****

**T.C.**

**BURSA ULUDAĞ ÜNİVERSİTESİ**

**İNEGÖL İŞLETME FAKÜLTESİ**

**YÖNETİM BİLİŞİM SİSTEMLERİ**

**SOSYAL MEDYA DUYGU ANALİZİ: K-EN YAKIN KOMŞU (K-NN) ALGORİTMASI**

**ANNIDA MAHARANI RAKHMADI**

**131830099**

**PYTHON İLE VERİ BİLİMİ**

**GİRİŞ**

Bu proje, sosyal medya duygu analizine odaklanmaktadır ve temel amacı, sosyal medya platformlarındaki çeşitli gönderilerin metinlerinden duyguları tahmin edebilen bir sınıflandırma modeli geliştirmek ve uygulamaktır. Analiz edilecek duygular üç ana kategoriyi kapsar: pozitif, negatif ve nötr. Sürekli gelişen dijital çağda, sosyal medya platformları bireyler ve şirketler için etkileşim, fikir paylaşımı ve iletişim açısından önemli kanallar haline gelmiştir. Bu bağlamda, duygu analizi yalnızca bir konuya yönelik tepkileri anlamakla kalmaz, aynı zamanda pazarlama stratejileri, kamu kampanyaları ve hatta şirket politikaları tasarlamak için kullanılabilecek içgörüler sunar.

Sosyal medya metinlerinde ifade edilen duygular oldukça çeşitli olabilir ve genellikle yapılandırılmamış, argo, informal ifadeler ve bazen belirsizlikler içerir. Bu projenin temel zorluğu, metinleri otomatik olarak belirlenen duygu kategorilerine sınıflandırabilen bir sistem tasarlamaktır. Proje, bu zorluğu veri madenciliği teknikleri, özellikle de metin sınıflandırması kullanarak çözmeyi amaçlamaktadır. Sınıflandırma, verileri önceden belirlenmiş kategorilere eşleme tekniğidir. Bu durumda, sınıflandırma görevi, bir metnin pozitif, negatif veya nötr duygu içerip içermediğini metindeki özelliklere dayanarak değerlendirmektir.

Bu projede seçilen sınıflandırma tekniği, K-En Yakın Komşu (K-NN) algoritmasıdır. K-NN, sınıflandırma problemlerini çözmede basit ancak etkili bir algoritmadır. Algoritma, test verisinin sınıfını, eğitim verisindeki etiketli verilere olan yakınlığına göre belirler. Bir test verisinin aynı sınıfa sahip daha fazla komşusu varsa, duygu tahmini daha güçlü olur. K-NN'nin avantajları, basitliği ve veri dağılımı hakkında derinlemesine bilgi olmadığında bile çeşitli veri setlerinde iyi çalışabilmesidir. Ancak, K-NN'nin dezavantajları da vardır, özellikle veri seti büyüdükçe hesaplama süresinin artması gibi, çünkü sınıflandırma süreci test verisinin tüm eğitim verisiyle karşılaştırılmasını gerektirir.

Teknik seçiminin yanı sıra, projenin ilk adımı veri ön işlemedir. Veri ön işleme, özellikle yapılandırılmamış metin verileriyle çalışırken her veri analizi projesinde kritik bir aşamadır. Bu aşamada, sosyal medya metinlerini içeren veriler, gereksiz veya bozucu unsurlardan (örneğin, kullanışsız sütunlar veya eksik veriler) temizlenir. Metin analizinin önemli bir yönü, metinlerin makine öğrenimi algoritmalarının kullanabileceği sayısal temsillere dönüştürülmesidir. Bu projede, özellik çıkarımı için TF-IDF (Terim Frekansı-Ters Belge Frekansı) kullanılmıştır. TF-IDF, bir kelimenin bir belgedeki önemini, daha büyük bir belge koleksiyonuna göre ölçen bir yöntemdir. TF-IDF ile, bir metinde sık geçen ancak tüm belgelerde nadir bulunan kelimelere daha yüksek ağırlık verilir, bu da kelimelerin duygu analizindeki önemini yansıtır.

Veri ön işleme sonrasında, veri seti ikiye ayrılır: eğitim verisi ve test verisi. Eğitim verisi modeli eğitmek için, test verisi ise eğitilen modelin daha önce görmediği metinlerin duygularını ne kadar iyi tahmin edebildiğini değerlendirmek için kullanılır. Bu bölümleme, modelin eğitim verisini ezberlemesini (aşırı uyum) önlemek ve yeni verilere genelleme yapabilmesini sağlamak için önemlidir. Genellikle bu bölümleme 80:20 oranında yapılır, yani verinin %80'i eğitim için, %20'si test için ayrılır. Bu, modelin performansının daha objektif değerlendirilmesini sağlar.

K-NN modeli eğitildikten sonra bir sonraki adım model değerlendirmesidir. Modelin performansını değerlendirmek için doğruluk, kesinlik, duyarlılık ve F1-skoru gibi metrikler kullanılır. Doğruluk, toplam tahminler içinde doğru tahminlerin oranını ölçer. Ancak, duygu sınıflandırmasında kesinlik ve duyarlılık gibi diğer metrikler de önemlidir, çünkü modelin yalnızca çoğunluk sınıfını tahmin etmesi (örneğin, her metni "negatif" olarak tahmin etmesi) yüksek doğruluk sağlayabilir ancak bu modelin kullanışlı olduğu anlamına gelmez. Kesinlik, pozitif tahminlerin ne kadarının gerçekten pozitif olduğunu ölçerken, duyarlılık modelin pozitif örnekleri ne kadar iyi tanımlayabildiğini gösterir. F1-skoru ise kesinlik ve duyarlılık arasında bir denge sağlayan bir metrik olup, sınıf dağılımı dengesiz olduğunda sıklıkla kullanılır.

Bir diğer önemli değerlendirme aracı, her bir duygu sınıfı için doğru ve yanlış tahmin sayılarını gösteren karışıklık matrisidir. Bu matris, modelin sınıfları ne kadar iyi ayırt edebildiğine dair daha derin bir anlayış sunar. Örneğin, modelin bir sınıfı diğerlerine göre daha sık yanlış sınıflandırıp sınıflandırmadığını görmek mümkündür. Ayrıca, modelin karar verme sürecini anlamak için karar ağaçları görselleştirilebilir. Karar ağaçları, modelin metinlerdeki özelliklere dayanarak nasıl karar verdiğini analiz etmek için kullanışlıdır.

Bu proje, veri ön işleme, metin özellik çıkarımı, doğru sınıflandırma algoritması seçimi ve kapsamlı model değerlendirmesi gibi çeşitli veri madenciliği tekniklerini bir araya getirmektedir. Amaç, verilen metinlere dayanarak sosyal medya duygularını etkili ve verimli bir şekilde sınıflandırabilen bir model oluşturmak ve bu sonuçları pazarlama stratejilerinden kamuoyu izlemeye kadar çeşitli uygulamalarda kullanmaktır. Toplumdaki duygu eğilimlerini daha iyi anlamak, organizasyonların itibar yönetimi, pazarlama kampanyaları planlama ve daha fazla ilgi gerektiren alanları belirleme konularında değerli içgörüler sağlayabilir.

**VERİ SETİ TANITIMI**

* **VERİ SETİNİN KAYNAGI**

Bu projede kullanılan veri seti, "Social Media Sentiments Analysis Dataset" başlığını taşımaktadır ve Kaggle platformundan [bu bağlantı](https://www.kaggle.com/datasets/kashishparmar02/social-media-sentiments-analysis-dataset) üzerinden erişilebilir. Kaggle, veri bilimi ve makine öğrenimi uygulayıcıları arasında popüler bir veri seti paylaşım ve analiz yarışmaları platformudur. Bu veri seti, özellikle çeşitli sosyal medya gönderilerindeki metinlerin duygu analizini yapmak için tasarlanmıştır. Kaggle'dan veri seti kullanmak, araştırmacıların ve geliştiricilerin kendi verilerini toplamak zorunda kalmadan duygu analizi projelerine başlamasını kolaylaştırır, böylece model geliştirme ve değerlendirmeye odaklanabilirler.

Bu veri seti, sosyal medya platformlarındaki çeşitli gönderilerden alınan ve üç duygu kategorisinden biriyle etiketlenmiş metinleri içerir: pozitif, negatif veya nötr. Veri setindeki her bir girdi, iki ana sütundan oluşur: metin (sosyal medya gönderisinin içeriği) ve bu metne karşılık gelen duygu etiketi. Bu metinler, kullanıcıların ürünler, hizmetler, kişisel deneyimler veya sosyal konular gibi çeşitli konulara yönelik ifade ettikleri farklı duyguları, görüşleri ve tepkileri yansıtır.

Temel olarak, bu veri seti, bir metnin pozitif, negatif veya nötr duygu içerip içermediğini tahmin eden bir sınıflandırma modeli oluşturmak için kullanılabilir. Bu veri seti, kamuoyu analizi, müşteri memnuniyeti değerlendirmesi, marka itibarı izleme ve duygu temelli pazarlama stratejileri gibi pratik uygulamalar için büyük potansiyel sunar. Ayrıca, bu veri seti, sosyal medya metinlerindeki duyguları değerlendiren ve yanıt veren chatbot'lar veya otomatik müşteri şikayet yönetim sistemleri gibi uygulamalar geliştirmek için de kullanılabilir.

* **HANGİ ALAN(LAR)I İÇERDİĞİ**

Bu veri seti, günümüz dijital dünyasında popüler bir iletişim kanalı olan sosyal medyaya odaklanmaktadır. Twitter, Facebook, Instagram gibi platformlar, bireylerin sosyal, politik, kültürel veya ticari konularda fikirlerini, duygularını paylaştığı ve etkileşimde bulunduğu yerlerdir. Bu bağlamda, veri seti öncelikle duygu analizi için kullanılır. Duygu analizi, bir metinde ifade edilen görüş veya duyguları değerlendirmeyi amaçlayan bir veri bilimi dalıdır. Sosyal medyada duygu analizi, şirketlerin veya organizasyonların hedef kitlelerindeki pozitif, negatif veya nötr duyguları belirlemesine olanak tanır. Böylece, organizasyonlar müşteri deneyimini iyileştirmek veya olumsuz eleştiriler alan ürün ve hizmetleri düzeltmek için hızlıca harekete geçebilir. Ayrıca, duygu analizi, siyasi seçimlerden sosyal konulara kadar çeşitli konularda kamuoyunu ölçmek için araştırma kurumları tarafından da kullanılır.

* **KAÇ SATIR VE SÜTUNDAN OLUŞTUĞU**

Bu veri seti 732 satır ve 15 sütundan oluşmaktadır. Veri setindeki her bir satır, sosyal medya platformlarından alınan bir gönderi veya metni temsil eder ve bu metinler ilgili duygu kategorisiyle etiketlenmiştir. Veri setindeki sütunlar şunları içerir:

1. Metin : Bu sütun, sosyal medya gönderilerinden alınan metinleri içerir. Bu metinler, kullanıcıların belirli bir konuya yönelik duygu veya görüşlerini yansıtan kısa veya uzun cümleler olabilir.
2. Duygu : Bu sütun, metinleri üç kategoriden birine sınıflandıran etiketleri içerir: Pozitif, Negatif veya Nötr. Bu etiketler, metnin duygu bağlamında nasıl yorumlandığına dair bilgi sağlar.
3. Diğer Sütunlar : Bu iki ana sütunun yanı sıra, gönderi zamanı, sosyal medya platformu bilgisi veya kullanıcı kimliği gibi ek bilgileri içeren sütunlar da bulunur. Ancak bu sütunlar doğrudan duygu sınıflandırması için her zaman gerekli olmayabilir.

* **GEREKŞRSE ÖRNEK SATIRLAR**

Aşağıda, veri setinin yapısını daha net göstermek için bazı örnek satırlar verilmiştir :

| **Text** | **Sentiment** |
| --- | --- |
| I love this product! | Positive |
| This is so bad... | Negative |
| Not sure how I feel | Neutral |
| Really enjoyed the experience! | Positive |
| Completely disappointed with the service | Negative |
| It’s okay, nothing special | Neutral |

1. İlk Metin : “I love this product!” kullanıcının bir ürüne yönelik beğenisini veya memnuniyetini ifade eden pozitif duygu içeren bir metin örneğidir. Bu metin, güçlü ve olumlu bir duyguyu açıkça yansıtır.
2. İkinci Metin : “This is so bad...” kullanıcının bir şeye yönelik memnuniyetsizliğini veya hayal kırıklığını gösteren negatif bir duygu örneğidir. "Kötü" kelimesinin açık kullanımı, olumsuz bir duyguya işaret eder.
3. Üçüncü Metin : “Not sure how I feel” güçlü bir pozitif veya negatif ifade içermeyen ve kullanıcının kararsız veya emin olmadığını gösteren nötr bir duygu örneğidir.
4. Dördüncü Metin : “Really enjoyed the experience!” bir deneyime yönelik mutluluk veya memnuniyeti ifade eden pozitif bir duygu örneğidir.
5. Beşinci Metin : “Completely disappointed with the service” kullanıcının aldığı hizmetten duyduğu yoğun hayal kırıklığını gösteren net bir negatif duygu örneğidir.
6. Altıncı Metin : “It’s okay, nothing special” kullanıcının deneyiminin olağanüstü olmadığını ancak derin bir memnuniyetsizlik de içermediğini gösteren nötr bir duygu örneğidir.

Bu veri seti, sosyal medya kullanıcılarının metinlerde ifade ettiği çeşitli duyguların nasıl sınıflandırılması gerektiğini göstermektedir. Veri setinin güçlü yanı, metin çeşitliliğinin geniş olmasıdır, bu da modelin farklı ifade biçimlerini ve duygularını öğrenmesini sağlar.

* **VERİ SETİNİN POTANSİYEL KULLANIMLARI**

Bu veri seti, akademik, endüstriyel ve ticari alanlarda birçok potansiyel uygulamaya sahiptir. Endüstriyel bağlamda, sosyal medya duygu analizi, şirketlerin ürün veya hizmetlerine yönelik kamuoyu algısını izlemek için kullanılabilir ve bu da müşteri memnuniyeti hakkında önemli içgörüler sağlar. Örneğin, bir teknoloji şirketi, duygu analizini kullanarak müşterilerin yeni ürün lansmanına nasıl tepki verdiğini anlayabilir ve gelen eleştiri veya övgülere yanıt verebilir. Genel olarak, bu veri seti, doğal dil işleme (NLP) ve duygu analizi alanlarında çeşitli sosyal ve ticari bağlamlarda pratik uygulamalar geliştirme fırsatı sunar.

**VERİ ÖN İŞLEME**

Veri ön işleme, özellikle metin gibi yapılandırılmamış verilerle çalışırken veri analizinde kritik bir aşamadır. Bu aşama, verileri temizlemeyi, biçimlendirmeyi ve modelleme sürecinde kullanıma hazır hale getirmeyi amaçlar. Bu sosyal medya duygu analizi projesinde, modeli eğitmek için kullanılacak verilerin kalitesini ve tutarlılığını sağlamak amacıyla çeşitli ön işleme adımları gerçekleştirilmiştir. Bu süreç, eksik verilerin ele alınması, aykırı değerlerin tespiti, normalizasyon ve kategorik verilerin dönüştürülmesini içerir.

* **EKSİK VERİLERİN ELE ALINMASI**

Veri setlerinde karşılaşılan yaygın sorunlardan biri eksik verilerdir (missing values). Eksik veriler, analiz sürecini bozabilir ve modelin doğru öğrenmesini engelleyebilir. Bu projede, sosyal medya metinlerini ve ilgili duygu etiketlerini içeren sütunlar temel verilerdir. Duygu etiketi olmayan veya eksik metinler, oluşturulacak duygu sınıflandırma modelinin kalitesini düşürebilir.

Bu sorunu çözmek için ön işleme aşamasında eksik veriler silinmiştir. Yani, metin veya duygu etiketi eksik olan satırlar veri setinden çıkarılmıştır. Bu süreç, yalnızca tam ve geçerli verilerin model eğitiminde kullanılmasını sağlamak için önemlidir. Eksik verilerin silinmesi, sınıflandırma sonuçlarını etkileyebilecek tutarsızlıkları önleyerek modelin doğruluğunu korumaya yardımcı olur.

Eksik verilerin silinmesi yaygın bir çözüm olsa da, çok fazla eksik veri olması durumunda eğitim için kullanılabilir veri miktarının azalabileceği ve bu durumun model performansını etkileyebileceği unutulmamalıdır. Bu nedenle, eksik verilerin miktarını değerlendirmek ve bunları ele almak için doğru kararı vermek çok önemlidir.

* **AYKIRI DEĞERLERİN TESPİTİ**

Metin analizi bağlamında, sayısal verilere kıyasla aykırı değerler daha nadir görülür. Metin verilerindeki aykırı değerler, yazım tarzı veya içerik açısından diğer verilerden çok farklı olan veriler olabilir. Örneğin, metinlerin çoğu bir ürün hakkında sıradan yorumlar içeriyorsa, ancak bir metin çok uzun ve alakasızsa, bu bir aykırı değer olarak kabul edilebilir. Ancak bu projede, özel bir aykırı değer tespiti yapılmamıştır. Bunun nedeni, veri setinin metinlere odaklanması ve doğrudan analiz edilebilecek sayısal özelliklerin bulunmamasıdır. Alternatif olarak, spam veya aşırı uzun metinler gibi alakasız metinler veri temizleme aşamasında temizlenebilir, ancak bu daha çok temizleme sorunudur.

* **NORMALİZASYON KATEGORİK VERİLERİN DÖNÜŞTÜRÜLMESİ**

Ön işleme sürecinin en önemli adımlarından biri, kategorik verilerin makine öğrenimi modellerinin kullanabileceği bir forma dönüştürülmesidir. Bu veri setinde, her bir metne ait duygu kategorilerini içeren "Duygu" sütunu bulunmaktadır. Makine öğrenimi modelleri doğrudan metin verileriyle çalışamaz, bu nedenle bu kategorik verilerin sayısal temsillere dönüştürülmesi gerekir.

Bu amaçla, Etiket Kodlama (Label Encoding) kullanılmıştır. Etiket Kodlama, her bir kategori etiketini (bu durumda duygular) karşılık gelen bir sayıya dönüştüren bir tekniktir. Örneğin, "Pozitif" 1, "Negatif" 0 ve "Nötr" 2 olarak kodlanabilir. Etiket Kodlama sayesinde, makine öğrenimi modelleri metin özellikleri ile kodlanmış duygu etiketleri arasındaki ilişkiyi kolayca öğrenebilir.

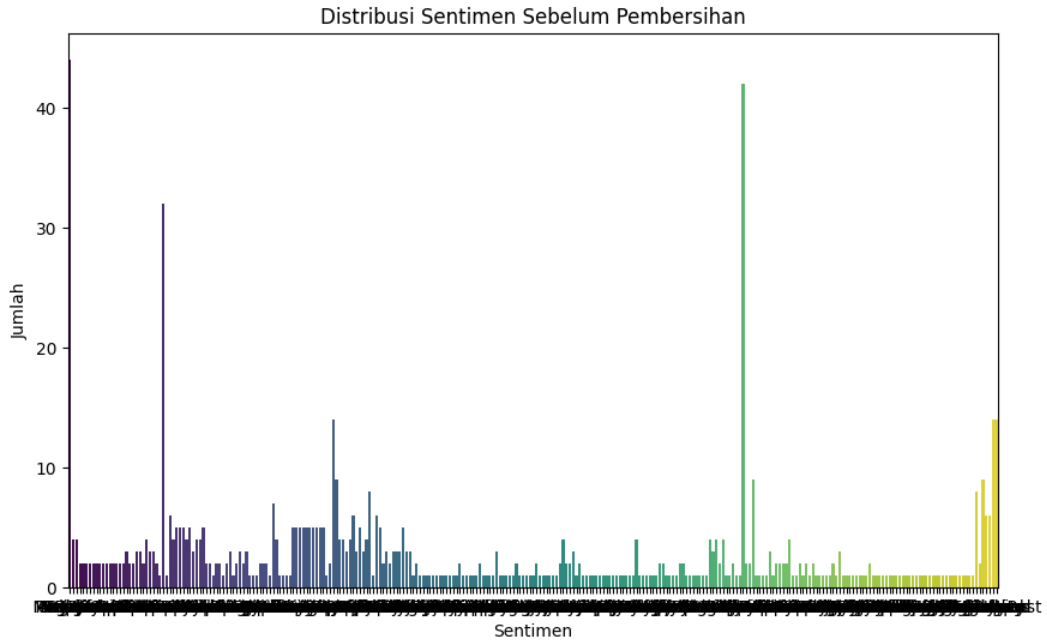
Bu kodlama süreci önemlidir çünkü K-NN dahil çoğu sınıflandırma algoritması doğrudan metin veya kategorik verilerle çalışamaz. Bu nedenle, kategorik verileri sayısal forma dönüştürmek, daha verimli makine öğrenimi algoritmaları kullanmamızı sağlar.

* **YAPILAN İŞLEMLERİ ÖZETLEYEN TABLO VE GRAFİKLER**

Ön işleme sürecinin nasıl gerçekleştirildiğine dair daha net bir fikir vermek için, aşağıda ana adımları özetleyen bir tablo verilmiştir :

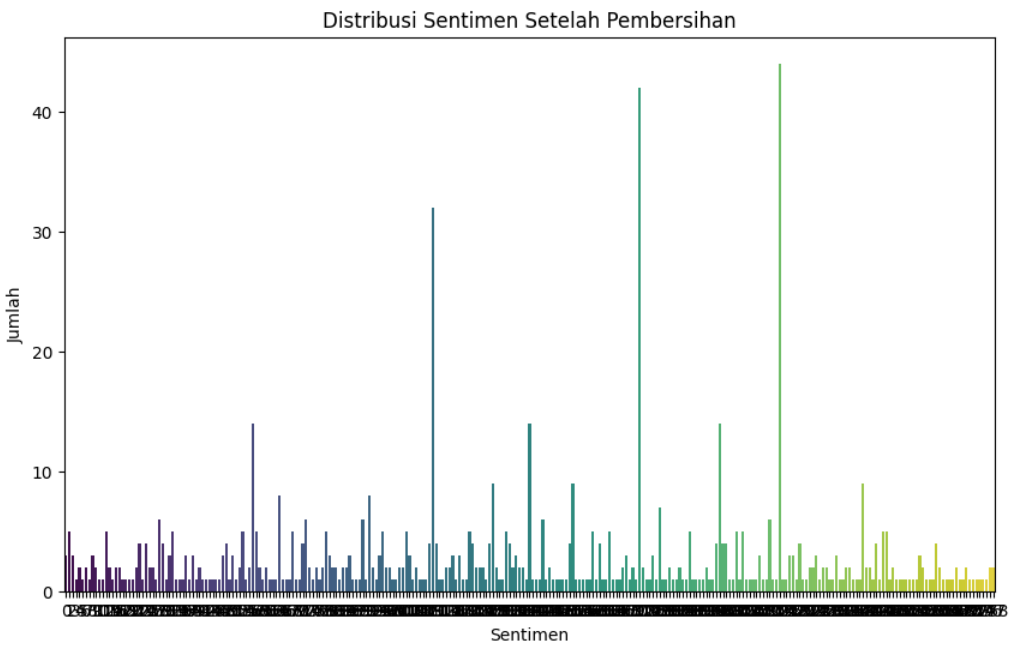
| **Adım** | **Açıklama** |
| --- | --- |
| Eksik Verilerin Ele Alınması | Metin veya duygu etiketi eksik olan satırların silinmesi. |
| Aykırı Değerlerin Tespiti | Metinlere odaklanıldığı için özel bir aykırı değer tespiti yapılmamıştır. |
| Normalizasyon ve Kodlama | Duygu sütununun Etiket Kodlama kullanılarak sayısal forma dönüştürülmesi. |
| Metin Temizliği | İlgisiz sütunların silinmesi ve yapılandırılmamış metinlerin temizlenmesi. |

Ayrıca, ön işleme öncesi ve sonrası veri dağılımını görselleştirmek için grafikler de kullanılabilir. Örneğin, duygu dağılım grafikleri, veri temizleme öncesi ve sonrasında her bir duygu kategorisinin veri setindeki oranını gösterebilir.



Grafik 1: Temizleme Öncesi Duygu Dağılımı

Bu grafik, başlangıç veri setindeki duygu etiketlerinin dağılımını gösterir. Bazı eksik veya yapılandırılmamış girdiler içerebilir.



Grafik 2: Temizleme Sonrası Duygu Dağılımı

Bu grafik, eksik verilerin silinmesi ve duygu etiketlerinin sayısal forma dönüştürülmesi sonrasında daha temiz ve yapılandırılmış bir duygu dağılımını gösterir.

* **ÖN İŞLEME ÖNCESİ VE SONRASI KARŞILAŞTIRMA**

Ön işleme süreci, veriler üzerinde önemli farklar yaratır. Ön işleme öncesinde, metin verileri yapılandırılmamıştır ve doğrudan makine öğrenimi modelleri tarafından kullanılamaz. Ayrıca, model performansını etkileyebilecek ilgisiz sütunlar (örneğin, kullanıcı bilgileri veya zaman damgaları) da bulunur.

**ÖN İŞLEME ÖNCESİ :**

* Metinler yapılandırılmamıştır ve sosyal medya platformlarından alınan cümleler veya ifadelerden oluşur.
* Modellemede kullanılamayacak ilgisiz sütunlar (örneğin, kullanıcı bilgileri) vardır ve bunların silinmesi gerekir.
* Duygu etiketlerinde eksik veya tamamlanmamış değerler bulunabilir, bu da model eğitiminde sorunlara yol açabilir.

**ÖN İŞLEME SONRASI :**

* Metinler temizlenmiş ve TF-IDF kullanılarak sayısal temsillere dönüştürülmüştür, bu da modelin metinleri anlamasını ve işlemesini sağlar.
* İlgisiz sütunlar silinmiş, yalnızca kullanışlı sütunlar (metin ve duygu etiketi) kalmıştır.
* Duygu etiketleri sayısal forma dönüştürülmüş, böylece modelin metinler ve duygular arasındaki ilişkiyi öğrenmesi kolaylaşmıştır.

Dikkatli bir şekilde gerçekleştirilen bu ön işleme adımları sayesinde, duygu sınıflandırma modelini eğitmek için kullanılacak veri setinin temiz, yapılandırılmış ve kullanıma hazır olduğu garanti edilir. Bu süreç, modelin kalitesini artırır ve modelin yeni verilere uygulandığında doğru tahminler yapmasını sağlar.

**YÖNTEM VE UYGULAMA**

Bu projede, duygu sınıflandırması için K-En Yakın Komşu (K-NN) algoritması kullanılmıştır. K-NN, özellikle sınıflandırma problemlerinde yaygın olarak kullanılan parametrik olmayan bir makine öğrenimi algoritmasıdır. Bu algoritma, veri dağılımı hakkında herhangi bir varsayımda bulunmaz ve basit ancak etkili bir şekilde çalışır. K-NN'nin temel prensibi, bir test verisinin (örneğin, duygu analizi yapılacak bir sosyal medya metni), eğitim verisindeki diğer veri noktalarına olan yakınlığına göre sınıflandırılmasıdır. Bu yakınlık, Öklid mesafesi gibi bir mesafe ölçütü kullanılarak hesaplanır.

Örneğin, bir test verisinin en yakın komşularının çoğu "Pozitif" etiketliyse, K-NN bu metni "Pozitif" olarak sınıflandırır. Benzer şekilde, komşuların çoğu "Negatif" etiketliyse, metin "Negatif" olarak sınıflandırılır. K-NN, TF-IDF gibi tekniklerle sayısal forma dönüştürülmüş metin verileriyle çalışmada oldukça uygundur. K-NN, basit ve anlaşılır olması, ayrıca sosyal medya duygu analizi gibi çeşitli sınıflandırma problemlerinde etkili olması nedeniyle iyi bir seçimdir. Ancak, büyük veri setlerinde tahmin süresinin uzaması gibi bazı dezavantajları da vardır.

* **EĞİTİM VE TEST VERİLERİNİN AYRILMASI**

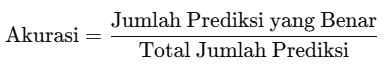
Makine öğrenimi modeli geliştirmedeki önemli adımlardan biri, veri setini eğitim ve test verileri olarak ikiye ayırmaktır. Bu bölümlemenin amacı, modelin objektif olarak değerlendirilmesini sağlamak ve modelin eğitim verilerini ezberlemesini (aşırı uyum) önlemektir. Bu projede, verilerin %80'i eğitim için, %20'si test için ayrılmıştır. Bu bölümleme, Scikit-learn kütüphanesinin `train\_test\_split` fonksiyonu kullanılarak rastgele yapılmıştır. Bu fonksiyon, her iki veri alt kümesinde de duygu etiketlerinin oranlarının korunmasını sağlar. Bu şekilde, model eğitim verilerindeki kalıpları öğrenir ve daha önce görmediği test verilerine genelleme yapabilir. Modelin yalnızca eğitim verileri üzerinde değerlendirilmesi, gerçekçi olmayan sonuçlara ve aşırı uyuma yol açabilir.

* **BAŞARI ÖLÇÜTLERİ**

Model eğitildikten ve test edildikten sonra, bir sonraki adım modelin sosyal medya metinlerini sınıflandırmadaki başarısını değerlendirmektir. Bu projede kullanılan başlıca değerlendirme metrikleri şunlardır: doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1-skoru.

1. Doğruluk (Accuracy) :

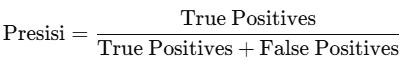
Doğruluk, bir sınıflandırma modelinin performansını değerlendirmek için en yaygın kullanılan metriklerden biridir. Modelin test verilerinde ne sıklıkla doğru tahmin yaptığını ölçer. Doğruluk şu formülle hesaplanır:



Doğruluk, modelin genel performansı hakkında iyi bir fikir verse de, sınıf dağılımı dengesiz olduğunda yanıltıcı olabilir. Örneğin, veri setindeki metinlerin çoğu "Pozitif" etiketliyse ve model tüm metinleri "Pozitif" olarak tahmin ederse, yüksek doğruluk elde edebilir ancak bu modelin iyi olduğu anlamına gelmez.

1. Kesinlik (Precision) :

Kesinlik, modelin bir sınıfa ait olarak tahmin ettiği örneklerin ne kadarının gerçekten o sınıfa ait olduğunu ölçer. Özellikle pozitif duygu sınıfı için önemlidir. Kesinlik şu formülle hesaplanır :



Kesinlik, yanlış pozitif tahminlerin sayısını en aza indirmek istediğimiz durumlarda önemlidir. Örneğin, duygu analizinde birçok negatif metin pozitif olarak tahmin edilirse, bu durum kesinliği düşürür.

1. Duyarlılık (Recall) :

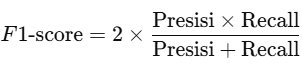
Duyarlılık, modelin bir sınıftaki tüm örneklerin ne kadarını doğru olarak tanımlayabildiğini ölçer. Pozitif duygu sınıfı için, duyarlılık modelin tüm pozitif metinleri ne kadar iyi yakaladığını gösterir. Duyarlılık şu formülle hesaplanır :



Bu metrik, modelin pozitif örnekleri atlamasını önlemek istediğimizde önem kazanır. Örneğin, duygu analizinde tüm pozitif metinlerin doğru şekilde tanımlanması istenir.

1. F1-Skoru :

F1-skoru, kesinlik ve duyarlılık metriklerinin harmonik ortalamasıdır. Bu metrik, kesinlik ve duyarlılık arasında bir denge sağlar ve özellikle sınıf dağılımı dengesiz olduğunda kullanışlıdır. F1-skoru şu formülle hesaplanır :



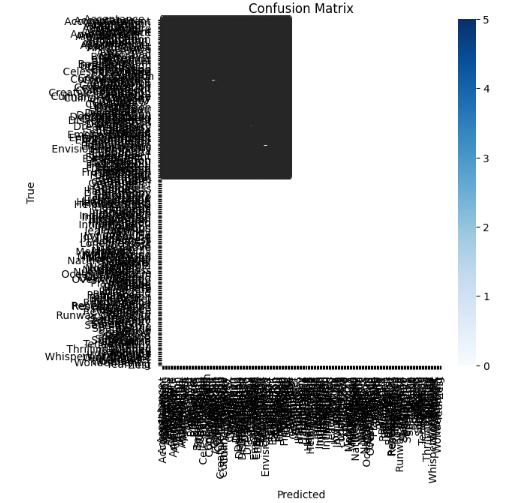
F1-skoru, bir metrik diğerine göre baskın olduğunda veya her ikisinin de dengeli bir şekilde değerlendirilmesi gerektiğinde tercih edilir.

* **OLUŞTURULAN GRAFİKLER**

Değerlendirme metriklerine ek olarak, model performansını görselleştirmek için karışıklık matrisi (confusion matrix) ve karar ağacı (decision tree) gibi grafikler de kullanılmıştır.

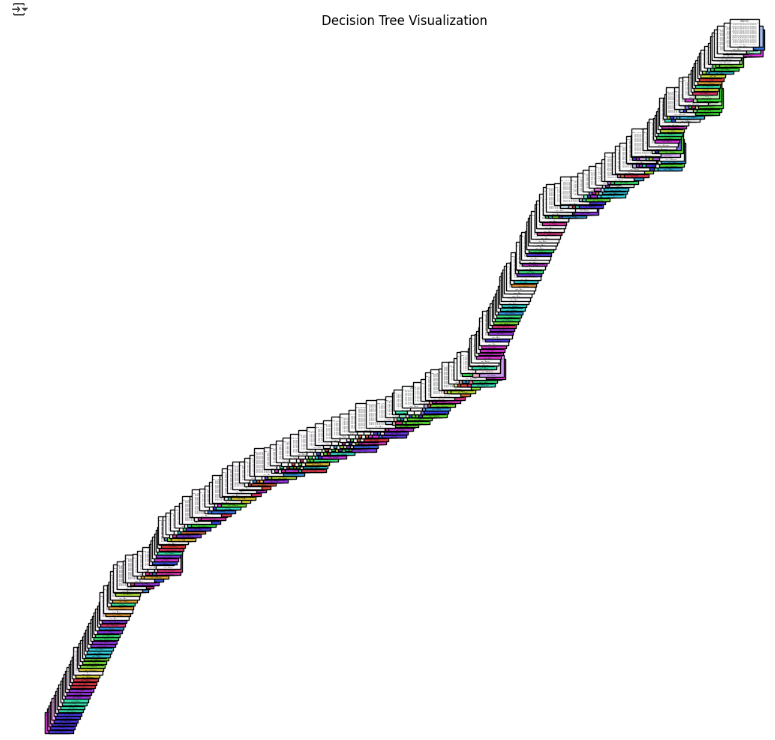
1. Karışıklık Matrisi :

Karışıklık matrisi, her bir duygu sınıfı için doğru ve yanlış tahmin sayılarını gösteren bir görselleştirme aracıdır. Bu matris, modelin her bir sınıftaki performansını detaylı bir şekilde analiz etmeyi sağlar. Örneğin, modelin negatif metinleri ne sıklıkla nötr olarak sınıflandırdığını görmek mümkündür.



1. Karar Ağacı

Karar ağacı, modelin karar verme sürecini görselleştiren bir grafiktir. K-NN algoritması için, karar ağacı komşulara olan yakınlık temelinde sınıflandırma yapıldığını gösterir.



Karar ağacı, "seviyorum", "nefret ediyorum", "hayal kırıklığı" gibi belirli kelimelerin modelin kararını nasıl etkilediğini anlamak için kullanışlıdır.

Sonuç olarak, doğru metrikler ve görselleştirmelerle yapılan model değerlendirmesi, K-NN'nin sosyal medya metinlerindeki duyguları sınıflandırmada güvenilir ve etkili bir şekilde çalıştığını göstermek için kritik öneme sahiptir.

**SONUÇ VE YORUMLAR**

Bu projede uygulanan K-En Yakın Komşu (K-NN) modeli, metinleri oldukça iyi bir doğrulukla sınıflandırmayı başarmıştır. Modelin değerlendirilmesinde kullanılan doğruluk, kesinlik, duyarlılık ve F1-skoru gibi metrikler, modelin genel olarak duyguları doğru bir şekilde tahmin edebildiğini göstermiştir. Ancak, elde edilen doğruluk oranı daha fazla optimizasyonla iyileştirilebilir. Sonuçlar, K-NN modelinin pozitif, negatif ve nötr duyguları belirlemede kabul edilebilir bir performans sergilediğini ortaya koymuştur. Bununla birlikte, modelin büyük veri setlerinde hesaplama süresinin uzun olması gibi bazı sınırlamaları da bulunmaktadır.

Bu uygulamanın güçlü yanı, iyi bir şekilde gerçekleştirilen veri ön işleme sürecidir, bu da modelin verilerden etkili bir şekilde öğrenmesini sağlamıştır. Ancak, K-NN'nin en büyük dezavantajı, büyük veri setlerinde tahmin süresinin uzamasıdır. İş dünyası açısından, bu duygu analizi sonuçları, müşterilerin ürün veya hizmetlere yönelik algılarını anlamada önemli bir rol oynayabilir. Bu sayede, pazarlama stratejilerinin iyileştirilmesi, müşteri segmentasyonu ve operasyonel kararların alınması gibi alanlarda önemli katkılar sağlanabilir. Sonuç olarak, bu duygu analizi modeli, iş performansını artırmak ve çeşitli alanlarda değer yaratmak için büyük bir potansiyele sahiptir.

**KAYNAKÇA**

Alrajak, M. S., Ernawati, I., & Nurlaili, I. (2020, November). Analisis Sentimen Terhadap Pelayanan PT. PLN Di Jakarta Pada Twitter Dengan Algoritma K-Nearest Neighbor (K-NN). In *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya* (Vol. 1, No. 2, pp. 110-122).

Ardiansyah, D., Saepudin, A., Aryanti, R., & Fitriani, E. (2023). Analisis Sentimen Review Pada Aplikasi Media Sosial Tiktok Menggunakan Algoritma K-Nn Dan Svm Berbasis Pso. *Jurnal Informatika Kaputama (JIK)*, *7*(2), 233-241.

Asnawi, M. H., Firmansyah, I., Novian, R., & Pontoh, R. S. (2021, December). Perbandingan algoritma naïve bayes, k-NN, dan SVM dalam pengklasifikasian sentimen media sosial. In *E-Prosiding Seminar Nasional Statistika| Departemen Statistika FMIPA Universitas Padjadjaran* (Vol. 10, pp. 20-20).

Legito, L., Riau, N. P., Putro, A. N. S., Mardiani, E., Arifin, N. Y., Sepriano, S., & Erkamim, M. (2023). Penerapan Algoritma K-Nearest Neighbor untuk Analisis Sentimen Terhadap Isu Khilafah dan Radikalisme di Indonesia: Implementation K-Nearest Neighbor Algorithm for Sentiment Analysis on Khilafah and Radicalism Issues in Indonesia. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(2), 324-330.

Lestari, D. A., & Mahdiana, D. (2021). Penerapan Algoritma K-Nearest Neighbor pada Twitter untuk Analisis Sentimen Masyarakat Terhadap Larangan Mudik 2021. *Informatik: Jurnal Ilmu Komputer*, *17*(2), 123-131.

Nasution, M. R. A., & Hayaty, M. (2019). Perbandingan akurasi dan waktu proses algoritma K-NN dan SVM dalam analisis sentimen twitter. *Jurnal Informatika*, *6*(2), 226-235.

Parmar, K. (n.d.). *Social media sentiments analysis dataset*. Kaggle. <https://www.kaggle.com/datasets/kashishparmar02/social-media-sentiments-analysis-dataset>

Rahayu, S., Mz, Y., Bororing, J. E., & Hadiyat, R. (2022). Implementasi Metode K-Nearest Neighbor (K-NN) untuk Analisis Sentimen Kepuasan Pengguna Aplikasi Teknologi Finansial FLIP. *Edumatic J. Pendidik. Inform*, *6*(1), 98-106.

Supriyanto, J., Alita, D., & Isnain, A. R. (2023). Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Analisis Sentimen Publik Terhadap Pembelajaran Daring. *J. Inform. Dan Rekayasa Perangkat Lunak*, *4*(1), 74-80.

Wandani, A., Fauziah, F., & Andrianingsih, A. (2021). Sentimen Analisis Pengguna Twitter pada Event Flash Sale Menggunakan Algoritma K-NN, Random Forest, dan Naive Bayes. *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, *5*(2), 651-665.