BALIKESİR ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

TWİTTER'DA VERİ MADENCİLİĞİ İLE TÜRKİYE'DEKİ EĞİTİM SİSTEMİNİN NETWORK VE DUYGU ANALİZİ

BİTİRME PROJESİ

Batuhan ÇİMEN
Selin Esin KÖKÇE
Umut SARGINCAN

DANIŞMAN: DR. KAMİL TOPAL

2021

ÖZET

TWITTER'DA VERİ MADENCİLİĞİ İLE TÜRKİYE'DEKİ EĞİTİM SİSTEMİNİN NETWORK VE DUYGU ANALİZİ

LİSANS TEZİ

Batuhan ÇİMEN
Selin Esin KÖKÇE
Umut SARGINCAN

BALIKESİR ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

DANIŞMAN: DR. KAMİL TOPAL

BALIKESİR, 2021

Günümüzde giderek daha fazla insan, herhangi bir konu hakkındaki düşüncelerini belirtmek için sosyal medyayı kullanıyor. Yayınladıkları bu düşünceler, bireylerin hoşlandığı veya hoşlanmadığı şeyler hakkında bazı yararlı bilgiler içerebilir. Dolayısıyla işlenmemiş bu verilerin analizi, değerli öngörüler oluşturulmasına yardımcı olabilir. Bu projenin amacı Twitter kullanan kişilerin, eğitim sistemi hakkındaki düşüncelerinin olumlu veya olumsuz duygu analizinin yapılmasıdır. Çalışma Ekim 2020 ile Şubat 2021 arasında atılan Türkçe tweetleri kapsamaktadır. Buna ek olarak, eğitim ve eğitim sistemi konusunda Twitter'da en etkili kişiyi bulmak amacıyla network analizi yapılmıştır. Duygu analizi kısmında Google Cloud API, network analizi kısmında ise Gephi programı kullanılmıştır. Veri madenciliği, doğal dil işleme ve makine öğrenmesi yöntemlerinden yararlanılmıştır.

Anahtar Kelimeler: Twitter, Duygu Analizi, Network Analizi, Veri Tabanı, Veri Madenciliği, Gephi

ABSTRACT

NETWORK AND SENTIMENT ANALYSIS OF THE EDUCATIONAL SYSTEM IN TURKEY WITH DATA MINING ON TWITTER

SENIOR THESIS

Batuhan ÇİMEN
Selin Esin KÖKÇE
Umut SARGINCAN

BALIKESIR UNIVERSITY FACULTY OF ENGINEERING

COMPUTER ENGINEERING DEPARTMENT

SUPERVISOR: DR. KAMİL TOPAL

BALIKESİR, 2021

More and more people nowadays use social media to express their thoughts on any topic. These thoughts they publish may contain some useful information about the likes or dislikes of individuals. Therefore, analysis of this raw data can help generate valuable insights. The aim of this project is to analyze the positive or negative sentiments of people using Twitter about the education system. The study includes Turkish tweets sent between October 2020 and February 2021. In addition, a network analysis was conducted to find the most influential person on Twitter about education and the education system. Google Cloud API was used in sentiment analysis and Gephi program was used in network analysis. Data mining, natural language processing and machine learning methods were used.

Keywords: Twitter, Sentiment Analysis, Network Analysis, Database, Data Mining, Gephi

ÖNSÖZ

Bu tez çalışmasında Twitter'da eğitim sistemi ile ilgili Türkçe atılmış olan tweetler ile toplumun eğitime genel bakışı duygu analizi ve sosyal ağ analizi yöntemleri ile incelenmiştir.

Öncelikle tez konusunu seçerken isteklerimizi göz önünde bulundurup bize tezimizi geliştirmek için yardımcı olan tez danışmanımız Dr. Kamil TOPAL a teşekkürlerimizi sunarız. Sosyal ağ analizi konusunda değerlendirme ölçütlerinin belirlenmesinde değerli tavsiyelerde bulunan Prof. Dr. Gökhan Barış BAĞCI hocamıza teşekkürlerimizi bir borç biliriz.

İÇİNDEKİLER

Sayfa

KISALTMALAR DİZİNİ

Kısaltmalar	Açıklama
ACID	Atomicity, Consistency, Isolation, Durability
AI	Artificial Intelligence
API	Application Programming Interface
CSV	Comma-Separated Values
DBMS	Database Management Systems
JSON	JavaScript Object Notation
KDD	Knowledge Discovery in Textual Databases
KDT	Knowledge Discovery in Textual Databases
KPSS-B	Kamu Personeli Seçme Sınavı-B Grubu
ML	Machine Learning
RT	Retweet
SMO	Sequential Minimal Optimization
YSA	Yapay Sinir Ağları

ŞEKİLLER DİZİNİ

Sayfa

TABLOLAR DİZİNİ

Sayfa

1. GİRİŞ

Günümüzde internetin hızla gelişmesi, interneti insanların hayatında vazgeçilmezlerden birisi yapmıştır. İnsanların kişisel fikirlerini, ilgi alanlarını paylaşmak istemesi sosyal medyayı önemli bir bilgi kaynağı haline getirmiştir. Paylaşılan bu düşünceler, bireylerin hoşlandığı veya hoşlanmadığı şeyler hakkında yararlı bilgiler içerebiliyor. Dolayısıyla işlenmemiş bu verilerin analizi, değerli öngörüler oluşturulmasına yardımcı olabilmektedir. Sosyal medya verilerinin işlenmemiş haliyle üzerinde çalışılması oldukça zordur ve bu veriler incelendiğinde çoğunluğunun hatalı yazılmış kelimeler, kısaltmalar ve günlük konuşma dilinde kullanılmayan sosyal medyaya özgü jargon sözcüklerden oluştuğu görülmektedir. Bu nedenle verilerin doğal dil işleme yöntemleri ile süzülmesi ve işlenmesi gerekmektedir. Biz de çeşitli yöntemler kullanarak verileri topladık, işledik ve öngörüler çıkardık.

Twitter'dan toplanan veriler ile doğal dil işleme ve veri madenciliği alanında pek çok çalışma yapılmaktadır. Bu çalışmalara örnek olarak:

Bahri Baran KOCAK ve arkadaşlarının Nisan-Mayıs 2016 tarihleri arasında atılmış olan tweetlerden Twitter kullanıcılarının havayolu pazarına yönelik duygu kutuplarının belirlenmesi çalışmasında 8672 kullanıcı yorumu olumlu, nötr ve olumsuz etiketlerle ayrıştırılmıştır. Elde edilen etiketler etiket bulutunda toplanmış ve sonuçlar makine öğrenmesi yöntemi, SMO sınıflandırmasında standart ve normalize kernel polinomları ile analiz edilmiştir. Analiz sonucu başarı ölçütleri incelendiğinde ise standart kernel polinomunun SMO algoritması içinde daha iyi bir performans gösterdiği; ancak genel olarak sınıflandırma başarısının düşük çıktığı görülmektedir (Koçak, Polat, & Koçak, 2016).

Beyzanur BOSTANCI ve arkadaşlarının yaptığı duygu analizi ile kişiye özel içerik önerme çalışmasında Facebook ve Twitter sosyal medya platformlarında paylaşılan kullanıcı yorumları duygu analizi teknikleri ile değerlendirilmiştir. 65 Facebook kullanıcısı ve 82 Twitter kullanıcısı profilinden alınan veriler ilk olarak veri ön işleme sürecine tabi tutulmuştur. Ardından gürültüsüz olan verilerin, belirlenen kategorilere göre sınıflandırılması yapılmıştır. Sınıflandırma yapılırken her kategori için 25 tane anahtar kelime belirlenmiştir. Seçilen bu kelimelerin tekrarlanma sıklığını ölçmek için TF – IDF

tekniği kullanılmıştır. Bu teknik sonucunda %77.82 ile dışadönük kategorisinin diğer kategorilere göre doğruluk oranı daha fazla olduğu gözlemlenmiş ve bu duygu analizinin sonucunda, reklam amaçlı Düzce üniversitesi için beş farklı afiş oluşturulmuştur (Bostancı & Albayrak, 2021).

Ayşe BEŞKİRLİ ve arkadaşlarının metin madenciliği yöntemleri ile Twitter verilerinden bilgi keşfi çalışmasında aşı kelimesinin İngilizcesi olan vaccine kelimesi ile ilgili Twitter platformu üzerindeki konuşmalardan veriler toplanarak vaccine ile ilgili aşı duyurusu öncesi ve sonrası olmak üzere iki farklı veri seti oluşturulmuştur ve bu veri setlerine analizler yapılmıştır. Yapılan analiz sonuçları şekiller ile görselleştirilip tablolarda tüm verilerin duygu analizleri kıyaslanmıştır. Kıyaslama sonucunda aşı çalışmaları sürecinde iken sosyal medya konuşmaları olumlu ve olumsuz düşünceler neredeyse aynı oranda iken nötr düşüncelerin daha fazla oranda olduğu görülmüştür (Beşkirli, Gülbandılar, & Dağ, 2021).

Buğra AYAN ve arkadaşlarının 2019 yılında yaptığı Twitter üzerindeki islamofobik twitlerin duygu analizi ile tespiti çalışmasında 162.000 tweet incelenmiş ve bunların islamofobik olup olmadığının belirlenmesi için duygu analizi yapılmıştır. Duygu analizi için Naive-Bayes sınıflandırma algoritması ve Ridge regresyonu kullanılmıştır. Bu çalışmada Ridge ile birlikte Naive-Bayes daha verimli çalışırken, bunun yanında sınıflandırma işlemi daha kısa sürmüştür (Ayan, Kuyumcu, & Ciylan, 2019).

Bu çalışmada eğitim sistemi hakkında olumlu ya da olumsuz bir yargıya varılması ve ağdaki etkili kişilerin belirlenmesi hedeflenmektedir.

Tezin ilerleyen bölümlerinde sırasıyla kavramlar, materyal, yöntem ve alınan sonuçlardan bahsedilecektir.

2. KAVRAMLAR

2.1 Yapay Zeka

En basit ifadeyle yapay zeka (AI), görevleri yerine getirmek için insan zekasını taklit eden ve topladıkları bilgilere göre yinelemeli olarak kendilerini iyileştirebilen sistemler veya makineler anlamına gelir. Yapay zeka pek çok biçimde kendini gösterir. Örneğin:

- Sohbet robotları, müşterilerin sorunlarını daha hızlı bir şekilde anlamak ve daha verimli cevaplar vermek için yapay zekadan yararlanır.
- Akıllı asistanlar, zamanlamayı iyileştirmek için büyük kullanıcı tanımlı veri kümelerinden kritik bilgileri çekmek için yapay zekadan yararlanır.
- Öneri motorları kullanıcıların izleme alışkanlıklarına göre TV programları için otomatik önerilerde bulunabilir.

Yapay zeka, herhangi bir özel biçim veya işlevden ziyade süper güçlendirilmiş düşünce ve veri analizi yeteneği ve süreciyle ilgilidir. Yapay zeka üst seviye işleve sahip insan benzeri robotların dünyayı ele geçirmesine ilişkin görüntüler sunsa da, yapay zekanın amacı insanların yerini almak değildir. Amaç insan yeteneklerini belirgin şekilde geliştirmek ve bunlara katkıda bulunmaktır. Bu nedenle oldukça değerli bir ticari varlıktır (Oracle, 2021).

2.2 Makine Öğrenmesi

Makine öğrenmesi (ML), bir bilgisayarın doğrudan yönergeler olmadan öğrenmesine yardımcı olmak için matematiksel modelleri kullanma işlemidir. Bu, yapay zekanın (AI) bir alt kümesi olarak kabul edilir. Makine öğrenmesi, verilerdeki kalıpları belirlemek için algoritmaları kullanır. Tahmin yapabilen bir veri modeli oluşturmak için de bu kalıplar kullanılır. Tıpkı insanların daha fazla alıştırma yaptıkça gelişmesi gibi, veri ve deneyim miktarı arttıkça makine öğrenmesinin sonuçları da daha doğru hale gelir.

Uyarlanabilirliği sayesinde makine öğrenmesi verilerin, isteklerin veya görevlerin sürekli değiştiği senaryolarda veya bir çözümün etkili bir şekilde kodlanmasının mümkün olmadığı durumlarda harika bir seçenektir (Microsoft, 2021).

2.3 Veri Madenciliği

Veri madenciliği, büyük miktarlardaki verinin içinden geleceği tahmin edilmesinde yardımcı olacak anlamlı ve yararlı bağlantı ve kuralların bilgisayar programlarının aracılığıyla aranması ve analizidir. Ayrıca veri madenciliği, çok büyük miktardaki verilerin içindeki ilişkileri inceleyerek aralarındaki bağlantıyı bulmaya yardımcı olan ve veri tabanı sistemleri içerisinde gizli kalmış bilgilerin çekilmesini sağlayan veri analizi tekniğidir. Bu işlemlerin uygulama alanı oldukça geniştir.

Veri madenciliği araçları kullanılarak, işletmelerin daha etkin kararlar almasına yönelik karar destek sistemlerinde gerekli olan eğilimlerin ve davranış kalıplarının ortaya çıkarılması mümkün olmaktadır. Geçmişteki klasik karar destek sistemlerinin kullanıldığı araçlardan farklı olarak, veri madenciliğinde çok daha kapsamlı ve otomatize edilmiş analizler yapmaya yönelik, birçok farklı özellik bulunmaktadır. Veri madenciliğinin işletmelere sunduğu en önemli özellik, veri grupları arasındaki benzer eğilimlerin ve davranış kalıplarının belirlenmesidir. Bu süreç aynı zamanda otomatize edilmiş bir biçimde hayata geçirilebilmektedir. Bu fonksiyon özellikle hedef pazarlara yönelik pazarlama faaliyetlerinde yoğun olarak kullanılmaktadır. Başka bir özelliği ise daha önceden bilinmeyen, veri ambarları içerisinde bulunan ancak ilk etapta görülemeyen bilgilerin ortaya çıkarılabilmesidir. Örneğin bir firma sattığı ürünleri analiz ederek, ilerideki kampanyalarını şekillendirebilir ya da sattığı ürünler arasındaki bağları keşfedebilir. Burada amaç daha önceden fark edilmeyen veri kümelerinin bulunabilmesidir. Günümüzün ekonomik koşulları ve yaşanan hızlı değişim ortamlarında, iş deneyimi ve önsezilere dayanarak alınan kararlarda yanlış karar alma riski çok yüksektir. Riski azaltmanın tek yolu bilgiye dayalı yönetimi öngören karar destek çözümleridir. Veri madenciliği teknikleri gerçek anlamda bir karar destek sistemi oluşturmada olmazsa olmaz araçlardır. Bu noktada bilgi teknolojilerinden yararlanmak kaçınılmaz olmuştur.

2.3.1 Veri Madenciliğini Etkileyen Etmenler

Veri madenciliği temel olarak 5 ana faktörden etkilenir:

- Veri: Veri madenciliğinin bu kadar gelişmesindeki en önemli faktördür.
- Donanım: Gelişen bellek ve işlem hızı kapasitesi sayesinde, birkaç yıl önce madencilik yapılamayan veriler üzerinde çalışmayı mümkün hale getirmiştir.
- Bilgisayar ağları: Yeni nesil internet, çok yüksek hızları kullanmayı sağlamaktadır. Böyle bir bilgisayar ağı ortamı oluştuktan sonra, dağıtık verileri analiz etmek ve farklı algoritmaları kullanmak mümkün olacaktır.
- Bilimsel hesaplamalar: Günümüz bilim adamları ve mühendisleri, simülasyonu, bilimin üçüncü yolu olarak görmekteler. Veri madenciliği ve bilgi keşfi, teori, deney ve simülasyonu birbirine bağlamada önemli bir rol almaktadır.
- Ticari eğilimler: Günümüzde, işletmeler rekabet ortamında varlıklarını koruyabilmek için daha hızlı hareket etmeli, daha yüksek kalitede hizmet sunmalı, bütün bunları yaparken de minimum maliyeti ve en az insan gücünü göz önünde bulundurmalıdır.

2.3.2 Veri Madenciliğini Karşılaşılan Problemler

Büyük hacimli verilerin bulunduğu veri ortamlarında büyük sorunlar ortaya çıkabilir. Bu nedenle küçük veri kümelerinde, benzetim ortamlarında hazırlanmış veri madenciliği sistemleri, büyük hacimli, eksik, gürültülü, boş, atık, aykırı veya belirsiz veri kümelerinin bulunduğu ortamlarda yanlış çalışabilir. Bu nedenle veri madenciliği sistemleri hazırlanırken bu sorunların çözülmesi gerekmektedir. Veri madenciliği uygulamalarında karşılaşılabilecek sorunlar şunlardır:

- Artık veri: Artık veri, problemde istenilen sonucu elde etmek için kullanılan örneklem kümesindeki gereksiz niteliklerdir. Bu durum pek çok işlem sırasında karşımıza çıkabilir.
- Belirsizlik: Yanlışlıkların şiddeti ve verideki gürültünün derecesi ile ilgilidir.

- Boş veri: Bir veri tabanında boş değer, birincil anahtarda yer almayan herhangi bir niteliğin değeri olabilir. Boş değer, tanımı gereği kendisi de dâhil olmak üzere hiçbir değere eşit olmayan değerdir.
- Dinamik veri: Kurumsal çevrim içi veri tabanları dinamiktir ve içeriği sürekli olarak değişir. Bu durum, bilgi keşfi metotları için önemli sakıncalar doğurmaktadır.
- Eksik veri: Veri kümesinin büyüklüğünden ya da doğasından kaynaklanmaktadır. Eksik veriler olduğunda yapılması gerekenler şunlardır:

  - Eksik veri içeren kayıt veya kayıtlar çıkarılabilir.
  - Değişkenin ortalaması eksik verilerin yerine kullanılabilir.
  - Var olan verilere dayalı olarak en uygun değer kullanılabilir.

  Eksik veriler, yapılacak olan istatistiksel analizlerde önemli problemler yaratmaktadır. Çünkü istatistiksel analizler ve bu analizlerin yapılmasına olanak veren ilgili paket programlar, verilerin tümünün var olduğu durumlar için geliştirilmiştir.
- Farklı tipteki verileri ele alma: Gerçek hayattaki uygulamalar makine öğreniminde olduğu gibi yalnızca sembolik veya kategorik veri türleri değil, fakat aynı zamanda tamsayı, kesirli sayılar, çoklu ortam verisi, coğrafi bilgi içeren veri gibi farklı tipteki veriler üzerinde işlem yapılmasını gerektirir.
- Gürültülü ve kayıp değerler: Veri girişi veya veri toplanması esnasında oluşan sistem dışı hatalara gürültü denir. Büyük veri tabanlarında pek çok niteliğin değeri yanlış olabilir. Veri toplanması esnasında oluşan hatalara ölçümden kaynaklanan hatalar da dâhil olmaktadır. Bu hataların sonucu olarak birçok niteliğin değeri yanlış olabilir ve bu yanlışlardan dolayı veri madenciliği amacına tam olarak ulaşmayabilir.
- Sınırlı bilgi: Veri tabanları genel olarak basit öğrenme işlerini sağlayan özellik veya nitelikleri sunmak gibi veri madenciliği dışındaki amaçlar için hazırlanmışlardır. Bu yüzden, öğrenme görevini kolaylaştıracak bazı özellikler bulunmayabilir.

- Veri tabanı boyutu: Veri tabanı boyutları büyük bir hızla artmaktadır. Veri tabanı algoritması çok sayıda küçük örneklemi ele alabilecek biçimde geliştirilmiştir. Aynı algoritmaların yüzlerce kat büyük örneklemlerde kullanılabilmesi için çok dikkat gerekmektedir (Savaş, Topaloğlu, & Yılmaz, 2011).

2.3.3 Veri Madenciliğinde Kullanılan Yöntemler

a. Karar Ağaçları (Decision Trees)

Veri madenciliğinde karar ağaçları, kurulmasının ucuz olması, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca entegre edilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip tekniktir. Karar ağacı, adından da anlaşılacağı gibi bir ağaç görünümünde, tahmin edici bir tekniktir (Berry Michael and Linoff Gordon, 1999). Ağaç yapısı ile, kolay anlaşılabilen kurallar yaratabilen, bilgi teknolojileri işlemleri ile kolay entegre olabilen en popüler sınıflama tekniğidir (Curtarolo Stefano and Morgan Dane, 2003). Karar ağacı karar düğümleri, dallar ve yapraklardan oluşur (Han Jiawei and Kamber Micheline, 2000). Karar düğümü, gerçekleştirilecek testi belirtir. Bu testin sonucu ağacın veri kaybetmeden dallara ayrılmasına neden olur. Her düğümde test ve dallara ayrılma işlemleri ardışık olarak gerçekleşir ve bu ayrılma işlemi üst seviyedeki ayrımlara bağımlıdır. Ağacın her bir dalı sınıflama işlemini tamamlamaya adaydır. Eğer bir dalın ucunda sınıflama işlemi gerçekleşemiyorsa, o dalın sonucunda bir karar düğümü oluşur. Ancak dalın sonunda belirli bir sınıf oluşuyorsa, o dalın sonunda yaprak vardır. Bu yaprak, veri üzerinde belirlenmek istenen sınıflardan biridir. Karar ağacı işlemi kök düğümünden başlar ve yukarıdan aşağıya doğru yaprağa ulaşana dek ardışık düğümleri takip ederek gerçekleşir.

a. Yapay Sinir Ağları (Artificial Neural Networks)

Yapay sinir ağları (YSA), temelde tamamen insan beyni örneklenerek geliştirilmiş bir teknolojidir. Bilindiği gibi; öğrenme, hatırlama, düşünme

gibi tüm insan davranışlarının temelinde sinir hücreleri bulunmaktadır. İnsan beyninde tahminen 1011 adet sinir hücresi olduğu düşünülmektedir ve bu sinir hücreleri arasında sonsuz diyebileceğimiz sayıda sinaptik birleşme denilen sinirler arası bağ vardır. Bu sayıdaki bir birleşimi gerçekleştirebilecek bir bilgisayar sisteminin dünya büyüklüğünde olması gerektiği söylenmektedir; ancak 50 yıl sonra bunun büyük bir yanılgı olmayacağını bu günden kimse söyleyemez. İnsan beyninin bu karmaşıklığı göz önüne alındığında, günümüz teknolojisinin 1.5 kg'lık İnsan beynine oranla henüz çok geride olduğunu söylemek yanlış olmaz (Edelstein Herbert, 1999). YSA'nın hesaplama ve bilgi işleme gücünü, paralel dağılmış yapısından, öğrenebilme ve genelleme yeteneğinden aldığı söylenebilir. Genelleme, eğitim ya da öğrenme sürecinde karşılaşılmayan girişler için de YSA'nın uygun tepkileri üretmesi olarak tanımlanır. Bu üstün özellikleri, YSA'nın karmaşık problemleri çözebilme yeteneğini gösterir.

b. Genetik Algoritmalar (Genetic Algorithms)

Genetik algoritmalar, doğada gözlemlenen evrimsel sürece benzer bir şekilde çalışan arama ve eniyileme yöntemidir. Karmaşık çok boyutlu arama uzayında en iyinin hayatta kalması ilkesine göre bütünsel en iyi çözümü arar. Genetik algoritmalar problemlere tek bir çözüm üretmek yerine farklı çözümlerden oluşan bir çözüm kümesi üretir. Böylelikle, arama uzayında aynı anda birçok nokta değerlendirilmekte ve sonuçta bütünsel çözüme ulaşma olasılığı yükselmektedir. Çözüm kümesindeki çözümler birbirinden tamamen bağımsızdır. Her biri çok boyutlu uzay üzerinde bir vektördür. Genetik algoritmalar problemlerin çözümü için evrimsel süreci bilgisayar ortamında taklit ederler. Diğer eniyileme yöntemlerinde olduğu gibi çözüm için tek bir yapının geliştirilmesi yerine, böyle yapılardan meydana gelen bir küme oluştururlar. Problem için olası pek çok çözümü temsil eden bu küme genetik algoritma terminolojisinde nüfus adını alır. Nüfuslar vektör veya birey adı verilen sayı dizilerinden

oluşur. Birey içindeki her bir elemana gen adı verilir. Nüfustaki bireyler evrimsel süreç içinde genetik algoritma işlemcileri tarafından belirlenir.

c. K-En Yakın Komşu (K-Nearest Neighbor)

Veri madenciliğinde sınıflama amacıyla kullanılan bir diğer teknik ise örnekseme yoluyla öğrenmeye dayanan k-en yakın komşu algoritmasıdır. Bu teknikte tüm örneklemler bir örüntü uzayında saklanır. Algoritma, bilinmeyen bir örneklemin hangi sınıfa dahil olduğunu belirlemek için örüntü uzayını araştırarak bilinmeyen örnekleme en yakın olan k örneklemi bulur. Yakınlık Öklid uzaklığı ile tanımlanır. Daha sonra, bilinmeyen örneklem, k en yakın komşu içinden en çok benzediği sınıfa atanır. K-en yakın komşu algoritması, aynı zamanda, bilinmeyen örneklem için bir gerçek değerin tahmininde de kullanılabilir.

d. K-Means

K-Means algoritması bir unsupervised learning(denetimsiz öğrenme) ve kümeleme algoritmasıdır. K-Means' teki K değeri küme sayısını belirler ve bu değeri parametre olarak alması gerekir. K adet özgün küme oluşturduğu ve her kümenin merkezi, kümedeki değerlerin ortalaması olduğu için K - Ortalamalar denmektedir. Algoritma istatistiksel olarak benzer nitelikteki kayıtları aynı gruba sokar. Bir elemanın yalnızca bir kümeye ait olmasına izin verilir. Küme merkezi kümeyi temsil eden değerdir.

e. Naive-Bayes

Naive-Bayes algoritmasında her kriterin sonuca olan etkilerinin olasılık olarak hesaplanması temeline dayanmaktadır. Veri Madenciliği işlemini en çok verilen örneklerden biri ile açıklayacak olursak elimizde tenis maçının oynanıp oynanmamasına dair bir bilgi olduğunu düşünelim. Ancak bu bilgiye göre tenis maçının oynanması veya oynanmaması durumu kaydedilirken o anki hava durumu, sıcaklık, nem ve rüzgâr durumu bilgileri de alınmış olsun. Biz bu bilgileri değerlendirdiğimizde varsayılan tahmin yöntemleri ile hava bugün rüzgârlı tenis maçı bugün oynanmaz şeklinde

kararları farkında olmasak da veririz. Ancak Veri Madenciliği bu kararların tüm kriterlerin etkisi ile verildiği bir yaklaşımdır. Dolayısıyla biz ileride öğrettiğimiz sisteme bugün hava güneşli, sıcak, nemli ve rüzgâr yok şeklinde bir bilgiyi verdiğimizde sistem eğitildiği daha önce gerçekleşmiş istatistiklerden faydalanarak tenis maçının oynanma ve oynanmama ihtimalini hesaplar ve bize tahminini bildirir (Ayık, Özdemir, & Yavuz, 2010).

2.4 Metin Madenciliği

Metin madenciliği ile yapısal olmayan veriler yapısal bir hale getirilerek, analitik analize uygun bir kaynak elde edilir. Yazılı metinlerin tamamı veya yazılı olmayan fotoğraf, harita gibi görseller metin madenciliğinin konusunu oluşturabilir. Bunun yanı sıra; makale, gazete, kitap, akademik yayınlar, internet siteleri gibi tüm topluma açık olan metinler, e posta, hastalar raporları, adli sicil kayıtları, mektuplar, gibi bireye özel metinler de, metin madenciliğinin kaynağı olabilirler. Günümüzde sosyal medyanın yaygınlığı nedeniyle metin madenciliği denilince ilk akla gelen şey sosyal medyadaki paylaşımlar olmaktadır. Gün içinde pek çok kişi sosyal medyadan o ana dair, güncel konular hakkında, izlediği bir film veya diziye ilişkin, arkadaşlarının yapmış oldukları paylaşımlara ilişkin, pek çok yorum ve paylaşımlarda bulunmaktadırlar. Bu paylaşımların tamamı metin madenciliği için muazzam bir kaynak oluşturmaktadır.
Metin madenciliği; çok fazla miktardaki metinlerden, yüksek kaliteli bilgileri ayıklamak için kullanılan hesaplama yöntem ve teknikleridir. Metin madenciliği ile metinlerde yer alan gereksiz kısımlar atılarak istenilen bilgiye erişmek amaçlanmaktadır. Yapılandırılmamış veri ham haliyle bize çok fazla bilgi vermezken, metin madenciliği sayesinde, bu veriler bizi çok önemli bilgilere ulaştıracak bir kaynak haline gelir. Metin madenciliği aslında büyük bir dağ kitlesinin altında yatan küçük bir pırlantayı elde etmek için sarf edilen çaba diye tanımlanmıştır. Verilerin çok büyük bir kısmı yapılandırılmamış veri yani dağ kitlesi halindedir, bu verilerin kullanılabilmesi için pek çok yöntem geliştirilmiştir. Metin madenciliği geliştirilmiş olan bu yöntemlerin bir bütünüdür.

Veri madenciliği büyük miktardaki veriyi analiz ederek bu veriden anlamlı, kullanılabilir bilgi elde etme sürecidir. Metin madenciliği ise veri madenciliği alanında

çok farklı ve yeni çığırlar açan, veri madenciliğine ilham ve yön veren bir alan olmuştur. Bu yüzdendir ki metin madenciliği ile veri madenciliği arasında birçok üst düzey mimari benzerlik vardır. Basit bir tanımla; veri madenciliği; veri ambarlarından hareketle, verilerden anlamlı bilgiler ve ilişkiler çıkaran, ancak bunu yaparken yapılandırılmış verileri kullanan, çok büyük miktarlardaki veri arasındaki ilişkiyi analiz etmeye çalışan tekniktir. Tanımdan da anlaşılacağı üzere veri madenciliği çalışma alanı yapılandırılmış veridir. Metin madenciliği ise yapısal olmayan verilerden hareketle veri madenciliğinin yaptığı işleri yapar. Aslında veri madenciliği ile metin madenciliği arasındaki temel fark kullanılan verilerin yapılandırılmış olup olmamasıdır. Metinler insanlar tarafından yazılıp okunabilirken, bilgisayarların yapılandırılmış verileri işlemesi için veri tabanları programlanmıştır. Metinleri insanlar gibi okuyup anlayabilecek bilgisayar programları henüz yoktur. Ancak bu alanda ciddi çalışmalar sürdürülmektedir. Gelecekte insanlar gibi metinleri okuyup anlayan bilgisayar programlarının geliştirilmesi sürpriz olmayacaktır.

Veri madenciliğiyle metin madenciliği arasındaki güçlü ilişki nedeniyle gelişim süreçleri de birbirlerine paraleldir. Bilgisayarların günlük hayatta hızla yaygınlaşması, veri depolamanın ve veriye ulaşmanın kolaylaşıp, ucuzlaması veri madenciliğini olduğu gibi metin madenciliğini de geliştiren temel neden olmuştur. 1960'lı yıllarda başlayan bu süreç 2000'den sonra hızlanarak gelişimini sürdürmektedir. 1990'lardan sonra metin veri tabanlarında bilgi keşfi (Knowledge Discovery in Textual Databases – KDT) alanında yoğun bir şekilde çalışılmaya başlanmış, 1995 yılında Knowledge Discovery in Textual Databases (KDD) konferanslarının birincisi düzenlemiş, metnin yapısal özelliği nedeniyle işlenmesi zor olduğundan, metin analizleri için algoritmaların geliştirilmesi 1990'ların sonlarını bulmuştur. 90'lardan sonra metin analizi üzerine çalışmalar hızlanmış ve günümüze kadar gelmiş, bu alanda çok çeşitli yöntemler geliştirilmiştir.

Sosyal medyanın, yaşamın bir vazgeçilmezi hale gelmesiyle, gün içinde özelikle internet ortamında, milyonlarca metin halindeki veri ortaya çıkmaktadır. Bu verilerin işlenmesi için geliştirilen metotlar metin madenciliğinin çok fazla gelişmesine yol açmıştır. Yapılan araştırmaya göre Twitter'da, 1 dakikada 347.222 adet tweet atılmaktadır. Facebook gibi Twitter'dan daha fazla kullanılan diğer sosyal mecralar da

düşünüldüğünde anlık üretilen verinin büyüklüğü daha da çok ortaya çıkmaktadır. Dolayısıyla bu metinler, araştırmacılar için büyük bir nimet haline gelmiştir.

Yapılandırılmış veya yapılandırılmamış halde bulunan veriler, metin madenciliği için birer girdidirler. Metin madenciliği sürecinin çıktısı ise; karar vermek için kullanılacak olan özel bilgilerdir. Süreç; yazılım donanım kısıtları, güvenlik sorunları ve dilbilim kısıtları gibi zorlukları içermektedir. Metin madenciliğinin temel amacı verilerden hareketle, metinden anlamlı sonuçlar çıkarmak için metni işlemektir. Bu amaçla metinler çeşitli araç ve alan uzmanlığı (istatistik ve makine öğrenme) ile işlenirler (Kızılkaya, 2018).

2.5 Doğal Dil İşleme

Metin Madenciliğinde veriden bilgi çıkarma yöntemlerinden biri olan doğal dil işleme disiplini ile bilgi çıkarımında daha anlamlı sonuçlar elde edilmeye başlanmıştır. Doğal dil işleme ana işlevi bir doğal dili çözümleme, anlama, yorumlama ve üretme olan bilgisayar sistemlerinin tasarımını konu alan bir mühendislik alanıdır. Doğal dil işleme çalışmaları sayesinde insan-bilgisayar etkileşiminin arttırılması başarılmıştır. Metin Madenciliğinde, belgelerin analizi için, içeriğinin anlamını taşıyan kavramların tespit edilmesi gerekmektedir. Bu kavramlar kelimeler veya kelime grupları ile ifade edilir ve terimler olarak adlandırılır. Belge içindeki terimlerin çıkarılması başlı başına bir konudur ve doğal dil işleme çalışmaları kapsamında incelenen bir alandır. Doğal dil işlemenin belge analizi sürecindeki en önemli faydası terimlerin yani kelimelerin ayrıştırılması, eklerinden arındırılarak anlamını kaybetmeyen en kısa biçimlerine dönüştürülmesidir. Çünkü aynı anlam için kullanılan kelimeler dilbilgisi kuralları gereği farklı biçimlerde bulunabilir ve bu farklı kullanım biçimleri ortadan kaldırılmadığı takdirde farklı anlam taşıyan terimler gibi işleme alınarak, belgelerin gerçek anlamına ulaşılmasını engelleyebilirler. Doğal dil işleme çalışmaları kapsamında yürütülen girişimler dört ana grup altında toplanabilir:

- Biçimbirimsel çözümleme (Morfolojik analiz)
- Sözdizimi çözümlemesi (Sentaktik çözümleme)
- Anlam çözümlemesi (Semantik çözümlemesi)
- Anlam kargaşasının giderilmesi

2.5.1 Morfolojik Analiz (Biçimbirim)

Biçimbirim sözcüklerin yapısıyla ile ilgilenir. Türkçe için sözcüklerin türetilmesi ve ekler çok önem taşır. Her dilde iki farklı şekilde sözcük oluşturulabilir. Bunlardan biri çekim, diğeri ise türetme yöntemidir. Çekim yoluyla sözcük oluşturulurken bir sözcüğün farklı şekilleri kullanılır. Türetme ise var olan eski sözcüklere yapım ekleri eklenmesi yoluyla yeni sözcük oluşturma yöntemidir.

2.5.2 Sözdizimsel Analiz (Sentaktik Analiz)

Bilgisayarla doğal dil modellemelerinde anlamsal analize geçmeden önce, kelimeler yığınının geçerli bir cümle yapısı oluşturup oluşturmadığı kontrol edilmelidir. Rasgele kelimelerin yan yana gelmesiyle geçerli bir cümle meydana gelmeyecektir. Geçerli bir cümle yapısı oluşturulamadığı zaman, buradan anlam çıkarılmasını beklemek yanlış olacaktır. “küçük koş mavi” kelimeler topluluğu anlamlı bir yapı oluşturmamaktadır.

Sözdizimsel analiz, cümlenin yapısal bir tanımını oluşturabilmek için morfolojik analizin sonuçlarını kullanır. Bu işlemi yapmanın amacı, ardı ardına gelen kelime yığınlarının bu kelimeler yığınının ifade ettiği cümle birimlerini tanımlayan bir yapıya dönüştürmektir. Cümle birimleri, kelimeler tamlamalar veya buna benzer cümle parçacıkları olabilir.

Türkçe’de cümleler en genel şekliyle özne, nesne ve yüklem unsurlarından oluşur. Cümle ile ifade edilmek anlam arttıkça, cümlelere yer tamlayıcısı, zaman tamlayıcısı gibi yeni unsurlar eklenecektir. Bunların yanı sıra, cümlelerin anlamlarını kuvvetlendirmek için edat, bağlaç gibi unsurlar da cümlelere eklenebilir. Dillerin sözdizimi açısından sınıflandırılması önem taşımaktadır. Çeşitli dillerde cümlenin temel öğeleri olan özne (Ö), nesne (N) ve yüklemin (Y) düz cümledeki temel dizilişinin Ö-N-Y, Ö-Y-N ve Y-Ö-N şeklinde olduğu görülmüştür. Bunların yanı sıra, kullanılan diğer unsurlar, zarf tümleci (Z), dolaylı tümleç (D) olarak kabul edilebilir. Türkçenin anlamsal analizinden dolayı, önemsenen unsurlar cümlenin yüklemine yaklaştırılır.

2.5.3 Anlamsal Analiz (Semantik Analiz)

Bir cümlenin ne demek istediğinin anlaşılması, diğer bir deyişle bir cümle ile ifade edilmek istenilen duygu veya düşüncenin ne olduğunun anlaşılması, anlamsal analiz yardımıyla yapılır.

Anlamsal analiz yapılırken, öncelikli olarak kelimelerin tek tek veri tabanından uygun nesnelerle eşleştirilme işleminin yapılması gerekir. Bu işlem, her zaman birebir eşleme olamayabilir. Diğer bir deyişle, kelimelerin ifade ettikleri anlamlar her zaman bir tane olmayabilir. Ayrık kelimelerin bir cümledeki doğru anlamını bulma işlemine "kelime anlam berraklaştırılması" denir. Bu işlem, cümle içinde geçen bir kelimenin sözlükteki anlamlarının belirlenip bunlardan uygun olanının seçilmesidir. Cümle içinde geçen her bir kelime, diğer kelimelerin doğru anlamlarının ortaya çıkarılması için önem taşımaktadır (Ergün, 2011).

2.6 Veri Tabanı

Veri tabanı en genel tanımıyla, kullanım amacına uygun olarak düzenlenmiş veriler topluluğudur. Birbirleriyle ilişkileri olan verilerin tutulduğu, mantıksal ve fiziksel olarak tanımlarının olduğu bilgi depolarıdır. Veri tabanları gerçekte var olan ve birbirleriyle ilişkisi olan nesneleri ve ilişkileri modeller. Bu modellerin bir sistemleri vardır. Veri tabanı yönetim sistemleri (DBMS) olarak adlandırılır, verilere aynı anda birden çok bağlantı sağlayabilme özelliği sağlar. Bu sistemler, veri tabanı yönetiminin bir parçası olarak, verinin nasıl depolanacağı, kullanılacağı ve erişileceğini mantıksal olarak yönlendiren bir kurallar sistemidir.

2.6.1 İlişkisel Veri Tabanı (Relational Database System)

Günümüzde en yaygın kullanılan veri tabanı sistemlerinden biridir. Satır ve sütunların meydana getirdiği tablolardan oluşur. Bu tablolar birbiri ile ilişkileri olan tablolardır. Dolayısıyla bir veri tabanında ilişkiden söz edebilmek için en az iki tablonun yer alması ve bu iki tablodaki verilerin birbiri ili bir şekilde ilişkilendiriliyor olması gerekir. Bu şekilde ilişkisel veri tabanları büyük dosyalardan oluşur. Her bir tablo, belli yapıya uygun verileri saklamak üzere

tasarlanır. ACID; klasik ilişkisel veri tabanı sistemlerinde sağlanan temel özellikler aşağıda sunulmuştur:

- Bölünmezlik (Atomicity)
- Tutarlılık (Consistency)
- İzolasyon (Isolation)
- Dayanıklılık (Durability)

2.6.2 İlişkisel Olmayan Veri Tabanı (Non Relational Database System)

İlişkisel olmayan (NoSQL) veri tabanı; 1998 yılında ilk olarak Carlo Strozzi tarafından öne sürülen bir kavramdır. NoSQL, ilişkisel veri tabanı sistemlerine alternatif bir çözüm olarak ortaya çıkmıştır. İlişkisel olamayan veri tabanları yatay olarak ölçeklendirilen bir veri depolama sistemidir. Dünya'da NoSQL örneklerini incelediğimizde; sosyal ağlarda Digg'in 3 TB'lık çözümü, Facebook'un gelen postaları arama için 50 TB ve eBay'ın bütün verileri için 2 PB'lık çözümleri vardır. Veri tabanlarına ilişkin problemlerden biri olan ölçek sorununa, diğer çözümlerin içinde en iyi cevap vereni NoSQL'dir. Günlük 7 TB'lık işlem hacmine sahip Twitter ve 10 TB'lık Facebook örneğindeki gibi, çok büyük verilerin depolanması ve yazılmasında ilişkisel veri tabanlarının eksik kaldığı hususlarda, yatay ölçekleme yapan dağıtık NoSQL çözümleri geliştirilmiştir. Amazon bu gereksinimi "DynamoDB", Google ise "Big Table" ismini verdiği NoSQL veri tabanı sistemi ile çözmektedir. İlişkisel veri tabanını yerine NoSQL veri tabanını tercihi, özellikle hız ve yatay büyüme ile gereksiz ek maliyetten kurtulmaya dayanmaktadır.

İlişkisel veri tabanlarının kullandığı "ACID" işlemselliğine karşın NoSQL "BASE" kısaltması ile ifade edilir.

- Kolay Ulaşılabilirlik (Basically Available): Veri erişim sorunlarını ortadan kaldırmak için kopyaları kullanır ve paylaşılmış ya da bölümlenmiş veriyi birçok sunucudan alır.
- Esnek Durum (Soft State): ACID mantığında veri tutarlılığının olmazsa olmaz bir gereklilik olduğu savunulurdu fakat NoSQL sistemler tutarsız ve süreksiz verilerin barınmasına da izin verir.

- Eninde Sonunda Tutarlı (Eventually Consistent): Uygulamalar anlık tutarlılıkla ilgili olmasına rağmen, NoSQL sistemlerin gelecekte bir zamanda tutarlı olacağı farz edilir.

  ACID'in zorunlu tuttuğu tutarlılığa karşın NoSQL'de tanımlanmayan bir zamanda tutarlılığın oluşacağı garanti edilir (Öztürk & Atmaca, 2017).

2.7 MongoDB

MongoDB 2009 yılında geliştirilmiş açık kaynak kodlu bir NoSQL veri tabanıdır. Mongo kelimesi "Humongous"tan türemiştir. Türkçe muazzam anlamına gelmektedir. MongoDB'de her kayıt bir doküman olarak ifade edilir. Bu veri tabanı uygulaması ile NoSQL verilerini belge biçiminde saklanabilir ve ölçeklenebilir olması sayesinde kolayca analiz edilebilir. NoSQL sistemler içerisinde en çok tercih edilenlerden biridir. MongoDB genel olarak hızın önemli olduğu gerçek zamanlı uygulamalar, oyunlar, büyük verilerin olduğu uygulamalarda kullanılır. Başlıca özellikleri şunlardır:

- Ölçeklenebilirdir (scalable).
- Verileri belge olarak saklar. JSON verileri kullanılabilir. Veriler JSON olarak saklandığı için gelen verilerin yapısı değişse dahi kaydetme konusunda herhangi bir sıkıntı yaşanmaz.
- Verilerin birden fazla kopyasını saklayabilirsiniz. Böylelikle veri kaybının da önüne geçmiş olursunuz.
- Veriler üzerinde index oluşturabilirsiniz. Böylelikle aradığınız tüm verilere hızlı ve kolay bir şekilde ulaşabilirsiniz.

2.7.1 Csv ve Json Dosyaları

CSV ve JSON, yapılandırılmamış veya yarı yapılandırılmış verileri almak, değiştirmek ve depolamak için kullanılan en yaygın biçimler olabilir.

2.8 Microsoft Azure

Microsoft Azure, geniş kapsamlı özellikler barındıran büyük bir bulut platformudur, Microsoft Azure bulut platformu size web sunucularını barındırma, e-mail sunucusu oluşturma, veri tabanı sunucusu, dosya depolama sunucusu, sanal makineler, kullanıcı dizinleri, web uygulamaları, mobil uygulamalar veya farklı servisler olarak hizmet

edebilirler. Satın aldığınız bulut platformunu kendi amaçlarınız için gizli olarak veya tüm kullanıcıların erişimine açık olarak kullanabilirsiniz. Microsoft Azure'un en büyük avantajı düşük ücretler karşılığında kapsamlı bulut bilgisayarlara sahip olmaktır.

2.9 Twitter

Mart 2006 yılında Jack Dorsey tarafından geliştirilen Twitter, kullanıcılarına en fazla 280 karakterden oluşan (2017 yılının son çeyreğine kadar karakter kısıtı 140 adetti ancak yapılan düzenleme ile artık kullanıcılar 280 adet karakter kullanabilmektedir) Türkçe cıvıldamak anlamına gelen mesaj (tweet) gönderme, fotoğraf, video paylaşma ve karşılıklı direk mesaj yollama olanağı sunmuştur.

2006 yılında ortaya konulan Twitter için başlıca iki sıçrama basamağı olduğu görülmektedir. Bunlardan ilki 2007 yılında Southwest'te düzenlenen interaktif konferansta katılımcıların konferans boyunca birbirleriyle iletişimde bulunabilmeleri ve panellere gerçek zamanlı yorum yapabilmeleri için Twitter kullanılmıştır. O zamana kadar günlük kullanım miktarı 20.000 adet tweetken söz konusu konferans sayesinde günlük atılına tweet sayısı 60.000'e yükselmiştir. Bir günde 3 katına çıkan kullanım trafiği Twitterın tarihinde önemli yapıtaşlarından biri olmuştur. O dönemde 111 farklı mikroblog uygulaması olmasına rağmen Twitter o zamana kadarki en popüler uygulama olmayı başarmıştır. Ertesi yıl ise; ABD'de düzenlenen başkanlık seçimlerinde adaylardan Barac Obama seçim kampanyası boyunca sosyal medya kullanımına özen göstermiş, Twitter üzerinden aynı anda 150.000 kişi ile eş güdümlü olarak haberleşmeye başlamıştır. Bu sürecin Obama'nın başkanlık seçimini kazanmasında önemli bir etkiye sahip olduğu düşünülmektedir. Bu durum aynı zamanda Twitter için de kuruluşundan itibaren ikinci yılında önemli bir sıçrama aşaması olarak görülebilir.

Twitter'da kullanıcılar ücretsiz olarak üye olup kendi profillerini oluşturur ve diğer üyelerle iletişime geçerler. Son derece basit bir ara yüze sahip olan site kullanım kolaylığı ile cezbedicidir. Twitter 50'ye yakın farklı dil seçeneğiyle faaliyetlerini sürdürmektedir. Üyeler profillerini oluştururken ilk önce kullanıcı adlarını belirlerler burada dikkat edilmesi gereken daha önceden alınmamış olan bir kullanıcı adının belirlenmesidir. Kullanıcının almak istediği ad önceden başka bir kullanıcı tarafından

edinilinmişse sistem o ismi başkasının almasına müsaade etmeyecektir. Kullanıcı adı kişilerin diğer kullanıcılar tarafından aranırken bulunmasını sağlayacağı için önemlidir. Ayrıca tüm kullanıcı adlarının önüne otomatik olarak @ sembolü gelecektir. Ancak kullanıcı adı zaman için de değiştirilebilmektedir. Üye olabilmek için doğrulayıcı olarak en az bir aktif mail adresinin, cep telefonu gibi erişilebilir olunan bir bilginin sunulması esastır.

Twitter'da hesap oluştururken ikinci aşama profilin oluşturulmasıdır. Kullanıcı profilini oluştururken fotoğraf, kişisel bilgiler, kendine ait web sitesi varsa web sitesinin adresi, konum gibi bilgileri isteğe bağlı olarak girebilirler. Özellikle konum bilgilerinin paylaşılması bir tweetin atıldığı dil ne olursa olsun hangi ülkeden atıldığını, aynı ülkeden atılan tweetlerin hangi bölgeden veya hangi ilden atıldığını göstermesi bakımından önemlidir. Özellikle araştırmacılar ülke, bölge veya şehir bazında analiz yaptıklarında bu bilgilere ihtiyaç duyabilmektedirler.

Açılan hesaplar tüm kullanıcıların görebilecekleri şekilde kamuya açık olabilecekleri gibi sadece onaylanmış takipçilerin görebilecekleri şekilde kısıtlı da olabilirler. Yaklaşık 320 milyon aktif kullanıcıya sahip olan Twitter hesapları içinde sadece %5'lik bir kısmı korumalı, geriye kalan %95'i ise kamuya açıktır (Kızılkaya, 2018).

2.10 WordCloud

Türkçe anlamı kelime bulutudur. Word cloud metni istediğimiz şeklin içinde kelimeleri dağıtarak bize sunar. Kelime listesinde ise girdiğimiz metnin ve metnin içinde geçen kelimelerin ağırlığı hakkında bilgi verir. Hangi kelime daha sık geçiyorsa o kelime daha büyük puntoyla belirtilir. Buradaki amaç metin içerisindeki vurgulanan kelimeleri göstermektir.

2.11 Duygu Analizi

Duygu analizi duyguların metinlerde hangi yollarla anlatıldığını ve bu anlatımlarda olumlu veya olumsuz durumların tespit etmeyi sağlayan bir analizdir. Duygu analizi ya da fikir araştırması, insanların görüşlerinin, değerlendirmelerinin, tutumlarının ve duygularının hesaplamaya dayalı olarak çalışılmasıdır. Bir duygu analizi programı kullanılan kelimelerin ve ifadelerin özelliklerine dayalı olarak metinlerin duygu

içeriğini tahmin etmeyi amaçlamaktadır. Sosyal duygu analizinde sosyal medyadaki yorumlar pozitif, negatif, etkisiz hatta biraz pozitif, çok pozitif, biraz negatif, çok negatif şeklinde sınıflandırılır. Bu sınıflandırmalar 1,0 ile -1,0 arasında değerler alır. Sonuç olarak elinizde tüm bu yorumların bu kategorilere göre bir dağılımı oluşur.

2.12 Gephi

Gephi NetBeans platformunda Java ile yazılmış açık kaynaklı bir ağ analizi ve görselleştirme yazılımı paketidir. Grafikleri gerçek zamanlı olarak görüntülemek ve keşfi hızlandırmak için bir 3B oluşturma motoru kullanır. Her tür grafiği keşfetmek, analiz etmek, mekansalleştirmek, filtrelemek, kümeleştirmek, değiştirmek ve dışa aktarmak için kullanılır. Gephi, akademi, gazetecilik ve araştırma projeleri gibi birçok alanda faydalı bir araçtır (Gephi, 2021).

2.12.1 Fruchterman Reingold

Fruchterman-Reingold algoritmasının özelliği düğümleri merkezilik özelliklerine göre yerleştirmesidir. Genellikle 1000'den az düğümlü ağları görüntülemek için kullanılmaktadır. Merkezi düğümler ağın ortalarına yerleştirilir. Düğümler arasında merkezilik anlamında yüksek farklılık içermeyen ağlarda görsel iç içe geçmiş bir şekilde olmaktadır.

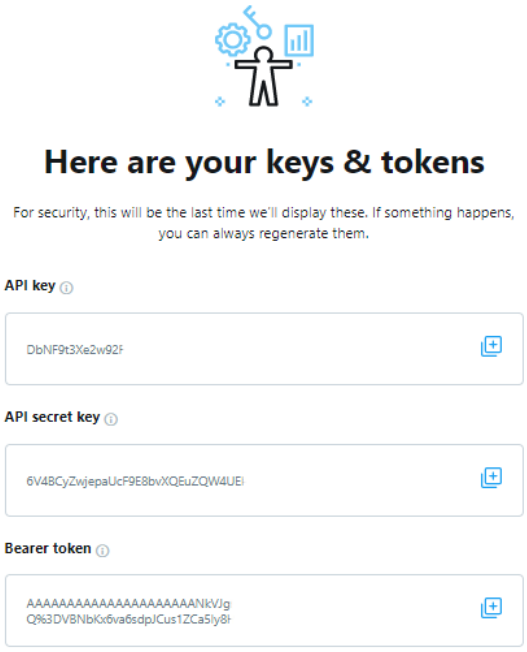
2.12.2 Force Atlas2

Birbirine ilişkili olan nodeları farklı kümeler olarak toplayan bir görselleştirme algoritmasıdır. Son düzen kalitesi ile hesaplama algoritmasının hızı arasında bir denge kurarak Force Atlas algoritmasının eksikliklerini gidermeye çalışır. Büyük ağlar için performansı, Force Atlas düzen algoritmasına kıyasla çok daha iyidir.

3. MATERYAL ve YÖNTEM

3.1 Twitter API Erişimi ve Veri Toplama

Twitter'dan veri çekebilmek için Twitter API* erişimine sahip olmalıyız. Bu erişim için öncelikle bir Twitter Developer hesabı oluşturmak ve API başvurusu yapmak gerekir. Elde ettiğimiz tecrübelere göre API alabilecek öncelikli hesaplar uzun süre önce Twitter'a katılmış hesaplar olmalıdır. Developer hesap başvurusu onaylanmasıyla birlikte API key ve token elde edildi.



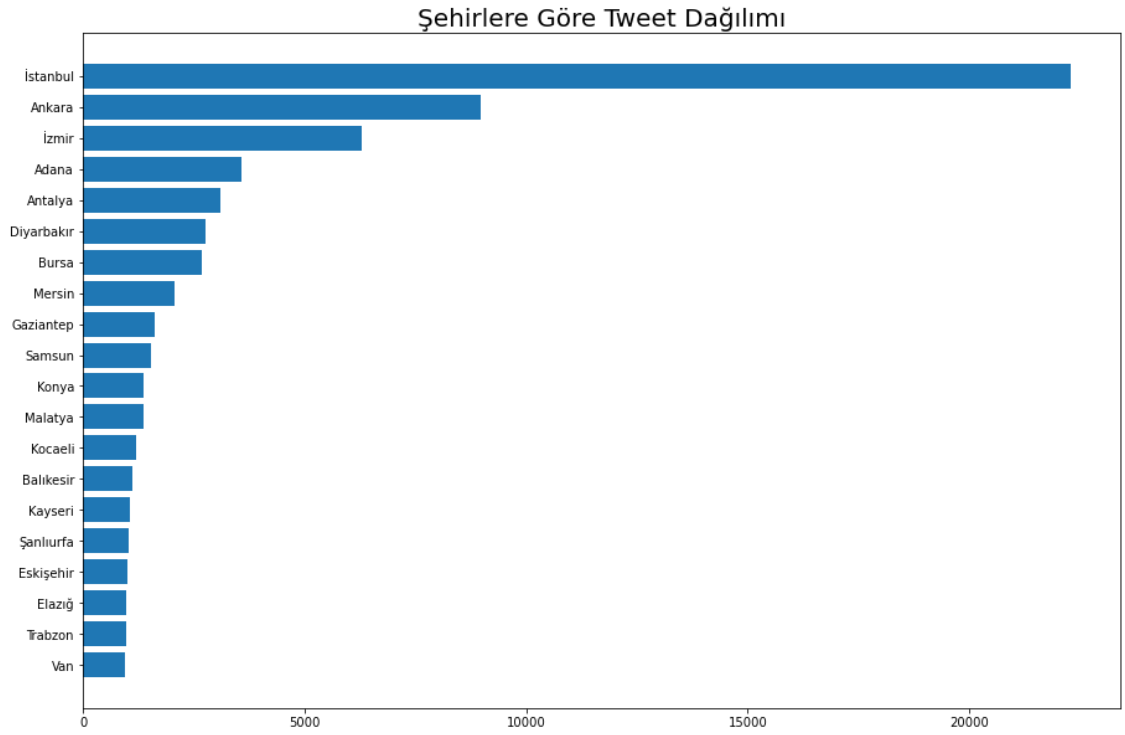
Şekil 1. API Key ve Token

API key'in elde edilmesiyle birlikte Tweepy kullanarak Microsoft Azure'den alınan sanal makine ile Ekim 2020- Şubat 2021 arası 670 bin tweet aşağıdaki keywordler kullanılarak mongoDB'ye çekildi.

Keywordler: Eğitim, öğretim, üniversite, ilkokul, ortaokul, anaokulu, sinav, temel yeterlilik test, alan yeterlilik test, ÖSYM, KPSS, TEOG, yüksek lisans, lisansüstü, öğrenci, Ziya Selçuk, öğretmen, dershane, lisesi, final haftasi, vize haftasi

3.2 Tweetlerin Konum Dağılımı

Twitter'dan veri çekme işleminin sona ermesiyle birlikte elimizdeki 670 bin tweet pandas kütüphanesi kullanılarak konum bilgileri satırından yapılan frekans sorgulaması ile elde edilen sayılar matplotlib kütüphanesi ile sütun grafiği haline getirildi. Grafik **Şekil 2**'de gösterilmiştir.



Şekil 2. Şehirlere göre tweet dağılımı

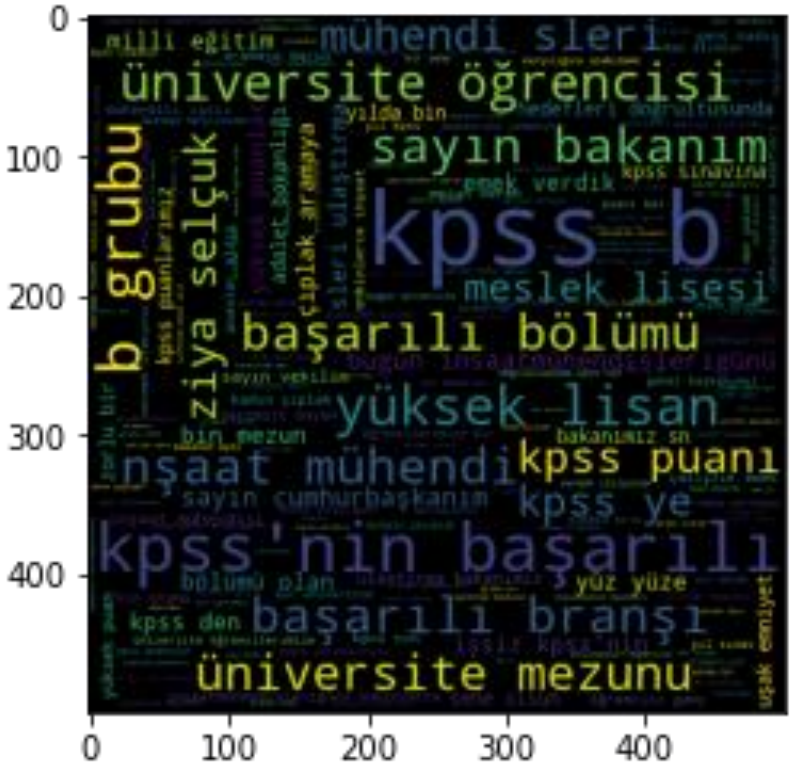
3.3 WordCloud

670 bin tweet üzerinde yapılan çalışmalarda, üzerinde çalışılan veriler twitter verisi olduğu için içerisinde çok fazla hashtag, rt, link, noktalama işaretleri gibi kullanılmayacak karakterler bulunmaktaydı. WordCloud kullanabilmek için öncelikle veriler bu karakterlerden arındırıldı ve kelimeler bir standarda (Kelimeler arası birden fazla boşluklar silindi. Tüm harfler küçük harflere çevrildi.) dönüştürüldü.

```
words = ' '.join([tweet for tweet in df['text']])
wordCloud = WordCloud(width=500, height=500).generate(words)
plt.imshow(wordCloud)
plt.show()
```

Şekil 3. WordCloud oluşturma

670 bin tweet ile yapılan wordcloud **Şekil 4**'te gösterilmiştir.



Şekil 4. WordCloud

3.4 Google Cloud Natural Language API Erişimi ve Duygu Analizi

Google Cloud, Google firmasının Google arama motoru ve Youtube gibi sitelerinde kullandığı sunucu altyapı hizmetlerini son kullanıcıya sunduğu bir bulut bilgi platformudur (Google, 2021).

Projemizin duygu analizi kısmında, Google Platformu tarafından sağlanan Yapay Zeka ve Makine Öğrenmesi ürünü olan Cloud Natural Language kullanılmasına karar verildi. Google Cloud'un hesap oluşturulurken öğrencilere özel vermiş olduğu 100 dolarlık krediyle API alındı. Duygu analizi için aylık ücretlendirme her bin birim başına; 5 bin kelimeye kadar ücretsiz, 5 bin – 1milyon kelime arası $1.00, 1 – 5 milyon kelime arası $0.50'dır.

Duygu analizi yapılırken sosyal medyadaki yorumlar pozitif, negatif, etkisiz hatta biraz pozitif, çok pozitif, biraz negatif, çok negatif şeklinde sınıflandırılır. Bu sınıflandırmalar 1,0 ile -1,0 arasında değerler alır. Sonuç olarak elinizde tüm bu yorumların kategorilere göre bir dağılımı oluşur.

100.000 random tweet üzerinde yaptığımız analiz sonucu genel sentiment score "0.0071" bulunmuştur. Bazı tweetlerin sentiment score örnekleri:

• Manisa lisesi pansiyonu televizyon odası... adı yok namı var. (-0.4)
• Yüzyüze eğitim olmadan gerçek bir egitim olmaz. (-0,4)
• Ortaokul mezunu birini koca bankaya yönetici mi yaptınız? Hmmm (-0.1)
• Üniversite mezunu cahilden daha kötü bi şey varsa o da üniversite hocası olmuş cahildir. (-0,6)
• Üniversite ne güzel şeymişsin sen (0,89)
• Sanırım ortaokul-lise-üni öğrencileri virüse yakalanmıyor bu kadar ısrar etmezlerdi yoksa #sınavagelmiyorum (-0,69)
• Doğru söyleyen sizler gibi akademisyenler gördükçe bizlerde sizlerin herzaman yanindayiz hocam (0,69)
• Umarim sinav cikisi okulun onunde show tv olurda ozaman videomu izleyin (0.0)

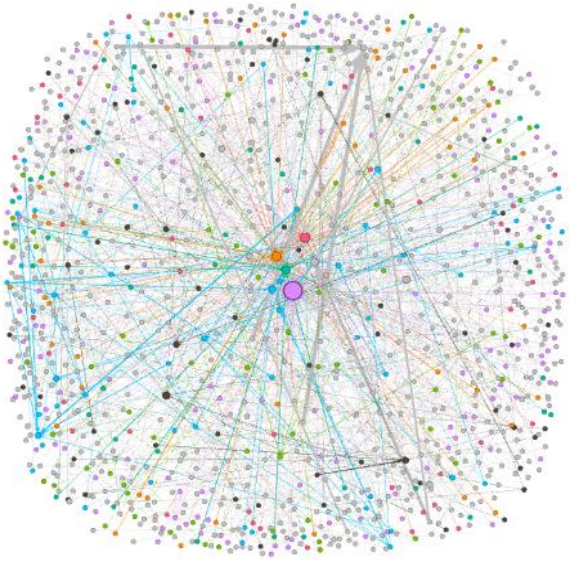
3.5 Gephi ile Network Analizi

Gephi'de kullanılan kuvvet yönelimli grafik çizim algoritmaları, estetik açıdan hoş bir şekilde grafik çizmek için kullanılan bir algoritma sınıfıdır. Amaçları, bir grafiğin düğümlerini iki boyutlu veya üç boyutlu uzayda konumlandırmaktır. Bu veri için Fruchterman Reingold, Force Atlas2, Modularity ve Eigenvector Centrality algoritmaları kullanıldı.

Bilgisayarlarımızın gücü Gephi'de 670 bin tweet ile çalışmaya elverişli olmadığı için rastgele 30 bin tweet sahibi ve bu tweetleri retweetleyen kişiler kullanılarak görselleştirmeler yapıldı.

3.5.1 Fruchterman Reingold

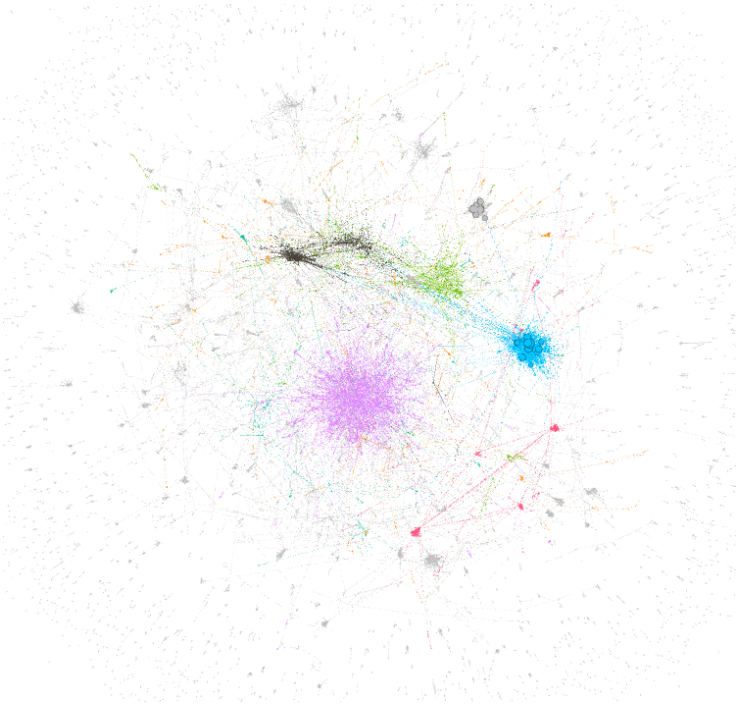
Fruchterman Reingold algoritması kullanılarak ağdaki en etkili kişiler **Şekil 5**'te gösterilmektedir.



Şekil 5. 30 bin tweet ile Fruchterman Reingold Algoritması

3.5.2 Force Atlas2

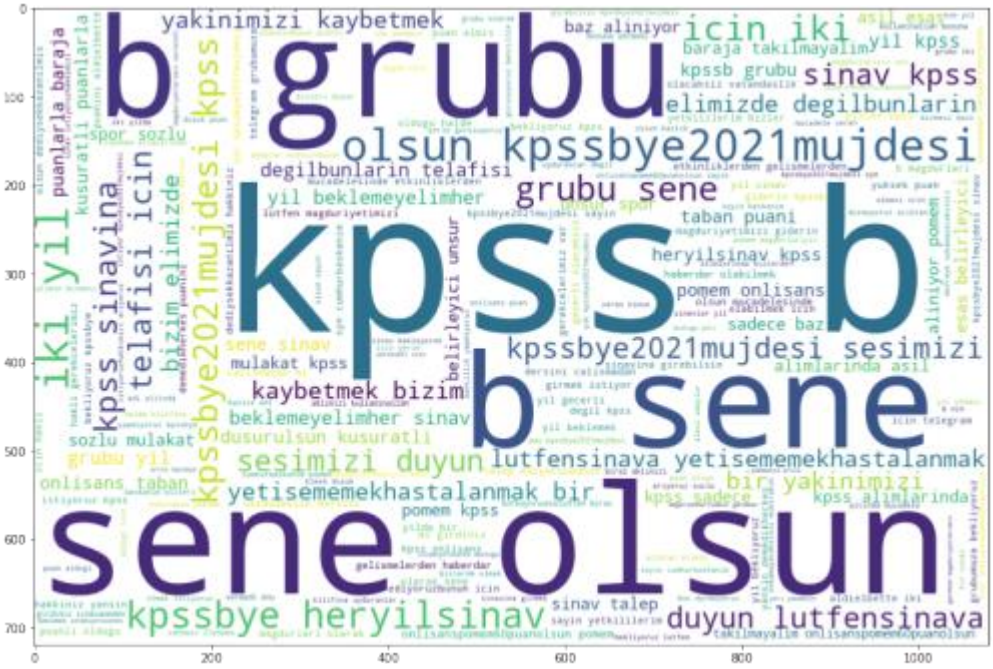
30 bin tweet ile Force Atlas2 algoritması **Şekil 6**'de gösterilmiştir.



Şekil 6. 30 bin tweet ile Force Atlas2 Algoritması

Şekil 6'da sağ alttaki mavi kümenin KPSS ile ilgili paylaşım yapanlar, sol alttaki mor kümenin öğretmen ağırlıklı atama bekleyenler olduğu saptandı. Sağ üstte bulunan yeşil kümenin sağlık ataması bekleyen paramedikçiler olduğu, sol üstte bulunan kahverengi kümenin ise eğitim ile ilgili paylaşım yapan inşaat mühendisleri olduğu görüldü.

Modularity ve Eigenvector Centerality kullanarak modellenen verinin kümelere göre wordcloudları oluşturuldu.



Şekil 7. Mavi küme WordCloud

Elde ettiğimiz sonuçlara göre mavi kümenin KPSS B ile ilgili konuştuğu görüldü. Küme içerisinden rastgele 20 kişinin profiline tek tek bakıldı ve şu sonuçları çıkarıldı:

Tweetleri çekilen kişilerin %5'i adalet öğretmeni, %5'i sosyolog, %5'i çevre mühendisi, %5'i ziraat mühendisi, %80'inde ise yeterli veri bulunmamaktadır.

Şekil 8. Mor küme WordCloud

Mor kümenin ise ağırlıklı olarak atama bekleyen öğretmenler olduğu görüldü. Küme içerisinden rastgele 20 kişinin profiline tek tek bakıldı ve şu sonuçları çıkarıldı:

Tweetleri çekilen kişilerin %15'i Türkçe öğretmeni, %10'u sosyal bilgiler öğretmeni, %10'u rehber öğretmen, %10'u matematik öğretmeni, %5'i ilahiyat fakültesi, %50'sinde ise yeterli veri bulunmamaktadır.

Şekil 9. Yeşil küme WordCloud

Yeşil kümenin genellikle paramedikçilerin bulunduğu, sağlık ataması bekleyen kişilerden oluştuğu görüldü. Küme içerisinden rastgele 20 kişinin profiline tek tek bakıldı ve şu sonuçları çıkarıldı:

Tweetleri çekilen kişilerin %65'i paramedikçi iken, %35'inde ise yeterli veri bulunmamaktadır.

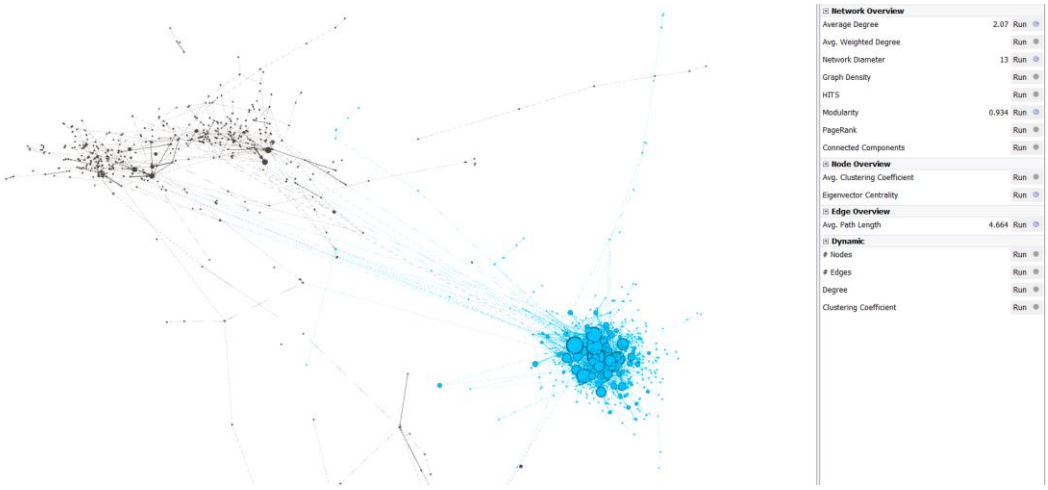
Şekil 10. Kahverengi küme WordCloud

Kahverengi kümenin inşaat mühendislerinden oluştuğu görülür. Küme içerisinden rastgele 20 kişinin profiline tek tek bakıldı ve şu sonuçları çıkarıldı:

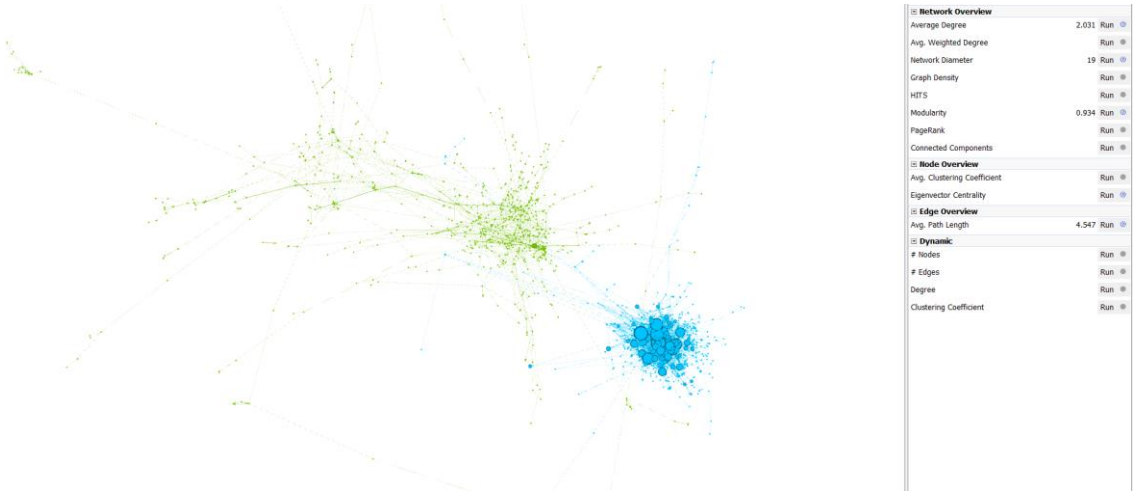
Tweetleri çekilen kişilerin %70'i inşaat mühendisi, %5'i mimar, %25'inde ise yeterli veri bulunmamaktadır.

3.5.3 Network Diameter, Avg Path Length ve Avg Degree

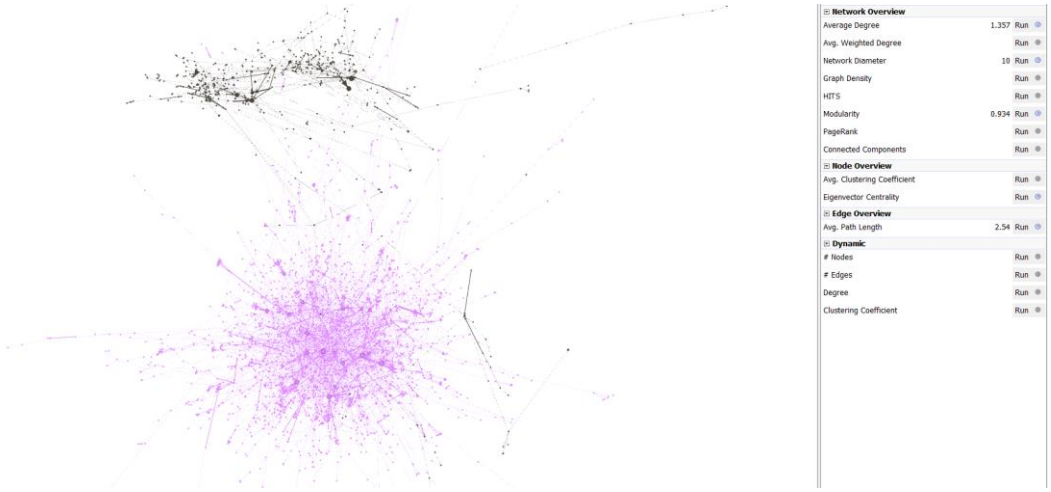
Kümelerin üzerinde durduğu konuların benzerliğini ölçmek için, her bir kümenin diğer kümeyle ilişkisi incelendi. Ağ çapı, ortalama yol uzunluğu ve ortalama derecesine bakıldı.



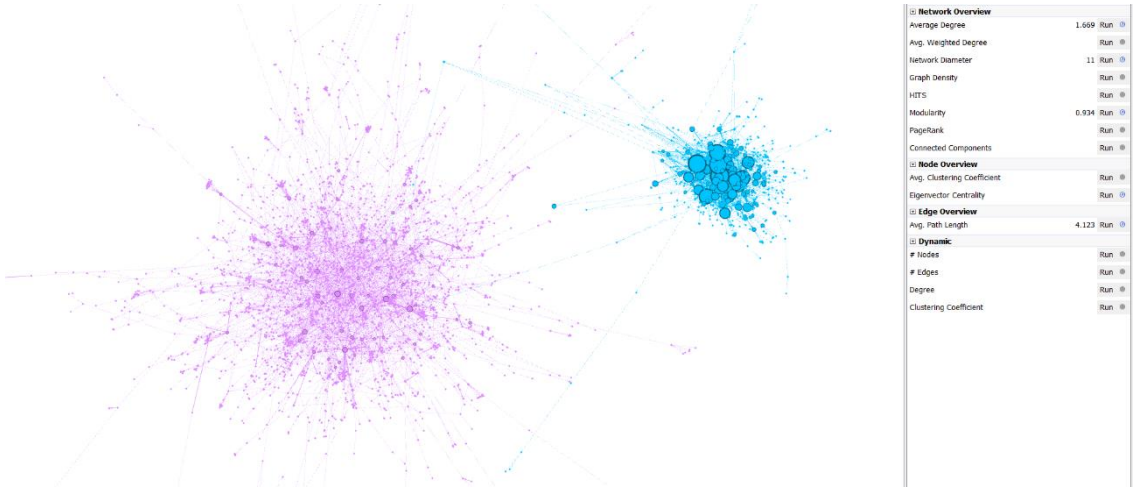
Şekil 11. Mavi ve kahverengi küme



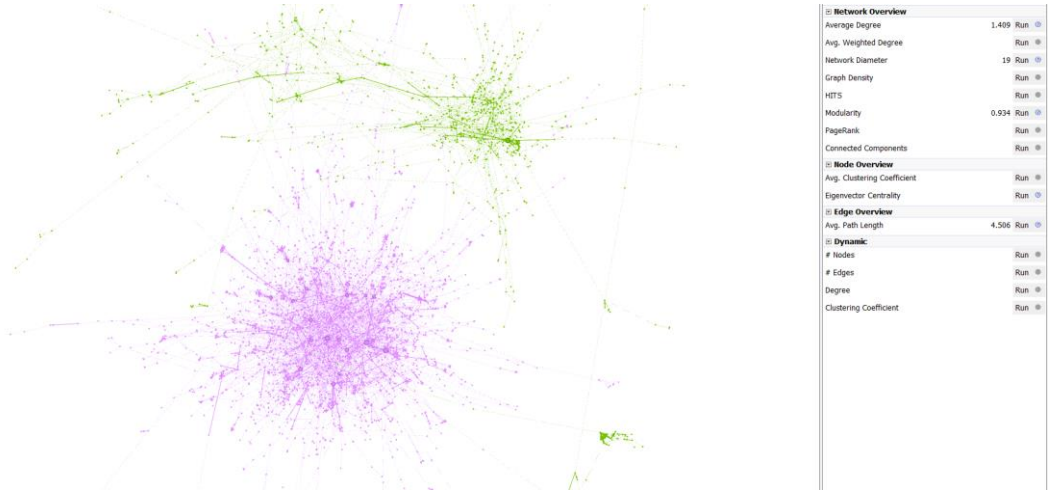
Şekil 12. Mavi ve yeşil küme



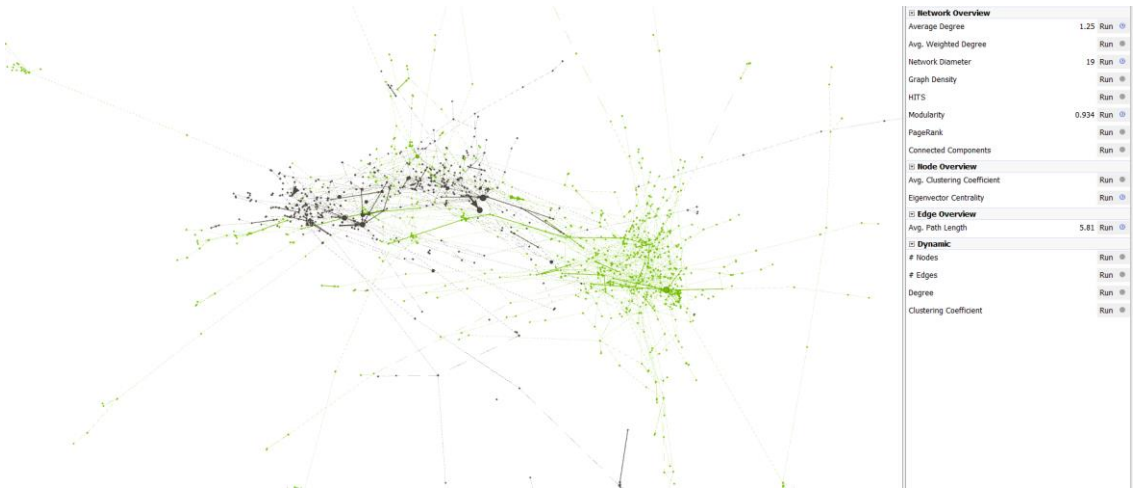
Şekil 13. Mor ve kahverengi küme



Şekil 14. Mor ve mavi küme



Şekil 15. Mor ve yeşil küme



Şekil 16. Yeşil ve kahverengi küme

- **Average Degree**

Ortalama derece, grafikteki düğüm başına ortalama kenar sayısıdır. Toplam kenarların toplam düğümlere bölünmesiyle bulunur.

*	Mor	Mavi	Yeşil	Kahverengi
Mor	0	1,669	1,409	1,357
Mavi	1,669	0	2,031	2,07
Yeşil	1,409	2,031	0	1,25
Kahverengi	1,357	2.07	1,25	0

Tablo 1: Average Degree

- **Network Diameter**

Ağdaki en uzak iki düğüm arasındaki en kısa mesafedir. Her düğümden diğer tüm düğümlere en kısa yol uzunluğu hesaplandığında, çap, hesaplanan tüm yol uzunluklarının en uzunu olur. Uzaklık arttıkça konular farklılaşır.

*	Mor	Mavi	Yeşil	Kahverengi
Mor	0	11	19	10
Mavi	11	0	19	13
Yeşil	19	19	0	19
Kahverengi	10	13	19	0

Tablo 2: Network Diameter

- **Average Path Length**

Ortalama yol uzunluğu, ağdaki tüm düğümler arasındaki mümkün olan en kısa yolu ölçerek, tüm ağ için iletişim verimliliğinin bir ölçüsünü sağlar. Tüm ağ için genel bir sayı hesaplanır, daha düşük sayılar ağın nispeten daha verimli olduğunu gösterir ve yüksek ortalama sayılar bilgi akışı için nispeten verimsiz bir grafiği belirtir.

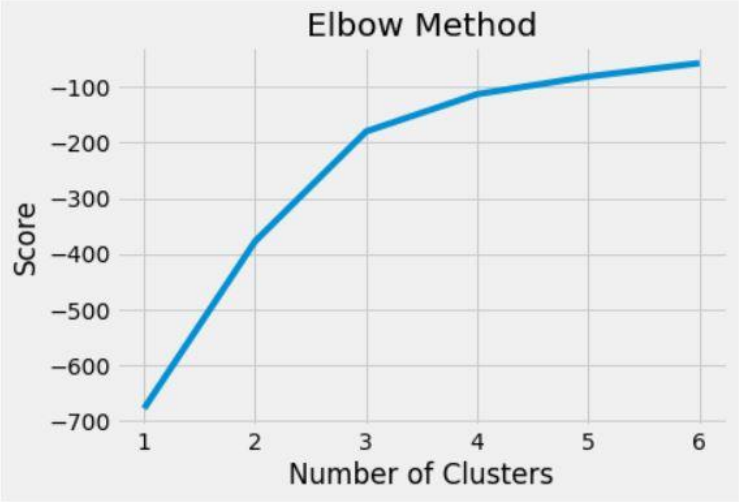
Bu sayı, düğümler arasındaki maksimum yol uzunluğunu temsil ettiğinden, mutlaka ağ çapından daha küçük olacaktır.

*	Mor	Mavi	Yeşil	Kahverengi
Mor	0	4,123	4,506	2,54
Mavi	4,123	0	4,547	4,664
Yeşil	4,506	4,547	0	5,81
Kahverengi	2,54	4,664	5,81	0

Tablo 3: Average Path Length

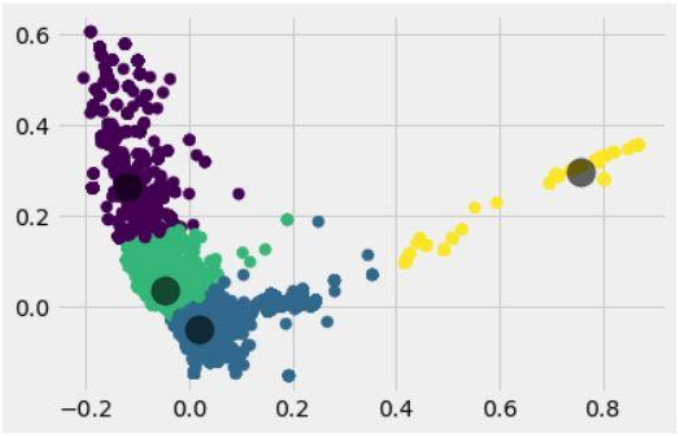
3.6 K - Means

K-Means'deki K küme değerinin belirlenmesi için Elbow Method uygulandı. Elbow Method, küme sayısının bir fonksiyonu olarak açıklanan varyansın yüzdesini dikkate alan bir yöntemdir (Çılgın & Kurt, 2021).



Şekil 17. Elbow Method

Şekil 17'ye göre kırılma noktalarına baktığımızda küme sayımızın 4 olması gerektiği görülür. Algoritma, istatistiksel olarak benzer nitelikteki kayıtları aynı gruba sokar. Küme merkezi kümeyi temsil eden değerdir. K-Means sonucu oluşan kümelerimiz **Şekil 18**'de gösterilmiştir.



Şekil 18. Kümeler

Küme 1	
Kelimeler	**Puan**
kpss	0.074765
istiyoruz	0.030791
2018	0.027252
bir	0.026733
bin	0.025397
sayin	0.025158
atama	0.024675
yil	0.023569
alim	0.019873
60	0.017748

Tablo 5: Küme 1

Küme 2	
Kelimeler	**Puan**
kadin	0.191588
oturun	0.187941
kalkin	0.187941
tutuldu	0.181532
cip***	0.180104
kul*******	0.180076
maruz	0.179906
emniyette	0.179626
cikarin	0.179354
aramaya	0.177023

Tablo 4: Küme 2

Küme 3	
Kelimeler	**Puan**
universite	0.036533
egitim	0.020328
bir	0.019598
var	0.011659
değil	0.011527
ilkokul	0.011004
sinav	0.010323
lisesi	0.010221
mezunu	0.010005
ortaokul	0.009585

Tablo 7: Küme 3

Küme 4	
Kelimeler	**Puan**
sene	0.151004
kpss	0.123029
grubu	0.103501
kpssbye2021mujdesi	0.099155
yil	0.098607
kpssb2021deolsun	0.077521
olsun	0. 076458
istiyoruz	0.076031
sinav	0.056084
carsmbacb60bintakvim	0.048873

Tablo 6: Küme 4

4. SONUÇLAR

Türkiye'de eğitim hakkında atılmış tweetlerden elde edilen veriler doğrultusunda çeşitli analizler yapılmıştır. Bu doğrultuda atılan tweetlerin konum bilgilerine bakıldığında en çok tweet atan ilk beş şehir sırasıyla; İstanbul, Ankara, İzmir, Adana ve Antalya'dır. İstanbul'un tüm tweetlerin %32'sini, Ankara'nın %14'ünü, İzmir'in %9'unu, Adana'nın %6'sını, Antalya'nın ise %5'ini oluşturduğu ortaya çıkmıştır. En az tweet atan şehirler ise Gümüşhane, Bayburt ve Ardahan'dır. Şehirlerin nüfusunun bu verilere etkisi açıkça görülmüştür.

Tweetler hakkında genel bir anlam çıkarmak amacıyla yapılmış olan kelime bulutuna bakıldığında en çok kullanılan kelimelerin; KPSS B, Ziya Selçuk, üniversite mezunu, yüksek lisans ve inşaat mühendisi kelimelerinin olduğu gözlemlenmektedir. KPSS B, Türkiye'de devlet kadrolarına personel almak amacıyla yapılan Kamu Personeli Seçme Sınavı sonucuna göre yapılan bir atama ve kadro türüdür. KPSS'de lise, ön lisans ve lise mezunları KPSS B sınavına katılırlar. Sınav iki senede bir olur. Tweetlerin içeriklerine bakıldığında genellikle KPSS B sınavının daha sık yapılmasını isteyenlerin bulunduğu görülmüştür. İnşaat mühendisi kelimesinin geçmesinin sebebinin ise inşaat mühendislerinin atama problemlerinden dolayı attığı tweetler olduğu görülmüştür. Diğer kelimelerin doğrudan eğitim ile ilgili olması beklenen sonucu karşılamaktadır.

Google Cloud API kullanarak yapılan duygu analizi sonucunda tweetlerin %39'unun pozitif, %34'ünün negatif, %27'sinin ise nötr olduğu görülmüştür. Buna göre atılan tweetler arasında bariz bir duygu baskınlığı bulunmadığı anlaşılmıştır. Tüm tweetlerden elde edilen duygu skorunun "0,0071" olması da duygu yoğunluğunun nötr bir çizgide olduğunu göstermektedir. Tweetlere karşılık Google Cloud tarafından hesaplanan duygu skorlarının doğruluk oranı, rastgele tweetlere bakarak kendi görüşümüzce %64 olarak hesaplanmıştır.

Retweet yapılan tweetin sahibi ve retweet yapan kişi ile oluşturulan sosyal ağın, gephi programı kullanılarak görselleştirilmesi sonucunda dört temel küme elde edildi. Her bir küme içerisinden rastgele elde edilen 20 tweet ile içeriklerine bakıldığı zaman; birinci kümenin KPSS ile ilgili paylaşımlar yapan kişilerden, ikinci kümenin öğretmen ağırlıklı olmak üzere atama bekleyen kişilerden, üçüncü kümenin inşaat mühendisleri ile ilgili tweet atan kişilerden, dördüncü kümenin ise sağlık atamaları ile ilgili tweet atan kişilerden

oluştuğu gözlemlenmiştir. Aynı doğrultuda kişilerin rastgele incelenmesi sonucunda birinci kümenin %5'inin adalet öğretmeni, %5'inin sosyolog, %5'inin çevre mühendisi, %5'inin ziraat mühendisi olduğu gözlemlenmiştir. %80'lik bölüm ise mesleği hakkında bilgi paylaşımı yapmamıştır. İkinci kümenin meslek bilgileri incelendiğinde %15'inin Türkçe öğretmeni, %10'unun sosyal bilgiler öğretmeni, %10'unun rehber öğretmen, %10'unun matematik öğretmeni, %5'inin ilahiyat fakültesinden olduğu gözlemlenmiştir. %50'lik bölüm ise mesleği hakkında bilgi paylaşımı yapmamıştır. Üçüncü kümenin meslek bilgileri incelendiğinde %70'inin inşaat mühendisi, %5'inin mimar olduğu gözlemlenmiştir. %25'lik bölüm ise mesleği hakkında bilgi paylaşımı yapmamıştır. Dördüncü kümenin meslek bilgileri incelendiğinde %65'inin paramedikçi, %35'inin ise mesleği hakkında bilgi paylaşımı yapmadığı gözlemlenmiştir.

Network diameter değerlerine bakıldığında ikinci küme ile üçüncü kümenin, yani öğretmen ağırlıklı kişilerin bulunduğu küme ile inşaat mühendisi olan kişilerin bulunduğu kümenin birbirlerine en yakın kümeler oldukları görülmektedir. Bunun sebebinin atama talebi içeren tweetler için iki grubun birbirine verdikleri destekler olduğu görülmüştür.

Tweetlerdeki konuşulan genel konuların bulunması amacıyla uygulanan Kmeans algoritması sonucunda insanların genel olarak dört farklı konu üzerinde kümelendiği gözlenmiştir. Bu kümelerden birincisi atamalar konusunda, ikincisi üniversite eğitimi ve sınavlar konusunda, üçüncüsü her sene KPSS B olmasını isteyenlerin ilettiği fikirler konusunda birleşmiştir. Dördüncü kümede ise ana konumuzdan bağımsız, siyasi veya politik olabilecek içerik bulunduğu için içeriği kısıtlanmıştır.

KAYNAKLAR

Ayan, B., Kuyumcu, B., & Ciylan, B. (2019). Detection of Islamophobic Tweets on Twitter Using Sentiment Analysis. *Gazi University Journal of Science Part C: Design and Technology*, 495-502.

Ayık, Y. Z., Özdemir, A., & Yavuz, U. (2010). Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği ile Analizi. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 441-454.

Beşkirli, A., Gülbandılar, E., & Dağ, İ. (2021). Information Discovery from Twitter Data with Text Mining Methods. *ESTUDAM Bilişim Dergisi, 2*(1), 21-25.

Bostancı, B., & Albayrak, A. (2021). Suggesting Individual Content with Sentiment Analysis. *Veri Bilimi Dergisi*, 53-60.

Çılgın, C., & Kurt, A. S. (2021). Dış Ticaret Verileri İçin Kümeleme Analizi: Türkiye, Azerbaycan ve Kazakistan Örneği. *Sosyoekonomi Dergisi*, 511-540.

Ergün, K. (2011). *Metin Madenciliği Yöntemleri ile Ürün Yorumlarının Otomatik Değerlendirilmesi.* Sakarya: Sakarya Üniversitesi.

Gephi. (2021). *Gephi*. https://gephi.org/about/ adresinden alındı

Google. (2021). *Google Cloud*. https://cloud.google.com/why-google-cloud?hl=tr adresinden alındı

Kızılkaya, Y. M. (2018). *Duygu Analizi ve Sosyal Medya Alanında Uygulama.* Bursa: Uludağ Üniversitesi Sosyal Bilimler Enstitüsü.

Koçak, B. B., Polat, İ., & Koçak, C. B. (2016). Determination of Twitter Users Sentiment Polarity Toward Airline Market in Turkey: A Case of Opinion Mining. *PressAcademia Procedia*, 684-691.

Microsoft. (2021). *Microsoft Azure*. https://azure.microsoft.com/tr-tr/overview/what-is-machine-learning-platform/ adresinden alındı

Oracle. (2021). *Oracle*. https://www.oracle.com/tr/artificial-intelligence/what-is-ai/ adresinden alındı

Öztürk, S., & Atmaca, H. E. (2017). İlişkisel ve İlişkisel Olmayan (NoSQL) Veri Tabanı Sistemleri Mimari Performansının Yönetim Bilişim Sistemleri Kapsamında İncelenmesi. *Bilişim Teknolojileri Dergisi*, 199-209.

Savaş, S., Topaloğlu, N., & Yılmaz, M. (2011). Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 1-23.