T.C.
SAKARYA ÜNİVERSİTESİ
İŞLETME ENTİTÜSÜ

TÜRKÇE SOSYAL MEDYA İÇERİKLERİNİN ANALİZİ İÇİN SANAL ASİSTAN TASARIMI

YÜKSEK LİSANS TEZİ

Meltem UZAVCI

Enstitü Anabilim Dalı : Yönetim Bilişim Sistemleri

Tez Danışmanı: Doç. Dr. Halil İbrahim CEBECİ

HAZİRAN – 2022

Meltem UZAVCI tarafından hazırlanan ‘Türkçe Sosyal Medya İçeriklerinin Analizi İçin Sanal Asistan Tasarımı’ başlıklı bu tez, 14/06/2022 tarihinde Sakarya Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliği'nin ilgili maddeleri uyarınca yapılan Tez Savunma Sınavı sonucunda başarılı bulunarak, jürimiz tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Danışman: Doç. Dr. Halil İbrahim CEBECİ
Sakarya Üniversitesi

Jüri Üyeleri: Dr. Öğretim Üyesi Alparslan KİBAR
Sakarya Üniversitesi

Dr. Öğretim Üyesi Mustafa YILMAZ
Sakarya Uygulamalı Bilimler Üniversitesi

SAKARYA ÜNİVERSİTESİ	**T.C.** **SAKARYA ÜNİVERSİTESİ** **İŞLETME ENSTİTÜSÜ** **TEZ SAVUNULABİLİRLİK VE ORJİNALLİK BEYAN FORMU**		**Sayfa : 1/1**

Öğrencinin		
Adı Soyadı	**:**	Meltem UZAVCI
Öğrenci Numarası	**:**	Y189054009
Enstitü Anabilim Dalı	**:**	Yönetim Bilişim Sistemleri
Enstitü Bilim Dalı	**:**	Yönetim Bilişim Sistemleri
Programı	**:**	☑ YÜKSEK LİSANS ☐ DOKTORA
Tezin Başlığı	**:**	Türkçe Sosyal Medya İçeriklerinin Analizi İçin Sanal Asistan Tasarımı
Benzerlik Oranı	**:**	**%4**

☐ Sakarya Üniversitesi İşletme Enstitüsü Lisansüstü Tez Çalışması Benzerlik Raporu Uygulama Esaslarını inceledim. Enstitünüz tarafından uygulama esasları çerçevesinde alınan Benzerlik Raporuna göre yukarıda bilgileri verilen tez çalışmasının benzerlik oranının herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi beyan ederim.

..... / ..... / 20....

☐ Sakarya Üniversitesi İşletme Enstitüsü Lisansüstü Tez Çalışması Benzerlik Raporu Uygulama Esaslarını inceledim. Enstitünüz tarafından uygulama esasları çerçevesinde alınan Benzerlik Raporuna göre yukarıda bilgileri verilen öğrenciye ait tez çalışması ile ilgili gerekli düzenleme tarafımca yapılmış olup, yeniden değerlendirilmek üzere gsbtez@sakarya.edu.tr adresine yüklenmiştir.

Bilgilerinize arz ederim. **..... / ..... / 20….**

Uygundur

Danışman
Unvanı / Adı-Soyadı: Doç. Dr. Halil İbrahim CEBECİ

Tarih:

İmza:

☐ KABUL EDİLMİŞTİR ☐ REDDEDİLMİŞTİR **EYK Tarih ve No:**	**Enstitü Birim Sorumlusu Onayı**

ÖNSÖZ

Bu tezin yazılması aşamasında, çalışmamı titizlikle takip eden danışmanım Doç. Dr. Halil İbrahim CEBECİ'ye değerli katkı ve emekleri için içten teşekkürlerimi ve saygılarımı sunarım. Yüksek lisans ders ve bilhassa tez döneminde destekleriyle katkıda bulunan değerli arkadaşım Sinan YILMAZ'a teşekkürlerimi iletmek isterim. Son olarak, hayatımın tüm aşamasında her açıdan yanımda olan sevgili annem, babam ve kardeşlerime şükranlarımı sunarım.

Meltem UZAVCI

14.06.2022

İÇİNDEKİLER

KISALTMALAR

API : Uygulama Programlama Arayüzü

BERT : Bidirectional Encoder Representations from Transformers

KNN : K-En Yakın Komşu

LDA : Latent Dirichlet Allocation (Gizli Dirichlet Ayrımı)

LSA : Saklı Anlamsal Analiz

NB : Naive Bayes

NLP : Natural Language Processing (Doğal Dil İşleme)

NLTK: Doğal Dil Araç Takımı

RF : Random Forests (Rastgele Orman)

SVM : Destek Vektör Makineleri

TDM : Terim Döküman Matrisi

XGB : XGBoost (Ekstrem Gradyan Arttırma)

TABLO LİSTESİ

ŞEKİL LİSTESİ

Tezin Başlığı: Türkçe Sosyal Medya İçeriklerinin Analizi İçin Sanal Asistan Tasarımı	
Tezin Yazarı: Meltem UZAVCI	**Danışman:** Doç. Dr. Halil İbrahim CEBECİ
Kabul Tarihi: 14.06.2022	**Sayfa Sayısı:** vi (ön kısım)+ 85 (tez)
Anabilim Dalı: Yönetim Bilişim Sistemleri	

Teknolojinin son yıllardaki hızlı gelişimi ile birlikte insanlar sosyal medya platformlarını sosyalleşme aracı olarak kullanmaya başlamıştır. Sosyal medya platformlarında insanlar hem bilgi üreten hemde bilgi tüketen konumdadır. Bu bilgiler alınan ürün, hizmet, veya gündem hakkında olabilmektedir. Bireysel kullanıcıların yanı sıra ticari platformlar sosyal medyayı hedef kitlelerine ulaşmak ve marka algılarını yönetmek için kullanabilmektedir. Sosyal medya analitiği bu noktada devreye girer ve sosyal medyadaki yapılandırılmamış verilerin içerisindeki kalıpların analiz edilerek ticari ve bireysel fayda oluşturulmasını sağlar. Sosyal medya analitiğinin arka planında ise metin madenciliği işlemleri yer almaktadır. Sosyal medya verisi konuşma dili tabanlı olması dolayısıyla kısatlamalar ve emojiler gibi yapılar barındırır ve metin madenciliğine göre daha hassas bir süreç gerektirir. Bu çalışmada sosyal medya analizlerinde sıklıkla kullanılan Twitter platformu üzerinden, veri toplama, ön işleme ve analiz sürecinin yönlendirmeler ile yapılmasını sağlayan bir arayüzü üzerinde çalışılmıştır. Bu arayüz aracılığıyla bahsedilen sosyal medya analitiği süreci Twitter verisi üzerinden adım adım gerçekleştirilebilmektedir. Burada araştırmacılar istedikleri herhangi bir konu, ürün veya hizmet için atılan Tweet'leri istedikleri tarih aralığında, dilde ve sayıda çekerek analiz sürecini başlatabilmektedir. Araştırmacı elde edilmek istenilen verinin boyutu hakkında veri çekme işleminden önce bilgilendirilerek yönlendirilir. Veri ön işleme adımında verinin birimleştirilmesi, standartlaştırılması ve filtrelenmesi çeşitli yönlendirmeler ile gerçekleştirilebilir. Örneğin standartlaştırma aşamasında araştırmacıya verisinde bulunan özel karakterlerin sayısal dağılımı verilmektedir. Araştırmacı temizlemek istenilen karakterleri bu şekilde temizleyebilir yada emoji ve kısaltmalar gibi konuşma dilinde bulunan kalıpları dönüştürebilir. Analiz aşamasında ise duygu analizi ve topik modelleme seçenekleri bulunmaktadır. Duygu analiz süreci için Turkish-BERT modeli kullanılmıştır. Topik modellemede ise etiketli kategori verisi üzerinde makine öğrenmesi işlemleri gerçekleştirilerek en başarılı model seçilerek bu modelle sınıflandırma işlemi gerçekleştirilmektedir. Tüm analiz sürecinde gerçekleştirilen işlemler, analiz sonuçları ve veri detay bilgileri kaydedilerek gelecek çalışmalarda program arayüzünün makine öğrenmesi destekli hale getirilmesi hedeflenmektedir.

Anahtar Kelimeler: Sosyal Medya Analitiği, Metin Madenciliği, Doğal Dil İşleme, Topik Modelleme, Duygu Analizi

Title of Thesis: Virtual Assistant Design for Analysis of Turkish Social Media Contents

Author of Thesis: Meltem UZAVCI **Supervisor:** Assoc.Prof. Halil İbrahim CEBECİ

Accepted Date:14.06.2022 **Np:** vi (pre text) + 85 (main body)

Department: Management Information Systems

With the rapid development of technology in recent years, people have started to use social media platforms as a means of socialization. In social media platforms, people are both in the position of producing and consuming information. This information may be about the product, service, or agenda received. In addition to individual users, commercial platforms can use social media to reach their target audiences and manage their brand perceptions. Social media analytics comes into play at this point and enables the creation of commercial and individual benefits by analyzing the patterns in the unstructured data in social media. In the background of social media analytics, there are text mining processes. Since social media data is speech-based, it contains structures such as abbreviations and emojis and requires a more sensitive process than text mining. In this study, a program interface was studied on the Twitter platform, which is frequently used in social media analysis, that enables data collection, preprocessing and analysis processes to be carried out with instructions. Through this interface, the mentioned social media analytics process can be carried out step by step over Twitter data. Here, researchers can start the analysis process by shooting Tweets for any topic, product or service they want, in the date range, language and number they want. The researcher is guided by being informed about the size of the data desired to be obtained before the data extraction process. In the data preprocessing step, data unitization, standardization and filtering can be performed with various orientations. For example, in the standardization phase, the numerical distribution of the special characters in the data is given to the researcher. In this way, the researcher can clean the desired characters or transform phrases in the spoken language such as emoji and abbreviations. In the analysis phase, there are sentiment analysis and topical modeling options. The Turkish-BERT model was used for the sentiment analysis process. In topic modeling, machine learning processes are performed on the labeled category data, and the most successful model is selected and classification is carried out with this model. It is aimed to make the program interface supported by machine learning in future studies by recording the operations performed during the entire analysis process, analysis results and data detail information.

Keywords: Social Media Analytics, Text Mining, Natural Language Processing, Topic Modelling, Sentiment Analysis

GİRİŞ

Teknolojinin son yıllardaki hızlı gelişimi sonucunda internete erişebilen cihazlar herkes tarafından kullanılarak günlük hayatın bir parçası haline gelmiştir. Sosyal bir varlık olan insan diğer insanlarla iletişim kurmak ve sosyalleşmek için bu cihazları bir araç olarak kullanmaktadır. Bu durum sosyal medya kavramının günlük yaşamın merkezine yerleşmesi sonucunu doğurmuştur. Sosyalleşmenin yanı sıra insanlar sosyal medyayı ürün, hizmet ve gündem hakkında görüş bildirmek için sıklıkla kullanmaktadır. Sosyal medyanın sağladığı anonim kalarak görüş paylaşma imkanı bir çok insanın görüşünü herhangi bir kaygı taşımadan paylaşmasını sağlamaktadır. Ayrıca ürün ve hizmet hakkında yapılan paylaşımlar işletmelerce ticari fayda sağlamak amacıyla kullanılabilmektedir. Ek olarak sosyal medya işletmeler ve ilgili markalara ürün ve hizmet hakkında müşteri farkındalığını arttırmak için kullanılabilecek yeni ve düşük maliyetli bir pazarlama kanalı sağlamaktadır.

Çevrimiçi sosyal medyada taraflar arası ortaklaşa oluşturulan içerik günümüzde hızla büyümeye devam etmektedir. Sosyal medya analitiği bahsedilen içeriğin toplanmasını, analizini ve yorumlanmasını içerir. Analiz ve yorumlama sürecinde bilinçli ve anlamlı karar vermeyi sağlamak için yarı yapılandırılmış ve yapılandırılmamış sosyal medya verilerindeki kalıpları analiz ederek kuruluşlara rekabet avantajı sağlamaktadır (Bekmamedova ve Shanks, 2014). Kuruluşlar bu sayede zamanında ve hedefli kampanyalar, duyarlı müşteri hizmetleri veya ilgili topluluklar oluşturarak müşterilerle bağlantı kurabilir ve onların marka algılarını şekillendirebilir. Kuruluşların sağladığı faydanın yanı sıra akademik çalışmalar içinde toplumun bir konu veya durum hakkında verdiği tepkileri Twitter gibi platformlar üzerinden analiz etmek mümkün hale gelmiştir.

Sosyal medya analitiği hem bilgi üreticileri hemde tüketicileri için nitelikli fırsat anlamına gelmekle birlikte zorluklarda barındırmaktadır. Ticari değeri sebebiyle sosyal medya platformları aracılığıyla sınırlı olarak paylaşılan veri kaynakları analiz edilecek örneklemin tamamını yansıtmama riski taşımaktadır. Gerekli veri elde edildiği varsayılırsa yapılandırılmamış olan metin verisinin temizlenme aşaması oldukça karmaşık ve uzun olabilmektedir. Veri çeşitliliğinin arttığı günümüzde sosyal medya verisini farklı kaynaklardaki veri ile de birleştirme süreci zahmetli olabilmektedir. Tüm bu analiz süreçleri

tamamlandıktan sonra anlamsal olarak veri detayına inildiğinde analiz edilen dile bağlı olarak kullanılan argo kelimeler, kısaltmalar, devrik cümleler, farklı dillerin bir arada kullanılması gibi detaylar veri içerisindeki anlamın ortaya çıkarılmasını zorlaştırmaktadır. Analiz sonucunda elde edilen büyük miktarda verinin görselleştirilmesi veri kaynağının yapılandırılmamış olması sebebiyle zorluklardan biridir. Sürekli büyümeye devam eden veri kaynakları, verinin görselleştirilerek özet halde sunulmasını gerektirdiği için görselleştirme araçları gelişmeye devam etmektedir.

Sosyal medyanın hızlı gelişimi ile veri işleme aşamasında da yenilikçi ve verimli teknikler gerektiren araçlara ihtiyaç ortaya çıkmıştır. Mevcuttaki araçlar incelendiğinde her birinin belirli bir alana yöneldiği görülmektedir. Araştırmacıların istedikleri alanda sosyal medya analitiği işlemlerini yapabilmesi için birden fazla araç kullanması gerekmektedir. Manuel ve otomatik analiz süreci yönetim seçenekleri bulunan bu araçlar çoğunlukla ek kodlama bilgisi gerektirmedir. Kodlama bilgisi olmadan yapılan işlemlerde sınırlama ile karşılaşmak mümkündür.

Sosyal medya analitiği olarak bahsedilen süreç, temelinde sosyal medya içerikleri üzerine yapılan metin madenciliği işlemleridir. Sosyal medya verisi genel kapsamı günlük konuşma dilinden oluşur dolayısıyla kısaltmalar, emojiler, devrik cümleler gibi yapıları barındırmaktadır. Bu da sosyal medya verisinin analizinde metin verisine göre daha özel yöntemlerde ilerlenmesi gerektiğini göstermektedir. Ayrıca sosyal medya analitiği kavramı ile sosyal ağ analizi de sıklıkla karıştırılmaktadır. Sosyal ağ analizi kişilerin, olayların ve durumların arasındaki ilişkinin tespit edilme sürecinde kullanılmaktadır ve web verisi gibi daha yapısal veriler üzerinde çalışmaktadır. Sosyal medya analitiği ise sosyal medya verileri üzerinde bir konu, ürün veya hizmet hakkında bilgi elde etme sürecini ifade eder.

Metin madenciliği süreci kelime ve anlamsal odaklı olarak iki türde ilerletilebilmektedir. Kelime türünde bakıldığında ilgili metinde geçen kelimelerin frekansı temel alınır. Daha basit analiz türlerinde dokumanları gruplamak için kullanılabilmektedir. Metin anlamsal olarak analiz edilmek istendiğinde ise doğal dil işleme süreçleri devreye girmektedir. Doğal dil işleme sürecinde kelimelerin cümle içerisindeki konumu, anlamı ve cümle içerisindeki görevi incelenerek anlamsal bir yaklaşımla analiz yapılmaya çalışılmaktadır.Doğal dil işleme

süreci sayesinde büyük miktardaki verinin kolayca özetlenmesi, makinelerin doğal dili algılayarak otomatik cevap oluşturması ve çeviri yapması gibi işlemler gerçekleştirilebilir.

Doğal dil işleme süreci veri kaynağının oluşum süreci ve diller arası farklılıklar sebebiyle birçok zorluk barındırmaktadır. Bu zorluklar; cümle içerisinde kullanılan kelimelerin anlam belirsizliği taşıması, cümlelerin sözdizimi belirsizliği, paragraf veya cümle içerisinde tekrarlardan kurtulmak için birbirleri yerine kullanılan kelimeler, devrik cümle yapıları, kısaltmalar ve farklı dillerde bulunan dile özgü cümle yapıları olarak özetlenebilir. Her ne kadar yıllar içerisinde farklı bakış açıları geliştirilse de doğal dil işleme konusunda hala önemli engellerin aşılması noktasında çalışmalara ihtiyaç duyulmaktadır.

Türkçe metin madenciliği sürecinde, genel metin madenciliği süreci zorlukları dışında dilin yapısına bağlı olarak bazı zorluklar bulunmaktadır. Bu zorluklardan en temeli Tükçe'nin sondan eklemeli bir dil olması sebebiyle kelimenin köküne ulaşılmasının zor olmasıdır. Bu noktada Türkçe kök analizi için büyük bir eğitim külliyatı veya geniş kapsamlı bir sözlük gerekmektedir. Kelimenin köküne inilmesi işleminin yanı sıra deyimler ve atasözleri, metaforlar ve ironiler açısından çok zengin bir dil olan Türkçe'nin Doğal Dil İşleme (NLP) ile analiz edilebilmesi için bahsedilen eğitim verisi ve sözlüğün geliştirilmesi gerekmektedir. Ayrıca Türkçe için duygu analizi yada konu modelleme gibi alanlarda yapılan çalışmalarda kullanılan eğitim veri setleri belirli alanlarla sınırlı kalmaktadır. Bu alanlar içinde büyük ve açık kaynak bir eğitim külliyatı oluşturularak daha sağlıklı analizler yapılabilir.

Türkçe'de NLP çalışmalarında bahsedilen kısıtlamaları aşmak için kullanılan yöntemlerden biri belgelerin İngilizce'ye çevirilerek analizlerin gerçekleştirilmesidir. Son yıllarda gelişen çeviri teknolojileri ve her geçen gün daha yüksek performansla çalışan İngilizce NLP yöntemleri araştırmacıları bu sürece yönlendirmektedir. Türkçe üzerinde ise doğrudan analiz yapabilmek adına python yazılım diline son yıllarda eklenerek önemli aşama kaydedilmesini sağlayan ve Google tarafından geliştirilen, derin öğrenme tabanlı BERT algoritmasının Türkçe versiyonu olan Turkish-BERT kullanılabilir.

Çalışmanın Önemi

Sosyal bilimler için en önemli veri kaynaklarından biri metinlerdir. Metinlerin işlenerek sahip oldukları anlamların ortaya çıkarılması süreci oldukça zahmetli olabilmektedir. Metin madenciliği olarak isimlendirilen bu süreç günümüzde çok önemli hale gelmiştir. Bu sürecin önemli hale gelmesinde metin verilerinin diğer veri türlerine göre daha erişilebilir olması ve örneğin yorumlar gibi kişilerin direk duygularını bildiren anlamlar taşıması etkili olmuştur. Bu alanda yapılan çalışmalar incelendiğinde bahsedilen metin madenciliği sürecinin her bir adımı için farklı programlar yada kodlama bilgisi ile ilerlendiği görülmektedir. Ayrıca kullanılan programlar kural tabanlı ilerlenmesini sağlayarak kullanıcılara süreç hakkında önerilerde bulunmamaktadır. Bu çalışmada hazırlanan program aracılığıyla metin verilerinin; ön işleme ve analiz adımlarının yönlendirmeler yardımıyla, programlama bilmeyen kişiler tarafından da kolayca yapılması sağlanacaktır. Sanal asistan tasarımı yardımı ile kulllanıcılar süreç hakkında yönlendirmeler alarak analizlerini ilerletirken aynı zamanda analiz sürecini öğrenme imkanı bulacaklardır.

Çalışmanın Amacı

Bu tezin amacı sanal asistan tasarımı yardımı ile sosyal medya kaynaklarından çekilen metin verilerinin öneri sistemi aracılığıyla bu alanda yetkin olmayan kişiler tarafından da analizinin yapılmasını mümkün kılmaktır. Metin verilerinin çekilmesi, temizlenmesi, işlenmesi ve analiz edilmesi süreci oldukça uzun olabilmektedir. Bu sürecin yeterince yetkin olmayan kişiler tarafından kolayca yapılabilmesi ve süreç içerisinde çeşitli önerilerle sürece dahil edilmesi temel hedeflerdendir. Ek olarak ön işleme ve analiz adımlarında kullanıcıya gösterilen bilgilendirme ekranları ile kullanıcının süreci öğrenmesi sağlanacaktır. Ayrıca kullanıcıların yaptıkları analiz sonuçları kaydedilerek benzer çalışmalara kaynak oluşturacak şekilde modellenmesi amaçlanmaktadır.

Çalışmanın Yöntemi

Sosyal bilimciler geleneksel yöntemler ile yaptıkları analiz sürecinde anket, odak grup vb. yöntemler ile metin verisi elde edebilmektedir. Günümüzde ise doğrudan kullanıcıların paylaştıkları veriye erişim mümkün hale gelmiştir. Veri toplama yönteminin değişmesi ile birlikte ilgili veriye erişmek için özel teknikler kullanılması gerekebilmektedir. Bu çalışmada veri kaynağı olarak Twitter verisi kullanılmıştır. Verinin çekilmesi adımında açık kaynak kodlu programlama dilinden yararlanılmıştır. Bu dilde bulunan kütüphaneler sayesinde Twitter üzerinden API'den bağımsız veri çekme işlemi gerçekleştirildiği için istenilen sayıda veri elde edilebilmektedir. Sanal asistan tasarımı bir arayüz olduğu için kullanıcı istediği konu hakkında, istediği tarih aralığında, istediği dilde ve istediği sayıda veri elde edebilmektedir. Veri elde etme aşamasından sonra kullanıcılar sanal asistan tasarımı yardımı ile veri ön işleme adımlarını gerçekleştirebilirler. Burada asistan temizlenmesi gereken noktaları kullanıcıya bilgilendirme notu olarak iletmektedir. Kullanıcı istediği temizleme adımlarını kullanarak analizi gerçekleştirebilir. Bu adımlar verinin analize hazır hale getirilmesine yardımcı olmaktadır. Analiz aşamasında ise duygu analizi ve topik modelleme yöntemleri kullanılmıştır. Kullanıcı yaptığı analizin içeriğine göre ön işleme adımlarından geçirdiği veriyi, duygu durumuna ve kategori dağılımına göre sınıflandırabilmektedir. Kategori dağılımının elde edilebilmesi için etiketlenmiş olarak hazır bulunan veri üzerinden makine öğrenmesi modelleri eğitilmiştir. En başarılı model üzerinden kullanıcının çektiği veri üzerinde tahmin yapılmaktadır. Bahsedilen analiz süreçlerinin tamamı asistan yardımı ile tamamlandıktan sonra yapılan analiz süreçlerinde kullanıcının yaptığı seçimler excel olarak kaydedilmektedir. Burada kaydedilen analiz verisi temizleme işlemlerini, duygu analizi ve topik modelleme sonuçlarını içermektedir. Bu veri sanal asistan süreçlerinin ve önerilerinin iyileştirilmesi için ileriki çalışmalarda makine öğrenmesi veri seti olarak kullanılacaktır.

Çalışmanın Organizasyonu

Metin madenciliğinin süreçlerinin detaylı olarak anlatıldığı; sosyal medya analitiğinin özellikle sosyal bilimler açısından incelendiği bu çalışmada sanal asistan yardımı ve yönlendirmesi ile kullanıcılar sosyal medya verisi kullanarak istedikleri konuda analiz yapma imkanı bulmaktadır. Bahsedilen süreçler bu çalışma içerisinde 6 bölüm başlığında organize edilmiştir.

- Giriş bölümünde sosyal medya analitiği süreci kullanım alanlarından, zorluklarından ve verinin işlenmesi ve analizi için kullanılan metin madenciliği süreçlerinden genel olarak bahsedilmiştir.
- Bölüm 1'de metin madenciliği sürecinin 4 temel adımından, süreçte karşılaşılabilecek zorluklardan ve bu alandaki uygulama yöntemlerinden bahsedilmiştir.
- Bölüm 2'te sosyal medya analitiğinin öneminden ve literatürde nasıl ele alındığından bahsedilmiştir. Ayrıca sosyal medya analitiği sürecindeki araştırma ve analiz süreçlerindeki zorluklara değinilmiştir. Son olarak ise sosyal medya analiz araçları özetlenmiştir.
- Bölüm 3'te sosyal medya analitiğinin literatürdeki çalışmaları ve kapsamı incelenerek süreçteki zorluklar irdelenmiştir.
- Bölüm 4'te sanal asistan tasarım sürecinde kullanılan metodolojilerden bahsedilerek tasarım ekranlarına ve süreç akış şemasına yer verilmiştir. Ayrıca tasarımın literatürdeki çalışmalarda bahsedilen boşlukları nasıl doldurduğu üzerinde durulmuştur.
- Bölüm 5'da sanal asistan kullanım süreci örnek bir veri çekilmesi ile başlayarak veri ön işleme ve analiz adımları da dahil adım adım aktarılmıştır.
- Çalışmanın son bölümünde ise sanal asistan süreci ile elde edilen kazanımlardan bahsedilmiş, bu alanda çalışma yapmak isteyen araştırmacılara ve bundan sonra yapılacak çalışmalara faydalı olacağı düşünülen öneriler ve rastlanılan zorluklardan bahsedilmiştir.

BÖLÜM 1: METİN MADENCİLİĞİ

Metin verisi yapılandırılmamış verilerin en temel örneklerinden biridir. Alışkın olunan ilişkisel veri yapısı düzeninden farklı olarak depolanması ve işlenmesi için farklı süreçler gerektirmektedir. Metin madenciliği yapılandırılmamış bilgileri işler, metinden anlamlı sayısal indeksler çıkarır ve metinde bulunan bilgileri çeşitli veri madenciliği (istatistiksel ve makine öğrenimi) algoritmaları için erişilebilir hale getirir (Agrawal ve Batra, 2013). Metin madenciliği ile veri madenciliği arasındaki fark incelendiğinde; metin madenciliğinde kalıplar doğal dil metninden çıkartılırken, veri madenciliğinde kalıplar veritabanlarından çıkartılır. Ayrıca web madenciliğinde analiz edilen web kaynakları yapılandırılmış olduğu için metin madenciliğinden bu noktada ayrışmaktadır.

Yapılandırılmamış veri kaynağı olan metinler 1990'ın sonlarında araştırmacılar tarafından veri kaynağı olarak kullanılmaya başlanmıştır (Dexter, 2017). İnsanların metinleri inceleyerek onları karar alma amaçlı kullanması teknolojininde gelişimi ile birlikte 2000'lerde daha yoğun görülmeye başlamıştır. Metin madenciliği günümüzde de çok farklı alanlarda ticari ve toplumsal fayda oluşturabilmek amacıyla kullanılmaya devam etmektedir. Aşağıda bu alanlardan bazıları özetlenmiştir (Sumathy ve Chidambaram, 2013).

- Telekomünikasyon, enerji ve diğer hizmet endüstrileri
- Bilgi Teknolojileri sektörü ve İnternet
- Yayıncılık ve medya
- Bankalar, sigorta ve finans piyasaları
- Siyasi kurumlar, siyasi analiz, kamu yönetim ve yasal belgeler
- İlaç ve araştırma şirketleri ve sağlık hizmetleri
- Biyo-Bilişim, İş Zekası ve ulusal güvenlik

Metin verisinin en temel yapı taşı kelimelerdir. Kelimeler cümleleri, cümleler paragrafları, paragraflar ise birleşerek dökümanları oluşturur. Bu yapının doğru şekilde anlaşılması ve analiz edilmesi dil bilimi, istatistik, makine öğrenmesi, veri madenciliği gibi bir çok disiplinin bir arada çalışmasını gerektirir.

Metin madenciliği, kelime frekansı odaklı ve anlamsal odaklı olmak üzere iki kapsamda değerlendirilebilir. Aşağıda bu yaklaşımlar özetlenmektedir.

Kelime odaklı yaklaşım: Kelime torbası olarak ifade edilen bu yaklaşımda kelimenin metinde geçme frekansı dikkate alınmaktadır. Kelime torbası yaklaşımı metin madenciliği sürecinde önemli faydalar sağlamaktadır ancak bu yaklaşımda kelimenin cümle içindeki konumunun ve kelimeler arası ilişkilerin göz ardı edildiğinin bilinmesi gerekir. Buda cümle bağlamının ve dolayısıyla belgedeki sözcüklerin anlamının yok sayılması demektir. Anlamsal kaygı taşımayan döküman sınıflandırılması gibi işlemlerde başarı ile kullanılabilir. Aşağıda dokumanlar üzerinden kelime torbası yöntemine örnek paylaşılmıştır.

Tablo 1: Kelime Torbası Örneği

Döküman	ekonomik	kriz	tüm	sektörü	etkiledi	ülke	etkisinde
Ekonomik kriz tüm sektörü etkiledi.	1	1	1	1	1	0	0
Ülke ekonomik krizin etkisinde.	1	1	0	0	0	1	1
Ekonomik belirsizlik fiyatlarıda etkiledi.	1	0	0	0	1	0	0

Doğal Dil İşleme: Bu yöntemde kelimelerin cümleler içerisindeki konumu, anlamı ve cümle içerisindeki görevi gibi birçok unsur incelenerek madencilik yapılmaktadır. NLP çalışmaları 1950'ler de başlamış ve bilgisayarların insanları anlamasına odaklanmıştır (Dexter, 2017). Teknolojinin de gelişimi ile birlikte günümüzde makine öğrenmesi ve derin öğrenme modelleriyle kullanılmaktadır. Doğal dil işleme bilgisayarların konuşulan dili anlaması, işlemesi ve yorumlaması olarak özetlenebilir. İnsan dili anlaşılması zor belirsizliklerle doludur. Eş anlamlı sözcükler, eşsesli sözcükler, deyimler, metaforlar, dilbilgisi ve kullanım istisnaları, cümle yapısındaki farklılıklar bu belirsizliklere örnek olarak verilebilir. NLP doğal dilde bulunan bu belirsizlikleri belirli ölçülerde aşarak sağladığı kolaylıklar sayesinde birçok alanda kullanılabilmektedir. Aşağıda bu alanlar özetlenmektedir.

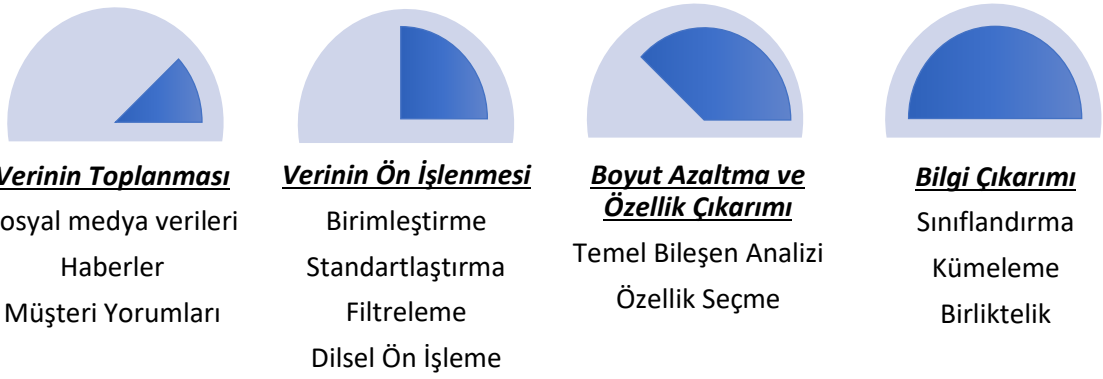
- Metin özetleme
- Makine çevirisi (Örn: Google Translate)
- Sohbet robotları – (Örn: Alexa, Siri)
- Bilgileri cümlelere çevirme
- Konuşma Tanımlama
- Metni konuşmaya dönüştürme

- Konu keşfi ve modelleme
- İçerik sınıflandırması

Yukarıda bahsedilen alanlar hem iş hem tüketici çevrelerinde birçok gerçek dünya uygulamasında kullanılmaktadır. Bu alanlarda genel amaç, doğal dil verisini alarak metni daha fazla değer sağlayacak şekle dönüştürmek veya zenginleştirmek için dilbilimini ve algoritmaları kullanmaktır.

1.1. Metin Madenciliği Süreci

Metin madenciliği süreci temelde 4 adımdan oluşur. Verinin (pdf, word, html, doc, css) toplanması, verinin ön işlenmesi, boyut azaltma ve özellik çıkarımı ve son olarakta metinden anlamlı veri çıkarılmasını sağlayan bilgi çıkarımı aşamasıdır.



Şekil 1: Metin Madenciliği Süreci

Kaynak: Tyagi, N. (2021). *Top 7 Text Mining Techniques.* Analytics Steps: https://www.analyticssteps.com/blogs/top-7-text-mining-techniques adresinden alındı

Sosyal medya, haber sitesi vb. platformlardan toplanan verinin temizlenmesi ve analize katkı sağlamayacak boyutların azaltılması ile devam eden süreç, bilgi çıkarım süreci ile sonuçlanmaktadır.

1.1.1. Veri Toplama

Bilgisayar ortamında bulunan bilgilerin, programlar tarafından işlenebilmesini sağlamak amacı ile derlenmiş ve formüle edilmiş şekline veri denir (Açiler, 2020). Metin madenciliği süreçleri verinin elde edilmesi ile başlar. Burada çok farklı veri kaynakları olabilir.

Müşterilerin e-ticaret sitesi yorumları, sosyal medya sitesi paylaşımları, haber siteleri, kişisel bloglar, raporlar, email, akademik yayınlar vb. Birçok kaynağa erişim sağlanabilir.

Veri kaynaklarının çeşitliliği, veri toplama aşamasında da farklılıklara yol açmaktadır. Örneğin Twitter gibi bir sosyal medya platformu üzerinden sitenin kendi API'si yada python'da bulunan çeşitli kütüphaneler yardımıyla veri çekilebilir. Bir haber sitesi içinse Web Kazıma işlemlerinin gerçekleştirilmesi gerekir.

1.1.2 Veri Ön İşleme

Metin madenciliğinin en önemli adımı olan veri ön işleme adımında çeşitli kaynaklardan toplanan verinin temizlenerek analize hazır hale getirilmesi hedeflenir. Veri kaynaklarının içeriklerinin çeşitliliği ve yapılan analizin içeriğine göre veri ön işleme adımları oldukça farklılaşabilir. Örneğin; duygu durumu değerlendirmesi önemli olmayan bir metnin analizinde emojilerin doğrudan temizlenmesi gerekirken tam aksine duygu durumuna odaklanılmış bir metinde emojilerin analize dahil edilmesi gerekir.

Veri ön işleme adımları genel olarak veri içerisindeki istenmeyen kelime ve ifadelerin temizlenerek, bilgi çıkarımının arttırılmasına odaklanır. Çünkü ön işleme adımından sonra, metin çeşitli yöntemler ile sayısallaştırılarak içerisinde saklı olan bilginin ortaya çıkarılması hedeflenir. Yanlış kurgulanmış bir ön işleme adımı sonrasında gerçekleştirilecek tüm adımların ve çıkarılan bilginin yanlış olmasına veya saklı kalmasına yol açacaktır. Ayrıca veri temizleme işleminin ilk adım olarak yapılması sonraki adımlarda ele alınacak veri boyutunu küçülterek kaynak açısından daha optimal sonuçlara ulaşılmasını sağlayacaktır. Sonuç olarak veri ön işleme aşamasının amacının, metni tek tek kelimelere ayırarak metni bir nitelik vektörü olarak sunmak ve elde edilen nitelikler ile metin arasında ilişki kurmak olduğu anlaşılmaktadır (Aksoy, Çelik ve Gülseçen, 2020).

Veri ön işleme adımları, Twitter gibi oldukça fazla sözcük çeşitliliği içeren veri kümelerinde sektör ve akademi verilerine göre daha uzun sürecektir. Sektör verilerinde sözcük çeşitliliği ve gramer dışı cümle sayısı oldukça azdır. Twitter verileri ise çok fazla kısaltma, büyük küçük harf kullanımı ve devrik cümle barındırır. Bu noktada analiz edilecek verinin kalitesi önem taşımaktadır. Verinin kalitesinin anlaşılmasını sağlayan bazı kriterler aşağıda özetlenmiştir.

a. ***Format:*** Dokumanların analize uygun benzer formatta olması gerekir.
b. ***Standart:*** Metin içerisindeki tüm cümlelerin benzer standartta olması, devrik cümle kullanımının düzenlenmesi
c. ***Tutarlılık:*** Veri seti içerisindeki belgelerin benzer cümle yapısı, dilbilgisi kullanılımı ve yazım diline sahip olması
d. ***Tam:*** Metin verilerinin eksiksiz ve tam bir şekilde alınması gerekmektedir.
e. ***Kesinlik:*** Analize kabul edilen her metnin hatasız olarak alınmış olması gerekir. Aksi takdirde çalışma sonucunun kesinliği tartışmaya açık olacaktır.
f. ***Tekrarlı:*** Analiz edilecek metinin tekrarlı veri içermesi sonuçların doğrulugunu etkileyecektir. Örneğin; Twitter verisi analiz edilirken retweet'lerin çoklamaya neden olmamak için temizlenmesi gerekir.

Metin verilerinin ön işlemesi aşamasında veri kalitesinin yukarıdaki metriklere bağlı olarak değişeceğinden bahsedildi. Bu noktada kalitenin arttırılması için aşağıdaki adımlar izlenerek daha doğru analiz sonuçlarının elde edilmesi sağlanabilir.

Şekil 2: Metin Verisi Kalite Arttırma Adımları

Kaynak: Kise, A. (2016). Veri Kalitesinin (Data Quality) Önemi ve Yönetimi. abdullahkise: http://www.abdullahkise.com/2016/07/veri-kalitesinin-data-quality-onemi-ve_19.html adresinden alındı

Veri kalitesini arttırma sürecinde yukarıdaki işlem adımlarını uygularken yapılacak analizin gerekliliklerine uygun olarak, işletme kurallarına hakim olmak, verinin elde edildiği kaynaklardan haberdar olmak, veri temizleme teknolojilerine ve yaklaşımlarına aşina olmak, geliştirme süresini kontrol altında tutabilmek ve her aşamada veri güvenliğini sağlamak gerekir.

Veri ön işleme adımları temelde 4 süreçten oluşur. Birimleştirme (Metnin kelime boyutuna indirilmesi), Standartlaştırma (Karakter ve biçim farklılıklarının düzeltilmesi), Filtreleme (Durdurma kelimelerinin temizlenmesi) ve Dilsel Önişleme (Dilbilgisi öğeleri detayına inilmesi)

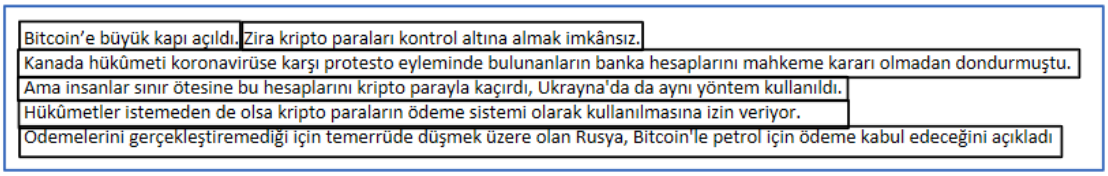
1.1.2.1. Birimleştirme

Metnin boşluk veya özel karakterler gibi istenen özelliklere göre parçalara ayrılması işlemine birimleştirme (tokenization), ayrılan kelime parçalarına ise token adı verilir. Temelde dokumanların, cümlelerin en küçük birimlerine; kelime ve kelime gruplarına ayrılması işlemidir. Bu aşamada dokumanda bulunan kelimeler veya kelime grupları bir liste halinde analize hazır hale getirilir. Ayrıştırma işlemi ise kullandıgımız birimleştirme yöntemine göre değişmektedir (Noyan, 2019).

a. ***Kelime Bölümleme:*** Cümleyi kelimelerine ayırma işlemi ve noktalama işaretlerinin ayıklanması
b. ***Cümle Bölümleme:*** Paragrafı cümlere ayırma işlemi
c. ***Ağaç Yöntemi ile Kelime Bölümleme***: Cümlelerdeki kelimelerin boşluklara ve noktalama işaretlerine göre ayrılması
d. ***Noktalama İşaretleri ile Bölümleme:*** Cümledeki noktalama işaretlerinin ayıklanması

Aşağıda kelime bölümleme ve cümle bölümleme için örnekler görülmektedir.

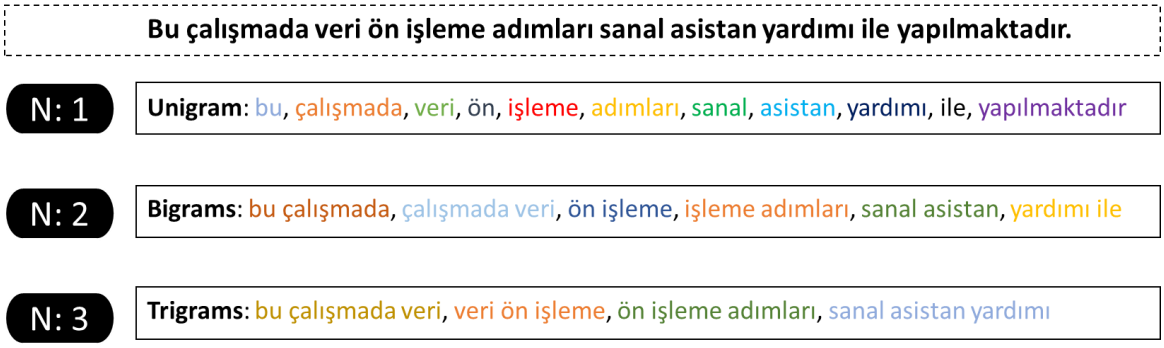
Şekil 3: Kelime Bölümleme Örneği



Şekil 4: Cümle Bölümleme Örneği

Birimleştirme aşamasında yapılan bir diğer adım n-gram'dır. N-gram modelleri, ardışık bir dizinin sonraki elemanını Markov zincirinden faydalanarak bulmakta kullanılır (Akıncı, 2020). Bu adımda 2'li ve 3'lü kelime gruplarının anlamlılığına bakılarak gerektiğinde grup haline bu kelime grupları da analiz dahil edilir. Burada dikkat edilmesi gereken nokta, 3'ten fazla kelime grubunun alınması veri setinin boyutunu arttırabilir. N-gram'a örnek olarak

telefonda mesaj yazarkan kullanılan sonraki kelime tahmini ve otomatik düzeltme verilebilir.

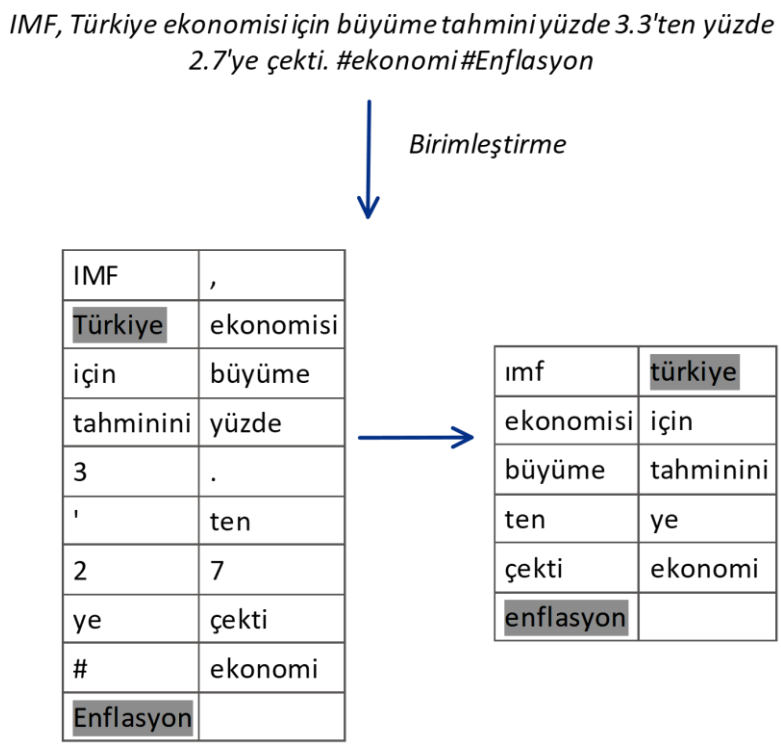


Şekil 5: N-Gram Çalışma Yöntemi

N-gram yöntemi görüldüğü gibi seçilen n parametresine göre değişmektedir. Kelime olarak birimleştirilen cümleler gerektiğinde anlamlılığa bağlı olarak kelime grupları olarak analiz edilebilir.

1.1.2.2. Standartlaştırma

Veri ön işleme sürecinin ikinci adımında birimleştirilen veri üzerinde standartlaştırma işlemleri gerçekleştirilir. Standartlaştırma işlemi veri kaynağına ve analiz ihtiyacına göre değişmekle birlikte genel olarak aynı süreçlerden oluşmaktadır. Büyük küçük harf, sayısal ifadeler, noktalama işaretleri, özel ifadeler, veri kaynağına özgü çıkarılması gereken kelime grupları (Twitter için rt, http vb.) metnin analize hazır hale getirilebilmesi için standartlaştırılmalıdır. Aşağıda standartlaştırma süreci örnek bir cümle üzerinden gösterilmektedir.

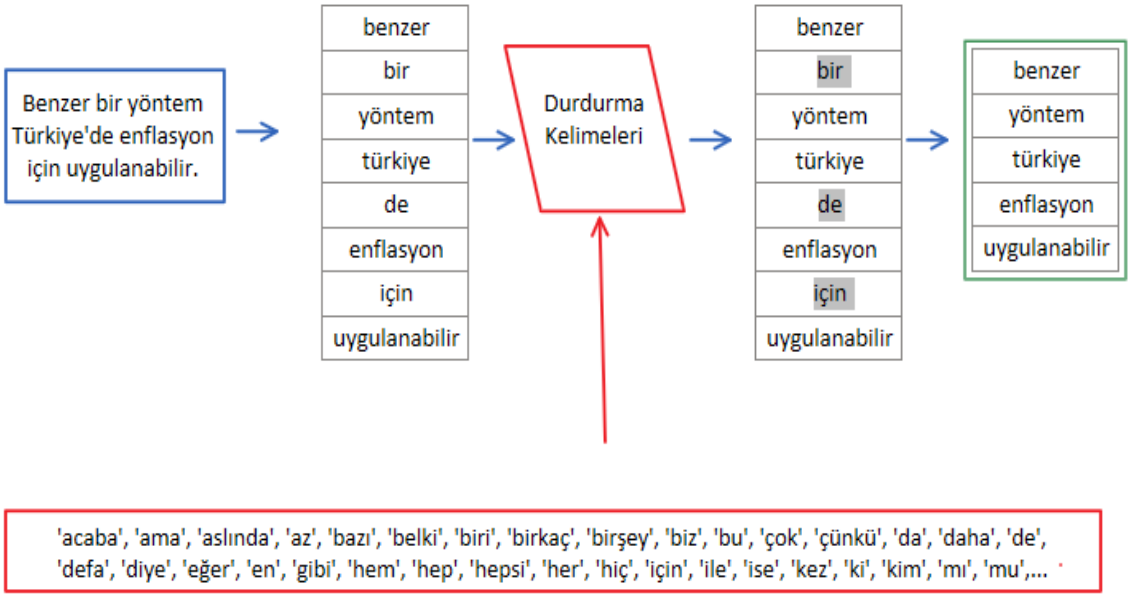


Şekil 6: Standartlaştırma Süreci Örneği

Örnekte görüldüğü üzere, büyük harflerin tamamı küçük harfe dönüştürülmüş ve noktalama işareti, sayısal ifadeler, özel karakterlerin hepsi için temizleme işlemi gerçekleştirilmiştir. Burada dikkat edilmesi gereken nokta örneğin emojilerin noktalama işaretleri temizlendiğinde kaybedileceğidir. Analiz içeriği ve süreci emojilerin ifade ettiği anlamları da gerektiriyor ise standartlaştırma aşamasında farklı bir süreç izlenmesi gerekir.

1.1.2.3. Filtreleme

Birimleştirme aşamasında kelime kelime ayrılan veri, standartlaştırma aşamasında içerdiği özel karakterlerden temizlendi. Ayrıca büyük küçük harf dönüşümü yapılarak harf hassasiyet durumunun önüne geçildi. Filtreleme aşamasına standart halde iletilen veri içerisinde bu aşamada kelime detayında inceleme yapılarak analize herhangi bir katkısı olmayan kelimelerin temizlenmesi gerekir. Bu aşama hem analiz çıktılarının doğrulugu hemde analiz sürecinin performansı için oldukça önemlidir.

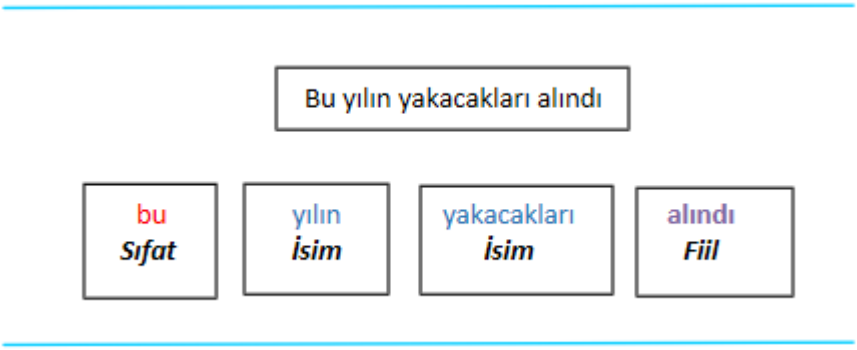


Şekil 7: Filtreleme Süreci Örneği

Filtreleme işleminde kullanılan durdurma kelimeleri python'da NLTK kütüphanesinde birçok dil için bulunmaktadır. Ayrıca farklı kaynaklarda farklı analiz türlerine hizmet eden listeleri de bulmak mümkündür. Burada durdurma kelimeleri dışında analizin içeriğine göre az geçen kelimelerde (sparsity) performans sebebiyle temizlenebilir. Ayrıca örneğin; ekonomi hakkında bir analiz yapılıyor ise analiz sonucunda Twitter verisinde yapılann etiketlemelerden dolayı en fazla geçen kelime ekonomi olarak belirlenecektir. Bu beklenen bir durum olduğu için veri içerisindeki daha derin anlamlara ulaşmak amacıyla bu tarz kelimelerde filtrelenebilir.

1.1.2.4. Dilsel Ön İşleme

Veri ön işleme adımlarında yapılan ilk üç adım biçimsel değişikliklere odaklanmaktaydı. Dilsel ön işleme adımı ise; verinin dilbilgisi kurallarına göre analiz edilmesi anlamına gelmektedir. Bu işlem NLP çalışmaları için yapılması gereken önemli adımlardan biridir. Dilsel ön işlemenin ilk adımı cümlelerin dilbilgisi kurallarına göre parçalanması sürecini içeren konuşma bölümü etiketleme (part of speech tagging) işlemidir.



Şekil 8: Konuşma Bölümü Etiketleme İşlemi

Konuşma bölümü etiketleme işlemi ingilizcede NLTK (Doğal Dil İşleme Araç Takımı) kütüphanesinde bulunan pos_tag fonksiyonu yardımıyla kolayca yapılabilir. Türkçe'de ise NLTK kütüphanesi içerisinde bir fonksiyon bulunmamaktadır. Bu alanda eğitilen çeşitli pre-train modeller kullanılarak etiketleme yapılabilir. Ancak Türkçe'de konuşma bölümü etiketleme için bazı zorluklar bulunmaktadır. Ana kısıtlama yeterince büyük bir eğitim külliyatının olmamasıdır. Eğitim külliyatı başarı üzerinde doğrudan etkilidir ve Türkçe için diğer diller ile kıyaslama yapıldığında eğitim külliyatı oldukça yetersiz kalmaktadır. Türkçe'nin özelliklerinden kaynaklanan ikinci kısıtlama ise İngilizce veride 'go' ve 'goes' kelimelerini bulmak yeterli iken; Türkçe'de etiketlemenin sağlıklı olması için 'gitmek', 'gittin', 'gitti', 'gittiniz', 'gittiler' kelimelerin hepsinin etiketlenmesi gerekir (Altunyurt ve Orhan, 2006).

Dilsel ön işleme adımında cümle yapısının dilbilgisi kurallarına göre ayrıştırılması sağlandıktan sonra, cümledeki kelimelerin köklerine inilerek analize devam edilir. Burada cümlelerin köküne inmek için iki farklı yaklaşım mevcuttur: Gövdeleme (Stemming) ve anlamsal köke inme (Lemmatizing).

Gövdeleme algoritmaları, bir kelimede bulunan ortak ön eklerin ve son eklerin bir listesini dikkate alarak kelimenin başlangıcını ve sonunu kesmeye çalışır (Özbek, 2019). Ayrım gözetmeksizin yaptığı bu kesim bazı durumlarda başarılı olabilir ancak kök ile kelime anlamı arasındaki ilişkiyi göz önüne almaması dezavantajıdır. Bu sebeple eksik sıkılama ve aşırı sıkılama adı verilen 2 farklı sorun ortaya çıkabilir. Aşırı sıkılama bir kelime çok fazla kırpıldığı zaman ortaya cıkarken, eksik sıkılama ise tam tersidir. Aşağıda 3 farklı stemming

algoritması listelenmektedir. Bu algoritmaların üçüne de python'un NLTK kütüphanesi üzerinden erişilebilir.

a. ***Porter Stemmer:*** Bulduğu kelimelerin ortak sonlarını çıkararak ortak kök elde etmeye çalışır.
b. ***Snowball Stemmer:*** Porter Stemmer'ın iyileştirilmiş hali, daha hızlı çalışır.
c. ***Lancaster stemmer:*** En agresif stemmer algoritması olarak bilinir. Bazı örneklerinde gerçek bir anlama gelmeyen kökler bulduğu belirlenmiştir.

Anlamsal köke inme ise kelimeleri morfolojik olarak inceler. Bu süreçte kelimenin çekimlenmemiş ilk haline kök (lemma) denir (Gidiyorlar fiili için gitmek). Anlamsal köke inme algoritmaları çalışmak için sözlüğe ihtiyaç duyar. NLTK kütüphanesi içerisinde WordNetLemmatizer'la kelimelerin kökleri bulunabilir (Noyan, 2019).

Kelimelerin köklerine inilerek ortak kökteki kelimelerin yakalanmasını sağlayan bu adımdan sonra tüm dokuman ve terimlerin bir arada bulunduğu Terim Döküman Matrisi (TDM) oluşturulur. Aşağıda TDM için örnek verilmiştir.

Tablo 2: Terim Dokuman Matrisi

	Tweet-1	Tweet-2	...	Tweet-N
Term-1	0	0	0	0
Term-2	1	1	0	0
...	0	0	1	1
Term-N	0	0	1	0

Matris içerisinde kelimelerin frekansları ya da yukarıda gösterildiği gibi varlık/yokluk (0,1) durumları bulunur. Örneğin Tweet-1 içerisinde Terim-1 geçmemektedir. Bu sebeple değeri 0 olarak atanmıştır.

1.1.3. Boyut Azaltma ve Özellik Çıkarımı

Veri kaynaklarının çeşitliliği ve gelişen teknoloji sayesinde erişilebilen veri miktarı gitgide artmaktadır. Bunun sonucu olarak ön işleme adımları ile analize hazır hale getirilen verinin boyutu çok yüksek, yönetilemez boyutlara ulaşabilmektedir. Bu sebeple bilgi çıkarımı aşamasına geçilmeden önce veriyi en iyi şekilde temsil edecek daha düşük boyutta bir veri elde edilmeye çalışılır. Özellik çıkarma işlemi, öğrenme algoritmasının doğruluğunu

iyiştirebilir ve çok daha hızlı çalışmasını sağlar. Veriyi daha küçük boyutlara indirebilmek için gerçekleştirilen işlem özellik çıkarımı olarak adlandırılır. Özellik çıkarımı yöntemleri aşağıda özetlenmiştir (Liang, Sun, Sun ve Gao, 2017).

a. ***Filtreleme Metodu:*** Ortak bilgi kazancı, Bilgi kazancı
b. ***Füzyon Metodu (Ağırlıklandırma):*** Ağırlıklı KNN (K En Yakın Komşu Yöntemi), Merkez vektör ağırlıklı yöntem
c. ***Haritalama Metodu:*** Saklı anlamsal analiz (LSA), En küçük kareler eşleme yöntemi
d. ***Kümeleme Metodu:*** CHI (chi-square) kümeleme, Konsept indeksleme, Derin öğrenme yaklaşımı

Özellik çıkarımı aşamasında görüldüğü gibi birçok yöntem bulunmaktadır. Aşağıda en sık kullanılan özellik çıkarımı yöntemleri hakkında detaylı bilgi verilmiştir.

Filtreleme metodu altında bulunan TF-IDF (Terim Sıklığı – Ters Döküman Sıklığı) en yaygın kullanılan özellik çıkarımı süreçlerinden biridir. TF-IDF sözcüğün bulunduğu dökümanı ne kadar temsil ettiğini gösteren bir istatistiksel değerdir (Durmuş, 2021). Buradaki TF ilgili kelimenin dökümandaki frekansını, DF dokuman frekansını ve IDF ise DF değerinin logaritması alınarak hesaplanır. TF-IDF ile metinde frekansı çok yüksek olan ve sık tekrar eden kelimelerin önemi düşürülebilir, frekansı cok düşük olan kelimelerin ise önemi attırılabilir. Bu sayede boyut azaltma işlemi gerçekleştirilmiş olur.

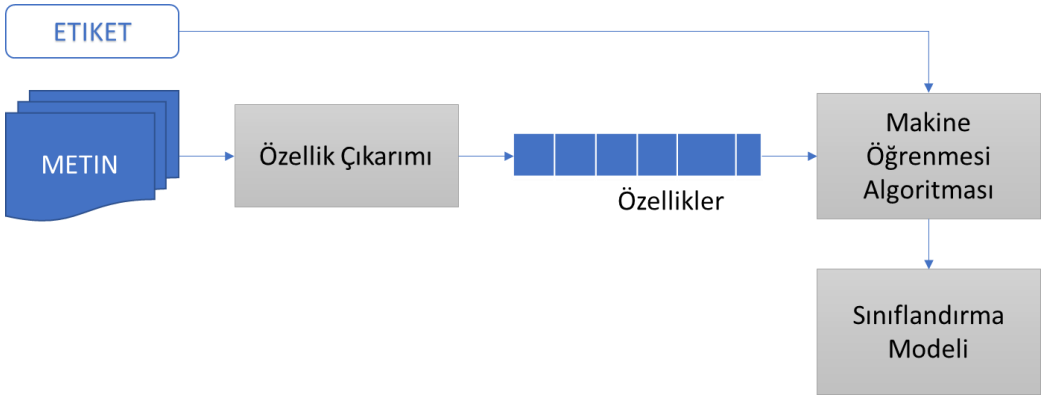
Saklı anlamsal analiz (LSA), metin kümesini analiz etmek için istatistiksel hesaplama yöntemi (SVD) kullanır. Böylece kelimeler arasındaki gizli semantic yapıyı çıkarır ve bu gizli yapıyı kelimeler ve metinleri temsil etmek için kullanır. Bunun sonucu olarakta kelimeler arasındaki korelasyonu ortadan kaldırır ve metin vektörlerini basitleştirerek boyutu azaltır. Tüm veri işleme adımları sonrasında analize hazır sayısallaştırılmış olan veri üzerinde analiz sürecine uygun olarak seçilen makine öğrenmesi yöntemleri kullanılabilir.

1.1.4. Bilgi Çıkarımı

Metin madenciliğinin son adımı olan bilgi çıkarımında sayısallaştırılmış olan metin verisi üzerinde sınıflandırma, kümeleme ve birliktelik analizi gibi analizler uygulanabilir. Bu analizlerin amacı metin verisi içerisindeki saklı bilgiyi çıkartarak yorumlamak ve karar desteği için kullanmaktır.

1.1.4.1. Sınıflandırma Algoritmaları

Sınıflandırma işleminde belirli dilsel kalıplar ile önceden belirlenmiş etiketler arasındaki ilişki araştırılır. Sınıflandırma sistemleri dilsel kurallara dayanmaktadır. Bu sebeple önceki adımlarda yapılan temizleme işlemlerinin doğruluğu oldukça kritiktir.



Şekil 9: Denetimli Öğrenme Süreci

Denetimli öğrenme kategorisi altında olan sınıflandırma algoritmalarına örnek olarak, K-En Yakın Komşu (KNN), Lojistik Resgresyon, Karar Ağaçları, Rastgele Orman, Gradyan Arttırma, XgBoost, Naive Bayes ve Destek Vektör Makineleri verilebilir. Bu algoritmalar birçok farklı alanda sınıflandırma için kullanılmaktadır. Örneğin; istenmeyen mesajların filtrelenmesi, Netflix gibi öneri sistemleri, teknoloji, politika veya sporla ilgili haber makalelerini sınıflandırma, yüz tanıma yazılımları vb.

K-En Yakın Komşu: Gözetimli bir makine öğrenmesi algoritması olan KNN, hem sınıflandırma hem de regresyon problemlerinin çözümünde kullanılabilir. Veri setine katılacak olan yeni verinin, mevcut verilere uzaklığı hesaplanıp k sayıda yakın komşusuna bakılır. Burada uzaklık hesabı için farklı ölçüm türleri mevcuttur. Bunlar; Euclidean, Manhattan ve Minkowski uzaklıklarıdır. KNN gürültülü verilere karşı dirençli olması sebebiyle popular sınıflandırma algoritmalarındandır (Ulgen E. , 2017). KNN modelinin başarısını ölçmek için, jaccard index (1'e yakın olması beklenir), F1 Score ve Logloss (0'a yakın olması beklenir) gibi metrikler kullanılır.

Lojistik Regresyon: Sınıflandırma da bağımlı değişken ikili yani yalnızca 1 (doğru) veya 0 (yanlış) olarak kodlanan verileri içerir. Lojistik regresyon, sınıflandırma yapmak için Sigmoid fonksiyonunu kullanır. Sigmoid fonksiyonu değerleri 0 ve 1 arasına sıkıştırır.

Varsayılan olarak sıkıştırılan değerlen 0,5'ten küçükse 0 (yanlış), büyükse 1 (doğru) olarak etiketlenir.

Karar Ağaçları: Sınıflandırmayı ağaç yapısı formunda gerçekleştiren bir algoritma türüdür. Karar ağacı veri setini entropi ve bilgi kazancı metrikleri sayesinde alt parçalara ayırır. Burada kritik nokta bilgi kazancının yüksek entropinin ise düşük olduğu noktadan ağacının bölünmesidir. Bölünmeler sonucu oluşan en son kırılımda ilgili karar sonucuna ulaşılmaktadır. Sınıflandırmanın yanı sıra sayısal verilerin kategorize edilmesi ile regresyon problemleri için de kullanılabilir.

Rastgele Orman: Karar ağaçları gibi hem regresyon hem de sınıflandırma problemlerinde kullanılabilir. Temelde birden fazla bağımsız karar ağacı kullanarak model doğruluğunu arttırmayı amaçlamaktadır. Karar ağacından farkı birden fazla karar ağacı oluşturulmasının yanı sıra rastgele orman için kök düğümün bulunması ve düğümlerin bölünmesi işleminin rastgele yapılmasıdır.

Gradyan Arttırma: Rastgele orman algoritmasında eğitilen bağımsız ağaçlardan sonra ortaya atılan bir yöntemdir. Her bir ağaç bir önceki eğitilen ağacın hataları üzerine kurulur. Bu şekilde zayıf öğreniciler güçlü öğrenicilere dönüştürülmeye çalışılır.

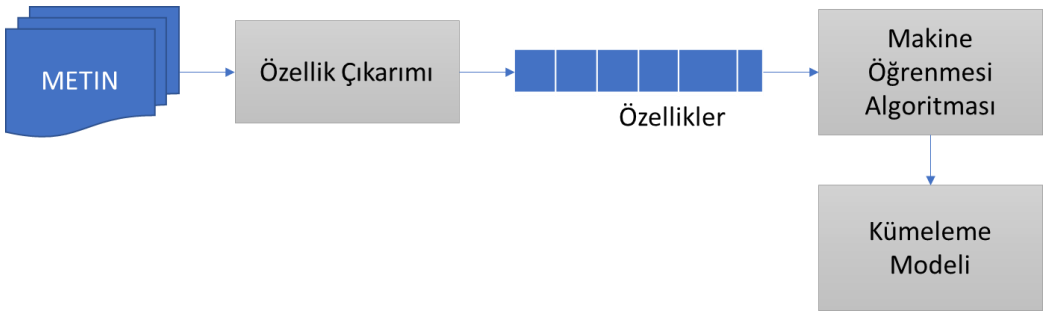
XgBoost: Gradyan arttırma algoritmasının çeşitli düzenlemeler ile performans açısından optimize edilmiş halidir. Başarı oranı diğer algoritmalara göre oldukça yüksek olabilmektedir.

Naive Bayes: Sınıflandırma algoritmaları içerisinde en eski yöntemlerden biridir. Temelinde bir eleman için her durumun olasılığını hesaplar ve olasılık değeri en yüksek olan sınıfa atama yapılır. Dengesiz veri kümelerinde de çalışabilir ancak eğitim kümesinde gözlemlenmeyen bir değişken oldugunda onun için tahmin yapamaz. (Sıfır frekans) (Hatipoğlu, 2018)

Destek Vektör Makineleri (SVM): Sınıflandırma ve regresyon problemleri için kullanılabilir. Koordinatı belirli olan her bir veri noktasını sınıflandırabilmek için en iyi ayrımı yapan hiper-düzlemi bulmaya çalışır.

1.1.4.2. Kümeleme Algoritmaları

Kümeleme ise denetimsiz algoritmalar grubundadır. Sınıflandırma da oldugu gibi modelleme öncesinde herhangi bir etiket verilmez. Modelin veri içerisindeki örüntüleri kendisinin keşfetmesi beklenir. Oluşan kümelerin kendileri içerisinde yüksek benzerliğe sahip olması, kümeler arasında ise düşük benzerliğe sahip olması beklenir. 4 tür kümeleme algoritması bulunmaktadır. Centroid Tabanlı Kümeleme, Yoğunluğa dayalı kümeleme, dağıtım tabanlı kümeleme, Hiyerarşik kümeleme.



Şekil 10: Denetimsiz Öğrenme Süreci

Kümeleme algoritmaları metinlerin içerisinde daha önce keşfedilmemiş örüntülerin bulunmasını sağlar. Bu örüntüler sayesinde metnin ayrıldığı kümelerin içerikleri incelenerek her kümeye isim verilebilir. Ayrıca kümeleme algoritmaları; sahte haberleri belirleme, spam algılama ve filtreleme, kitapları veya filmleri türüne göre gruplandırma gibi alanlarda da kullanılmaktadır. Kümeleme algoritmaları temelde ikiye ayrılır; Hiyerarşik ve hiyerarşik olmayan (Örn: K-Means, K-Medoids).

K-Means: Veri etiketine ihtiyaç duymadan veriyi kümelerine ayırabilen gözetimsiz algoritma türüdür. Örneklemin k adet kümeye bölünmesi ile; kümelerin kendi içinde yüksek benzerlik, kümeler arası ise düşük benzerlik içermesi amaçlanır. Buradaki benzerlik uzaklığa göre belirlenmektedir. K parametresi veri içeriğine göre dirsek yöntemi yardımı ile belirlebilir. Temelde uzaklık hesaplamasına dayanması sebebiyle gürültüye duyarlı bir algoritmadır.

K-Medoids: K-Means'in gürültü duyarlılığı dezavantajını gidermek için geliştirilmiştir. K-Medoids algoritması kümenin merkez noktasındaki elemanı yeni küme merkezi olarak belirlemektedir (Ulgen K. , 2018).

Hiyerarşik Kümeleme: Bu yöntemde iki adet yaklaşım bulunmaktadır: Agglomerative, Divisive. Agglomerative yönteminde her veri kendine ait küme oluşturur ve ardından birbirine yakın iki küme birleşerek yeni kümeyi oluşturur. Bu işlem tek bir büyük küme oluşana kadar devam etmektedir. Divisive'de ise başlangıçta tek bir veri kümesi vardır ve bu veri kümesi her bir veri noktası küme olana dek bölünmeye devam eder (Harman, 2020).

1.1.4.3. İlişkilendirme Algoritmaları

Metinlerin bir veri kümesinde bir arada olma olasılığını keşfetmek için kullanılan denetimsiz algoritmalardandır. En sık kullanılan ilişkilendirme algoritması Apriori'dir. Apriori, veri kümesinde ortak öğe kümelerini arayarak çalışır ve daha sonra bunlar üzerinde ilişki kurar. Örneğin bir eticaret sitesinde en sık satan ürün olan ayakkabıyı tespit etmemizi sağlayan apriori algoritması sonraki adımda ayakkabı alanların hangi olasılıkla tişörtte alacağının bulunmasını sağlar. Apriori'de tüm bu sonuçlara ulaşılmasını sağlayan metrikler; destek, güven ve lift değerleridir.

Denetimli (sınıflandırma) ve denetimsiz (kümeleme) öğrenme algoritmaları ayrı ayrı incelendiğinde her ikisinin de avantajları ve dezavantajları bulunmaktadır. Denetimli öğrenme için hızlı analiz süreci gerektiren durumlarda etiketlenmiş veri seti oluşturulması zaman alacaktır. Denetimsiz öğrenme sürecinde ise hızlı sonuç almak mümkündür ancak çok dağınık ve anlamsız kümeleme sonuçları elde edilebilir. Bu durumda kümelerin isimlendirmesinin zor olmasına sebep olur. Denetimli öğrenme de bunun aksine önceden belirlenen grupların sınıflandırılmasını sağladığı için daha net sonuçlar elde edilecektir.

1.2. Metin Madenciliği Sürecindeki Zorluklar

Veri kaynağının kolay erişilebilirliği, analiz edilmek istenen alanda özellikle sosyal medya analizlerinde ilgili kişilerin görüşlerini yansıtması sebebiyle metin madenciliği işlemleri sıklıkla tercih edilmektedir. Ancak metin madenciliği süreci çeşitli zorluklar barındırmaktadır ve bu zorluklar analiz sonucunda karar verme verimliliğini etkilemektedir. Bahsedilen zorluklar analiz edilecek dilin yapısına göre değişmektedir.

Doğal dilin kendisine has karmaşıklığı ve kalıpları bulunmaktadır. Farklı dillerde aynı kelimeler farklı anlamlarda kullanılabilmektedir. Bunun yanı sıra anlamsal olarak aynı dilde aynı kelimelerin birden fazla anlamı olabilmektedir. Bu kelimelerin denetimli veya

denetimsiz algoritmalar ile anlamlarının tespit edilmesi oldukça maliyetli olabilmektedir. Bu aşamada kelimelerden daha genel kapsama gidilerek kelimelerin farklı anlamlarda kullandığı kelime gruplarının etiketlenmesi ya da kümelenmesi gereklidir. Buda analiz süresini ve doğruluğunu oldukça etkilemektedir. Örneğin; Türkçe gibi sondan eklemeli dilllerde kelimenin köküne ulaşılması mevcut araçlarla zor olduğu gibi aynı zamanda ulaşılan kökün anlamsal karşılığını belirlemekte zor bir süreçtir.

Metin madenciliği sürecindeki zorluklar metnin elde edilmesi aşamasında başlamaktadır. Verinin değerinin oldukça arttığı günümüzde veri kaynaklarına erişim veri sahipleri tarafından belirlenen ölçüde gerçekleştirilebilmektedir. Örneğin; Facebook, Instagram gibi sosyal medya platformalarında programların kendi API'leri aracılığıyla belirli sayıda veri elde edilebilmektedir. Bazı veri kaynakları ise ücretli olarak araştırmacılara açılmaktadır. Bu engeller aşılarak alınan veri kaynağının depolanması da araştırmacı açısından kapasite ve güvenlik sebebiyle sorun teşkil etmektedir. Bunun yanı sıra yapısal olmayan metin verisi farklı site kaynaklarında farklı formatlarda olabilmektedir. Bu sitelerden doğrudan veri alınması mümkün olsa dahi veri formatı sebebiyle verinin toplanma aşaması cok uzun sürebilmektedir.

Veri elde edildikten sonra ön işleme adımında metni standartlaştırarak verimli hale getirmek için çeşitli işlemler uygulanır. Örneğin; Twitter verisinde bulunan özel karakterlerin (http, #, @ vb.) temizlenmesi veriyi analize hazır hale getirmenin ilk adımlarından biridir. Veride bulunan analiz sürecine katkı sağlamayacak kelimelerin temizlenmesi de önemli bir adımdır. Burada hem özel karakterlerin hemde çıkarılacak kelimelerin belirlenme süreci analizin içeriğine göre değiştiği için zaman alan bir aşamadır.

Türkçe gibi sondan eklemeleri dillerin analizinde bir diğer karşılaşılan zorluk dilsel ön işleme aşamasıdır. Bu aşamada kelimeler köklerine indirgenerek anlamlarının ortaya cıkarılması amaçlanır. Ancak farklı anlamlara gelen kelimeler köklerine indirgendiklerinde aynı anlama ulaşabilir. Buda analiz sürecinin doğruluğunu etkiler. Aşağıda metin madenciliği sürecinde karşılaşılan zorluklar özetlenmiştir.

- Veriye ulaşımda veri sahiplerinin kısıtlamaları
- Doğal dilin karmaşıklığı
- Bir kelime veya cümlenin birden fazla anlam ifade etmesi

- Anlamsal analiz yöntemlerinin pahalılığı
- Metin madenciliği aracına alan bilgisinin entegre edilmesi
- Mevcut metin madenciliği araçlarının uzmanlar için tasarlanmış olması

Metin madenciliği süreçleri bahsedildiği gibi veriden anlamlı bilginin çıkartılması sürecinde kullanılmaktadır. Yukarıda bahsedilen zorluklar belirli ölçülerde aşılarak çeşitli çalışmalar gerçekleştirilebilmektedir. Ek olarak derin öğrenme gibi yöntemlerin gelişimi ve makine disk kapasitelerindeki artışla birlikte dil bazında özgünleşen bu problemlerin büyük öğrenme setleri ile aşılması beklenebilir.

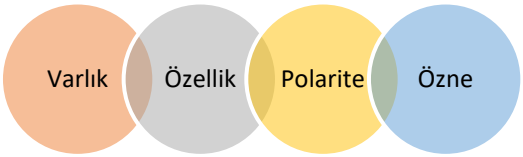
1.3. Metin Madenciliği Uygulamaları

Metin madenciliği akademik, ticari ya da bireysel manada oldukça geniş kullanım alanına sahiptir. Son yıllarda sosyal medya kullanımının da artışı ile metin madenciliği yaklaşımlarının önemide artmıştır. Bu bölümde 3 temel metin madenciliği uygulama alanı hakkında bilgi verilecektir.

1.3.1. Duygu Analizi

Akademik ve ticari çalışmalar için kaynak oluşturan insanların görüşlerinin incelenmesi, artan veri kaynaklarının etkisiyle önem kazanmıştır. Sosyal medya, kişisel siteler, bloglar vb. platformlar kişilerin herhangi bir konuda görüşlerini kolayca paylaşmasını sağlamaktadır. Duygu analizi ve fikir madenciliği insanların ürünler, hizmetler, organizasyonlar, olaylar gibi farklı konular hakkında görüşlerini ifade ettikleri metinler içinde saklı olan duygu, fikir ve düşünceleri ortaya çıkaran çalışmalara denilmektedir (Ozyurt, Akcayol, 2017). Duygu analizi çalışmaları sayesinde kişilerin ilgili olaya verdikleri olumlu, olumsuz, nötr tepkilerin ölçülmesi sağlanır. Bu sayede büyük miktarda veri kaynağının sınıflandırılması sağlanarak ticari fayda elde etmenin yanı sıra akademik açıdan da kitlesel eğilim izlenebilir.

Duygu analizinin temelinde bulunan duygu ifadesinin dört bileşenden oluşması beklenir. Bu bileşenler aşağıdaki şekilde görülebilir.



Şekil 11: Duygunun Dört Temel Bileşeni

Duygu analizi çalışmaları yapılacak analiz hedefine ve kapsamına göre çeşitli seviyelerden oluşur. Bu seviyeler aşağıda özetlenmiştir (Ozyurt ve Akcayol, 2017).

a. ***Döküman Seviyesinde Duygu Analizi:*** Analiz edilecek metni bir bütün olarak ele alma sürecidir. Burada metin olarak kastedilen veri sosyal medya verisi, haber metinleri, eticaret site yorumları vb olabilir. Metin içerisinde birden fazla konu hakkında yorum yapılması analiz sonuçlarının doğruluğunu etkileyecektir.
b. ***Cümle Düzeyinde Duygu Analizi:*** Döküman seviyesinde bir dezavantaj olarak bahsedilen metin içerisindeki konu kırılımları, cümle bazında bölünerek ayrıştırılır. Cümlenin duygu durumunun incelenmesine olanak sağlayan analiz türüdür.
c. ***Özellik Tabanlı Düzeyde Duygu Analizi:*** Duygu sınıflandırmasını dokuman düzeyde yapmak duyguların dağılımını belirsizleştirdiği gibi cümle seviyesinde yapmakta bazı saklı anlamların ortaya çıkmasını engellemektedir. Örneğin; e-ticaret sitesinde yorum yapan bir kullanıcı ayakkabının 3 özelliğinden olumlu, 1 özelliğinden olumsuz bahsettiğinde dokuman genel olarak olumlu gözükecektir. Burada olumsuz özelliğin ayırt edilebilmesi için özellik seviyesinde duygu analizi yapılması gerekir.

1.3.1.1. Duygu Analizi Yöntemleri

Duygu analizi bir metin madenciliği sürecidir. Metin olarak elde edilen ve temizlenen veri sayısallaştırılarak duygu analizine hazır hale getirilir. Temelde duygu analizi işlemleri makine öğrenmesi ve indeks tabanlı yöntem aracılığıyla yapılabilir (Medhat, Hassan ve Korashy, 2014). Makine öğrenmesinde daha önceki bölümde bahsedildiği gibi denetimli ve denetimsiz öğrenme algoritmaları bulunur. Sayısallaştırılmış metin verisi için amaca uygun bir etiketleme mevcutsa denetimli öğrenme algoritmaları (Naive Bayes, SVM, Karar Ağacı,

KNN vb.), veri içerisindeki örüntüler ve kümeler algoritma aracılığıyla bulunmak isteniyorsa denetimsiz öğrenme algoritmaları (K-Means, hiyerarşik kümeleme vb.) kullanılabilir.

İndeks tabanlı duygu analizi yöntemi duyguların etiketli olarak bulunduğu kelime sözlükleri aracılığıyla gerçekleştirilir. Analiz sürecine uygun her bir kelimenin duygu karşılığının yer aldığı bir sözlük kullanılması gerekir. Ayrıca dolaylı olarak ifade edilen duyguları tespit etme noktasında makine öğrenmesi yöntemine göre başarısız olmaktadır. Ek olarak bu yöntemde ingilizce için kapsamlı sözlükler bulunabilmektedir ancak Türkçe için yapılacak analiz türüne göre başlangıçta duygu sözlüğü oluşturulması gerekebilir. Aşağıda duygu analizi sözlüğüne örnek bir çalışmaya yer verilmiştir (Sağlam, Genç ve Sever, 2019).

```
adalet;0,172454336;1
adalet sarayı;-0,58918;-1
adalet yapmak;0,161517439;1
adaletli;0,198981161;1
adaletsiz;-0,240971817;-1
adaletsizce;0,286517439;1
adaletsizlik;-0,070095558;-1
adalır;0,225013952;1
adali;0,354477439;1
adalî;0,284771606;1
adalmış;0,232092524;1
adam;0,136030102;1
adam cuma;0,411517439;1
adam düşmek;0,411517439;1
adam eti yemek;-0,20442685;-1
adam etmek;0,081967439;1
```

Şekil 12: Duygu Analizi Sözlük Örneği

Duygu analizinde literatürde yapılan çalışmalar yöntem detayında incelendiğinde, makine öğrenmesine dayalı yöntemlerin daha başarılı sonuçlar verdiği gözlemlenmiştir (Ozyurt ve Akcayol, 2017). Bu sebeple makine öğrenmesi algoritmaları duygu analizi süreci için tercih edilebilir ancak indeks tabanlı yöntemde olduğu gibi yüksek boyutta veri seti ile algoritmanın beslenmesi gerekir. Bu da analiz sürecini uzatmaktadır. Ayrıca teknik yöntem kısıtların yanı sıra analiz sürecini etkileyen bazı zorluklar bulunmaktadır. Bu zorluklar aşağıda özetlenmeye çalışılmıştır (Seker, 2016).

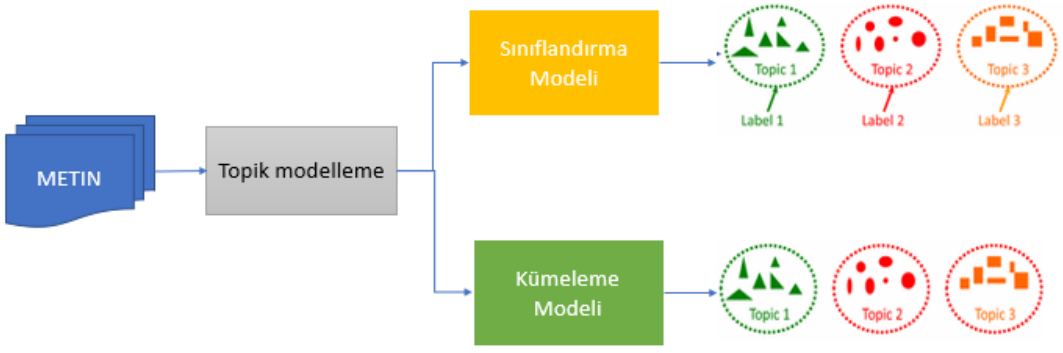
- Bazı cümlelerde ima yolu ile kapalı olarak görüş bildirilmesi,
- Aynı kelimelerin farklı anlamlara gelme durumu,
- Derlemin büyük kısmında olumlu bahsedilen konunun sonuç cümlesinde olumsuza dönmesi,

- Kültürler özelinde kullanılan cümle kalıplarının farklılaşması,
- Duygu ifadesinin cümledeki nesne dışında özneyle ilişkili olması,
- Bir cümlede birden fazla varlıktan bahsedildiğinde duygu durumu ayrıştırma zorluğu,
- Kelimenin olumsuz ek almasına rağmen olumsuz olmadığı durumlar olarak özetlenebilir.

Duygu analizi bahsedildiği gibi metinde saklı halde bulunan duygunun ortaya çıkarılma sürecidir ve birçok alanda kullanımı mevcuttur. Günümüzde en temelde sosyal medya verilerinin analiz edilerek ilgili araştırma konusu hakkında duygu dağılımının ortaya çıkarılmasında kullanılmaktadır. Bunun yanında e-ticaret siteleri sağladıkları ürün veya hizmet hakkında müşterilerin yorumlarının duygu dağılımı takibini yapabilmektedir. Bu sayede aynı ürünü satan farklı satıcılar içerisinden; olumsuz yorum alan satıcıların sayfa sıralaması değiştirilebilir. Olumlu yorum alan satıcılar ise arama sonuçlarında üst sıralara çıkarılarak yorumlar üzerinden müşteri memnuniyeti arttırılabilir.

1.3.2. Konu Modelleme

Günümüzde elde edilebilen veri miktarının artması ile metin verilerinin duygu dağılımının yanı sıra içerdiği konu dağılımlarının belirlenmesi de önem kazanmıştır. Konu modelleme, akademik, ticari yada bireysel yapılan tüm çalışmalar için yapılandırılmamış verinin içerdiği konu ve tema dağılımının tespitinde kullanılabilir. Sınıflandırma ve kümeleme olarak iki makine öğrenmesi yöntemi ile gerçekleştirilebilir.

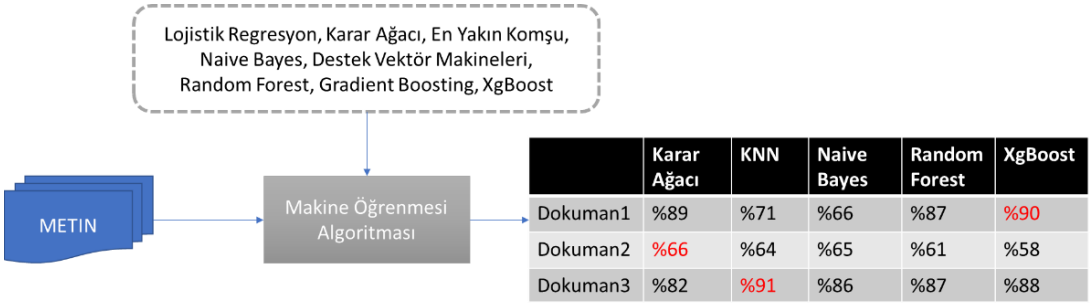


Şekil 13: Topik Modelleme Yöntemi

Gizli Dirichlet Ayrımı (LDA), metinleri bir konu koleksiyonu olarak kabul edilerek, belgedeki her kelimenin konulardan birine karşılık geldiği bir topik modelleme örneğidir (Gökmen, 2020). Temel mantığında kümeleme süreci bulunmaktadır. LDA verilen metni temel alarak her konu grubunu içerdiği kelimelere göre kümeler. Sonraki adımda ise elde edilen veri kümelerinin incelenerek analiz içeriğine göre isimlendirilmesi gerekir.

Sınıflandırma algoritmaları kullanılarak yapılan topik modelleme de ise farklı yöntemler bulunmaktadır. Buradaki yöntemler etiketli veri üzerinden öğrenme gerçekleştiren algoritmaların sonuçlarının kullanılma durumu ve sırası ile ilişkilidir. Aşağıda bu yöntemler özetlenmeye çalışılmıştır.

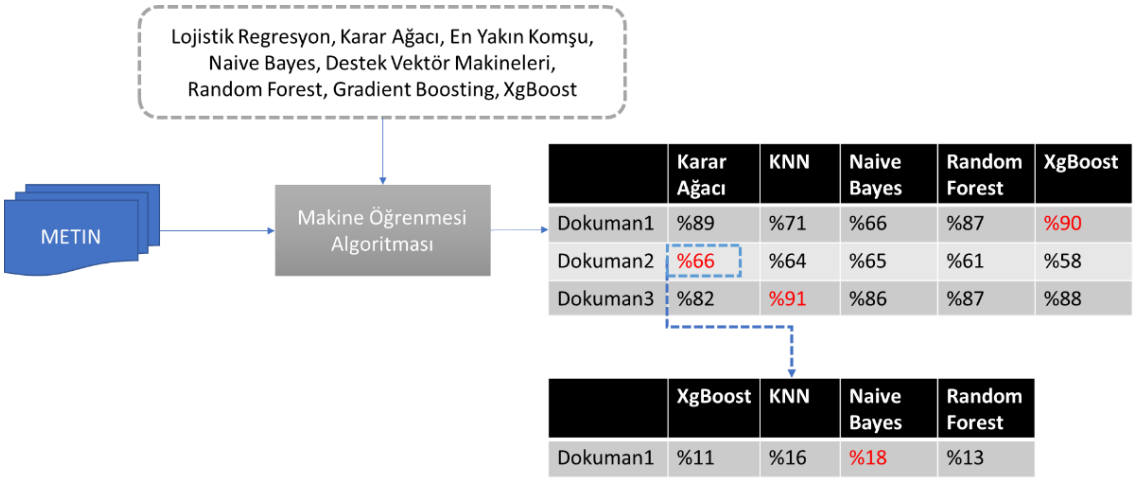
Başarılı algoritma seçimi: Etiketli veri üzerinde öğrenme gerçekleştiren algoritmalardan en başarılı sonucu veren algoritmanın seçilerek konu etiketlemesinde kullanılması sürecidir.



	Karar Ağacı	KNN	Naive Bayes	Random Forest	XgBoost
Dokuman1	%89	%71	%66	%87	%90
Dokuman2	%66	%64	%65	%61	%58
Dokuman3	%82	%91	%86	%87	%88

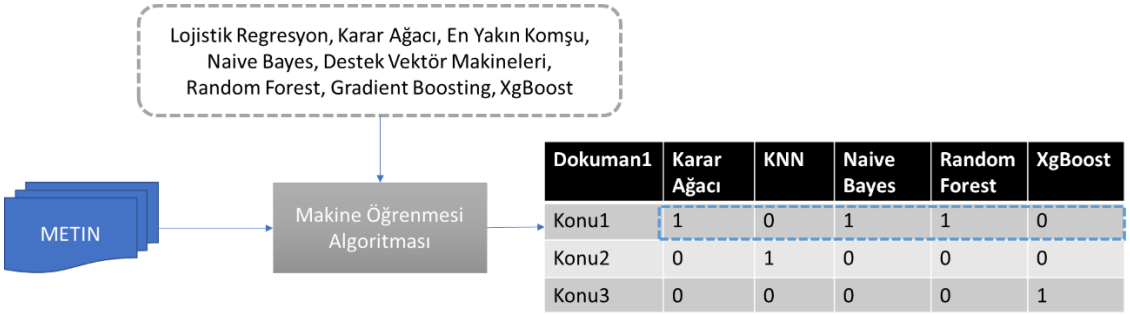
Şekil 14: Konu Modellemede Başarılı Algoritma Seçimi

Algoritmaların sıralı kullanılması: En başarılı öğrenme gerçekleştiren algoritmanın sınıflandırma olasılığının düşük olduğu metinlerin diğer algoritmalar yardımı ile tekrar sınıflandırılmasıdır. Örneğin aşağıdaki tabloda Dokuman2 için en iyi sınıflandırmayı yapan Karar Ağacı algoritmasının başarı oranı %66'dır. Bu algoritmanın metnin %34'ünü doğru tahminlemediği görülmektedir. Başarısız sınıflandırılan %34'lük kısım için diğer algoritmalar tekrar çalıştırılarak genel başarı oranı arttırılabilir.



Şekil 15: Topik Modellemede Algoritmaların Sıralı Kullanımı

Topluluk öğrenmesi yöntemi: Metin verisi üzerinde uygulanan algoritmaların sonuçlarının oylama yöntemi ile konu etiketine karar vermesinin sağlanmasıdır. Bu yönteme makine öğrenmesinde topluluk öğrenimi (ensemble learning) adı verilmektedir. Genel performansın iyileştirilmesi için birden fazla algoritmanın verdiği karar sonuçları birleştirilir. Örneğin dokuman1 için çalıştırılan algoritmaların çoğunluğu, döküman 1'in konu1'e ait olduğu sonucuna ulaşmıştır.



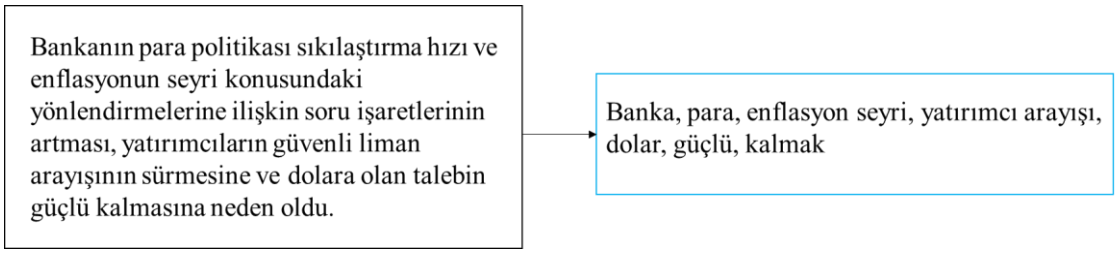
Şekil 16: Topik Modellemede Toplu Öğrenme Yöntemi

Sınıflandırma da kullanılan yöntemler temelde eğitim verisinin tahmin oranının doğruluğuna bağlıdır. Tahmin doğruluğu arttırılmak istenen algoritmalar bahsedilen yöntemler ile sıralı olarak ya da birleştirilerek kullanılabilir.

1.3.3. Özetleme

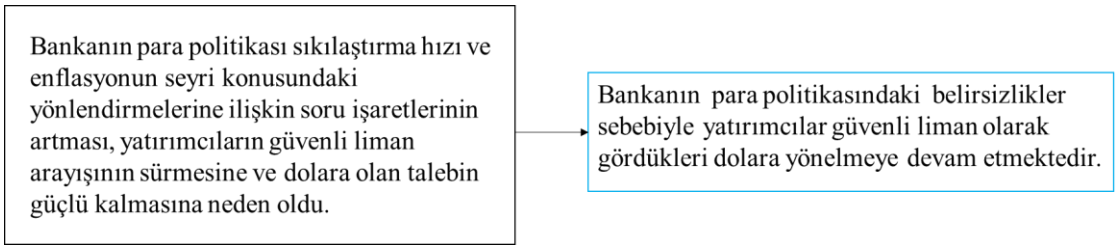
Teknoloijk gelişimin sonucu olarak günümüzde yapısal ve yapısal olmayan veri kaynaklarında önemli bir artış olmuştur. Yapısal olmayan metin verileri de bu artışın büyük bir kısmını oluşturmaktadır. Dijital alanda dolaşan büyük miktar verinin içindeki uzun metinleri otomatik olarak kısaltarak özetler halinde sunabilen yöntemlere ihtiyaç olmuştur. Bu yöntemlerde temel amaç metindeki konuların ana hatlarını tutarlı bir şekilde özetlemektir. (J.Garbade, 2018) Aşağıda özetleme için NLP'ye dayalı iki yöntem paylaşılmaktadır.

Çıkarıma Dayalı Özetleme: Bu yöntemde kaynak belgede bulunan anahtar sözcükler alınarak bu sözcükler özet için birleştirilir. Özetleme metinlerde bir değişiklik yapılmadan gerçekleştirilir bu sebeple bazen dilbilgisi açısından anlamsız özetler ortaya çıkabilmektedir. Aşağıda çıkarım tabanlı özetlemeye örnek verilmiştir.



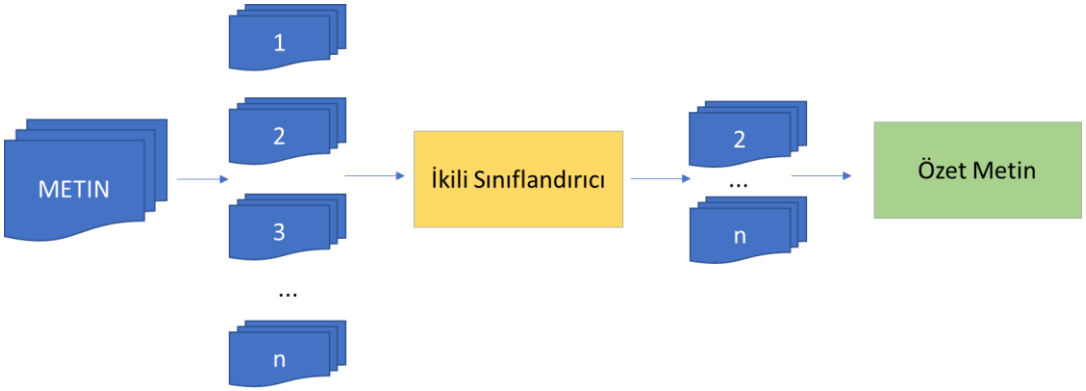
Şekil 17: Çıkarıma Dayalı Özetleme Yöntemi

Soyutlamaya Dayalı Özetleme: Soyutlama tekniği, kaynak belge bölümlerinin başka sözcüklerle ifade edilerek değiştirilmesi ve kısaltılmasını gerektirir. Çıkarma yönteminde bahsedilen dilbilgisi tutarsızlıklarının üstesinden gelebilir ancak tıpkı insanların yaptığı gibi orjinal metinden en faydalı bilgileri aktaran yeni ifadeler ve cümleler oluşturduğu için soyutlamaya dayalı metin özetleme algoritmalarının geliştirilmesi zordur. Aşağıda soyutlaya özetlemeye örnek verilmiştir.



Şekil 18: Soyutlamaya Dayalı Özetleme Örneği

Metin özetleme süreci yukarıdaki teknikler yardımı ile araştırma sürecini kısaltarak bilgi edinme sürecini hızlandırmaktadır. Bu süreçte genel olarak denetimli makine öğrenmesi algoritmaları kullanılmaktadır. Denetimli makine öğrenmesi yardımı ile ikili sınıflandırma (özetlemeye uygun / özetlemeye uygun değil) modeli kurularak özetleme gerçekleştirilebilir. İkili sınıflandırma modeli kurulurken metnin uzunluğu, metin içindeki kelimelerin sıklığı, metin içinde en çok tekrarlanan kelimeler gibi özniteliklerin kullanılması önerilir. Aşağıda ikili sınıflandırıcı süreci gösterilmektedir.



Şekil 19: İkili Sınıflandırma Süreci

Metin özetleme sürecinde denetimli öğrenme süreci kullanılabildiği gibi denetimsiz öğrenme süreci de kullanılabilmektedir. Kümeleme yapılarakta (Örneğin; LDA) metnin içerisindeki özet bilgiye ulaşmak mümkündür.

BÖLÜM 2: SOSYAL MEDYA ANALİTİĞİ

Sosyal medya, teknolojinin gelişmesi ile birlikte insanlara yeni bir sosyalleşme ortamı olarak ortaya çıkmıştır. Sosyal medya, kullanıcı tarafından oluşturulan içeriği paylaşmak, birlikte oluşturmak, tartışmak ve değiştirmek için yüksek düzeyde etkileşimli platformlar veya bireyler ve topluluklar oluşturmak için mobil ve diğer Web tabanlı teknolojilere bağlıdır (Sharda, Delen, Turban, Aronson ve Liang, 2014). Sosyal medya kullanımının oldukça arttığı günümüzde, sosyal medya kaynakları da her geçen gün artmaktadır. Bu kaynaklara örnek olarak; Wikipedia, ekşisözlük gibi bilgi paylaşım plaftormları, görsel medya kaynağı olarak youtube, dailymotion gibi video paylaşım siteleri ve temelde sosyal ağ olarak kullanılan Facebook, Linkedin, Twitter gibi platformlar verilebilir.

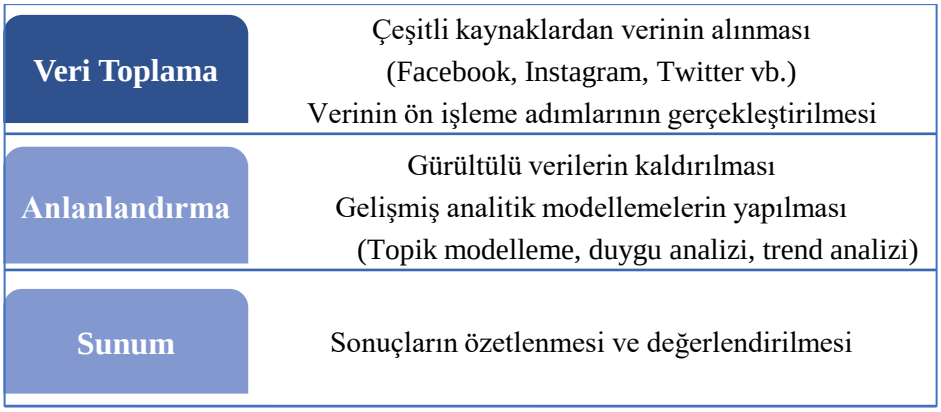
2.1. Sosyal Medya Analitiği Nedir?

Sosyalleşme insan hayatının önemli parçalarından biri olarak değerlendirilebilir. Sosyal bir varlık olan insan sosyalleşme ihtiyacını günümüzün teknolojik koşullarının da etkisiyle internet üzerinden karşılayabilmektedir. Sosyal medya kavramı bu noktada devreye girer. İnsanların hem haberleşmek hemde yakın çevresi ile iletişim kurmak için kullandığı sanal ağ ortamı sosyal medya olarak adlandırılır. Ancak tanım bununla sınırlı değildir. Sosyal medya, kullanıcısının aldığı herhangi bir hizmet yada ürün hakkında veya ülke gündemindeki herhangi bir konu hakkında görüşlerini anonim bir şekilde paylaşmasını mümkün kılar. Bu sayede insanlar herhangi bir kaygı taşımadan görüşlerini paylaşabilmektedir. Özellikle gündem hakkında yogunlukla paylaşım yapılan sosyal medya platformu Twitter en önemli veri kaynaklarından birini oluşturmaktadır. Yalnızca gündem değil bahsedildiği gibi ürün ve hizmet için buradan paylaşılan görüşlerde marka yöneticileri tarafından ticari fayda sağlamak için kullanılabilir.

Bahsedildiği gibi sosyal medya içerikleri insanların ve toplulukların bireysel görüşlerini içermesi sebebiyle önemli bir inceleme alanı oluşturmuştur. Siyasi, toplumsal ya da ekonomik bir olaya verilen tepkilerin doğrudan büyük bir veri üzerinden incelenebilmesi günümüz teknoloji gelişiminin bir sonucudur. Ayrıca bu teknoloji ile doğru orantılı olarak veriye erişim için kullanılacak araçlar ve sistem kapasitelerinde de önemli bir artış olmuştur.

Bu sayede milyonlarca kişinin görüşü gerekli teknik donanımlar ve sosyal medya analitiği kullanılarak analiz edilebilir.

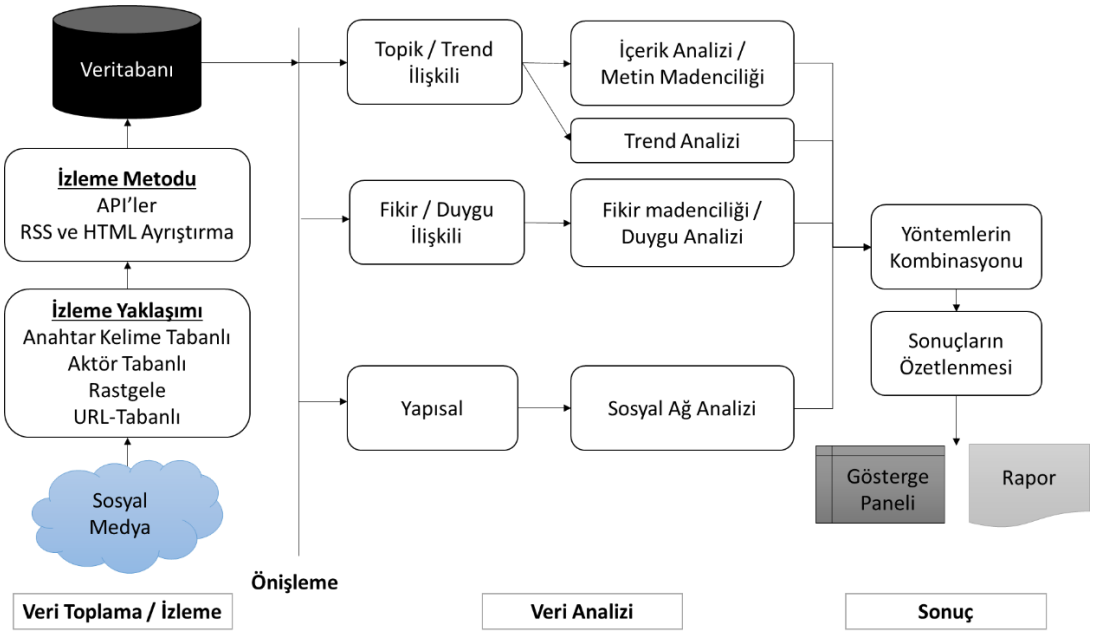
Literatürde sıklıkla kullanılan sosyal medya analiği çalışmalarının arka planında metin madenciliği işlemleri yer almaktadır. Bunun sebebi ulaşımının, depolanmasının ve analizinin daha kolay süreçlerden oluşmasıdır. Metin verisi dışında resim, video ve ses verileri sosyal medya araçları sayesinde gitgide artmaktadır. Konuşma bölümü etiketleme (Speech to text) denilen süreç ile birlikte ses kayıtları ve video verileri üzerinden elde edilen metin verisi de metin madenciliği aşamalarından geçirilerek analiz edilebilir. Bunun yanı sıra çeşitli görüntü işleme teknolojileri de görsel verinin yazıya dökülerek kullanılmasını mümkün kılmaktadır. Dolayısıyla tüm bu süreçlerin altında temelde yer alan ve sosyal medya analitiğinin temelini oluşturan metin madenciliği süreçleridir. Metin madenciliği ve sosyal medya analitiğinin farklı iki kavram olarak değerlendirilmesinin altında yatan neden ise sosyal medya verisinin yapısal değişikliğinden kaynaklanmaktadır. Kastedilen yapısal değişiklik; sosyal medya verilerinin içerdiği özel jargonlar, emojiler, gramer bağımsız yazım dilleri, içerdiği devrik cümleler gibi yüksek gürültü içeren bir veri seti olmasıdır. Bu da sosyal medya verisinin analizinde daha hassas ve özel yöntemlerle ilerlenmesi gerektiğini göstermektedir. Sosyal medya analitiği süreçleri aşağıda genel olarak özetlenmiştir (Fan ve Gordon, 2014).



Şekil 20: Sosyal Medya Analitiği Adımları

Sosyal ağ analizi ile sosyal medya analitiği birbirlerine çok yakın iki kavram olmaları sebebiyle sıklıkla karıştırılmaktadır. Sosyal medya analitiği yapılandırılmamış sosyal medya verilerine (Twitter, Facebook, Instagram vb.) uygulanan raporlama, görselleştirme, arama ve metin madenciliği işlemlerinin tamamı kapsamaktadır. Sosyal ağ analizi ise bireyler ve

gruplar arasındaki bağlantıları, ilişkileri ve etkiyi belirlemeye odaklanan gelişmiş ağ analitiğine verilen isimdir. Bu noktada farklılaştıkları nokta sosyal ağ analizi kavramının yapısal veri yapısı ile çalışması ve analiz sürecinde insanlar ve olaylar arasındaki ilişkiyi araştırmasıdır. Aşağıda sosyal medya analiğinin tüm sürecinin akarıldığı akış şeması paylaşılmaktadır.



Şekil 21: Sosyal Medya Analitiği Süreci

Kaynak: Stieglitz, S., Dang-Xuan, L., Bruns, A., ve Neuberger, C. (2014). Social Media Analytics. *Business and Information Systems Engineering*, 89-96.

Sosyal medya kaynağından konu etiketi, kişi ve url tabanlı veri elde edilebilmektedir. Yapılandırılmamış formatta elde edilen bu veri üzerinde konu modelleme, duygu analizi gibi işlemler uygulanabilir. Araştırmacı analiz sonucu elde ettiği anlamı ortaya çıkararak sunmak için ise son adımda gösterge panoları ve raporlardan yararlanır.

2.2. Sosyal Medya Analitiği Zorlukları

Sosyal medya bireylerin ve toplumların anlaşılmasını sağlayan zengin bir veri kaynağı içerir. Bu veri kaynağı yapılandırılmamış veri olarak sınıflandırılmaktadır. Gelişen teknoloji ile birlikte yapılandırılmamış verinin analizinde oldukça ilerleme sağlanmıştır. Ancak sosyal medya verisinin elde edilmesinde ve analizinde hala bazı zorluklar bulunmaktadır. Bu zorluklar aşağıda iki başlık altında özetlenmiştir.

2.2.1. Araştırma Zorlukları

Sosyal medya akademik ve ticari açıdan oldukça zengin bir kaynak sağlamaktadır. Ancak gerçekleştirilen analiz sürecinde çeşitli zorluklar bulunur. Aşağıda bahsedilen zorluklar özetlenmektedir (Batrinca ve Treleaven, 2015).

a. ***Veri Kazıma:*** Sosyal medya verileri erişilebilir durumda olmasına rağmen ticari değerleri dolayısıyla ham verilere erişim giderek zorlaşmaktadır. Çok az sayıda veri sağlayıcı veri kaynağına uygun fiyatlı yada ücretsiz erişim sunmaktadır. Örneğin; Twitter kendi API'si aracılığıyla herkese açık olarak paylaşılan Tweet'lere erişimi belirli sınırlar dahilinde mümkün kılmaktadır.
b. ***Veri Temizliği:*** Yapılandırılmamış verilerin temizlenme aşaması oldukça fazla zaman alabilmektedir. Özellikle gerçek zamanlı erişilebilen veri kaynaklarının barındırdığı temizleme dolayısıyla analiz zorlukları devam etmektedir. Günümüzde özellikle bazı sektörler de gerçek zamanlı verinin analizi oldukça önem taşımaktadır. Bu sebeple sosyal medya verilerinde gerçek zamanlı veri temizleme ve analiz süreci kritiktir.
c. ***Bütünsel Veri Kaynakları:*** Veri kaynaklarının çeşitliliğinin artması ile birlikte farklı kaynaklardaki verilerin bir araya getirilerek analiz edilmesi ihtiyacı doğmuştur. Bu da farklı kaynakların bir araya getirilme zorluğunu ortaya cıkarmaktadır. Örneğin; Bir e-ticaret firmasına sosyal medya da yapılan yorumlardaki duygu durumu dağılımı ile satışlar ve mevsimsellik arasındaki korelasyon incelenmek istendiğinde, yapısal olan ve yapısal olmayan veri kaynaklarının bir araya getirilmesi gerekmektedir.
d. ***Elde Edilen Verinin Korunması:*** Çeşitli kaynaklardan alınan verilerin tek bir kaynak altında birleştirilmesi aşamasından sonra verilerin güvence altına alınarak saklanması gerekir. Saklanan veri kaynağında kullanıcılara farklı yetkilendirme seviyeleri verilmez ise veri farklı kişiler tarafından elde edilebilir.
e. ***Veri Analitiği:*** Sosyal medyadan elde edilen veri içerisinde bulunan yabancı kelimeler, argo kullanımlar, yazım hataları vb. içerikler veri analitiği sürecini zorlaştırmaktadır.
f. ***Analitik Gösterge Panoları:*** Sosyal medya platformları kullanıcıların veri kaynağına erişmesi için çeşitli API'ler yazmasını gerektirir. Buda kullanıcıların bilgisayar

bilimlerinde yetkin olmalarını gerektirir. Araştırmacılar için programlama gerektirmeyen arayüzler gereklidir. Örneğin; Verinin çekilmesi, ön işleme adımlarının gerçekleştirilmesi ve verinin modellenmesi aşamasında kullanıcıyı yönlendiren bir sanal asistan arayüzü.

g. ***Verinin Görselleştirilmesi:*** Elde edilen büyük miktarda verinin görselleştirilerek içerisindeki anlamların ortaya çıkartılması veri kaynağının yapılandırılmamış olması sebebiyle zorluklardan biridir. Günümüzde sürekli olarak büyüyen veri kaynakları düşünüldüğünde görselleştirme yöntemlerinin önemi gitgide artmaktadır.

Sosyal medya analitiği ilk aşamasından son aşamasına kadar tüm süreçte zorluklar barındırmaktadır. Teknolojik gelişim ve veri kaynaklarının artan önemi nedeniyle bu zorlukların gitgide daha kolay aşılacağı söylenebilir.

2.2.2. Analiz Süreci Zorlukları

Sosyal medya verilerinin analizi sürecinde verinin temel yapısı ve toplanma şekli sebebiyle bazı zorluklar bulunmaktadır. Bu zorluklar aşağıda kısaca özetlenmeye çalışılmıştır. (Lee, 2017)

- Sosyal medya platformları yalnızca belirli bir müşteri grubuna erişebilmektedir. Bu nedenle sosyal medya verileri genel kitlenin tahmin edilmesini engelleyen, temsili olmayan bir örneklemi baz alabilir. Ayrıca verilerini paylaşan kullanıcılar düşündükleri ve yaptıklarının yalnızca bir kısmını paylaşmış olabilirler buda verinin eksiksiz ve doğru olarak kabul edilmesini zorlaştırır.
- Sosyal medya analitiğinin yeni bir disiplin olması sebebiyle, veriyi değerlendirirken iyi sosyal medya metriklerini seçmek ve analizi doğru şekilde ilerletmek uzmanlık ve deneyim eksikliği nedeniyle zordur.
- Sosyal medya verilerinde spam, sahte incelemeler, yanlış hesaplar ve yinelenen içerik dahil olmak üzere birçok gürültü kaynağı vardır. Verileri filtreleme ve temizleme, sosyal medya analizlerinin kalitesi için önemlidir. Buna rağmen,mevcut sosyal medya analizleri, sınırlı otomatik filtreleme ve temizleme yetenekleri sunar.
- E-posta, haber makaleleri, kişisel bloglar, resimler, ses ve videolar dahil olmak üzere birçok sosyal medya içeriği yapılandırılmamıştır. Yapılandırılmamış verilerin analiz

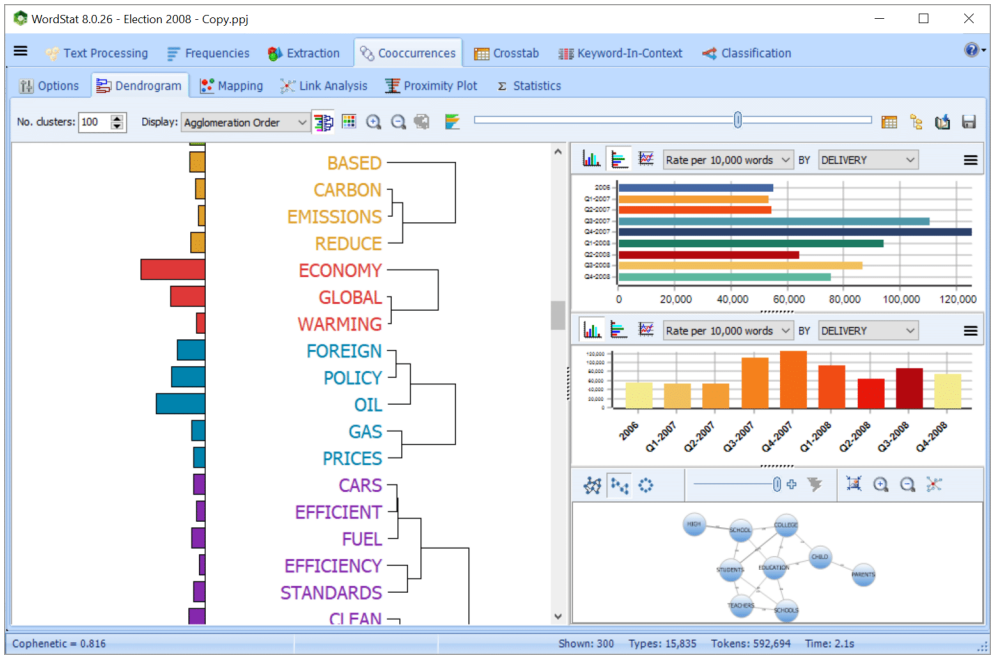
edilmesi bireysel faydanın yanı sıra ticari olarakta büyük kazançlar sağlayacaktır ancak işlenmesi oldukça zordur.

2.3. Sosyal Medya Analiz Araçları

Sosyal medya verilerinin özetlenmesi ve amaca uygun olarak gerekli bilgilerin elde edilmesi sürecinde çeşitli otomatik araçlardan destek alınabilir. Aşağıda bu araçlar özetlenmektedir (Stieglitz ve Dang-Xuan, 2012). Burada dikkat edilmesi gereken nokta sosyal medya verilerinin içeriğinin ve kaynaklarının farklılığı nedeniyle standart bir analiz sürecinin bulunmamasıdır. Programlama bilen kişiler analiz için kendi özel süreçlerini kodlama ile aşağıdaki programlara entegre edebilirler.

2.3.1. WordStat

Metin madenciliği sürecinin ortak adımlarının bazı paket programlar aracılığıyla yürütülmesi mümkündür. Bu paket programlara örnek olarak; WordStat, LIWC, General Inquirer, vb verilebilir. Wordstat temel ve ileri düzeyde metin madenciliği uygulamalarını özellikle sosyal bilimciler başta olmak üzere kamu ve özel kurumların kullanabileceği düzeyde bir paket program olarak sunmaktadır (Akbıyık, 2019).

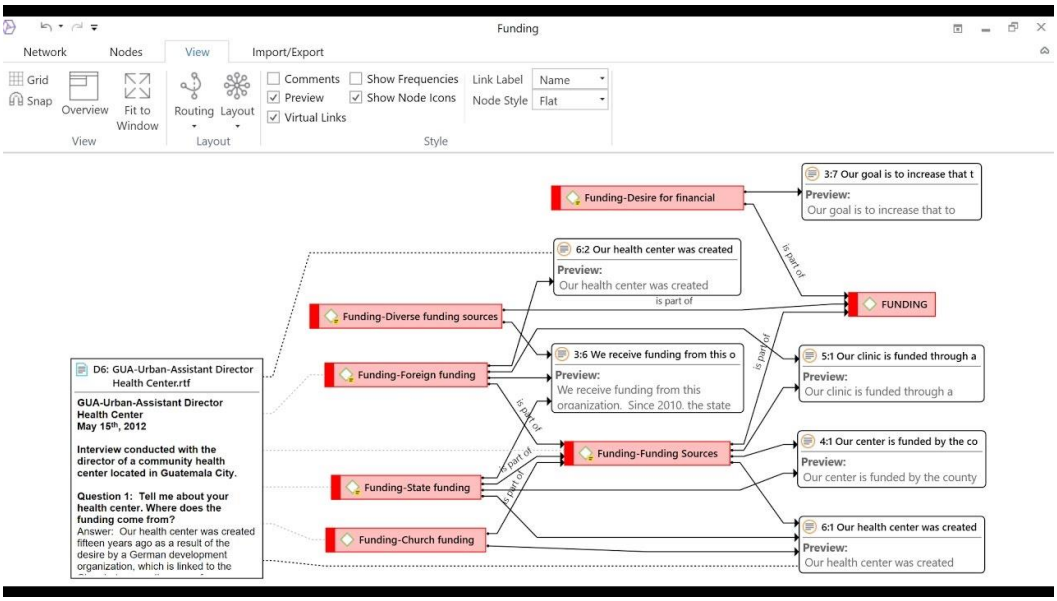


Şekil 22: Wordstat Program Arayüzü

Wordstat kural tabanlı bir analiz süreci sağlar. Kullanıcı metin madenciliği süreçlerini oromatik olarak gerçekleştirebilir ancak yönlendirme yapılmamaktadır.

2.3.2. Atlas.ti

Sosyal medya analiz sürecinde veri kaynağının özgünlüğü sebebi ile otomatik analiz yapılmasını sağlayan programlar yetersiz kalabilmektedir. Sürecin manuel olarak yürütülebileceği bazı programlar bulunmaktadır. Bu programlara örnek olarak ATLAS.ti, QDAMiner, The Ethnograph, vb. verilebilir. Manuel analiz sürecinin avantajı veriye özgü çözümler sunmasıdır ancak kodlama bilgisi gerektirir.



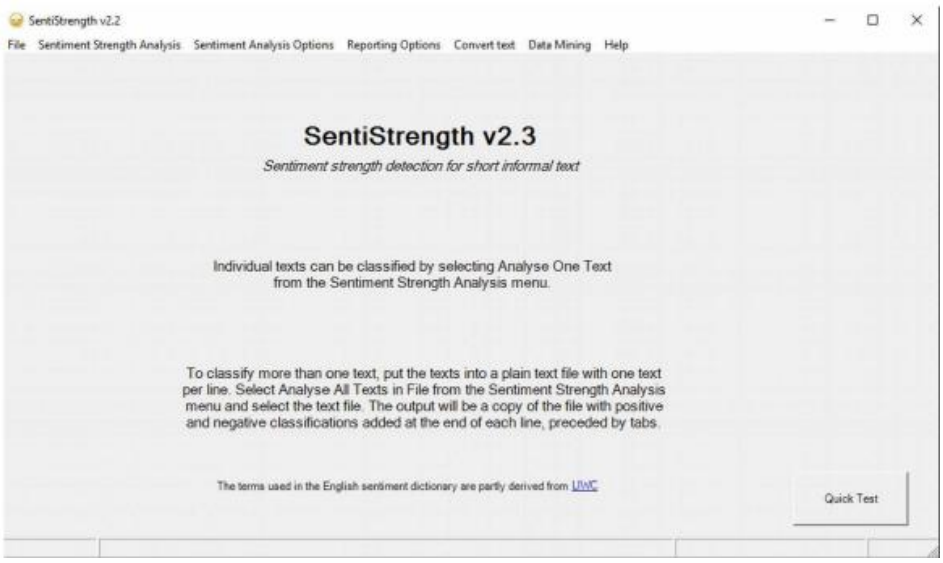
Şekil 23: Atlas.ti Program Arayüzü

Atlas.ti içerik analizi programı metin, ses, grafik, video verilerinin analizi için bir arayüz sunmaktadır. İçerik analizleri sonucunda veriler arası ilişkilerin yorumlanıp rapor haline getirilmesi için bilgisayar desteği sunmaktadır.

2.3.3. SentiStrength

Sosyal medya analitiğinde temiz veri üzerinde çeşitli analizler uygulanabilir. Bu analizlerden biride duygu analizidir. Duygu analizi / fikir madenciliği süreçlerinin otomatik olarak gerçekleştirilmesini sağlayan araçlara örnek olarak; SentiStrength, PolArt, SentiWordNet

verilebilir. SentiWordNet fikir madenciliği için sözlük yapısında bir kaynaktır. Kelimelerin pozitif, negatif ve nötr olarak sınıflandırılmasını sağlar (Esuli, 2021).

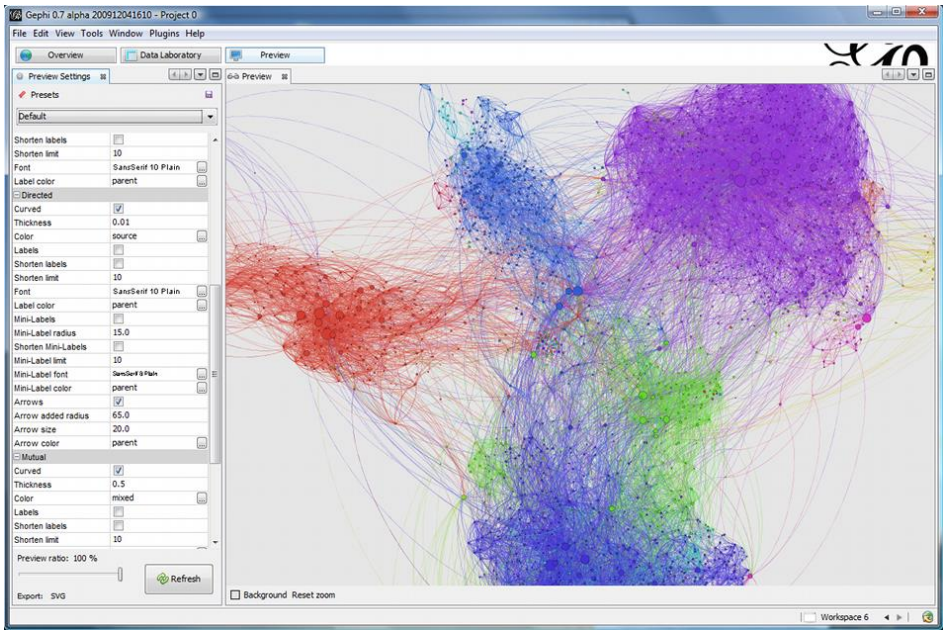


Şekil 24: SentiStrength Program Arayüzü

SentiStrength akademik araştırmalar için ücretsiz olarak kullanılabilen duygu analizi arayüzüdür. Bir veya daha fazla metni sınıflandırmasını sağlamak için her satırda bir metin olacak şekilde metinlerin programa yüklenmesi gerekir. Çıktıda her metnin yanına pozitif ve negatif sonuç etiketleri eklenir. Negatif etiket -1 ile -5 arasında, pozitif etiket ise 1 ile 5 arasında verilmektedir.

2.3.4. Gephi

Sosyal ağlardaki katılımcılar arasındaki etkileşimlerin analiz edilmesi sayesinde ticari ve akademik olarak çok çeşitli ve yararlı bilgi çıkarımları elde etmek mümkündür. Bir ağdaki en etkili kişi, kişiler arası roller, bilginin yayılma şekli vb konular sosyal ağ analizi ile incelenebilmektedir (Biber, 2020). Sosyal ağ analizinde kullanılan programlara örnek olarak Gephi, UCINET, Pajek verilebilir. Aşağıda açık kaynaklı sosyal ağ analizi programı olan Gephi'nin görselleştirme sürecine örnek resim paylaşılmıştır.



Şekil 25: Gephi Program Arayüzü

Yukarıdaki tüm özellikler üzerinden değerlendirme yapıldığında sosyal medya verilerinin analizinin her aşamasında özel yaklaşımlar gerekmektedir. Bu bağlamda toplanan veri üzerinde analize uygun metin madenciliği işlemlerinin gerçekleştirilmesi ve sonrasında veri içeriğinin değerlendirilmesini ve sınıflandırılmasını sağlayan duygu analizi ve topic modelleme gibi başlıklar ön plana çıkmaktadır.

BÖLÜM 3: ÇALIŞMANIN KAPSAMI VE LİTERATÜR İNCELEMESİ

Sosyal medya topluluklar arasında konuşma, paylaşılabilen içerik oluşturma ve yayma işlemlerini ifade etmektedir. Sosyal medya, geleneksel ve endüstriyel medyadan farklı olarak yazarlık ve okurluk arasındaki sınırları yıkmış, temelde bilgi tüketim ve yayma sürecinin bilgi üretme ve paylaşma süreci ile iç içe geçmesini sağlamıştır (Zeng, Chen, Lusch ve Li, 2010). Geleneksel süreç incelendiğinde firmalar tüketicilerin markalarını nasıl algıladıklarını daha iyi anlamak için görüşmeler, odak grupları ve anketler gibi maliyetli ve zaman alıcı pazar araştırması yöntemleri kullanmak zorundaydı (Moe, Netzer ve Schweidel, 2017). Şuan ise tüketiciler gönüllü olarak sosyal medyaya yönelerek fikirlerini hem müşteriler hem de marka yöneticilerinin erişimine sunarak halka açık olarak paylaşmaktadır. Bu da firmalara pazar araştırması için önemli bir kaynak oluşturmaktadır. Ek olarak kamu yönetimi, endüstri, finans, turizm, eğitim ve sağlık gibi alanlarda da sosyal medya platformalarında paylaşılan kullanıcı görüşleri, süreçlerin iyileştirilmesi ve karar desteği için kullanılabilmektedir (Rathore, Kar ve Ilavarasan, 2017). Sosyal medya analitiği çeşitli sektörler için bahsedilen faydalarının yanında doğası gereği belirli zorluklar barındırmaktadır. Verilerin elde edilmesi adımından başlayan bu zorluklar sonuçların yorumlanması aşamasına kadar devam edebilmektedir.

Sosyal medya verileri içeriklerinde yapılandırılmış ve yapılandırılmamış veri yapısı bulundururlar. Yapılandırılmış veriler kullanıcı profil özellikleri, mekan ve zaman vb. bilgilerdir. Yapılandırılmamış veriler ise kullanıcı tarafından oluşturulan metin, ses, video olarak düşünülebilir. Sosyal medya platformları her iki yapıda da çok büyük miktarda veri üretilirler ve oldukça dinamik ve karmaşık bir yapıya sahiptirler. Bu sebeple geleneksel veritabanı yönetim araçları, veri işleme uygulamaları ve görselleştirme paketleri kullanılarak kolayca işlenemezler. (Stieglitz, Dang-Xuan, Bruns ve Neuberger, 2014)

Sosyal medya platformlarında bulunan veri, ilgili konu hakkında bir çok insanın görüşünü barındırmaktadır. Bu açıdan önemli bir veri kaynağı olsa da verilerin doğruluğu ve tarafsızlığı tartışma konusudur. İnternet kaynaklarından gelen büyük veri kümeleri, özellikle birden fazla veri kümesi birlikte kullanıldığında potansiyel eksiklikleri ve tutarsızlıkları nedeniyle genellikle güvenilmezdir (Boyd ve Crawford, 2012). Ayrıca içerikler belirli bir alanda genel görüşten ziyade içerik oluşturan kullanıcıların öznel görüşünü yansıtacak

şekildedir ve paylaşım ağ eksenli olması sebebiyle sosyal ilişkilere bağlıdır. Sosyal medya içerikleri tutarlı olsa dahi ilgili zamanda trend olan alanlar ile alakalı olduğu için etkileri çoğu zaman geçicidir (Zafarani, Abbasi ve Liu, 2014).

Sosyal medya platformlarında bulunan karakter sınırlamaları sebebiyle sosyal medya verileri diğer metin verilerine göre daha kısadır. Genellikle günlük konuşma dilinden oluşması sebebiyle devrik cümle kullanımı ve kısaltmaları da yaygın olarak içermektedir. Günlük konuşma dilinde kullanılan argo, alay ve ironi bildiren ifadeler de kelimelerin birincil manasına çok daha farklı anlamlar yükleyebildiği için analiz sonucunu doğrudan etkilemektedir (Pradhan, Vala ve Balani, 2016). Ek olarak farklı dillerin beraber kullanılması sebebiyle düz metinlere göre daha fazla gürültü barındırırlar.

Sosyal medya verileri emojiler, linkler, resimler, gramer kurallarına uymayarak yazılmış ifadeler, tekrarlı ifadeler ve karakterler gibi sorunlardan dolayı yüksek gürültülü veri olarak değerlendirilebilirler. Bu şekilde gürültülü verilerin, analiz öncesinde temizleme ve düzeltme işlemlerinin doğru şekilde gerçekleştirilmesi sonuçların doğruluğu açısından kritik önem taşımaktadır. Bahsedildiği gibi doğası gereği bir çok zorluk barındıran sosyal medya verisi temizleme aşaması, teknik olmayan araştırmacılar için daha zor olmaktadır. Bu zorlukları aşmak için manuel olarak temizleme işlemi gerçekleştiren araştırmacılar zaman ve maliyet açısından oldukça zorlanmaktadır. (Susenoa, Laurell ve Sick, 2018)

Sosyal medya analitiği bahsedilen tüm zorlukların yanı sıra Türkçe için ek zorluklar barındırmaktadır. Özellikle Türkçe dilinin sondan eklemeli ve zengin biçimsel yapısı analizlerin başarı durumunu etkilemektedir (Çetin ve Eryiğit, 2018). Analiz sürecinde gerçekleştirilen cümle ve kelime bölümleme işlemleri, Türkçe metinlerde yer alan öğelerin cümle içerisindeki dizilimlerinin farklı olması sebebiyle analiz sürecini zorlaştırmaktadır. Bu zorluk zaten biçimsel olarak problemli ve oldukça gürültülü olan sosyal medya mesajlarında daha da ön plana çıkmaktadır. Tüm zorluklar değerlendirildiğinde sosyal medya analitiğinin her aşamasında özel yaklaşımlar gerektiği görülmektedir. Bahsedilen özel yaklaşımlar araştırmacıları her adımda farklı programlar kullanmaya itmektedir. Aşağıdaki tabloda literatürde analiz süreçleri farklı programlar aracılığıyla gerçekleştirilen çalışmalara yer verilmiştir.

Tablo 3: Sosyal Medya Analitiği Literatür İncelemesi

	Yazar (Yıl)	Alan	Açıklama	Yöntem	Döküman Kaynağı
1	Wang, Ye (2017)	Toplum	Doğal afetlerin yönetilmesi için sosyal medya analizi	Topik Modelleme, Duygu Analizi	Twitter, Flickr, Facebook, Weibo
2	Stieglitz, Dang-Xuan (2012)	Siyaset	Siyasal iletişimde sosyal medya analitiğinin kullanılması	Topik Modelleme, Duygu Analizi	Twitter, Facebook
3	Thelwall (2017)		Youtube video yorumları üzerinden sosyal medya analitiği stratejisinin tanıtılması	Duygu Analizi	Youtube
4	Suseno, Laurell, Sick (2018)		Storytel kullanıcı etkileşimleri üzerinden sosyal medya analiği ile değer yaratma süreci incelemesi	Topik Modelleme, Duygu Analizi	Storytel
5	He, Tian, Tao, Zhang, Yan, Akula (2017)	Turizm	Otel yorumları üzerinde sosyal medya analiği uygulaması	Duygu Analizi	Tripadvisor.com
6	He, Shen, Tian, Li, Akula, Yan, Tao (2015)	Perakende	Walmart ve Costco perakende zincirleri üzerinden sosyal medya analitiği ile rekabetin incelenmesi	Duygu Analizi, Topik Modelleme	Twitter

Literatürde incelenen sosyal medya analitiği çalışmalarının her birinin farklı analiz programı bağımlılıkları bulunmaktadır. Doğal afetlerin yönetilmesi için yapılan çalışma içerisinde Twitter'dan çekilen veri bir program arayüzü aracılığıyla elde edilmiştir. Siyasal iletişimde sosyal medya analitiğinin kullanıldığı çalışmada, Twitter ve Facebook üzerinden veri elde etmek için programların kendi API'leri kullanılmıştır. Bu durumda sınırlı sayıda veriye erişim sağladığı için analizi sınırlandırmaktadır. Ayrıca bu çalışmada veri ön işleme adımında yalnızca durdurma kelimelerinin temizlenmesi ve lemmatization işlemleri gerçekleştirilmiştir. Twitter gibi kullanıcıların kısaltma,emoji, noktalama işaretleri ve sayıların metin içerisinde kullanıldığı platformlarda daha detaylı ön işleme adımları gerçekleştirilmelidir. Ek olarak kelime bulutu oluşturmak için Wordle.net adı verilen yazılım, sosyal ağ analizi içinse Gephi yazılımı kullanılarak görselleştirme yapılmıştır. Tek bir analiz sürecinin her adımının farklı programlar üzerinden ilerletilmesi araştırmacılar için temel zorluklardan biridir.

Youtube yorumları üzerinden yapılan duygu analizi çalışmasında otomatik duygu analizi işlemleri gerçekleştiren SentiStrength programı kullanılmıştır. Benzer bir çalışma olan storytel kullanıcı etkileşimleri çalışmasında ise analizin en temel adımlarından olan veri ön işleme adımında manuel ilerlenerek istenmeyen veriler kaldırılmıştır. Otel yorumları üzerinde yapılan sosyal medya analitiği çalışmasında ise veri ön işleme adımları için Nvivo yazılımı ve R programlama dili ile ilerlenmiştir. Duygu analizi için ise Google Prediction API kullanılmıştır. Perakende alanında yapılan bir diğer çalışmada Walmart ve Costco ile ilgili Tweet'ler API üzerinden çekilerek rekabet açısından analiz edilmiştir. Veri kümesindeki her bir tweet'in duygularının tespiti için Lexalytics adlı duygu analiz aracı kullanılmıştır. Ayrıca ürün bazında müşteri dağılımının analiz edilmesi ve ilgili tweet'lerin çıkartılması için Leximancer adında metin madenciliği aracı kullanılmıştır.

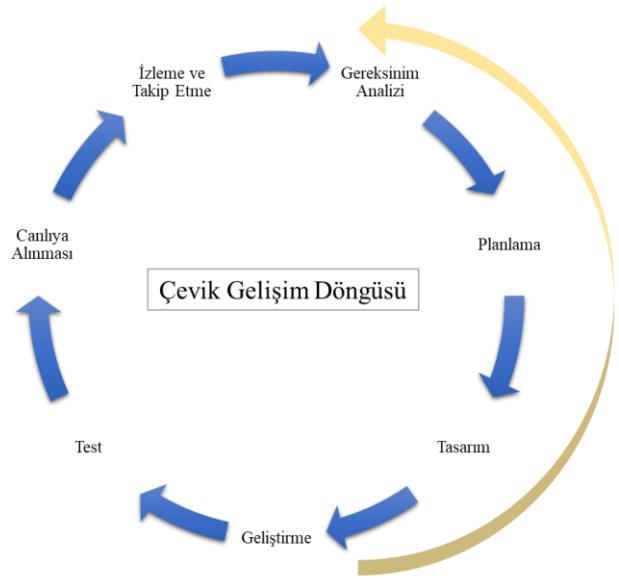
Sosyal medya analitiği üzerine görüldüğü gibi farklı alanlarda yapılan birçok çalışma bulunmaktadır. Bu çalışmaların genel çerçevesi analiz ihtiyacına uygun olarak verinin elde edilmesi ile başlayarak, verinin temizlenmesi ve analizi ile devam etmektedir. Örnek çalışmalarda görüldüğü gibi genel olarak veri toplama aşamasında ilgili programın API'sinin sınırlarına bağımlı kalınmaktadır. Bu da istenilen sayıda ve içerikte veriye erişimi sınırlamaktadır. Veri temizleme aşamasında ise hazır programlar yada açık kaynak

programlama dilleri kullanılmaktadır. Hazır program arayüzleri yüksek maliyetin yanı sıra özelleştirilmiş analiz adımları için ek kodlama bilgisi gerektirir. Ek olarak programlarda kural tabanlı ilerlenmektedir. Analizi yapan kullanıcı analiz hakkında yönlendirme almaz, analiz sürecini bildiği varsayılarak standart adımlarla analizi tamamlaması sağlanır. Tüm bu zorluklar düşünüldüğünde araştırmacılara analiz aşamalarının her birinde ihtiyaç duyduğu yönlendirmelerin sağlanması ve tüm analiz sürecinin tek bir program arayüzü üzerinden gerçekleştirilebilmesi önem kazanmaktadır.

BÖLÜM 4: YÖNTEM

Veri kaynakları sosyal medya kullanımına ve çeşitliliğine bağlı olarak her geçen gün artmaktadır. Kullanıcıların ve toplulukların görüşlerini yansıtan bu kaynakların artması ve metin verilerinin sayısallaştırılarak analizinin mümkün hale gelmesi sosyal bilimler analiz süreçleri içinde önemli kaynak oluşturmaktadır. Literatür taraması bölümünde bahsedildiği gibi metin verisinin analiz sürecinde bir çok hazır program bulunmaktadır. Ancak bu programlar standart bir düzende olmaları sebebiyle kodlama bilmeyen kişiler tarafından esnek şekilde kullanılamamaktadır. Aynı zamanda veri toplama, ön işleme, analiz ve görselleştirme adımlarının hepsinin bir arada yapılarak analizin uçtan uca tamamlandığı bir program yapısı bulunmamaktadır. Her bir analiz aracı bir işlem adımını gerçekleştirerek diğer adımlar için farklı programlar kullanmayı zorunlu kılmaktadır. Ek olarak uzun ve karmaşık metin madenciliği ön işleme adımında araştırmacının daha önce benzer alanda analiz yapmış kullanıcıların tercihlerinden haberdar edilmesi analiz sürecini kolaylaştıracaktır. Bu çalışmada tasarlanan metin madenciliği sanal asistanı verinin toplanmasından analiz sürecine kadar olan tüm süreçte araştırmacıyı yönlendirerek analiz sürecini kolaylaştırmaktadır. Tüm analiz sürecinin bir arada yapılabilmesi, araştırmacıya önceki kullanıcılardan elde edilen bilgilerin sunulması ve analiz sürecinin öğretilerek ilerlenmesi sanal asistan tasarımının özgün yönlerindendir.

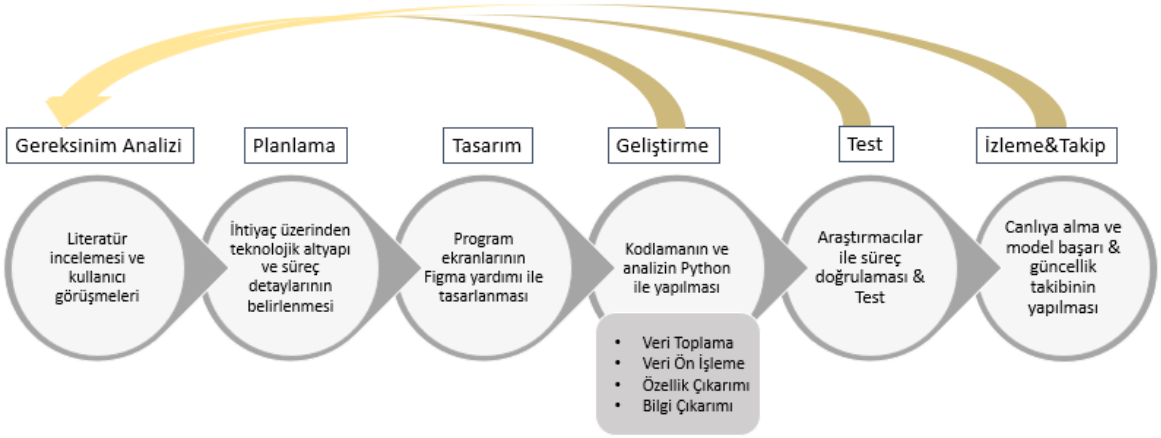
Sanal asistan süreci tasarlanırken çevik yazılım geliştirme metodolojisinden yararlanılmıştır. Çevik yöntemin kullanılmasının sebebi sürekli gelişen ve yeni kaynaklar eklenen sosyal medyanın sürecin temelini oluşturmasıdır. Çevik yöntem esneklik, sürekli iyileştirme ve hıza odaklanır. Kısa planlama döngüleri ile, süreç sırasında herhangi bir değişiklik yapmak ve kabul etmek kolaydır. Bu yöntemin kullanılması sayesinde sanal asistan tasarımına herhangi bir güncelleme ya da yenilik kolay şekilde adapta edilecektir. Ayrıca çevik süreçler projelerin tüm yaşam döngüsü boyunca kullanıcılardan sürekli geri bildirim almayı destekler. Bu sayede sanal asistanı araştırmacıların bildirimleriyle sürekli olarak beslemek mümkün hale gelecektir. Aşağıdaki şekilde çevik yöntem döngüsü gösterilmektedir.



Şekil 26: Çevik Geliştirme Döngüsü

Kaynak: Yoldash, R. (2018, 04 17). *Medium.* Agile (Çevik) Yazılım Geliştirme nedir ve nasıl uygulanır?: https://medium.com/@ryoldash/agile-%C3%A7evik-yaz%C4%B1l%C4%B1m-geli%C5%9Ftirme-nedir-ve-nas%C4%B1l-uygulan%C4%B1r-93e85ffc866 adresinden alındı

Bahsedildiği gibi çevik yöntemlerde programı kullanıcak olan araştırmacıların süreç gelişimine katılımı oldukça yüksektir. Bu şekilde gereksinimlerin doğru olarak belirlenmesi ve sürece adapte edilmesi kolaylaştırılmış olur. Ayrıca ortaya çıkan son ürünün kabul görmesi ihtimalini arttırır. Sanal asistan tasarımı için çevik proje yaklaşımı ile yürütülen süreç aşağıdaki şekilde özetlenmiştir.



Şekil 27: Sanal Asistan Tasarımı Çevik Geliştirme Süreci

Çevik yöntemler, sanal asistan tasarım sürecinde de görüldüğü gibi gereksinim belirleme ve planlama adımlarını atlamaz. Yalnızca bu adımları süreç içerisine yayarak her adımda gereksinimlerin belirlenmesini ve bu doğrultuda sürecin beslenmesini mümkün kılar.

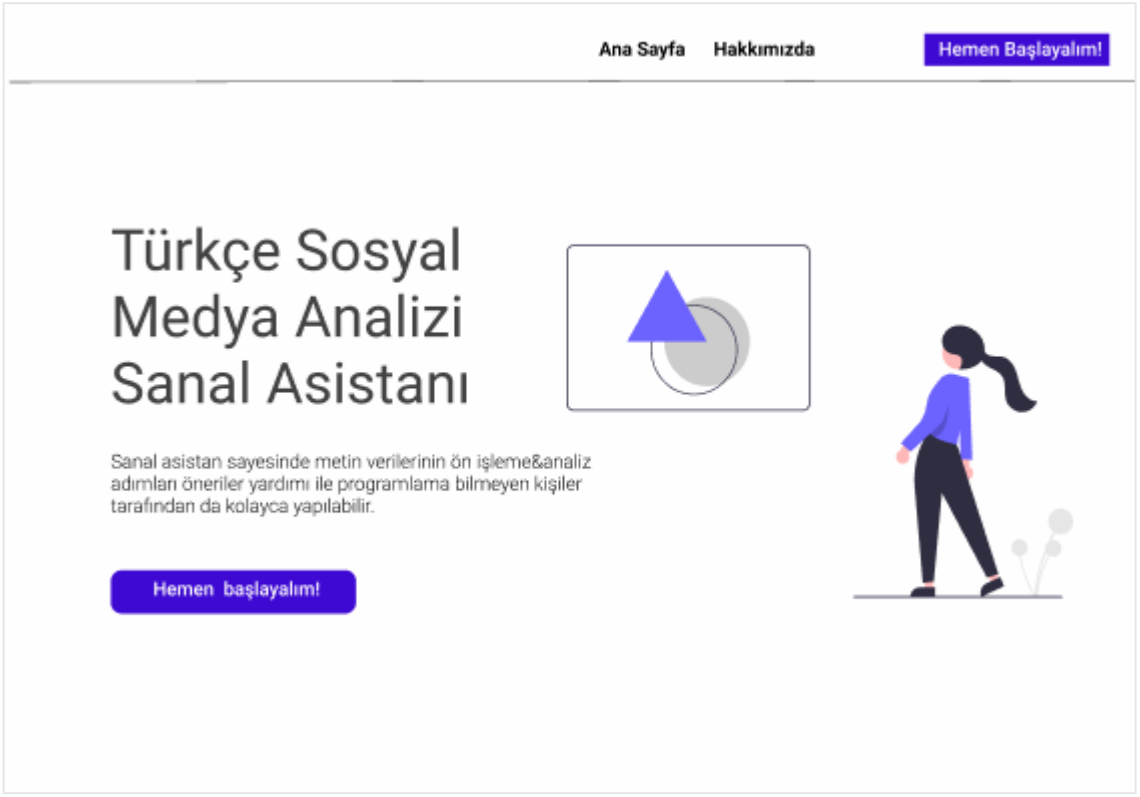
Gereksinim Analizi: Sosyal medya analitiği üzerine yapılan çalışmalar incelendiğinde tüm analiz sürecinin bir arada ilerletildiği ve yönlendirmeler içeren bir program arayüzü bulunmamaktadır. Bu noktada ilk ihtiyac sosyal bilimciler için yönlendirmeler ile tüm sürecin yürütüldüğü bir arayüz oluşturmaktır. Ayrıca yapılan çalışmalar incelendiğinde büyük bir çoğunluğunda Twitter API'sine bağımlı kalındığı görülmüştür. Kodlama bilgisi gerektiren ve Twitter'dan sınırsız veri çekilmesini sağlayan araçların bi arayüz üzerinden doğrudan kullanılabilmesi sosyal bilimciler için gereksinimlerden biridir.

Planlama: Sanal asistan olarak düşünülen sosyal medya analitiği arayüzü için veri toplama, ön işleme ve analiz adımlarında gerekli teknoloji ve alt yapının belirlendiği aşamadır. Öncelikli olarak tüm sürecin temelinde bulunan veri kaynağının seçilmesi gerekmektedir. Günümüzde yaygın olarak kullanılan Twitter kişilerin görüşlerine erişimi mümkün kıldığı için literatürde sosyal medya analizlerinde sıklıkla tercih edilmektedir. Veri içeriği ve erişim kolaylığı açısından sanal asistan tasarımı içinde Twitter veri kaynağı kullanılması planlanmıştır. İleride yapılacak geliştirmelerde ise Facebook ve Instagram gibi sosyal medya uygulamalarıda dahil edilebilir.

Twitter üzerinden veri toplanması için en uygun yöntemin açık kaynaklı programlama dili kullanılarak Twitter API'sinden bağımsız veri çekilmesi olduğu belirlenmiştir. Ayrıca kullanıcıya belirli formatta metin verisi yükleme seçeneğide sunulacaktır.

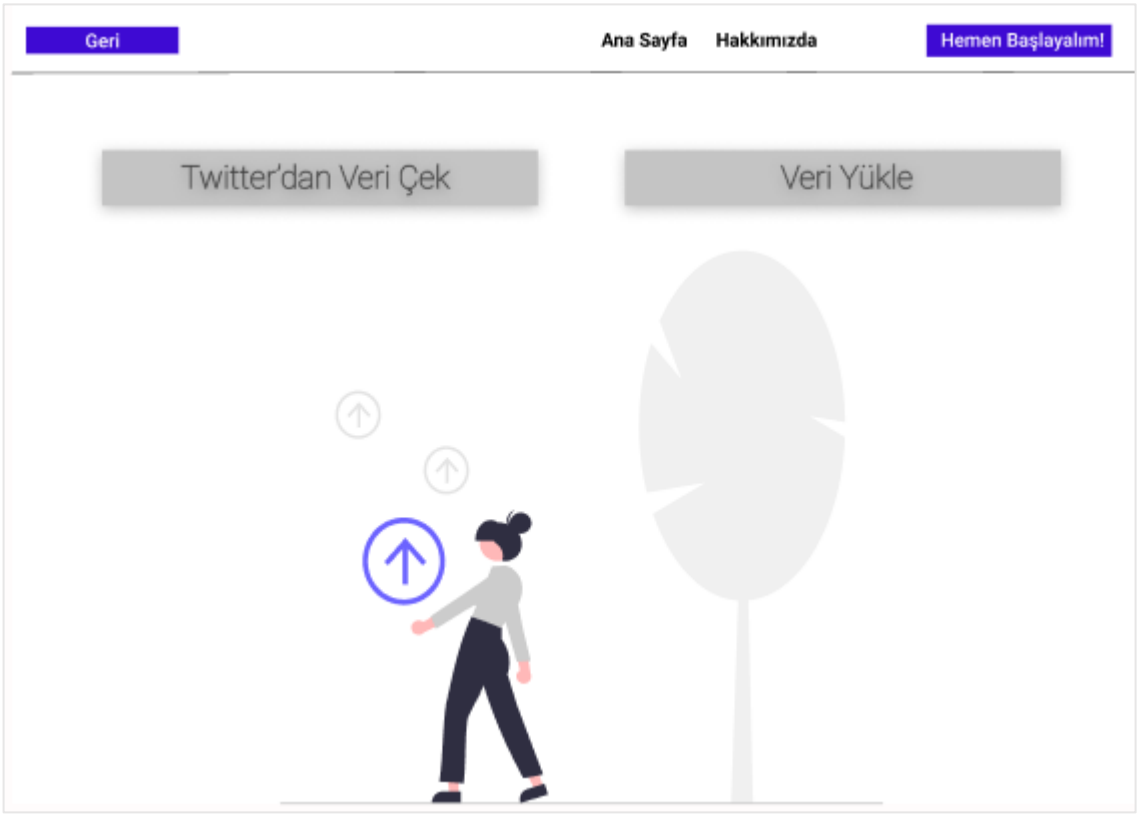
Veri toplama aşamasından sonra veri ön işleme adımları Python programlama dili kullanılarak gerçekleştirilecektir. Temizlenen veri üzerinde duygu analizi işlemleri ve topik modelleme yapılması planlanmaktadır. Duygu analizi için oldukça geniş bir veri seti ile egitilmiş Turkish-BERT modeli kullanılması planlanmıştır. Topik modelleme süreci içinse 7 farklı kategori barındıran etiketli veri seti üzerinden makine öğrenmesi modellerinin eğitilmesi sonucu en başarılı model kullanılacaktır.

Tasarım: Sanal asistan ön yüzü planlanan adımlara bağlı olarak gerçekleştirilmiştir. Burada ön yüz tasarlanırken figma tasarım programından yararlanılmıştır. Araştırmacıyı karşılayan sanal asistan giriş ekranı aşağıda görülmektedir.



Şekil 28: Sanal Asistan Giriş Ekranı Tasarımı

Araştırmacı “Hemen başlayalım!” butonuna tıkladıktan sonra veri çekme seçeneklerini içeren ekran ile karşılaşmaktadır. Burada Twitter’dan veri çekme ekranı ile yada uygun formatta veri yükleyerek analize devam edebilir.

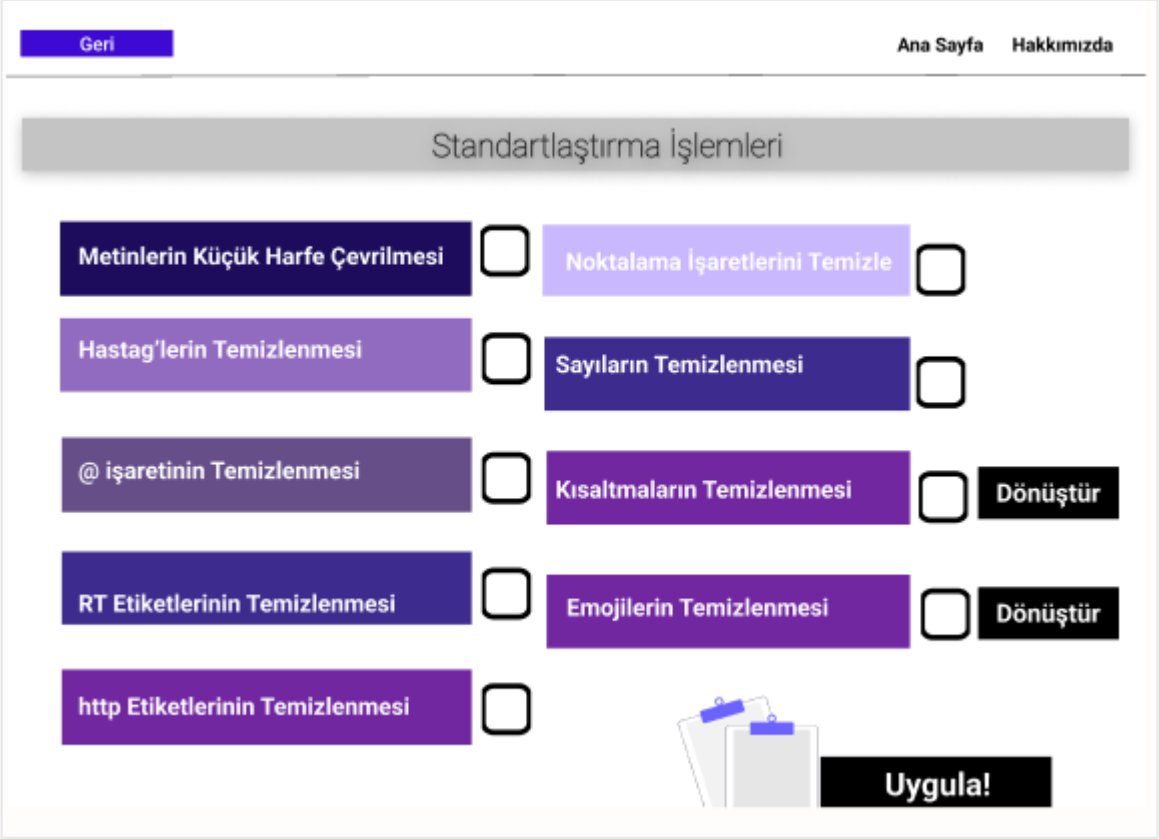


Şekil 29: Sanal Asistan Veri Çekme ve Yükleme Ekran Tasarımı

“Twitter’dan Veri Çek” butonu aracılığıyla ilerlendiğinde veri çekme ekranı aşağıdaki şekilde tasarlanmıştır. Araştırmacı hangi dilde, hangi tarihler arasında, hangi anahtar kelimeleri ve kaç adet çekmek istediği girişini yaparak veriyi elde edebilir. Burada çekilmek istenen veri etiketine göre bir günlük veri sayısı üzerinden kullanıcının yönlendirilmesi hedeflenmektedir. İlgili etiketin çok sayıda veri içermesi durumunda araştırmacının daha az veri çekmeye yönlendirilmesi gerekebilir.

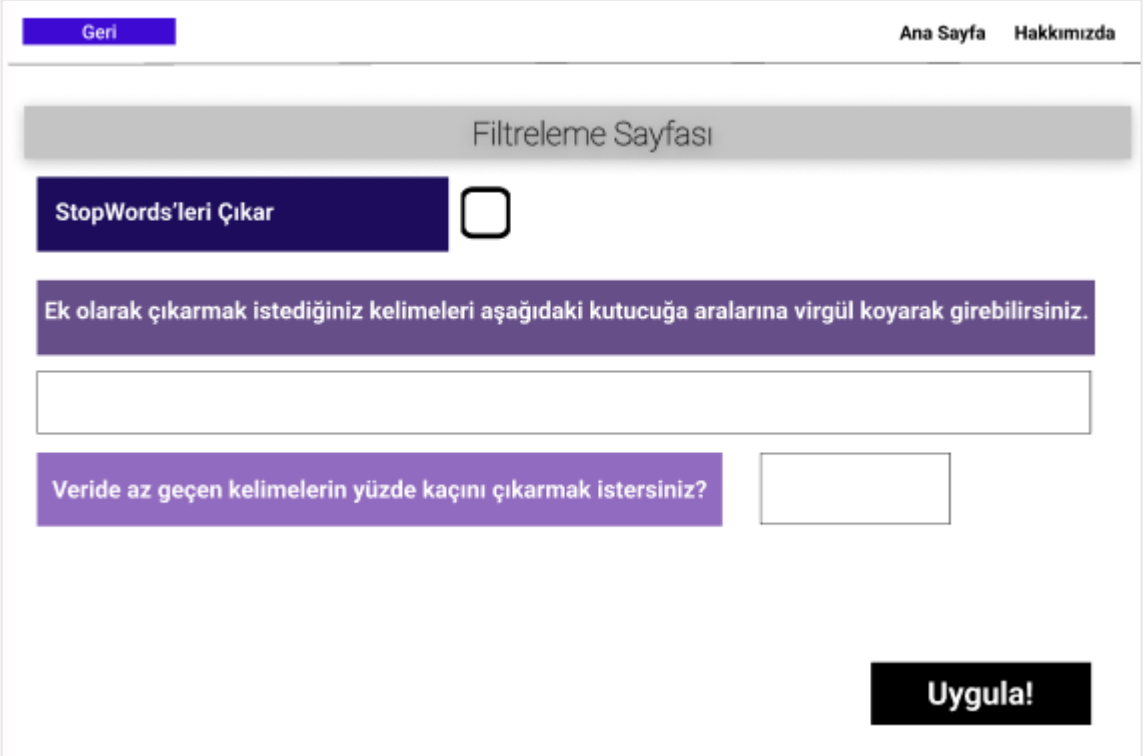
Şekil 30: Twitter'dan Veri Çekme Ekran Tasarımı

Veri çekme işleminin ardından verinin standartlaştırılması adımı gelmektedir. Bu adımda araştırmacı Twitter verisinde bulunan http, @, #, rt gibi kalıpları, noktalama işaretleri ve sayısal ifadeleri analiz sürecinin gerekliliklerine uygun olarak seçerek temizleyebilir. Ek olarak araştırmacıya verisinde oransal olarak kaç adet sayı, noktalama işareti ve diğer özel karakterlerden olduğu analiz esnasında bu ekranda gösterilmektedir. Araştırmacı bu sayede temizleme sonucunda veri boyutunun nasıl etkileneceği konusunda bilgi edinebilmektedir.



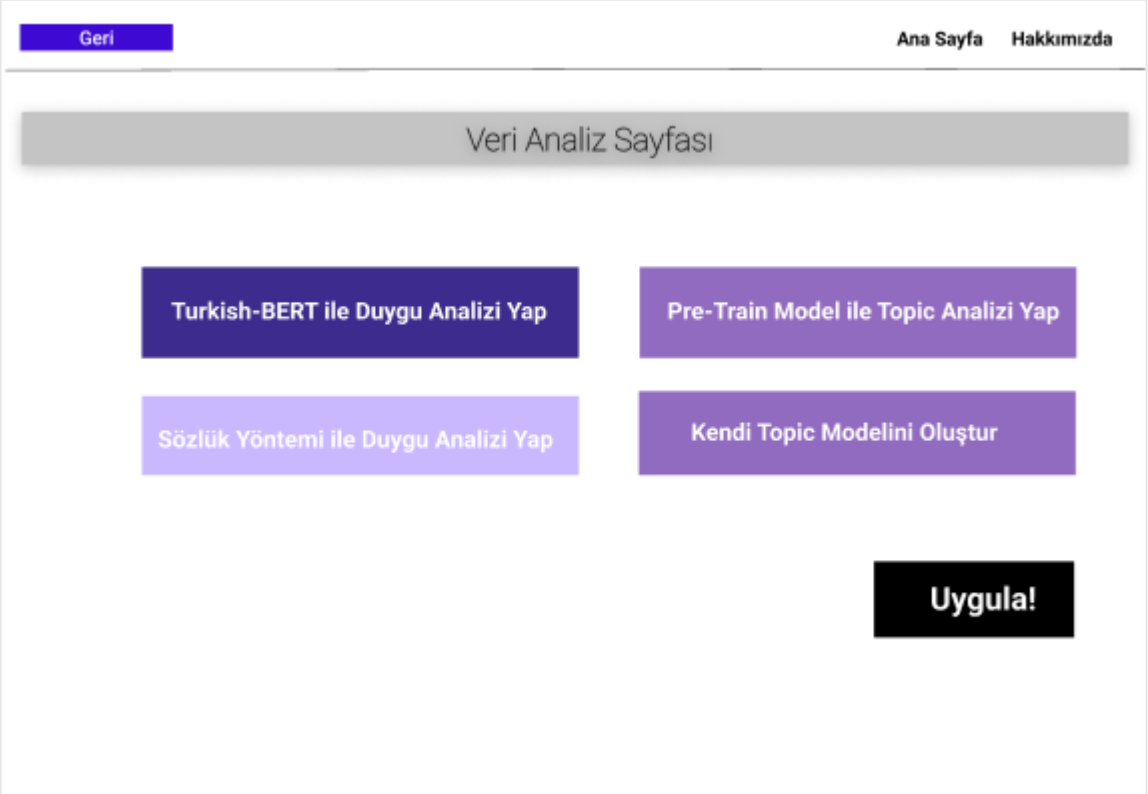
Şekil 31: Standartlaştırma Ekranı Ekran Tasarımı

Filtreleme ekranında ise veride bulunan durdurma kelimeleri ve araştırmacı tarafından temizlenmek istenen kelimelerin çıkartılması işlemi gerçekleştirilir. Bu ekrandan hemen önce bir veri örneği üzerinden kullanıcıya durdurma kelimelerinin bir listesi sunularak filtreleme hakkında bilgi verilmektedir. Ayrıca araştırmacı veride az geçen kelimeleri oransal olarak ya da sayısal olarak temizleyebilmektedir. Örneğin; Veride geçen kelimelerin %10'unu silinmesi veya veride az geçen son 100 kelimenin silinmesi. Bu aşamada araştırmacıya silmek istediği %10'unun sayısal olarak kaç kelimeyi ifade ettiği yada silmek istediği en az geçen son 100 kelimenin toplam verinin yüzde kaçına denk geldiği bilgisi verilmektedir. Kullanıcı bu doğrultuda karar alarak filtreleme işlemini gerçekleştirebilir.



Şekil 32: Filtreleme Sayfası Ekran Tasarımı

Son tasarım sayfası olan veri analizi adımı araştırmacının yapacağı analiz seçeneklerini barındırmaktadır. Burada kullanıcı analizlerin herhangi birinin üzerine tıkladığında öncelikle o analiz hakkında detaylı bilgi verilerek analiz süreci öğretilmektedir. Kendi topic modelini oluştur seceneğine tıkladığında araştırmacıdan etiketlenmiş eğitim verisini yüklemesi istenir. Yükleme ekranından sonra ise arka planda asistan yardımı ile eğittiği model üzerinden istediği veri üzerinde sınıflandırma işlemini gerçekleştirebilmektedir. Diğer seçeneklerde ise doğrudan tahminleme işlemi gerçekteştirilmektedir.



Şekil 33: Veri Analiz Sayfası Ekran Tasarımı

Geliştirme: Tasarım aşamasında oluşturulan ekranların kodlanma sürecidir. Bu çalışmada kodlama için açık kaynak kodlu programlama dili olan Python tercih edilmiştir. Python programlama dili yazılım geliştirme, altyapı yönetimi ve veri analizinde kullanılabilmektedir. Çalışma süreci yazılım geliştirme ve veri analizi adımlarını bir arada içermektedir. Aşağıda veri çekme aşamasında kullanılan Python kodu görülmektedir.

```
for i,tweet in enumerate(sntwitter.TwitterSearchScraper((anahtar_kelime_li
st[0] or

     anahtar_kelime_list[1]) +

     ' lang:{0} since:{1} until:{2}

     -filter:links -
filter:replies'.format(language,since_,until_)).get_items()):

     if i > data_row :

           break

     user = user + [tweet.username]

     date = date + [tweet.date]

     content = content + [tweet.content]

veri_seti = pd.DataFrame({"user":user, "date":date, "content":content})
```

Veri çekme aşamasında Python kütüphanesinde bulunan snscrape kütüphanesi kullanılmıştır. Bu kütüphane içerisinde bulunan TwitterSearchScraper fonksiyonu sayesinde istenilen tarih aralığı, istenilen anahtar kelimeler ve istenilen sayıda veri elde edilebilir. Veri çekme aşamasından sonra veri ön işleme adımları gelmektedir. Aşağıda ön işleme adımları için kod örneği paylaşılmıştır.

```
temiz_v1['content'] = veri_seti['content'].apply(lambda x: "
".join(x.lower() for x in x.split()))

temiz_v2['content'] = temiz_v1['content'].str.replace(r'#\S+', '')
temiz_v3['content'] = temiz_v2['content'].str.replace(r'@\S+', '')
temiz_v4['content'] = temiz_v3['content'].str.replace(r'http\S+', '')
temiz_v5['content'] = temiz_v4['content'].str.replace(r'[^\w\s]', '')
temiz_v6['content'] = temiz_v5['content'].str.replace(r'\d', '')
```

Python'da bulunan replace fonksiyonu metinler içerisinde değişiklik yapmak için kullanılmaktadır. Metinde bulunan temizlenmek istenilen ifadeler (@, http, #) yukarıda görüldüğü gibi temizlenmektedir.

Veri analiz seçeneklerinden biri olan Turkish-BERT modelinin çalışmaya dahil edilebilmesi için Python'da bulunan transformers kütüphanesinin yüklenmesi gerekmektedir. Aşağıda bu kütüphanenin yüklenme süreci ve model sürecinin oluşturulması görülmektedir.

```
from transformers import AutoModelForSequenceClassification, AutoTokenizer
,
pipeline

model = AutoModelForSequenceClassification.from_pretrained("savasy/bert-
base-turkish-sentiment-cased")

tokenizer = AutoTokenizer.from_pretrained("savasy/bert-base-turkish-
sentiment-cased")
```

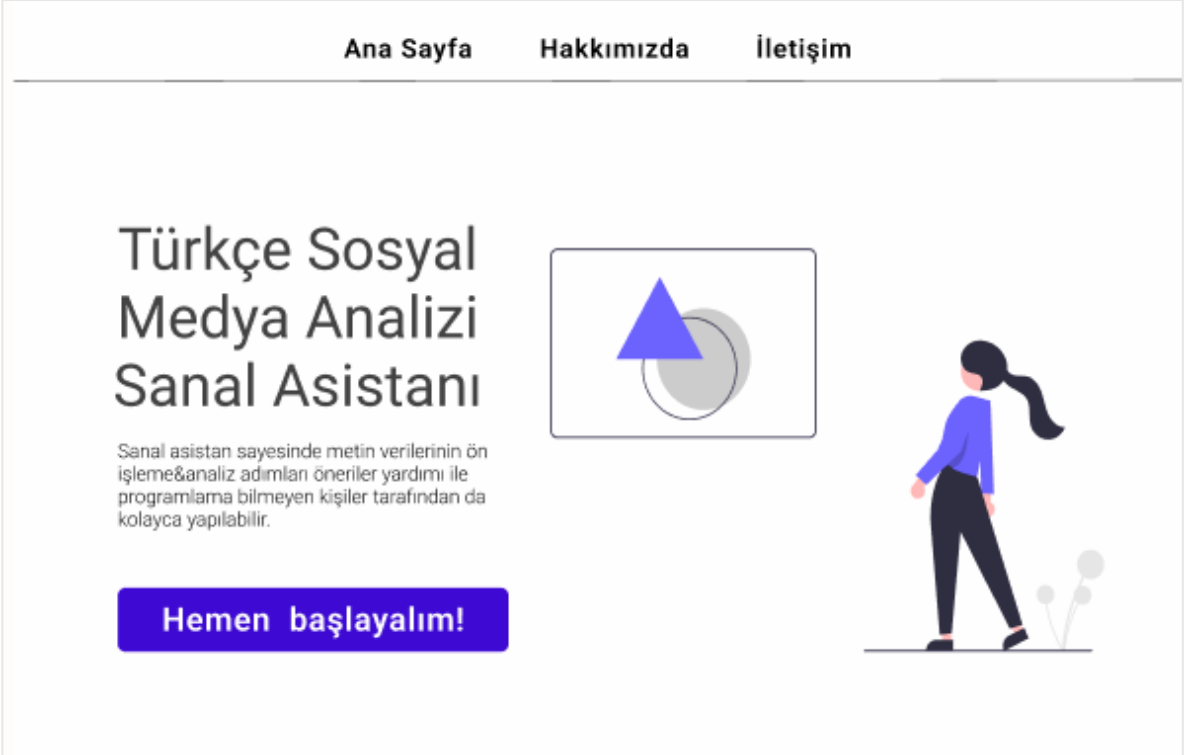
Veri ön işleme ve model sonuçlarının görselleştirilmesi için Python'ın seaborn ve matplotlib kütüphaneleri kullanılmıştır. Ayrıca figma ile tasarlanan arayüzler canlıya alınırken Python'da bulunan tkinter kütüphanesi kullanılmıştır. Bu kütüphane python veri analiz sonuçlarının basit şekilde kullanıcı arayüzüne dönüştürülmesini sağlamaktadır.

Test: Test aşaması ortaya çıkarılan ürünün müşteri ihtiyaçları ile kesişiminin ve farklarının incelendiği aşamadır. Sosyal bilimcilerden destek alınarak program arayüzünün farklı veri setleri ile test edilmesi sağlanmıştır. Bu aşamada çevik yöntemin esnekliğinden faydalanılarak kullanıcıların yönlendirilmesi doğrultusunda sürecin başına dönülerek gereksinim analizi, tasarım ve geliştirme süreçleri tekrarlanarak süreç iyileştirilmesi gerçekleştirilmiştir. Örneğin; araştırmacılar veride az geçen kelimeleri sayısal olarak cıkartırken bu sayının toplam verideki oranını merak ettiklerini belirtmişlerdir. Bu doğrultuda ilgili ekranlara bilgilendirme ekranları eklenmesi sağlanmıştır.

Canlıya Alma ve Takip: Proje sürecinin son adımlarından olan canlıya alma aşamasında hazırlanan ürün müşterilere teslim edilmektedir. Ancak müşteriye teslim sürecin sonlanmasını ifade etmemektedir. Müşteri kullanmaya başladığında tekrar ele alınması gereken detaylar ortaya çıkabilir. Takip aşamasında ise kullanıcılardan geri bildirim toplamanın yanı sıra, Turkish-BERT modeli ve önceden eğitilmiş olarak kullanılan topik modellemenin zaman içerisinde başarı oranı düşebilir. Bu modellerin izlenerek güncellenmesi ve gerektiğinde farklı modellerle değiştirilmesi sağlanacaktır.

BÖLÜM 5: UYGULAMA

Metin madenciliği süreçleri veri kaynağının içeriği ve yapılmak istenen analiz türüne göre oldukça uzun ve zahmetli olabilmektedir. Bahsedildiği gibi en temelde ve erişilmesi kolay olan Twitter verisi için ön işleme adımları genel olarak aynı süreçleri içermektedir. Bu bölümde sanal asistan yardımı ile Twitter'dan kullanıcının belirlediği özellikler doğrultusunda veri çekme işlemi, analiz ihtiyacına göre seçilen ön işleme adımları ve sonucunda temizlenen veri üzerinden yapılabilecek analiz süreçleri (duygu analizi, topik modelleme) bulunmaktadır. Aşağıda sanal asistan tasarımının giriş ekranı görülmektedir.



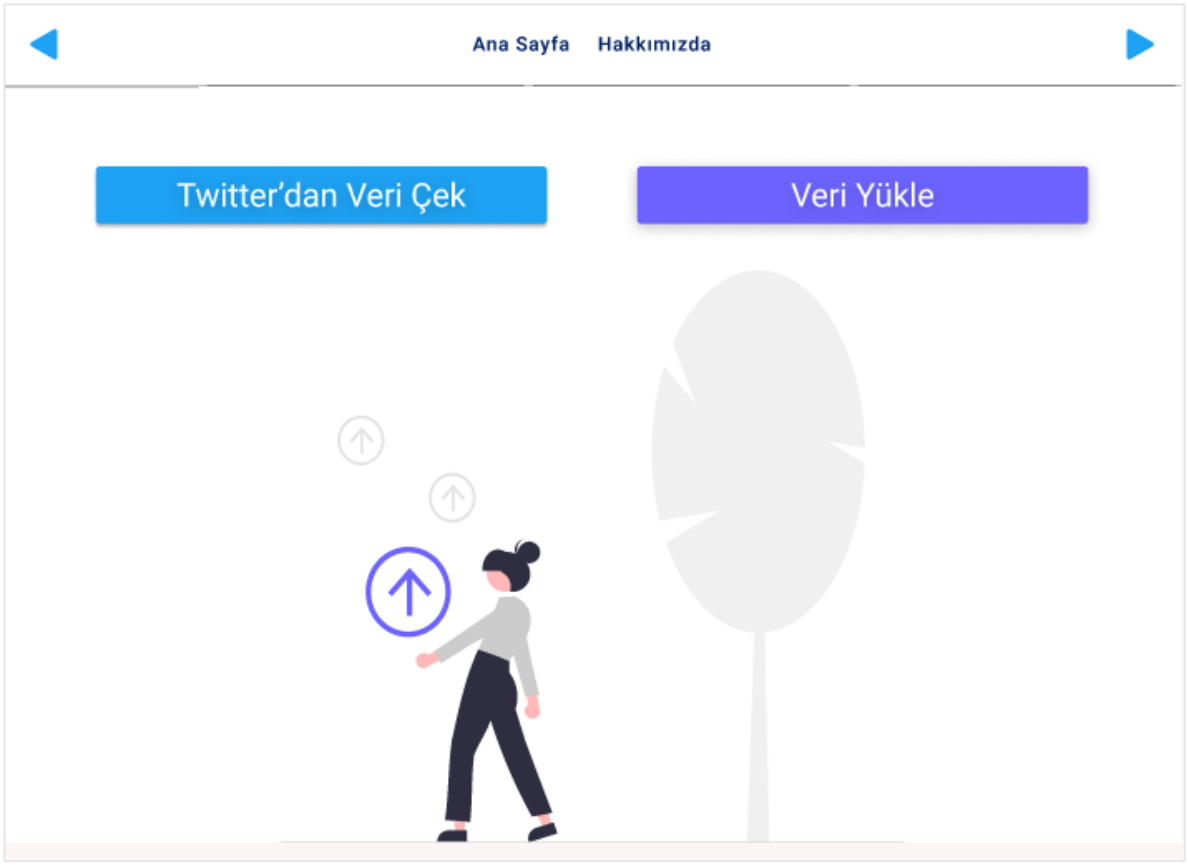
Şekil 34: Sanal Asistan Giriş Ekranı

5.1. Verinin Elde Edilmesi

Sosyal medya kaynaklarının artması ile birlikte çok farklı kaynaklardan metin verisine ulaşarak analiz yapmak mümkün hale gelmiştir. Facebook, Instagram, Twitter gibi sosyal medya platformlarını yanı sıra haber, blog ve eticaret siteleri örnek olarak verilebilir. Bu çalışma Twitter verisi üzerinden gerçekleştirilmiştir. Verinin elde edilmesi aşamasında açık kaynak kodlu programlama dili olan Python kullanılmıştır. Python ile Twitter üzerinden veri

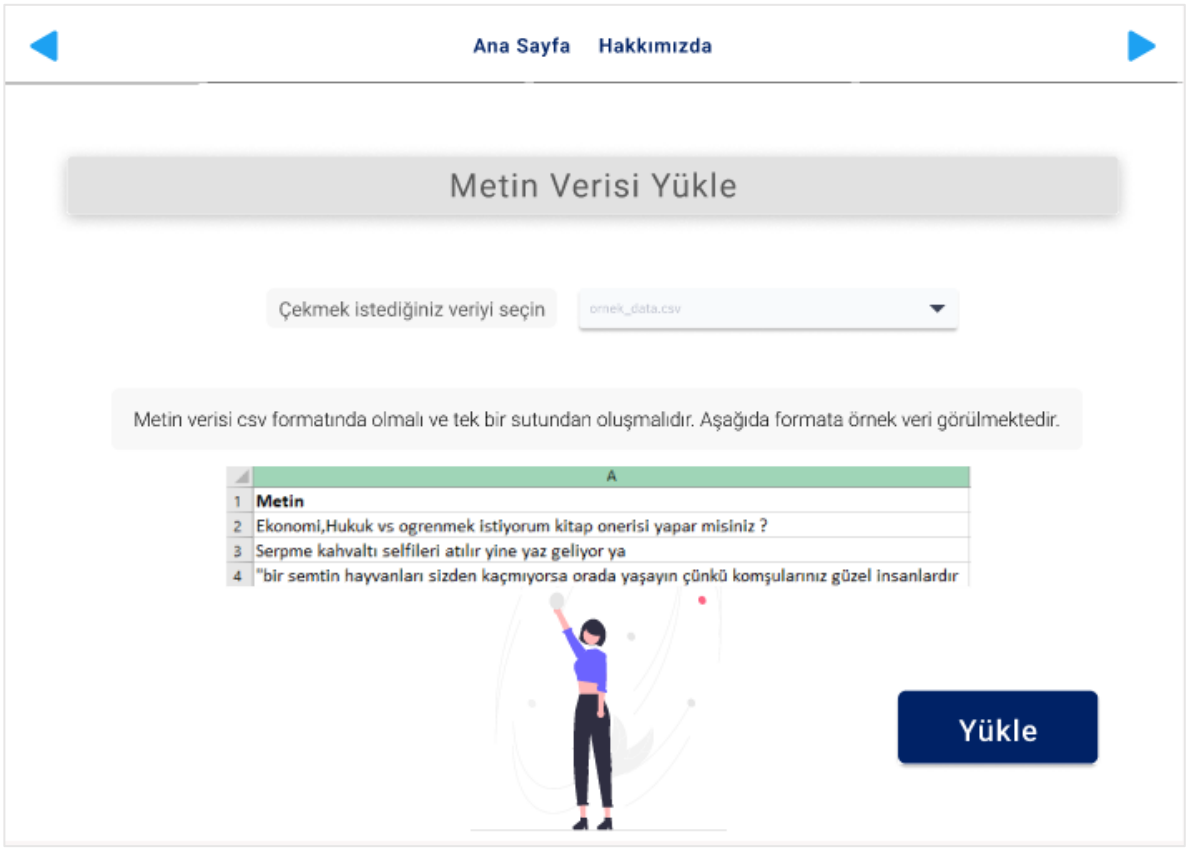
çekme işlemi sırasında snscrape kütüphanesi kullanılmıştır. Bu kütüphane sosyal ağlardan kullanıcı profilleri, hashtag'ler gibi bilgilerin toplanmasını sağlar. Burada yalnızca Twitter'dan veri almak için kullanılmıştır ancak Facebook üzerinden; kullanıcı profilleri, gruplar ve topluluklar (ziyaretçi gönderileri), Instagram üzerinden; kullanıcı profilleri, hashtag'ler ve konumlar vb. bilgilerin alınması için kullanılabilir.

Verinin elde edilmesi aşamasında sanal asistan aracılığıyla kullanıcı Twitter'dan veri çekebileceği gibi elinde hazır bulunan metni programa yükleyerekte kullanabilmektedir. Aşağıda ön yüzde bu seçimin yapıldığı ekran görülmektedir.



Şekil 35: Sanal Asistan Veri Çekme ve Yükleme Ekranı

Dışarıdan yapılan veri yüklemelerinde yüklenen belge csv formatında olmalıdır. Arayüz ekranında kullanıcı format hakkında bilgilendirilerek yönlendirilmektedir.



Şekil 36: Sanal Asistan Dış Kaynak Veri Yükleme Ekranı

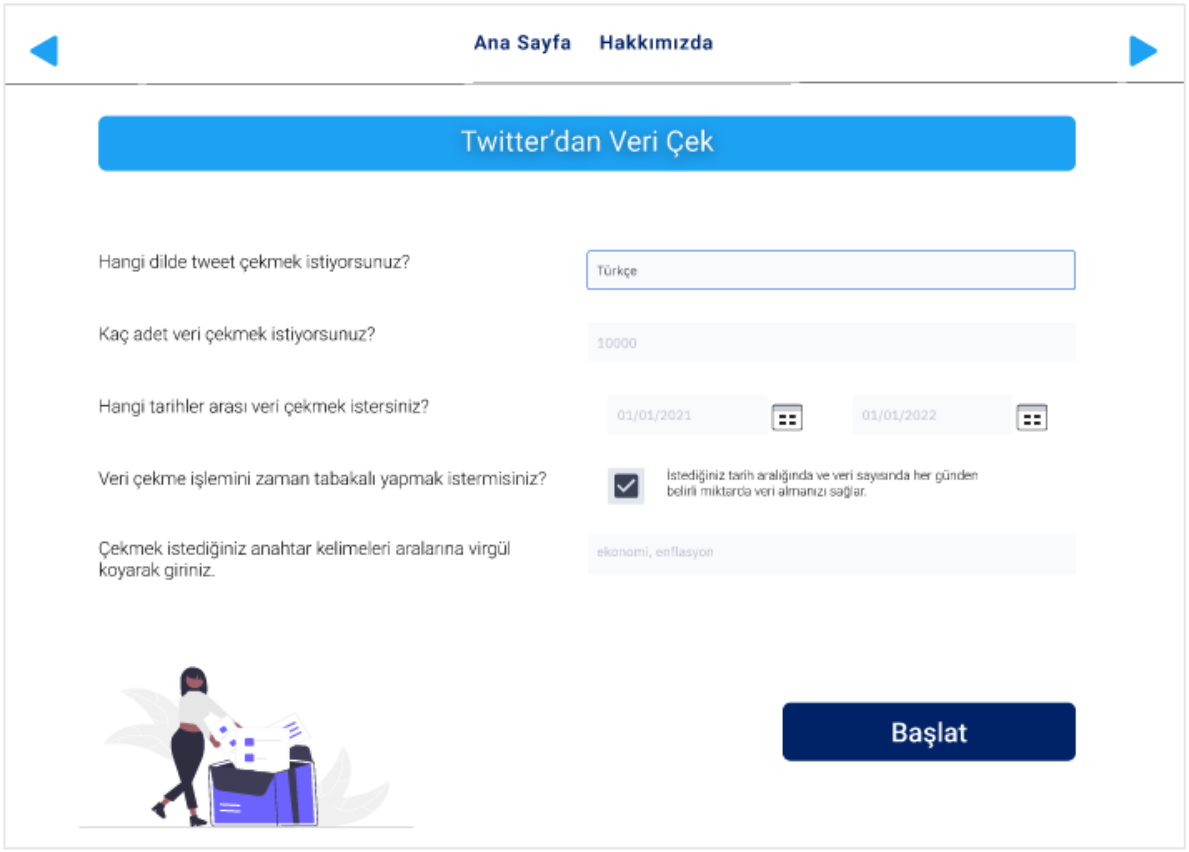
Kullanıcı Twitter ile veri çekme adımından devam ederse, veri çekme işlemi sırasında ilk adım arka planda gerekli kütüphanelerin kurulumunun yapılmasıdır.

```
import pandas as pd
import snscrape.modules.twitter as sntwitter
```

Gerekli kütüphaneler import edildikten sonra kullanıcıdan çekmek istediği verinin detayları istenir. Bu detaylar aşağıda maddelenmiştir.

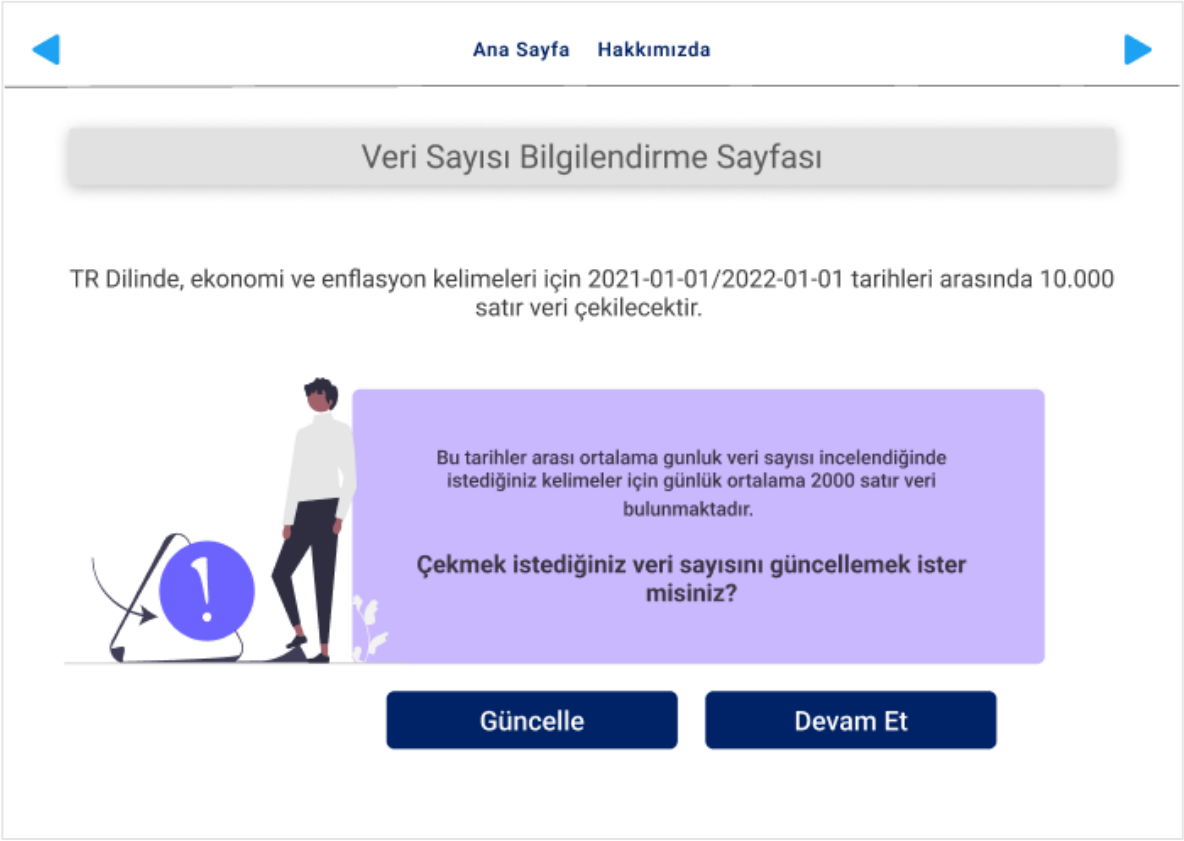
- Dil detayı (tr/en)
- Çekilmek istenen veri miktarı
- Analiz tarih aralıkları
- Analiz edilecek anahtar kelimeler

Yukarıdaki maddeler ön yüzde kullanıcıya aşağıdaki şekilde sorulmaktadır. Bu çalışma içerisinde sürecin anlatımında 10.000 satırlık "ekonomi" etiketi altında olan Türkçe Tweet'ler kullanılmıştır.



Şekil 37: Twitter'dan Veri Çekme Ekranı

Kullanıcı "Başlat" butonuna tıkladıktan sonra arka planda istenilen kelimeler ve tarih aralığı için günlük bazda veri çekilerek günlük ortalama veri sayısı hesaplanır. Kullanıcı istediği veri sayısından az ya da fazla veriye ulaşma durumunu bu yönlendirme sayesinde görebilmektedir. İstediği veriden daha az veriye erişimi olan kullanıcılar veri çekme ekranına dönerek daha fazla veya farklı anahtar kelime ile aramayı tekrarlayabilir. Fazla veri elde edilebilen kullanıcılar ise veri sayısını güncelleyerek analiz kapsamını genişletebilir.



Şekil 38: Çekilen Veri Sayısı Hakkında Bilgilendirme Ekranı

Kullanıcıdan gerekli bilgiler alındıktan sonra bir döngü yardımı ile Twitter'dan veri çekme işlemi gerçekleştirilir. Bu işlem sırasında snscrape kütüphanesindeki TwitterSearchScraper fonksiyonu kullanılmaktadır. Burada verinin gelme süresi çekilmek istenen veri sayısı ve girilen tarih aralığına göre değişkenlik gösterir. Programa giriş yapan her kullanıcının çektiği etiket ve veri sayısına göre veri çekme süresi arka planda kaydedilmektedir. Bu veri eğitim seti için yeterli boyuta geldiğinde etiket ve veri miktarına göre veri çekme süresi tahmin edilerek kullanıcıya makine öğrenmesi destekli yönlendirme sağlanacaktır. Aşağıda verinin çekilmesi sırasında kullanılan Python kod parçası gösterilmektedir.

```
for i,tweet in enumerate(sntwitter.TwitterSearchScraper((
anahtar_kelime_list[0] or anahtar_kelime_list[1]) +
    ' lang:{0} since:{1} until:{2}
    -filter:links -
filter:replies'.format(language,since_,until_)).get_items()):
    if i > data_row :
          break
    user = user + [tweet.username]
    date = date + [tweet.date]
    content = content + [tweet.content]
veri_seti = pd.DataFrame({"user":user, "date":date, "content":content})
```

Veri çekme işlemi başarılı olarak tamamlandıktan sonra verinin üzerinde herhangi bir temizlik işlemi yapılmadan kullanıcıya veri hakkında özet bilgi sağlanır. Bu özet bilgi kullanıcıya veri temizleme adımında rehberlik etmesi için hazırlanmıştır.

Şekil 39: Veri Hakkında Özet Bilgilendirme Ekranı

Kelime bulutunun oluşturulması için aşağıdaki kod parçasında görüldüğü gibi gerekli kütüphanelerin import edilmesi gerekir. Ayıca yıllık ve aylık bazda tweet sayısının dağılımı için Python'un görselleştirme kütüphaneleri olan seaborn ve matplotlib kullanılmıştır. Aşağıda gerekli kod parçası görülmektedir.

```
import seaborn as sns
import matplotlib.pyplot as plt
from wordcloud import WordCloud
wordcloud = WordCloud(background_color = "white").generate(text)
plt.imshow(wordcloud, interpolation = "bilinear")
plt.axis("off")
plt.tight_layout(pad = 0)
plt.show()
```

En sık ve en az geçen kelime dağılımlarını kelime bulutu üzerinde gören kullanıcı sonraki adımlarda analizin içeriğine özel kelime temizleme işlemleri gerçekleştirebilir.

5.2. Verinin Hazırlanması

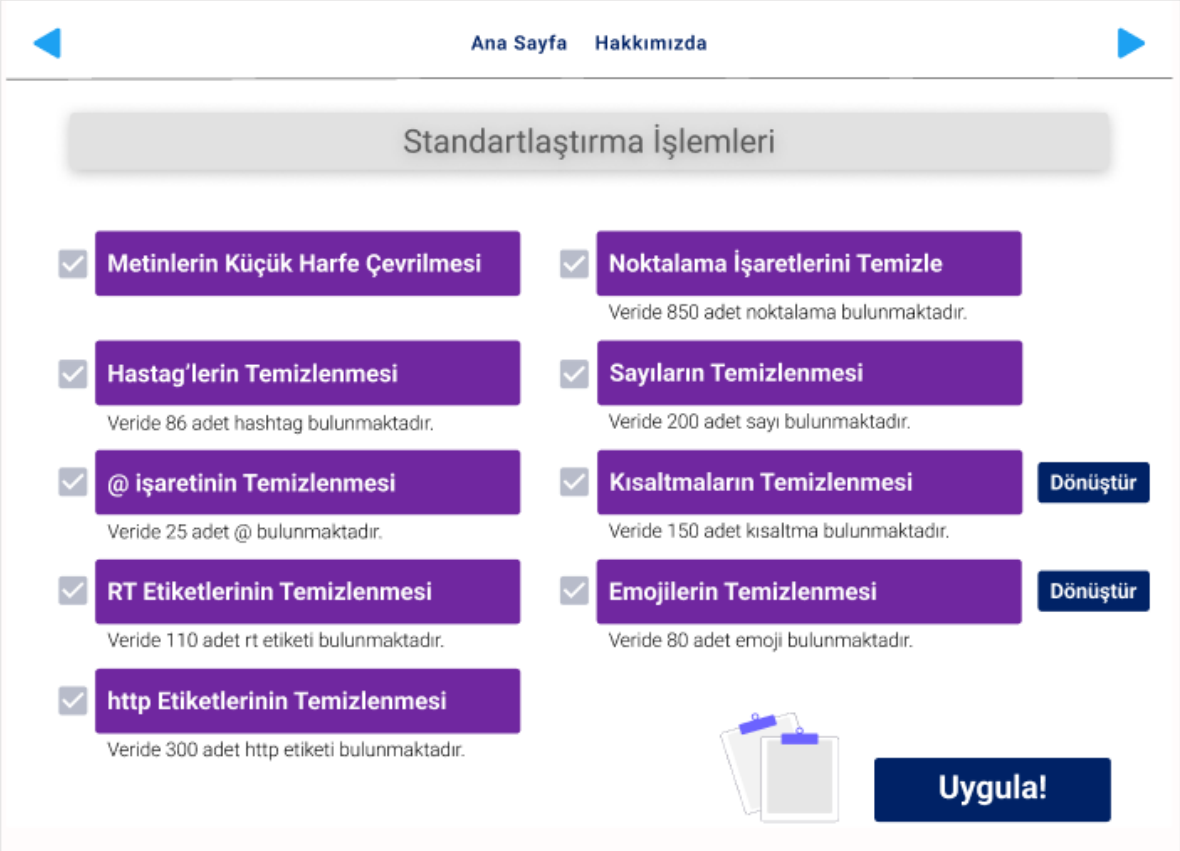
Veri özet bilgilerini gören kullanıcı sonraki adımda metin verilerinin standartlaştırılması adımına yönlendirilir. Standartlaştırma adımında kullanıcı aşağıda madde halinde verilen işlemleri gerçekleştirebilir.

- Metinlerin küçük harfe çevrilmesi
- #, @, rt, http gibi etiketlerin temizlenmesi
- Noktalama işaretlerinin temizlenmesi
- Sayıların temizlenmesi
- Kısaltmaların temizlenmesi ve dönüştürülmesi
- Emojilerin temizlenmesi ve dönüştürülmesi

Önişleme süreçlerinde Python'ın programlama dilinde bulunan "replace" ve "lambda" fonksiyonları kullanılmıştır.

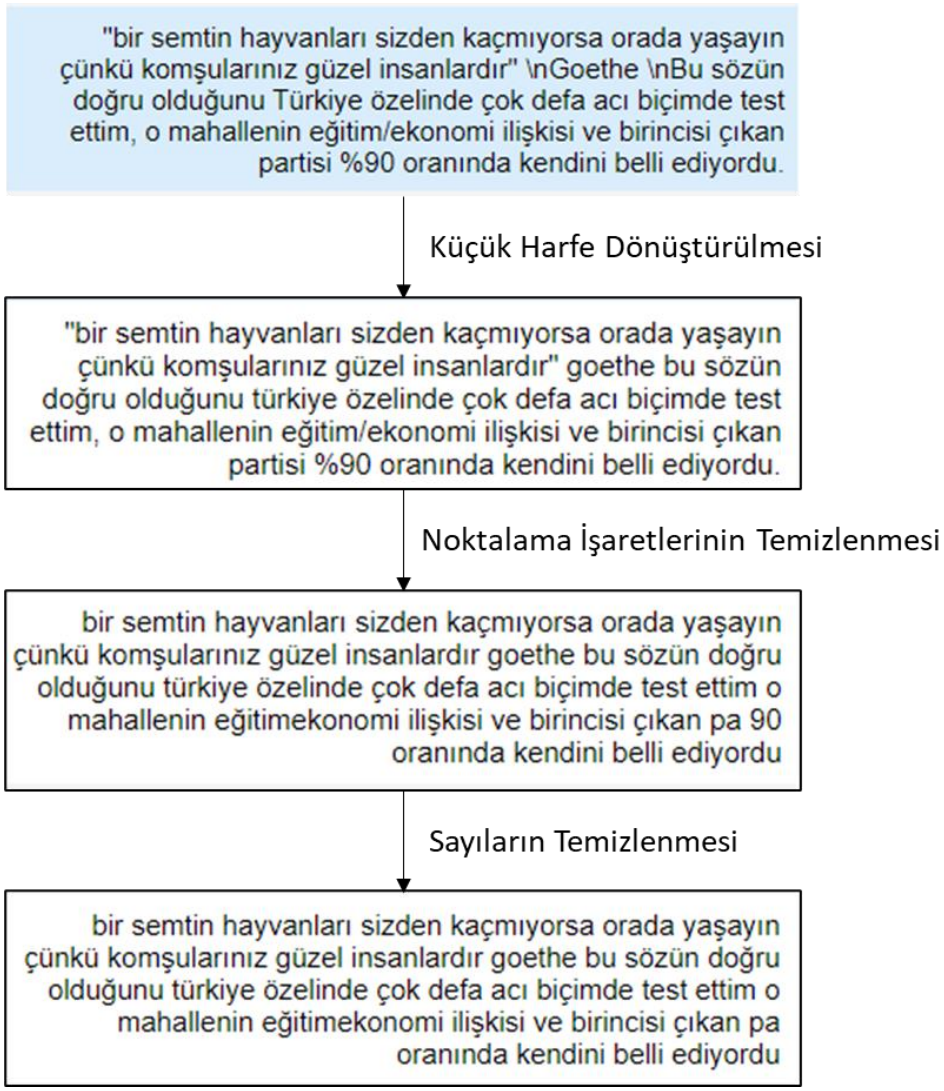
```
temiz_v1['content'] = veri_seti['content'].apply(lambda x: "
".join(x.lower() for x in x.split()))

temiz_v2['content'] = temiz_v1['content'].str.replace(r'#\S+', '')
temiz_v3['content'] = temiz_v2['content'].str.replace(r'@\S+', '')
temiz_v4['content'] = temiz_v3['content'].str.replace(r'http\S+', '')
temiz_v5['content'] = temiz_v4['content'].str.replace(r'[^\w\s]', '')
temiz_v6['content'] = temiz_v5['content'].str.replace(r'\d', '')
```

Kullanıcı aşağıdaki şekilde önyüz aracılığıyla seçim yaparak istediği standartlaştırma işlemlerini gerçekteştirebilir. Bazı durumlarda kullanıcının analiz sürecine bağlı olarak temizlemek istemediği özel işaretler olabilir. Bu durumlarda seçim yapmadan ilerlenmesi gerekir. Temizleme sürecinde kullanıcı çektiği veya yüklediği veriye özgü özel karakterlerin sayısal dağılımı konusunda bilgilendirilir. Örneğin; ekonomi için çekilen veri içeriğinde 850 adet noktalama işareti bulunmaktadır. Oldukça fazla olan bu oran kullanıcı tarafından silinebilir. Ayrıca kullanıcı verisinde sayısal olarak fazla olan kısaltma ve emojileri metne dönüştürerek kullanabilmektedir.



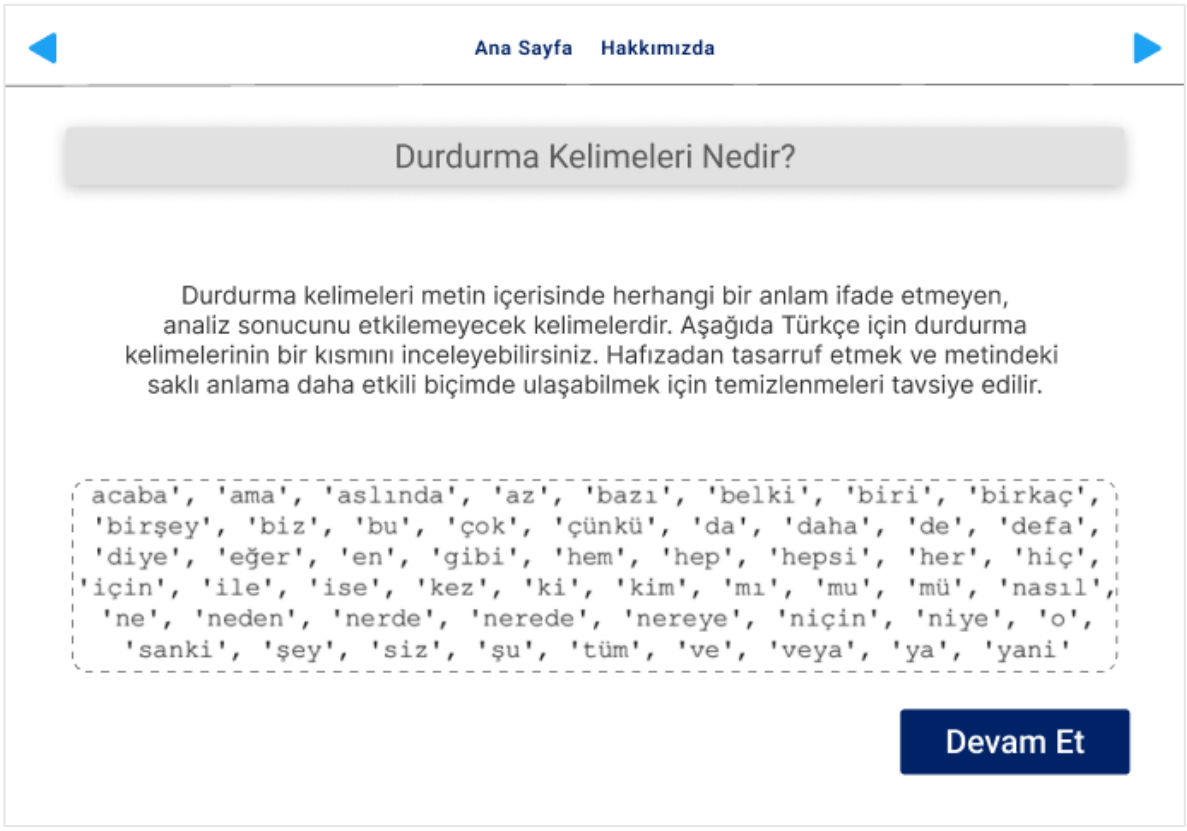
Şekil 40: Veri Standartlaştırma Ekranı

Seçimler sonrasında arka planda metinlerin küçük harfe dönüştürülmesi, özel ifadelerin, noktalama işaretlerinin ve sayısal ifadelerin temizlenmesi çekilen veri örneği üzerinden gösterilmektedir.



Şekil 41: Veri Standartlaştırma Süreci Örneği

Standartlaştırma işleminden sonra veri ön işlemenin başka bir adımı olan filtreleme adımı gelmektedir. Burada Tweet'ler içerisinde geçen durdurma kelimeleri python içerisinde bulunan NLTK kütüphanesindeki geniş listeden standart olarak temizlenebilir. Ayrıca kullanıcının analiz sürecine göre belirlediği kendi özel durdurma kelimelerini çıkarabilmesi de bu ekran üzerinden mümkündür. Kullanıcı filtreleme ekranına yönlendirilmeden önce durdurma kelimeleri hakkında bilgi alması için bilgilendirme ekranına yönlendirilmektedir.



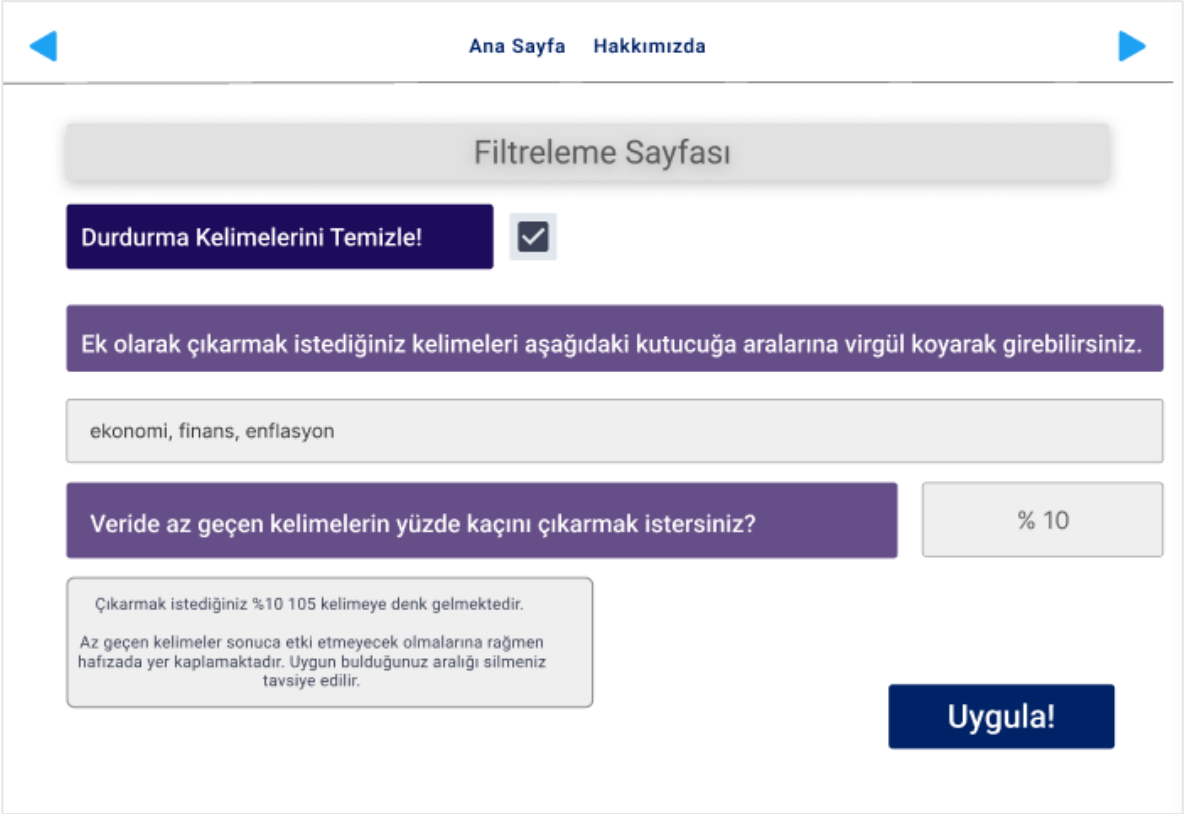
Şekil 42: Durdurma Kelimeleri Bilgilendirme Ekranı

Kullanıcı "Devam Et" butonu ile ilerlediğinde filtreleme ekranına ulaşacaktır. Aşağıda arka planda durdurma kelimelerinin çıkartılması için gereken python kod parçası görülmektedir. NLTK kütüphanesi aracılığıyla indirilerek sw adlı değişkenin içerisine alınan durdurma kelimeleri listesi lambda fonksiyonu aracılığıyla silinmiştir.

```
import nltk

nltk.download("stopwords")

from nltk.corpus import stopwords

sw = stopwords.words("turkish")

temiz_v7['content'] = temiz_v6['content'].apply(lambda x: " ".join(x
for x in x.split() if x not in sw))
```

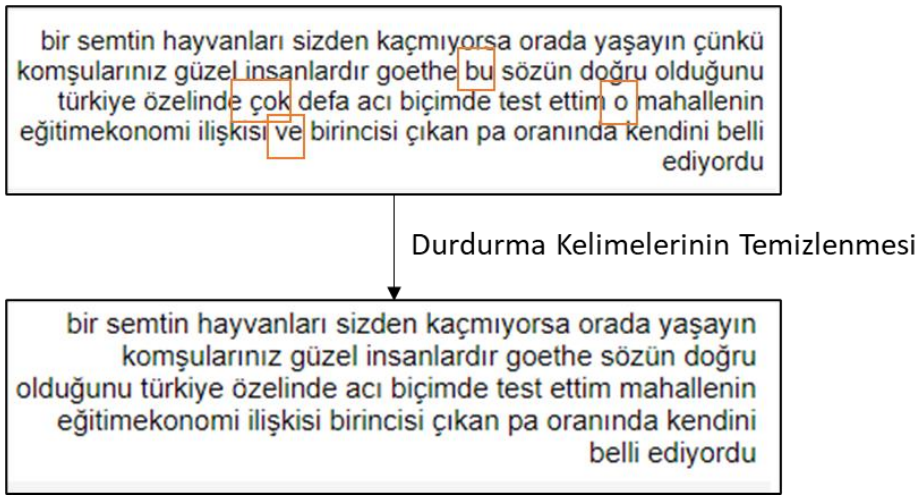
Kullanıcı arayüz üzerinden aşağıdaki şekilde seçim yaparak durdurma kelimeleri temizleme işlemini gerçekleştirebilmektedir. Ayrıca bu ekran aracılığıyla yalnızca standart belirlenen

kelimeler değil, analiz sürecine uygun çıkartılmak istenen kelimeler de aralarında virgül olacak şekilde yazılarak çıkartılabilir. Örneğin; ekonomi üzerine yapılan metin madenciliği sürecinde, sık geçen kelimeler üzerinden hazırlanan kelime bulutunun "ekonomi" kelimesini büyük şekilde görselleştirilmesi beklenen bir durumdur. Bu kelime kaldırılarak saklı olan diğer anlamlar kelimeler aracılığıyla ortaya çıkarılabilir. Ek olarak veride az geçen kelimeler sonuca etki etmedikleri halde boyutsal olarak yer kaplamaktadır. Kullanıcı yüzdesel olarak giriş yaparak az geçen kelimelerin istediği miktarda silinmesini sağlayabilir.



Şekil 43: Veri Filtreleme Ekranı

Aşağıda örnek olarak çekilen veri üzerinden durdurma kelimelerinin çıkarılması işlemi gösterilmektedir.

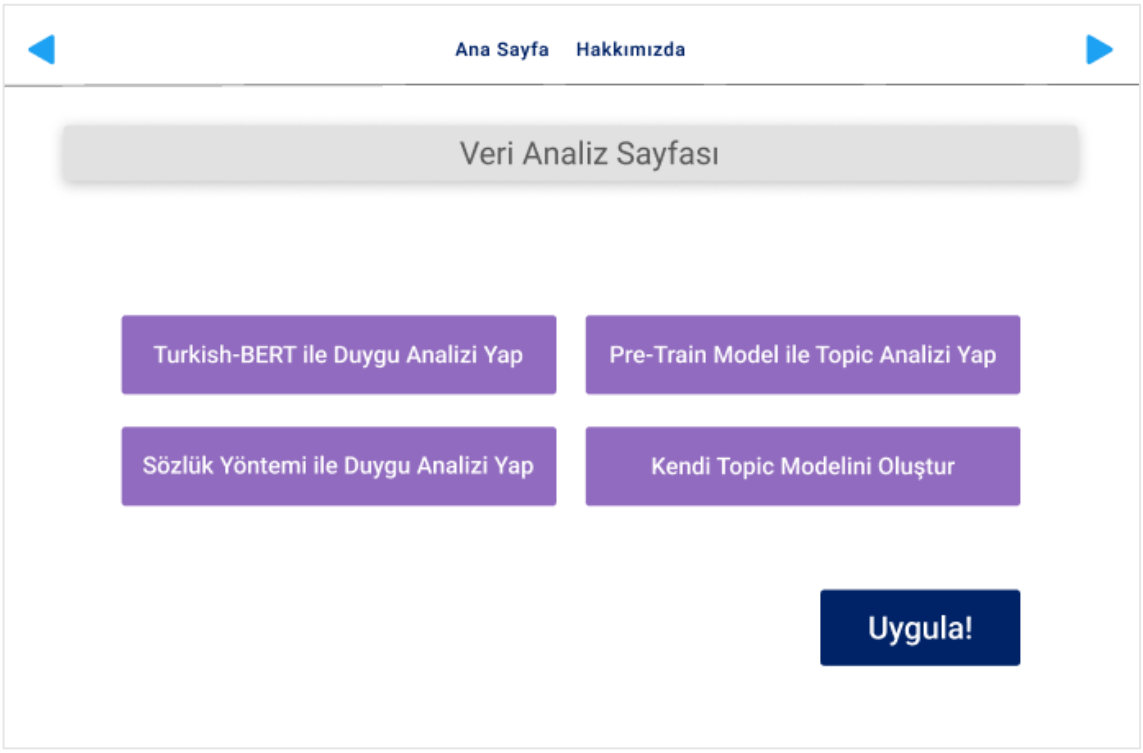


Şekil 44: Filtreleme Süreci Örneği

Durdurma kelimeleri içerisinde bulunan "çok", "bu", "ve", "o" gibi anlamsal olarak analiz sonucuna etkisi olmayan kelimelerin filtreleme adımında cıkarılması sağlanmıştır.

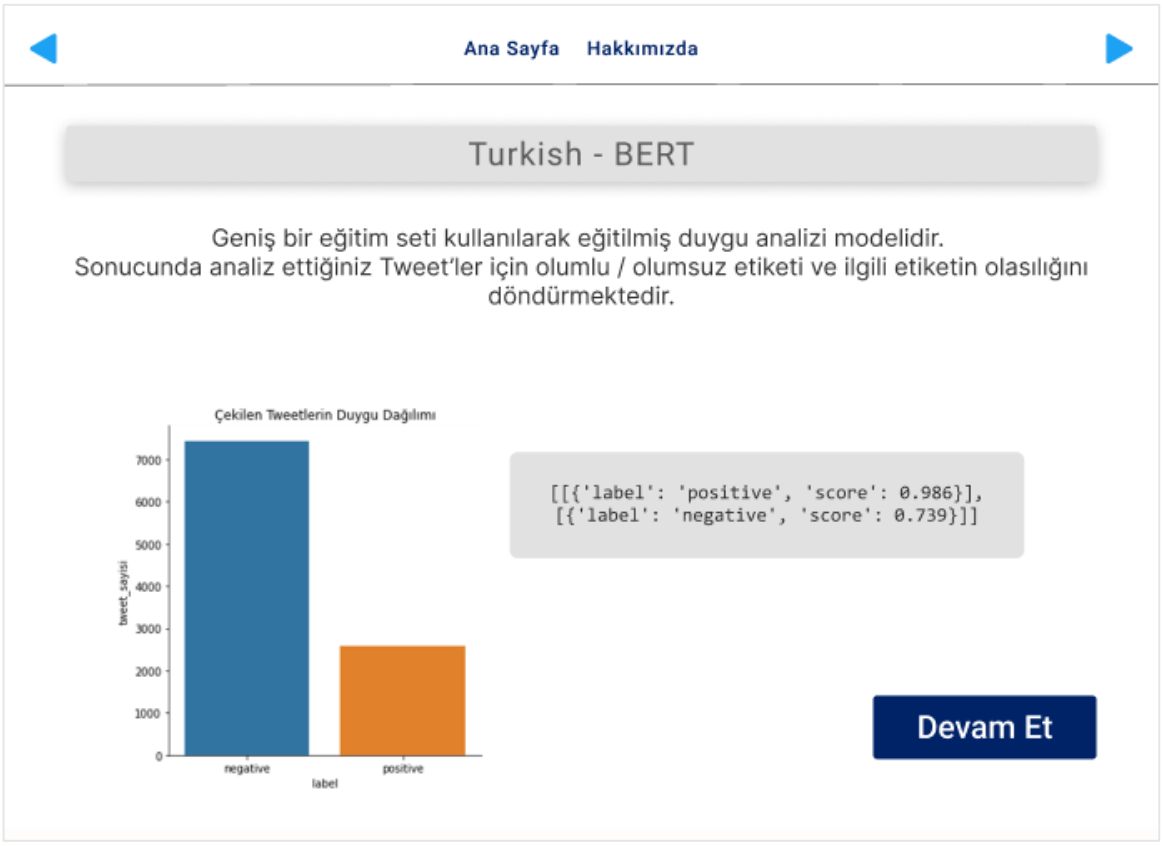
5.3. Bilgi Çıkarımı

Birimleştirme, standartlaştırma ve filtreleme veri ön işleme adımlarından sonra temizlenen veri analize hazır hale getirildi. Bu aşama da temiz veri üzerinde tekrar en sık geçen kelimelerin dağılımını kelime bulutu yardımıyla incelemek faydalı olacaktır. İstenmeyen kelime ve kelime grupları için standartlaştırma ve filtreleme adımına dönülebilir. Sanal Asistan önyüzünde aşağıdaki ekran üzerinden istenilen analiz türü seçilerek devam edilebilir.



Şekil 45: Veri Analiz Ekranı

Kullanıcı bu analizlerden herhangi biri ile devam etmek istendiğinde ilgili analiz hakkında detaylı bilgi alması için bilgilendirme ekranı açılmaktadır. Aşağıda bu ekranlardan Turkish-BERT örneği paylaşılmaktadır.



Şekil 46: Veri Analiz - Turkish-BERT Bilgilendirme Ekranı

Kullanıcı analiz ihtiyacına göre analizlerin bilgilendirmelerini inceleyerek seçim yapabilmektedir.

5.3.1. Turkish-BERT ile Duygu Analizi

Veri analiz sürecinin ilk adımı olarak, Google tarafından yönetilen BERT (Bidirectional Encoder Representations from Transformers) adlı çalışmanın Türkçe versiyonu TurkishBERT ile duygu analizi işlemleri gerçekleştirilecektir. BERT algoritması temelde cümleleri sağdan sola ve solda sağa doğru tarayarak kelimeler arası ilişkiyi iki yönlü aramayı amaçlar. Tarama işlemi sırasında ilişkiyi sentezlemek için ise transformer denilen temelinde İleri Beslemeli Sinir Ağları (Feed Forward Neural Networks) olan bir fonksiyon kullanır. TurkishBert modelinin mevcut sürümü, Türk OSCAR külliyatının filtrelenmiş ve cümle bölünmüş bir versiyonu, yakın tarihli bir wikipedia dökümü, çeşitli OPUS külliyatları ve Kemal Oflazer tarafından sağlanan özel bir külliyat üzerinde eğitilmiştir. Temel öğrenme

veri setinin de genişliği sayesinde (yaklaşık 35gb Türkçe metin) TurkishBERT yüksek başarım elde edebilmektedir.

Turkish-BERT modelinin hazırlanan temiz veri üzerinde uygulanabilmesi için öncelikle gerekli kütüphanelerin çalışmaya dahil edilmesi gerekir. Python'da bulunan transformers kütüphanesi kullanılarak model pipeline süreci aşağıdaki şekilde oluşturulmuştur.

```
from transformers import AutoModelForSequenceClassification, AutoTokenizer, pipeline

model = AutoModelForSequenceClassification.from_pretrained("savasy/bert-base-turkish-sentiment-cased")

tokenizer = AutoTokenizer.from_pretrained("savasy/bert-base-turkish-sentiment-cased")

sa= pipeline("sentiment-analysis", tokenizer = tokenizer, model = model)
```

Oluşturulan model pipeline'ı (sa) yardımıyla aşağıdaki görüldüğü gibi duygu sonuçları elde edilebilir. Burada dikkat edilmesi gereken nokta analiz için hazırlanan temiz verinin string veri türünde olması gerektiğidir. Aşağıda bu dönüşüm gösterilmektedir.

```
data["content"] = data["content"].apply(lambda r: str(r))

sentiment_list = []

for i in data["content"]:

     sentiment_list.append(sa(i))
```

Turkish-BERT model sonucunda aşağıdaki şekilde çıktı üretmektedir. Çıktıda görüldüğü gibi metinlerin pozitif, negative etiketlemesinin yanı sıra duygu skor olasılıkları bulunmaktadır. Bu olasılıklar yardımı ile nötr duygu dağılımına sahip cümlelerde belirlenebilir. Bu çıktı python içerisinde veri çerçevesine dönüştürülerek görselleştirmeler de kullanılabilir.

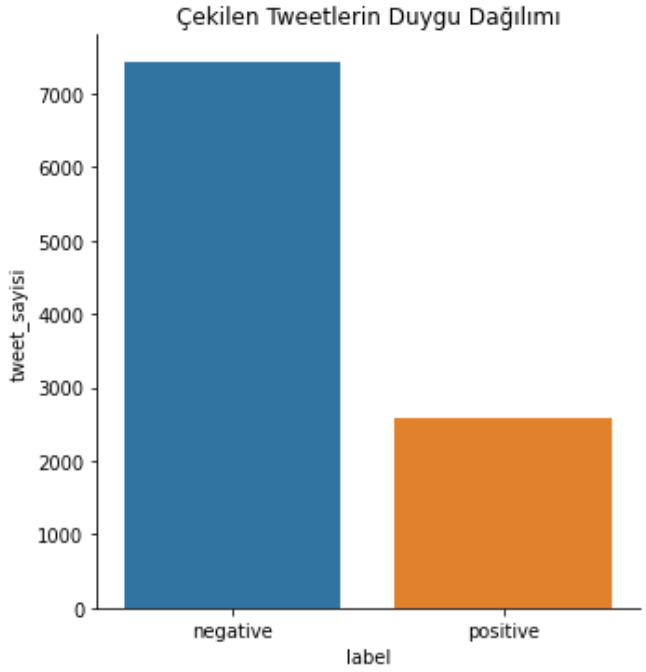
```
[[{'label': 'positive', 'score': 0.9860413670539856}],

[{'label': 'negative', 'score': 0.7391416430473328}],

[{'label': 'positive', 'score': 0.5551885366439819}],

[{'label': 'negative', 'score': 0.9220163226127625}],
```

[{'label': 'negative', 'score': 0.997983455657959}],

[{'label': 'positive', 'score': 0.9592059254646301}]]

Ekonomi etiketi baz alınarak çekilen Twitter verisinin duygu skor dağılımı seaborn kütüphanesi aracılığyle görselleştirilmiştir. Python kodu ve duygu dağılım grafiği aşağıda görülmektedir.

```
sns.catplot(x = "label", y = "tweet_sayisi", kind = "bar", data =
duygu_sonuc).set(title='Çekilen Tweetlerin Duygu Dağılımı')
```



Şekil 47: Turkish-BERT Duygu Analizi Sonuçları

Görüldüğü gibi "ekonomi" etiketi ile çekilen 10.000 satırlık verinin duygu dağılımının yaklaşık %70'i negatif olarak etiketlenmiştir. Kalan %30'luk kısım ise pozitif olarak görülmektedir. Tweet paylaşımında bulunan kullanıcıların genel olarak olumsuz yorumlarda bulunduğu görülse de bu duygu dağılımının ilgili konu başlıkları altında nasıl dağıldığını görmek yorumlama da bulunmak için kritiktir.

5.3.2. Pre-Train Model ile Konu Analizi

Araştırmacılar konu modellemesi için internette hazır bulunan etiketlenmiş veri setlerini kullanabilecekleri gibi kendi ihtiyaçlarına uygun olarak veri seti etiketlemelerini gerçekleştirebilirler. Aşağıda metin belgeleri üzerinde konu etiketleme örneği paylaşılmıştır.

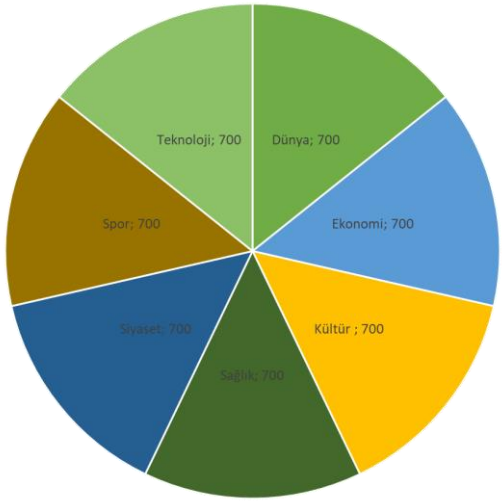
Tablo 4: Topik Modelleme - Etiketli Veri Örneği

Yorum	Kategori Etiketi
Ürün çok güzel herkese tavsiye ederim ancak kargo elime çok geç ulaştı	Kargo
Ürünü severek kullanıyorum site arayüzünden zor sipariş versemde en uygun fiyatlı buradaydı	Fiyat, Site arayüzü
Ürün kullanışlı fakat hala aynı ürün için kampanya mesajları almaya devam etmekteyim bıktım gerçekten	Kampanya
Kargo elime çok geç ulaştı	Kargo

Etiketlenmiş veri kullanmanın dışında gözetimsiz öğrenme algoritmaları yardımı ile araştırmacının veri setini kümelere ayırması ve küme içeriklerine göre kategorize etmesi mümkündür. Aşağıda metin verisi kümeleme örneği gösterilmektedir.

Şekil 48: Topik Modelleme - Kümeleme Örneği

Bu çalışmada bahsedilen yöntemler içerisinden analiz içeriğine göre yeterlilik sağlayan, hazır veri seti ile model eğitim süreci gerçekleştirilecektir. Sanal asistan içerisinden kullanıcıların kendi etiketlemelerini yaptıkları veri setini yükleyerek eğitmesi ve bunun sonucunda yeni gelen verileri kategorize etmesi mümkündür. Aşağıda internetten hazır olarak alınan veri seti içeriğinin kategori dağılımı görülmektedir.



Şekil 49: Eğitim Verisi Kategori Dağılımı

Veri seti 7 kategori için toplamda 4900 satır veriden oluşmaktadır. Her bir kategori için grafikte görüldüğü gibi 700 adet veri bulunmaktadır. Konu modellemesi için bu veri setinin eğitim ve test veri setine bölünerek çeşitli algoritmalar üzerinden en başarılı modelin seçilmesi gerekmektedir.

Modelleme aşamasına geçilmeden önce veri etiketlerinin ve yorum verisinin sayısal hale getirilmesi gerekmektedir. Veri etiketleri için Python'da bulunan factorize fonksiyonu kullanılmıştır. Verilen etiketler: Siyaset:0, Dünya: 1, Ekonomi: 2, Kültür: 3, Sağlık: 4, Spor: 5, Teknoloji: 6

Yorum verilerinin sayısallaştırılması içinse, python'ın makine öğrenmesi kütüphanesi sklearn içerisindeki (TfidfTransformer, TfidfVectorizer) belirli fonksiyonlar kullanılarak TF-IDF dönüşümü ile sayısallaştırma işlemi gerçekleştirilmiştir. Aşağıda sayısallaştırma sonucu gösterilmektedir.

```
(0, 32751)	0.0716785418867453
(0, 40510)	0.0254416775637408
(0, 46768)	0.1183630916579576
(0, 10348)	0.1162063265539641
(0, 56957)	0.0721996111311338
(0, 21838)	0.0449074766303545
(0, 52560)	0.1385974006164784
(0, 3126)	0.0545436076990602
(0, 90019)	0.1208528942465008
(0, 40822)	0.0359772823901719
(0, 11511)	0.0612038920597334
(0, 92436)	0.1208528942465008
```

Sayısallaştırma işlemlerinin ardından bahsedilen veri kaynağı eğitim (%80) ve test (%20) olmak üzere iki parçaya ayrılmıştır. Bunun sebebi elde hazır olarak bulunan etiketlenmiş verinin modelin başarısını test etmek içinde kullanılabilmesidir. Bu başarı oranına göre uygulanan algoritmalar arasından en iyi sonucu veren model başlangıçta temizlenen veriye uygulanacaktır.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(model_df["text"],
model_df["labels"], test_size = 0.2, random_state = 4)
```

Eğitim ve test olmak üzere ikiye ayrılan veri seti üzerinde bağımlı değişken olarak etiketlenen metin ve bağımsız değişken olarak ise etiketler (kategoriler) kullanılmıştır. Aşağıda eğitim ve test olarak ayrılan veri seti üzerinde uygulanan algoritmalar listelenmiştir.

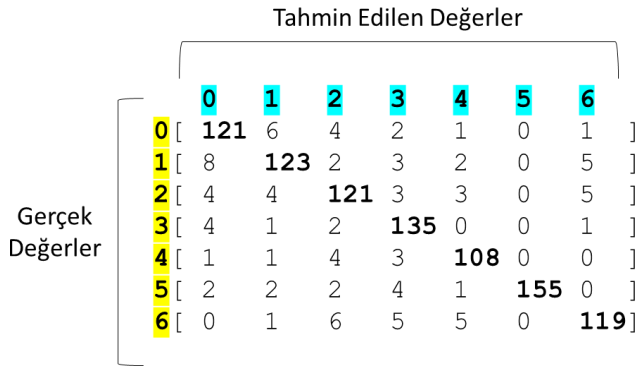
- Naive Bayes
- Lojistik Regresyon
- Karar Ağacı

- Rasgtele Orman
- Gradyan Arttırma
- XgBoost

Algoritmaların uygulanması aşamasında Python'un sklearn kütüphanesi kullanılmıştır. Tüm modellerin nesnelerinin oluşturulması ile başlayan süreç, verinin modele uydurulması (fit edilmesi) ve sonucunda tahminlemenin yapılması ile devam etmektedir. Modelin öğrenme sonucunda tahmin ettiği kategori dağılımı ile gerçek kategori değerleri arasındaki farklılıkların incelenmesini sağlayan karışıklık matrisi (confusion matrix) yardımıyla model başarısı değerlendirilebilir. Aşağıda model kurma sürecine örnek olarak Lojistik Regresyon Python kodu gösterilmektedir.

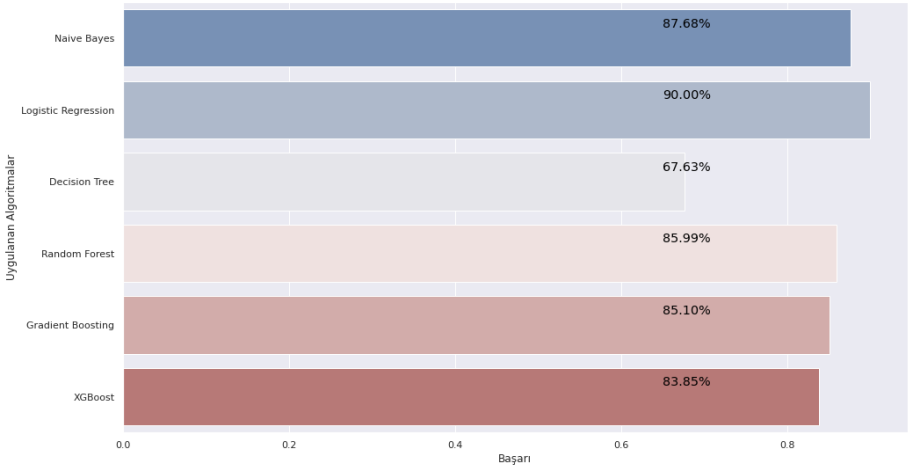
```
LogicReg = LogisticRegression()

LogicReg.fit(train_vectors, y_train)

prediction = LogicReg.predict(test_vectors)

print("Logistic Regression ::\n", confusion_matrix(y_test,
prediction),"\n")

print(accuracy_score(y_test, prediction))
```

Uygulanan tüm algoritmaların başarısının değerlendirilmesinde en temelde aşağıdaki confusion matrix çıktısından yararlanılabilir. Bu matris sayesinde ilgili kategori bazında tahmin edilen değerlerin doğruluk oranı belirlenebilir. Örneğin; 0 etiketine sahip olan Siyaset kategorisi için test veri setinde yaklaşık 140 adet gözlem bulunmaktadır. (Test veri setinde 4900'ün %20'si 980 adet kayıt bulunmaktadır. 7 kategori için 980/7 üzerinden yaklaşık 140'ar adet kayıt bulunması beklenir.) Bu gözlemlerin 121 tanesi gerçekte Siyaset iken siyaset olarak doğru tahmin edilmiştir. 6 tanesi ise gerçekte Siyaset iken 1 etiketine sahip dünya kategorisi olarak yanlış tahmin edilmiştir.



Şekil 50: Karışıklık Matris Sonucu

Confusion matris üzerinden hesaplanan doğruluk değeri toplamda doğru olarak sınıflandırılanların tüm sınıflandırılan kategorilere oranı olarak hesaplanır. En temel başarı metriği olarak kullanılmaktadır. Veri üzerinde uygulanan tüm algoritmaların başarı oranı aşağıdaki grafikte görülmektedir.

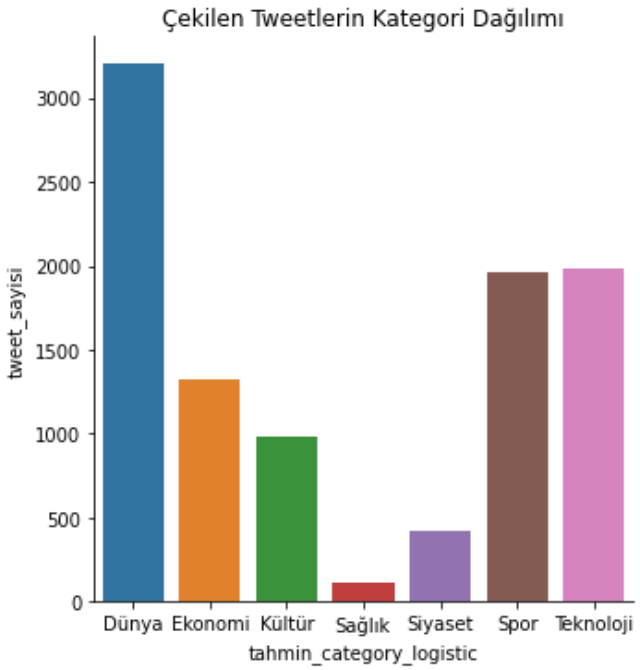


Şekil 51: Topik Modelleme Algoritma Başarı Dağılımı

Grafikte görüldüğü gibi en başarılı algoritma Lojistik Regresyon olarak belirlenmiştir. Burada dengeli bir veri seti kullanıldığı için accuracy değeri üzerinden algoritma seçimi gerçekleştirilebilir. Ancak dengesiz dağılan veri setlerinde örneğin; 100 adet işlem içerisinden %1 oranında yapılan aykırı (fraud) işlemi bulabilmek için doğruluk oranına bakılması yetersiz kalacaktır. Çünkü veri seti aykırı olmayan işlem sayısı daha fazla olduğu için bu sınıfı doğru olarak %99 oranında tespit edebilecektir oysaki analizin temel amacı %1'lik kısmı tespit edebilmektir. Bu durumda farklı başarı metrikleri ile değerlendirme yapılması gerekir. Bu metrikler; kesinlik (precision), duyarlılık (recall) ve f1 skoru olarak

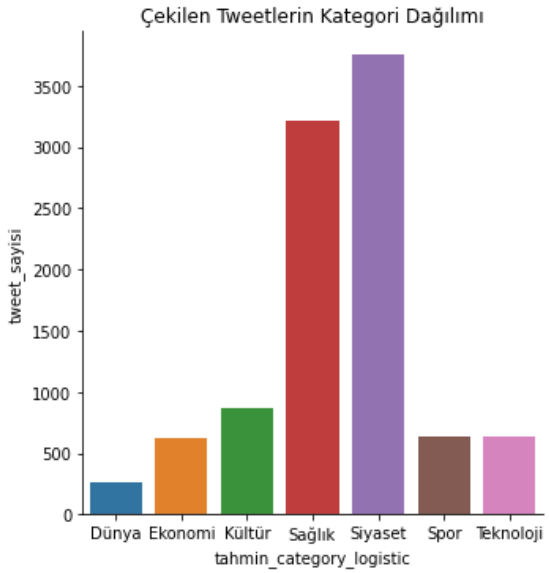
özetlenebilir. Duyarlılık aykırı işlem üzerinden açıklanacak olursa; gerçekte sahtekarlık yapan işlemlerin kaç tanesinin aykırı olarak etiketlendiğinin sayısının hesaplanmasında kullanılır. Kesinlik ise aykırı olarak tahmin edilen değişkenlerin kaçının gerçekte doğru tahmin edildiği sayısının hesaplanmasında kullanılır. Son olarak f1 skoru ise kesinlik ve duyarlılık metriklerinin harmonik ortalamasının alınması ile hesaplanır. Genel değerlendirme için bu metriğin kullanılması tavsiye edilmektedir.

Aşağıda "ekonomi" etiketi için çekilen 10.000 satırlık veri için yapılan kategori dağılım sınıflandırmasının sonuçları görülmektedir.



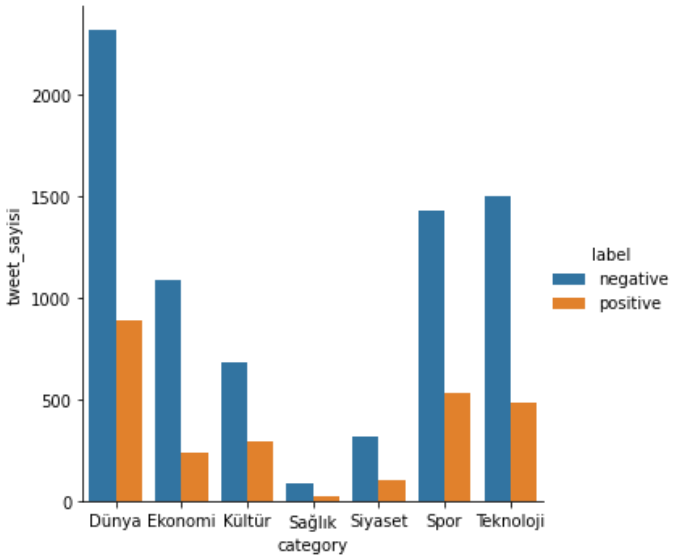
Şekil 52: Ekonomi Verisi Kategori Dağılımı

Verinin kategori dağılımı incelendiğinde büyük bölümünün Dünya olarak etiketlendiği görülmektedir. İkinci sırada ise Teknoloji ve spor kategorisi gelmektedir. Ekonomi terimi günlük yaşamda her alanı genel olarak etkilediği için bu analiz doğrultusunda beklenen bir sonuçla karşılaşıldığı yorumu yapılabilir. Sanal asistanı kullanan kişiler yapacakları analiz doğrultusunda farklı anahtar kelimeler seçerek analizlerine özgü kategori dağılımlarını inceleyebilirler. Örneğin aşağıda "korona" kelimesi için oluşturulan kategori dağılımı gösterilmektedir.



Şekil 53: Korona Kelimesi Kategori Dağılımı

Ekonomi etiketi için gerçekleştirilen analizin kategori bazında duygu dağılımı aşağıda görülmektedir. Her kategoride ağırlıklı olarak negative duyguların yer aldığı söylenebilir.



Şekil 54: Topik Modelleme ve Duygu Analizi Sonuç Dağılımı

Kullanıcılar için özetlenen örnek süreç "ekonomi" anahtar kelimesi üzerinden gösterilmiştir. Kategori ve duygu dağılımı çekilen veriye özgü olarak değişecektir.

5.3.3. Analiz Sonuçlarının Kaydedilmesi

Sanal asistan verinin toplanması, ön işlemesi ve analiz aşamaları sonrasında kişilerin yaptıkları analizlerin bir kaydını tutmaktadır. Tüm analiz sürecinde çekilen etiketler, veri sayısı, veri üzerinde yapılan işlemler ve analiz adımlarının sonucu (duygu, kategori) sonraki çalışmalarda yapılacak makine öğrenmesi analizlerine eğitim verisi olarak verilebilecektir. Bu sayede benzer etikette veri çeken kişilere yapacağı temizleme adımlarında nelere dikkat etmesi gerektiği ve bunun sonucunda duygu ve kategori olarak nasıl sonuçlar alacağı henüz analizi tamamlamadan gösterilebilecektir. Ek olarak bu kayıtların metinlerin ilgili etikette hangi duygu ve kategori dağılımında olduğunu içermesi farklı modellemelere de kaynak olabilir. Aşağıda bahsedilen kayda örnek paylaşılmıştır.

Tablo 5: Analiz Sonuçları Kayıt Tablosu

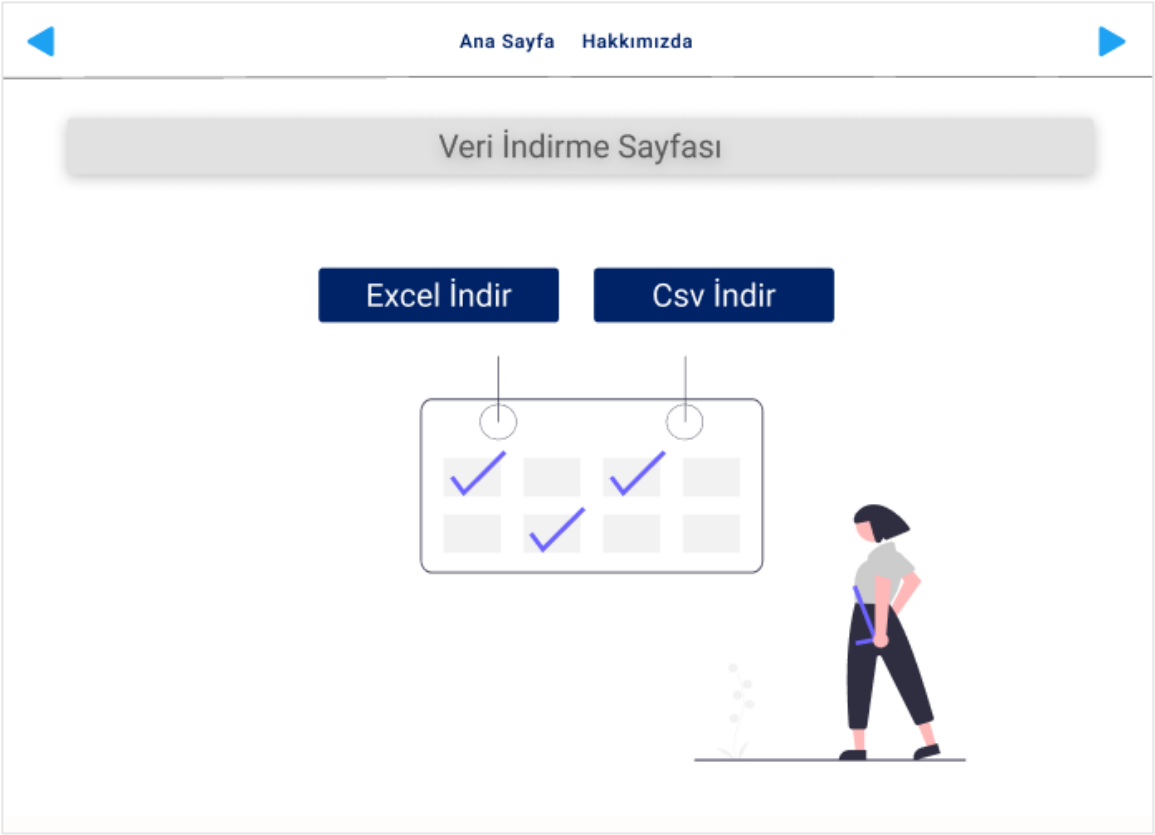
Dil	Satır Sayısı	Baslangıc Tarihi	Bitis Tarihi	Anahtar Kelime Sayısı	Anahtar Kelime Listesi	Alınan Satır Sayısı
TR	1000	2022-01-01	2022-04-01	2	[‘ekonomi’, ‘para’]	1000

http var mı?	http Sayısı	Noktalama var mı?	Noktalama Sayısı	Sayı var mı?	Sayıların Sayısı	Std Durdurma Kelimeleri Temizlendi mi?	Özel Durdurma Kelime Listesi
0	15	1	712	1	487	1	[‘para’, ‘piyasa’]

Kategori dünya	Kategori ekonomi	Kategori siyaset	Kategori sağlık	Kategori kültür	Kategori teknoloji	Kategori spor	Duygu pozitif	Duygu negatif
24	482	356	78	29	24	8	190	810

Görüldüğü gibi tablo içerisinde Twitter’dan çekilen veri detaylarının yanı sıra temizleme işlemleri hakkında detaylar bulunmaktadır. Örneğin; sütunlarda ilgili temizleme kullanıcı tarafından yapıldı ise 1, yapılmadı ise 0 olarak etiketlenmektedir. Kategori ve duygu durumu dağılımı içinse örnek olarak burada 1000 satır verinin duygu durumu bazında dağılımı görülmektedir. (190: Pozitif, 810: Negatif)

Kullanıcılar program aracılığıyla elde ettikleri, temizledikleri ve analiz ettikleri verileri de yine program aracılığıyla indirebilirler. Aşağıda indirme işlemini yapılabilen önyüz ekranı paylaşılmıştır.



Şekil 55: Analiz Sonuçları İndirme Ekranı

Excel veya csv formatında indirilen temiz veri üzerinden araştırmacılar çalışmalarını sürdürebilmektedir. Ayrıca başlangıcta elde edilen ham veriye arka planda atılan index sayesinde eski ve yeni veri karşılaştırması ve eşleştirilmesi kolaylıkla yapılabilmektedir.

SONUÇ

Sosyal medya analitiği, günümüzde artan sosyal medya platformlarının da etkisi ile gitgide önem kazanmaktadır. Çevrimici kaynaklardan elde edilebilen büyük miktardaki yapılandırılmamış verinin karar süreçlerine dahil edilebilmesi oldukça önemlidir. Bu verinin işlenmesi ve içerisinden anlamlı sonuçların çıkarılması ise daha karmaşık analiz süreçlerine gereksinimi ortaya çıkarmaktadır. Metin madenciliği bu gereksinim üzerinden ortaya çıkarak yarı yapısal ve yapısal olmayan formdaki metinlerin içlerindeki örtülü bilgiyi ortaya çıkarmayı hedefleyen bilişim destekli bir süreçtir. Sosyal medya analitiği ise arka planda metin madenciliği süreçlerini içerir ve günlük konuşma dili ağırlıklı olması sebebiyle daha hassas yaklaşımlar gerektirir.

Bu çalışmada en yaygın kullanılan sosyal medya platformlarından biri olan Twitter üzerinden, veri toplama, veri ön işleme ve analiz adımlarında araştırmacıya analiz sürecinin öğretilmesini ve yönlendirmeler ile desteklenmesini sağlayan bir sanal asistan arayüz tasarımı gerçekleştirilmiştir. Bu tasarım ile araştırmacı veri toplama aşamasından başlayarak analiz sürecinin tamamında desteklenmektedir. Türkçe için ek zorluklar barındıran sosyal medya analitiği sürecini sosyal bilimciler için kolaylaştırmak çalışmanın temel motivasyonlarındandır.

Metin madenciliği sürecinin yönlendirmeler ile yapılmasını sağlayan arayüz tasarımı, veri toplama aşamasında araştırmacıya elde etmek istediği verinin boyutu hakkında bilgi vermektedir. Bu sayede araştırmacı çekmek istediği veri boyutuna ilgili anahtar kelimeler ile ulaşamazsa aradığı anahtar kelime sayısını arttırabilir. Ayrıca analiz edeceği anahtar kelime üzerinden çekilen veri boyutu çok yüksek ise performans sorunları yaşamamak adına daha kısa tarih aralıklı veri çekebilir. Ek olarak kullanılan açık kaynak kodlu programlama dilinde bulunan kütüphane aracılığıyla ilgili sosyal medya platformunun API'sine bağımlı kalınmadan veri elde edilmektedir. Bu da araştırmacıya veri boyut ve içeriği konusunda esneklik sağlamaktadır.

Süreçte ikinci adım olan veri ön işleme aşaması yapılandırılmamış metin verileri için oldukça karmaşık olabilmektedir. Bu aşamada kullanıcıya standartlaştırması gereken özel karakterlerin sayısı hakkında da destek verilmektedir. Kısaltmalar ve emojiler gibi veri

içerisinde bulunan ve dönüştürüldüğünde anlamsal olarak değerli olabilecek özel ifadelerin ise araştırmacının isteğine bağlı olarak dönüştürülmesi sağlanır. Bu sayede yapılacak duygu analizi vb. işlemlerde emojilerin ve kısaltmaların barındırdığı anlamların kaybedilmemesi sağlanır. Bu tarz dönüştürme işlemlerinin her biri arka planda farklı bir program ya da kod parçası kullanmayı gerektirir. Bu çalışmada gerçekleştirilen tasarım sayesinde teknik olmayan araştırmacıların analiz süreci kolaylaştırılmaktadır.

Filtreleme aşaması da metin madenciliğinde verinin boyutunun yönetilmesi için önemli adımlardan biridir. Analiz çıktısına etki etmeyecek durdurma kelimelerinin temizlenmesi için öncesinde kullanıcıya bu kelimeler hakkında bilgi verilir sonrasında ise temizlemesi sağlanmaktadır. Ek olarak kullanıcı kendi analizi doğrultusunda çıkarmak istediği kelimeleri bu ekran üzerinden filtreleyebilmektedir. Bunun yanı sıra araştırmacının performans sorunları yaşamaması için veride az geçen kelimelerin çıkarılması bu adımda gerçekleştirilebilir.

Son aşamada sosyal medya içeriklerinde saklı, temel unsur olan duygular için araştırmacıya duygu analiz modeli olan Turkish-BERT hakkında bilgi verilmektedir. Analiz aşamasında ayrıca topik modelleme yapılarak ilgili metnin kategorilere ayrılma işlemi gerçekleştirilebilmektedir. Ek olarak kullanıcı sözlük yöntemi ile duygu analizi gerçekleştirebilir ve kendi topik modelini oluşturabilir. Özetle araştırmacı, tasarlanan arayüz aracılığıyla belirlediği kriterlere uygun olarak verinin toplanmasını, temizleme işlemlerini ve analizini tek bir program üzerinden yönlendirmeler aracılığıyla gerçekleştirebilmektedir.

Gelecek çalışmalar için yapılabilecekler aşağıda maddeler halinde listelenmiştir.

- Veri toplama aşamasına farklı sosyal medya platformları entegre edilerek veri çeşitliliği arttırılabilir. Farklı veri kaynakları veri ön işleme ve analiz adımlarının da çeşitlendirmesini gerektirir.
- Veri toplama aşamasından hemen sonra çekilen veri Türkçe'den İngilizce'ye çevirilerek ingilizce için hali hazırda bulunan kapsamlı analiz fırsatlarından yararlanılması sağlanabilir.
- Veri toplama aşamasına dahil edilecek makine öğrenmesi süreci ile kullanıcıya çektiği anahtar kelimelere bağlı olarak ilişkili olduğu kelimeler önerilerek analiz kapsamını genişletmesi sağlanabilir.

- Kullanıcıların filtreleme aşamasında özel olarak çıkarmak istediği kelimeler kaydedilerek, benzer anahtar kelime analizi yapmak isteyen araştırmacılara bu kelimeleri filtrelemesi önerilebilir.
- Tüm analiz sürecinin yapılan işlemler, temizleme detayları ve analiz sonuçları bu çalışma sonucunda kaydedilmektedir. Yapılan analiz sayısı arttıkça burada kayda değer bir metin madenciliği eğitim seti oluşacaktır. Gelecek çalışmalarda bu veri üzerinden eğitilen makine öğrenmesi modeli tüm sürecin yönlendirmelerini besleyebilir.

Çalışmanın kısıtları aşağıda maddeler halinde listelenmiştir.

- Veri kaynaklarının ticari değeri sebebiyle birçok sosyal medya platformu veriyi sınırlı yada ücretli olarak paylaşmaktadır. Bu sebeple çalışmada veri kaynağı olarak sınırsız erişim imkanı sağlayan Twitter kullanılmıştır.
- Metin verilerinin yapılandırılmamış olmaları sebebiyle temizleme adımları oldukça karmaşık olabilmektedir. Veri kaynağına göre farklılık gösteren temizleme işlemlerini bir standart altında toplayarak tek bir süreç haline getirmek temel zorluklardandır.
- Türkçe gibi sondan eklemeli dillerde kelimenin köküne inilerek anlamının ortaya çıkarılması için kapsamlı eğitim setlerine veya sözlüklere ihtiyaç duyulmaktadır. Türkçe'de bu şekilde kapsamlı bir çalışmanın olmaması kısıtlamalardan biridir.
- Çalışmadaki topik modelleme analizinin 7 kategori barındıran hazır bir eğitim seti üzerinden gerçekleştirilmesi çalışma kapsamını sınırlandırmaktadır. Farklı konu dağılımları içeren metin verileri için bu veri setinin genişletilerek eğitilmesi gerekmektedir.

KAYNAKÇA

Açiler, S. (2020, 01 03). *Veri Nedir?* Enstitü Herkes İçin Eğitim: https://www.iienstitu.com/blog/veri-nedir adresinden alındı

Agrawal, R., & Batra, M. (2013). A Detailed Study on Text Mining Techniques. *International Journal of Soft Computing and Engineering*, 118-121.

Akbıyık, A. (2019). Sosyal Bilimlerde Metin Madenciliği WordStat Uygulamaları. *Sakarya Yayıncılık*.

Akıncı, M. (2020, 06 08). *Ngram Algoritması Nedir?* Starlang: https://starlangyazilim.com/ngram-nedir/ adresinden alındı

Aksoy, T., Çelik, S., & Gülseçen, S. (2020). Data Pre-Processing In Text Mining. *Who Runs The World: Data* (s. 123). içinde Istanbul - Turkey: Istanbul University Press.

Altunyurt, L., & Orhan, Z. (2006). Part Of Speech Tagger For Turkish. *Computer Engineering*, 7.

Arslan, E. (2020). *Makine Öğrenmesi — KNN (K-Nearest Neighbors) Algoritması Nedir?* Medium: https://arslanev.medium.com/makine-%C3%B6%C4%9Frenmesi-knn-k-nearest-neighbors-algoritmas%C4%B1-bdfb688d7c5f adresinden alındı

Ashish K. Rathore, A. K., & Ilavarasana, P. V. (2017). Social Media Analytics: Literature Review and Directions for Future Research. *Decision Analysis*, 229-249.

Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, 89-116.

Bekmamedova, N., & Shanks, G. (2014). Social Media Analytics and Business Value: A Theoretical Framework and Case Study. *Hawaii international conference on system sciences*, 3728-3737.

Biber, K. (2020). *Sosyal Ağ Analizi Nedir?* Sosyalag: http://sosyalag.com.tr/sosyal-ag-analizi-nedir/#:~:text=Bir%20a%C4%9Fdaki%20en%20etkili%20ki%C5%9Finin,sosyal%20a%C4%9F%20analizi%20ile%20cevaplanabilmektedir. adresinden alındı

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 662-679.

Brownlee, J. (2020). *Feature Selection For Machine Learning in Python.* Machine Learning Mastery: https://machinelearningmastery.com/feature-selection-machine-learning-python/ adresinden alındı

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. *In Proceedings of the first workshop on social media analytics*, 115-122.

Çetin, F. S., & Eryiğit, G. (2018). Türkçe Hedef Tabanlı Duygu Analizi İçin Alt Görevlerin İncelenmesi – Hedef Terim, Hedef Kategori Ve Duygu Sınıfı Belirleme. *Bilişim Teknolojileri Dergisi*, 43-56.

Derici, E. (2020). *Duygu Analizi Nedir, Kullanım Alanları ve Zorlukları*. Artiwise: https://www.artiwise.com/2020/07/13/sentiment-analysis/#:~:text=Duygu%20analizi%20temel%20olarak%20bir,g%C3%B6z%C3%BCndeki%20yerine%20anlamak%20i%C3%A7in%20kullan%C4%B1l%C4%B1r. adresinden alındı

Dexter, S. (2017, 1 24). *Greenbook.* Text Analytics: A Primer: https://www.greenbook.org/mr/market-research-leaders/text-analytics-a-primer/#:~:text=In%20the%20late%201990s%2C%20researchers,has%20a%20much%20longer%20history. adresinden alındı

Durmuş, M. (2021). *TF-IDF Nedir?* Github: https://mdurmuss.github.io/tf-idf-nedir/ adresinden alındı

Esuli, A. (2021). *SentiWordNet*. Github: https://github.com/aesuli/SentiWordNet adresinden alındı

Fan, W., & Gordon, M. D. (2014). The Power of Social Media Analytics. *Communications of the ACM*, 74-81.

Garg, B. R. (2018). *7 Types of Classification Algorithms*. Analytics in Diamag: https://analyticsindiamag.com/7-types-classification-algorithms/ adresinden alındı

Gökmen, B. (2020). *Latent Dirichlet Allocation(LDA) Kullanarak Nasıl Topic Modeling Yapılır?* Medium: https://medium.com/@busragokmen67/latent-dirichlet-allocation-lda-kullanarak-nas%C4%B1l-topic-modeling-yap%C4%B1l%C4%B1r-75fe8dddcdd2 adresinden alındı

Grubmüller, V., Götsch, K., & Krieger, B. (2013). Social media analytics for future oriented policy making. *European Journal of Futures Research*, 1-9.

Harman, S. (2020). *Makine Öğrenmesi | Clustering (Kümeleme) Teknikleri.* Medium: https://samed-harman.medium.com/makine-%C3%B6%C4%9Frenmesi-clustering-k%C3%BCmeleme-teknikleri-bd1b59a0a177 adresinden alındı

Hatipoğlu, E. (2018). *Machine Learning — Classification — Naive Bayes — Part 11.* Medium: https://medium.com/@ekrem.hatipoglu/machine-learning-classification-naive-bayes-part-11-4a10cd3452b4 adresinden alındı

He, W., Tian, X., Tao, R., Zhang, W., Yan, G., & Akula, V. (2016). Application of social media analytics: a case of analyzing online hotel reviews. *Online Information Review*.

J.Garbade, M. (2018). *Makine Öğreniminde Metin Özetlemeye Hızlı Bir Giriş*. Medium: https://towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning-3d27ccf18a9f adresinden alındı

K.L.Sumathy, & M.Chidambaram. (2013). Text Mining: Concepts, Applications, Tools and Issues – An Overview. *International Journal of Computer Applications*, 4.

Karaca, İ. (2020). *Python — GUI Programlama (Tkinter)*. Medium: https://medium.com/@ilyaskaraca/python-gui-programlama-tkinter-d63a99b43179 adresinden alındı

Kise, A. (2016). *Veri Kalitesinin (Data Quality) Önemi ve Yönetimi*. abdullahkise: http://www.abdullahkise.com/2016/07/veri-kalitesinin-data-quality-onemi-ve_19.html adresinden alındı

Lakhiwal, A., & Kar, A. K. (2016). Insights from Twitter Analytics: Modeling Social Media Personality Dimensions and Impact of Breakthrough Events. *In Conference on e-Business, e-Services and e-Society*, 533-544.

Lee, I. (2017). Social media analytics for enterprises: Typology, methods, and processes. *Business Horizons*, 199-210.

Liang, H., Sun, X., Sun, Y., & Gao, Y. (2017). Text feature extraction based on deep learning: a review. *EURASIP journal on wireless communications and networking*, 1-12.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 1093-1113.

Mesevage, T. G. (2020). *Best Text Classification APIs – Automatically Organize Data.* Monkey Learn: https://monkeylearn.com/blog/text-classification-apis/ adresinden alındı

Moe, W. W., Netzer, O., & Schweide, D. A. (2017). Social Media Analytics.

Moe, W. W., Netzer, O., & Schweidel, D. A. (2017). Social Media Analytics. *In Handbook of marketing decision models*, 483-504.

Nanda, P., University, S., & Noida, G. (2019). Social Media to Social Media Analytics:Ethical Challenges. *International Journal of Technoethics*, 57-70.

Noyan, M. (2019). *Doğal Dil İşleme (Natural Language Processing).* Medium: https://merveenoyan.medium.com/do%C4%9Fal-dil-i%CC%87%C5%9Fleme-natural-language-processing-2d7c72daf245 adresinden alındı

Ozyurt, B., & Akcayol, M. A. (2017). Fikir Madenciliği Ve Duygu Analizi, Yaklaşımlar, Yöntemler Üzerine Bir Araştırma. *Selçuk Üniversitesi Mühendislik, Bilim ve Teknoloji Dergisi*, 668-693.

Öğündür, G. (2019). *Python ile Sınıflandırma*. Medium: https://medium.com/@gulcanogundur/python-ile-s%C4%B1n%C4%B1fland%C4%B1rma-algoritmalar%C4%B1-74797c9c98a9 adresinden alındı

Özbek, A. (2019, 01 12). *Stemming ve Lemmatization.* Anıl Özbek: https://anilozbek.blogspot.com/2019/01/stemming-ve-lemmatization.html adresinden alındı

Park, D., Kim, W. G., & Choi, S. (2018). Application of social media analytics in tourism crisis communication. *Current Issues in Tourism*, 1810-1824.

Pradhan, V., Vala, J., & Balani, P. (2016). A Survey on Sentiment Analysis Algorithms for Opinion Mining. *International Journal of Computer Applications*, 7-11.

Przybyla, M. (2021). *How to Clean Text Data*. towardsdatascience: https://towardsdatascience.com/how-to-clean-text-data-639375414a2f adresinden alındı

Rathore, A. K., Kar, A. K., & Ilavarasan, P. V. (2017). Social Media Analytics: Literature Review and Directions. *Decision Analysis*, 229-249.

Sağlam, F., Genç, B., & Sever, H. (2019). Extending a Sentiment Lexicon with Synonym-Antonym Datasets: SWNetTR++.

Seker, S. E. (2015). Metin Madenciliği (Text Mining). *YBS ansiklopedi*, 30-32.

Seker, S. E. (2016). Duygu Analizi (Sentimental Analysis). *YBS Ansiklopedi*, 21-36.

Stieglitz, S., & Dang-Xuan, L. (2012). Social media and political communication: a social media analyics framework. *Social network analysis and mining*, 1277-1291.

Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social Media Analytics. *Business & Information Systems Engineering*, 89-96.

Sumathy, & Chidambaram. (2013). Text Mining: Concepts, Applications, Tools and Issues - An Overview. *International Journal of Computer Application*, 4.

Susenoa, Y., Laurell, C., & Sick, N. (2018). Assessing value creation in digital innovation ecosystems: A Social Media Analytics. *The Journal of Strategic Information Systems*, 335-349.

Şimşek, H. K. (2018). *Makine Öğrenmesi Dersleri 6: NLP'ye Giriş*. Medium: https://medium.com/data-science-tr/makine-%C3%B6%C4%9Frenmesi-dersleri-6-do%C4%9Fal-dil-i%CC%87%C5%9Fleme-nlp-453c3c6b062a adresinden alındı

Şirin, E. (2017). *Ensemble Yöntemler (Topluluk Öğrenmesi): Basit Teorik Anlatım ve Python Uygulama*. Veri Bilimi Okulu: https://www.veribilimiokulu.com/ensemble-

yontemler-topluluk-ogrenmesi-basit-teorik-anlatim-ve-python-uygulama/#:~:text=%C3%96zetle%20s%C3%B6yleyecek%20olursak%20ensemble%20y%C3%B6ntemler,%C3%B6%C4%9Frenme%20ba%C5%9Far%C4%B1s%C4%B1%20ve%20birbirlerinden%20 adresinden alındı

Thelwall, M. (2017). Social media analytics for YouTube comments: potential and limitations. *International Journal of Social Research Methodology*, 303-316.

Tyagi, N. (2021, 05 03). *Analytics Steps.* What is Text Mining? Process, Methods and Applications: https://www.analyticssteps.com/blogs/what-text-mining-process-methods-and-applications adresinden alındı

Tyagi, N. (2021, 05 03). *Analytics Steps.* What is Text Mining? Process, Methods and Applications: https://www.analyticssteps.com/blogs/what-text-mining-process-methods-and-applications adresinden alındı

Tyagi, N. (2021). *Top 7 Text Mining Techniques.* Analytics Steps: https://www.analyticssteps.com/blogs/top-7-text-mining-techniques adresinden alındı

Ulgen, E. (2017). *Makine Öğrenimi Bölüm-2 (k-En Yakın Komşuluk).* Medium: https://medium.com/@k.ulgen90/makine-%C3%B6%C4%9Frenimi-b%C3%B6l%C3%BCm-2-6d6d120a18e1 adresinden alındı

Ulgen, K. (2018). *Python ile Kümeleme Algoritmaları (Makine Öğrenimi Bölüm-8).* Medium: https://medium.com/@k.ulgen90/python-ile-k%C3%BCmeleme-algoritmalar%C4%B1-makine-%C3%B6%C4%9Frenimi-b%C3%B6l%C3%BCm-8-8204ffa702f2 adresinden alındı

Uslu, M. (2018). *Birliktelik Kuralları Analizi (Association Rules Analysis)*. veribilimiokulu: https://www.veribilimiokulu.com/associationrulesanalysis/#:~:text=Birliktelik%20kural%C4%B1%20analizi%2C%20t%C3%BCm%20s%C4%B1k,en%20pop%C3%BCler%20ve%20klasik%20algoritmad%C4%B1r. adresinden alındı

Ye, Z. W. (2017). Social media analytics for natural disaster management. *International Journal of Geographical Information Science*, 49-72.

Yoldash, R. (2018, 04 17). *Medium.* Agile (Çevik) Yazılım Geliştirme nedir ve nasıl uygulanır?:https://medium.com/@ryoldash/agile-%C3%A7evik-yaz%C4%B1l%C4%B1m-geli%C5%9Ftirme-nedir-ve-nas%C4%B1l-uygulan%C4%B1r-93e85ffc866 adresinden alındı

Zafarani, R. A. (2014). Social media mining: an introduction. Cambridge University Press.

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social media mining: an introduction.* Cambridge University Press.

Zeng, D., Chen, H., Lusch, R., & Li, S.-H. (2010). Social Media Analytics. *IEEE Intelligent Systems*, 13-16.

Zhao, X., Yeung, K., Huang, Q., & Song, X. (2015). Gaining Competitive Intelligence from Social Media Data: Evidence from Two Largest Retail Chains in the World. *Industrial management & data systems*.

ÖZGEÇMİŞ

Adı Soyadı : Meltem UZAVCI

ÖĞRENİM DURUMU

Derece	Eğitim Birimi	Mezuniyet Yılı
Yüksek Lisans	Sakarya Üniversitesi/ İşletme Enstitüsü/ Yönetim Bilişim Sistemleri	Devam ediyor
Lisans	Sakarya Üniversitesi/ Bilgisayar ve Bilişim Bilimleri Fakültesi/ Bilişim Sistemleri Mühendisliği	2018
Lise	Sakarya Atatürk Anadolu Lisesi	2013

İŞ DENEYİMİ

Yıl	Yer	Görev
2021-Halen	Boyner Büyük Mağazacılık	Veri ve Analitik Yöneticisi
2020-2021	FLO Mağazacılık	Veri Bilimci
2019-2020	FourDotOne Teknoloji	Jr. Veri Bilimci

YABANCI DİL

İngilizce

ESERLER

1. Uzavcı, M. ve Yılmaz, S. (2021). Giyilebilir Teknolojilerde Önde Gelen İsimlerden Birisi: Akıllı Saatler. *Uluslararası Bilimsel Araştırmalar ve Yenilikçi Çalışmalar Sempozyumu*, 835-848.