve iletileri kısmen yönlendirerek Twitter Streaming API'yi kullanmayı kolaylaştırır.

Listening kütüphanesinde StreamListener metoduyla akışın bağlantılı olacağı API belirtilir. Sonrasında tweepy kütüphanesinin Stream metoduyla tweet akışını dinleyecek nesne oluşturulur. Tweepy aracılığıyla birçok twitter akışı bulunmaktadır ve filter metoduyla belirlediğimiz kelime için veri akışı süzülür.

```

import tweepy
from tweepy import OAuthHandler
import listening

#twitter api keys to authanticate
consumer_key = "Consumer Key"
consumer_secret = "Consumer Secret"

access_token = "Access Token"
access_token_secret = "Access Token Secret"

#authentication sections
auth = OAuthHandler(consumer_key, consumer_secret)
print("First auth done")
auth.set_access_token(access_token, access_token_secret)
print("2nd auth done")

#tracked word
WORDS = ['iphone']

#from listening library, StreamListener object created
listener = listening.StreamListener(api=tweepy.API(wait_on_rate_limit=True))
#from tweepy library, Stream begins
streamer = tweepy.Stream(auth=auth, listener=listener)

print("Tracking: " + str(WORDS))
streamer.filter(track=WORDS, languages=["en"]) #filtering keywords

```

Şekil 8.2: Authentication.py dosyası

Bu kodlar çalıştırıldığında ekrana aşağıdaki gibi bir çıktı verir:

```

First auth done
2nd auth done
Tracking: ['iphone']
Tweet streaming begin.
865920952626028546 Sat May 20 13:23:37 +0000 2017 bitcoinlovers2
-- WIN IPHONE 7 PLUS: Paid our Co Pay for Doctors appointment for our little bun in the oven with #Bitcoin
Thanks BitPay Our first child i...
Sentiment Result: polarity = 0.265625, subjectivity = 0.35833333333333334

865920954118979584 Sat May 20 13:23:38 +0000 2017 thexeon
There's a handy shortcut menu hidden on your iPhone or iPad https://t.co/029VPaBrA2 #thexeon
Sentiment Result: polarity = 0.21666666666666667, subjectivity = 0.6166666666666667

865920958598701057 Sat May 20 13:23:39 +0000 2017 showmesolutions
There's a handy shortcut menu hidden on your iPhone or iPad - ZDNet https://t.co/fb0mXoNSAo #iPad
Sentiment Result: polarity = 0.21666666666666667, subjectivity = 0.6166666666666667

865920964697182209 Sat May 20 13:23:40 +0000 2017 devitaaas
allkpop: SHINee Key's 'Nylon' cover pictorial was shot with an iPhone ??
https://t.co/aCU3u63mJQ https://t.co/cujkk13Uhn
Sentiment Result: polarity = 0.0, subjectivity = 1.0

865920972150509568 Sat May 20 13:23:42 +0000 2017 woanvo_Vuomdo
6 3FT 30PIN USB SYNC DATA POWER CHARGER ORANGE CABLE CORD IPHONE IPOD TOUCH IPAD https://t.co/w5yNjw0lcu h
tps://t.co/4e40rIp9us
Sentiment Result: polarity = 0.0, subjectivity = 0.0

865920979196928000 Sat May 20 13:23:44 +0000 2017 woanvo_Vuomdo
3 6FT 30PIN USB SYNC DATA POWER CHARGER ORANGE CABLE CORD IPHONE IPOD TOUCH IPAD https://t.co/CerhR1dZp1 h
tps://t.co/x4Ire2kvIC
Sentiment Result: polarity = 0.0, subjectivity = 0.0

865920984339054593 Sat May 20 13:23:45 +0000 2017 christyg3131
@paulapoundstone I just watched Colbert (on my iPhone). Awesome as usual! Great to see you on Late Night a
gain. #sentfrommyiphone :(
Sentiment Result: polarity = 0.0875, subjectivity = 0.72

865921007000981506 Sat May 20 13:23:50 +0000 2017 dougabels
In case you missed it, an amazing essay... John McCain: Why We Must Support Human Rights - NYTi https://
t.co/7xmsPOgfnk
Sentiment Result: polarity = 0.30000000000000004, subjectivity = 0.5

```

Şekil 8.3: Toplanan tweetlerin ekrandaki çıktısı

Ayrıca gelen tweetler ve analiz sonuçları tweets.txt dosyasında aşağıdaki gibi saklanmıştır:

```
1 857638636732514306 Thu Apr 27 16:52:39 +0000 2017 Goutja_Qaiqde
2 Glossy 3D Pearl Gorgeous Lace Cover Case Shell For Apple iPhone 4G 4S 4th 4FEE https://t.co/WlqpmCVXBh https://t.co/0NNcFfj09Y
3 Sentiment Result: polarity=0.7, subjectivity0.9
4
5 857638638011723777 Thu Apr 27 16:52:40 +0000 2017 pregcitygirl
6 iPhone Cases - Luxury at your Fingertips https://t.co/ofQTVYMTZK on @bloglovin
7 Sentiment Result: polarity=0.0, subjectivity0.0
8
9 857638636883464192 Thu Apr 27 16:52:39 +0000 2017 Yaukba_Luimwe
10 Glossy 3D Pearl Gorgeous Lace Cover Case Shell For Apple iPhone 4G 4S 4th 4FEE https://t.co/8jNyod97xK https://t.co/MZVU02UawD
11 Sentiment Result: polarity=0.7, subjectivity0.9
12
13 857638638405943296 Thu Apr 27 16:52:40 +0000 2017 SolisFarid
14 Apple iPhone 7 Plus - 32GB - Black (Sprint) Smartphone Giveaway RT & Follow https://t.co/xpv45qHQLj
15 Sentiment Result: polarity=-0.16666666666666666, subjectivity0.43333333333333335
16
17 857638640327020548 Thu Apr 27 16:52:40 +0000 2017 Xeehsi_Luokci
18 Glossy 3D Pearl Gorgeous Lace Cover Case Shell For Apple iPhone 4G 4S 4th 4FEE https://t.co/KXikhyJH8N https://t.co/x4ITYzs8FB
19 Sentiment Result: polarity=0.7, subjectivity0.9
20
21 857638639144239104 Thu Apr 27 16:52:40 +0000 2017 Xoutva_Gioyya
22 NIB LIFEPROOF iPhone 4 & 4S Black Case https://t.co/qlhuM2x2eB https://t.co/TOzTZ0VYP7
23 Sentiment Result: polarity=-0.16666666666666666, subjectivity0.43333333333333335
24
25 857638651433496576 Thu Apr 27 16:52:43 +0000 2017 Xouwqo_Yoejwe
26 NIB LIFEPROOF iPhone 4 & 4S Black Case https://t.co/w0qZyke7c0 https://t.co/U3onetrG3s
27 Sentiment Result: polarity=-0.16666666666666666, subjectivity0.43333333333333335
28
29 857638652855320577 Thu Apr 27 16:52:43 +0000 2017 Offylia
30 I really want to buy an iphone 7, i can afford it but i dont think it's a good economic decision.
31 Sentiment Result: polarity=0.36666666666666664, subjectivity0.3333333333333333
32
33 857638654038220800 Thu Apr 27 16:52:43 +0000 2017 Xoutva_Gioyya
34 Glossy 3D Pearl Gorgeous Lace Cover Case Shell For Apple iPhone 4G 4S 4th 4FEE https://t.co/154M9aJ5cS https://t.co/oKbdHv15dC
35 Sentiment Result: polarity=0.7, subjectivity0.9
```


Şekil 8.4: Toplanan tweetlerin tweets.txt dosyasındaki hali

9. ANALİZ SONUCU

Textblob gelen veriyi işleyerek analiz sonucunu öznellik ve kutupsallık özelliklerine göre belirli değerler verir. Fakat bazı verileri yanlış analiz edebilmektedir. Filtrelenen sözcük hakkında kinaye içeren cümleler veya bu sözcüklerin olduğu fakat hakkında herhangi bir yorum yapılmayan cümleleri yanlış analiz etmiştir.





9.1. Doğru Analiz Örnekleri

- iPhone 7 cameras are dark sided and I don't know how to work FaceTune
 - Sentiment Result: polarity = -0.15, subjectivity = 0.4
- iPhone dumb with the battery life
 - Sentiment Result: polarity=-0.375, subjectivity0.5

- my iphone 7 plus is stupid af
 - Sentiment Result: polarity=-0.7999999999999999, subjectivity1.0
- This is the next iPhone that Apple should release <https://t.co/7nrndHOcjQ>
<https://t.co/FNssQkG4vv>
 - Sentiment Result: polarity=0.0, subjectivity0.0
- Share your iPhone home screen!
 - Sentiment Result: polarity = 0.0, subjectivity = 0.0
- I so want to #win this new 128GB Limited Edition Red iPhone 7 from @Gleamapp 
<https://t.co/39QZ4CfqOh>
 - Sentiment Result: polarity=0.21623376623376625, subjectivity0.24935064935064932
- Sometimes the iPhone 7+ camera looks better than real life
 - Sentiment Result: polarity = 0.35, subjectivity = 0.4
- glad iphone stock photos #silvercsgo
 - Sentiment Result: polarity=0.5, subjectivity1.0
- I'm now a B-List+++ celebrity in Kim Kardashian: Hollywood. You can be famous too by playing on iPhone! <https://t.co/1XntuGUsfX>
 - Sentiment Result: polarity=0.625, subjectivity1.0

9.2. Yanlış Analiz Örnekleri

9.2.1. Pozitif Olması Gereken Örnekler

- Can someone please buy me an iPhone 7+   
 - Sentiment Result: polarity=0.0, subjectivity0.0
- #Top3Apps for #GautengAfricaTrade Twitter for iPhone 51% Twitter Web Client 32%
Twitter for Android 10%
 - Sentiment Result: polarity=0.0, subjectivity0.0
- My iPhone battery lies to me on a daily basis 
 - Sentiment Result: polarity=0.0, subjectivity0.0

9.2.2. Negatif Olması Gereken Örnekler

- All we want is group FaceTime calls & unlimited battery fam....
<https://t.co/siPtg8EWV2>
 - Sentiment Result: polarity=0.0, subjectivity0.0
- broke another iphone how cool
 - Sentiment Result: polarity=0.35, subjectivity0.65

9.2.3. Nötr Olması Gereken Örnekler

- Want to create better video content for your business? join my workshop
<https://t.co/ssDGGM4Ecq> <https://t.co/5stqniuLFH> #iphone #android
 - Sentiment Result: polarity=0.5, subjectivity0.5
- The rumored iPhone 5SE could come in hot pink, not rose gold
<https://t.co/iW1gyQ2GV4> <https://t.co/55kqrVP4Lw>
 - Sentiment Result: polarity=-0.049999999999999996, subjectivity0.7000000000000001
- NAZTECH Vault Waterproof Cover for iPhone SE/ 5 / 5s - Black
<https://t.co/yEoHtAEfXI>
 - Sentiment Result: polarity=-0.16666666666666666, subjectivity0.43333333333333335
- People with bad acne using the iPhone 7+ camera is tooo damn funny ☹️
 - Sentiment Result: polarity=-0.22499999999999992, subjectivity0.8333333333333333
- 1 awesome users just followed me. Via @FindUnfollower <https://t.co/IOQc6cuueC>. #iPhone #App
 - Sentiment Result: polarity=1.0, subjectivity1.0

10. TÜRKÇE DUYGU ANALİZİ

TextBlob kullanarak Türkçe tweetlerle analiz yapıldığında doğru hiçbir sonuç alınamıyor. Gelen tüm verilerin kutupsallık ve öznellik değeri 0.0 olarak çıkıyor. Aşağıda bunun örnekleri yer alıyor:

- iPhone yere düşmüş ahizesi çalışmıyormuş diye konuşamıyoruz saçmalığa bakar mısınız?

- Sentiment Result: polarity=0.0, subjectivity0.0
- Telefonum iPhone 7plus diyemi Twitter bende hata vermiyor? 😊😊
 - Sentiment Result: polarity=0.0, subjectivity0.0
- iPhone kullanıcıları çok sadık! <https://t.co/eanprfAcxs> <https://t.co/87nE8qwm5j>
 - Sentiment Result: polarity=0.0, subjectivity0.0
- Baya güzel iPhone için duvar ekran
 - Sentiment Result: polarity=0.0, subjectivity0.0
- @XezaleGecece @yakkadin 5 yaşındaki çocuk bile iphone kullanıyor. O aradığınız zor biraz işte 😊
 - Sentiment Result: polarity=0.0, subjectivity0.0

11. KAYNAKÇA

- <http://www.morganclaypool.com/doi/pdfplus/10.2200/S00416ED1V01Y201204HLT016>
- http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=7359041#ref_12
- <http://www.sciencedirect.com/science/article/pii/S2090447914000550>
- <http://crowdsourcing-class.org/assignments/downloads/pak-paroubek.pdf>
- <http://www.clips.ua.ac.be/sites/default/files/ctrs-001-small.pdf>
- <https://www.irjet.net/archives/V4/i3/IRJET-V4I3581.pdf>
- <https://www.digitalocean.com/community/tutorials/how-to-create-a-twitter-app>
- <http://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/>
- <http://socialmedia-class.org/twittertutorial.html>
- <http://www.json.org/>
- <https://dev.twitter.com/overview/api/tweets>
- <https://textblob.readthedocs.io/en/dev/>
- <https://dev.twitter.com/oauth>
- <https://oauth.net/>
- http://textblob.readthedocs.io/en/dev/advanced_usage.html
- http://textblob.readthedocs.io/en/dev/api_reference.html#textblob.en.sentiments.PatternAnalyzer
- <http://www.clips.ua.ac.be/pattern>
- <https://dev.twitter.com/streaming/overview/connecting>