Kümeleme Tabanlı Metin Sınıflandırma: Cinsiyet Çıkarımı Örneği

Ensar Şehitoğlu

Atatürk Üniversitesi Bilgisayar Mühendisliği Fakültesi, Erzurum, Türkiye

Özet

Makine öğrenimi, günümüzde birçok sektörde etkili bir şekilde kullanılmaktadır. Bu çalışmada, kümeleme tabanlı yaklaşımlar kullanılarak Türkçe ve İngilizce metinler üzerinde cinsiyet tahmini probleminin çözülmesi amaçlanmıştır. Türkçe ve İngilizce metinler üzerinde cinsiyet çıkarımı yapmak amacıyla kümeleme tabanlı bir metin sınıflandırma yaklaşımı geliştirilmiştir. Veri ön işleme kısmında metinlerde gereksiz karakterler, sayılar, URL’ler ve emojiler temizlenmiş; durak kelimeler çıkarılmış, metinler standart bir forma dönüştürülmüş ve her iki dil için dil işleme kuralları uygulanmıştır. Geleneksel metin temsil yöntemleri (TF-IDF, Bag of Words ve n-gram özellikleri) ile modern gömülü kelime temsil modelleri (Word2Vec, GloVe ve FastText) kullanılmıştır ayrıca Transformer tabanlı modern dil modelleri (örneğin, BERT, RoBERTa) ile metinlerin derin bağlamsal temsilleri elde edilerek sınıflandırma performansını artırmak için kullanılmıştır. Kümeleme algoritmalarıyla sınıflandırma özellikleri zenginleştirilmiştir. Sınıflandırma modelleri arasında Naive Bayes, Lojistik Regresyon, Destek Vektör Makineleri, Karar Ağaçları ve Rastgele Orman yöntemleri denenmiştir ve performans değerlendirmesi için çapraz doğrulama ve çeşitli metrikler (doğruluk, F1-skoru, duyarlılık ve kesinlik) kullanılarak kapsamlı bir analiz yapılmıştır.

*Anahtar Kelimeler:* Kümeleme tabanlı sınıflandırma, metin işleme, cinsiyet çıkarımı, transformer modelleri, dil işleme

1. Giriş

Sosyal medya platformları, kullanıcıları hakkında büyük miktarda veri üreten çevrimiçi iletişim ve topluluk oluşturma araçları olarak giderek daha önemli hale geldi. Bu veriler, pazarlama ve reklamcılık, kamuoyu araştırmaları ve sosyal bilimler araştırmaları dahil olmak üzere çeşitli alanlarda değerli bilgiler sağlayabilir. Kullanıcıların demografik özellikleri hakkında bilgi edinmek, bu platformlarda kullanıcı davranışlarını anlamak için çok önemlidir. Bu bağlamda, cinsiyet tahmini, kullanıcıların sosyal medya etkinliklerine dayanarak cinsiyetlerini otomatik olarak belirlemeyi amaçlayan önemli bir görev haline gelmiştir [1] Twitter, basit bir veri modeline ve anlaşılır veri erişim API'sine sahip olduğundan, çevrimiçi davranış kalıplarını, duygu analizini, cinsiyet tesbiti ve spam, botlar, sahte haberler ve nefret söylemi gibi tehditleri analiz eden sosyal ağ çalışmaları için idealdir. Bu sebeble, bu çalışmada cinsiyet belirlemeyi sınıflandırmak için sosyal medya ortamı Twitter üzerinden elde edilen mesajlar kullanılmıştır [59]. Son yirmi yıl içerisinde, bilgisayar ortamında üretilen belgelerin sayısı sürekli bir artış göstermektedir. Özellikle internetin yaygınlaşması, kişisel bilgisayarların, akıllı telefonların ve tabletlerin daha erişilebilir hale gelmesi, teknolojinin gelişmesi ve toplumun dijitalleşmesi, dijital ortamda üretilen dokümanların sayısının artmasına neden olmuştur. Ancak, bu artışın beraberinde getirdiği birçok fayda ile birlikte çeşitli sorunların da ortaya çıkmasına neden olmuştur. Bu bağlamda, karşılaşılan zorluklardan biri de dijital ortamda yer alan metinlerin sınıflandırılması problemidir. Metin sınıflandırma, genel olarak bir metnin önceden belirlenmiş sınıflardan hangisine ya da hangilerine ait olduğunun belirlenmesi anlamına gelir. Bu sorun, belge sınıflandırma veya metin kategorilerinin belirlenmesi gibi farklı isimlerle de anılmaktadır. Bu zorlukların çözülmesi, elektronik metinlerin etkili bir şekilde yönetilmesi ve kullanılmasını sağlamak adına önemli bir adımdır [2]. Metin kategorizasyonu, Doğal Dil İşleme alanında temel bir görevdir. Metin verilerinin hızla artması ve otomatik etiketleme ihtiyacı, metin kategorizasyon probleminin önemini artırmaktadır. [3]. Metin sınıflandırma süreçlerinde karşılaşılan önemli bir zorluk, belli bir formata sahip olmadıkları için, doğrudan sınıflandırılamazlar. Belirli bir formatı olmayan metinlerin başarılı bir sınıflandırma için, öncelikle metinlerin önişlemden geçirilmesi gereklidir [4]. Bu çalışmada, talep verileri makine öğrenimi teknikleri kullanılarak sınıflandırılmıştır. Bu amaçla, öncelikle ham metin verileri, anlamsız karakterlerden ve tekrarlardan arındırılmıştır [5]. Özellik çıkarımı, makine öğrenmesi ve özellikle de yüksek boyutlu verilerin analizinde kritik bir adımdır. Mikroarray verileri gibi yüksek boyutlu veri kümelerinde, özellik çıkarımı, verilerin karmaşıklığını azaltmak, hesaplama maliyetini düşürmek ve sınıflandırma doğruluğunu artırmak için kullanılır [6]. Kelime gömme modelleri, kelimeleri sayısal vektörler olarak temsil ederek, anlamsal ve sözdizimsel ilişkilerini yakalamayı amaçlayan güçlü araçlardır. Doğal dil işleme (NLP) görevlerinde yaygın olarak kullanılırlar ve anlamsal analiz, bilgi alma, bağımlılık ayrıştırma, soru cevaplama ve makine çevirisi gibi alanlarda önemli rol oynarlar [7]. Transformer mimarisi, büyük metin külliyatlarında ön eğitim için özellikle uygundur ve metin sınıflandırması, dil anlama, makine çevirisi, ortak referans çözünürlüğü, sağduyu çıkarımı ve özetleme gibi aşağı akış görevlerinde doğrulukta büyük kazanımlar sağlamıştır [8]. Büyük Dil Modelleri (LLM'ler), yapay zeka alanında özellikle dil işlemede son yıllarda çığır açan bir gelişmeyi temsil etmiştir. Bu modeller, Turing Testi'nin 1950'lerde ortaya konulmasından bu yana dil zekasını makinelerde ustalıkla kullanma arayışımızda önemli bir adım niteliğindedir [9]. Kümeleme, denetimsiz öğrenme tekniklerinden biridir ve etiketlenmemiş veriler içindeki kalıpları belirlemede oldukça etkilidir [10]. Sınıflandırma sürecinde, verilerin modelin eğitim ve test aşamaları için ayrılması gereklidir. Bu bağlamda, veri kümesi rastgele olarak eğitim ve test kümelerine bölünmektedir. Ancak, bu işlem veri kümesinin yapısına bağlı olarak modelin eğitim ve test süreçlerinde sapma ya da aşırı öğrenme gibi sorunlara neden olabilir [11].

1. Literatür Taraması

Makalenin bu bölümü bu çalışmanın konusu olan metin sınıflanrıdma, kümeleme , makine öğrenimi algoritmaları kullanarak sınıflandır ve cinsiyet çıkarımını tahmin etmek veya tespit etmek için yapılan bazı araştırma çalışmalarına adanmıştır.

Bu çalışmada, sosyal ağların hızlı büyümesiyle elde edilen büyük kullanıcı verilerinin, yazarlık analizi ve cinsiyet gibi bilgilerin çıkarılmasında önemli fırsatlar sunduğu vurgulanmıştır. Twitter, kullanıcı verilerini API aracılığıyla sunarak bu veriler için değerli bir kaynak oluşturmuştur. Çalışmada, tweet metninden seçilen 1 ila 5 gramlık n-gram özellikleriyle Perceptron ve Naive Bayes algoritmaları kullanılarak Twitter kullanıcılarının cinsiyeti belirlenmeye çalışılmıştır. N-gram özelliklerinin, tweetlerin hızlı akışını ve hacmini ele almak için akış uygulamalarında kullanıldığı, ayrıca geleneksel sözlük yöntemleriyle değerlendirilemeyen resmi olmayan tweetlerin n-gramlarla temsil edildiği belirtilmiştir. Seçilen en iyi n-gram özellikleri ile Naive Bayes ve Perceptron algoritmaları, %99'un üzerinde doğruluk, dengeli doğruluk ve F1-score elde etmiştir [12]. Bu araştırma, Twitter'da cinsiyet tahmini doğruluğunu artırmak amacıyla çeşitli yöntemler kullanarak mevcut veri setini genişletmiştir. Çalışma, Twitter Kullanıcı Cinsiyet Sınıflandırması veri setini 296.108 ek tweet ile genişletmiş ve veri ön işleme işlemleriyle modelin eğitilmesine uygun hale getirilmiştir. Ayrıca, GloVe, BERT, GPT2 ve Word2Vec gibi gelişmiş kelime gömme modelleri kullanılarak kelimeler arasındaki anlamsal ilişkiler yakalanmıştır. Özellikle GloVe modeli, 2 milyar tweet üzerinde eğitilen versiyonuyla en yüksek doğruluk oranını elde etmiştir. Profil açıklamalarının dahil edilmesi de doğruluğu artırmış ve cinsiyet tahmininde önemli bir katkı sağlamıştır. Çalışmada, Naive Bayes, Rastgele Orman, Karar Ağacı ve XGBoost gibi farklı makine öğrenimi algoritmalarının performansları değerlendirilmiş ve GloVe modeliyle birleştirilen Rastgele Orman algoritması %70 doğrulukla en iyi performansı göstermiştir [13]. Araştırma, Twitter'da cinsiyet sınıflandırması için profil özelliklerinin kullanımını incelemiştir. Kullanıcı profillerinden elde edilen veriler, cinsiyet tahmini için etkili bir kaynak olarak değerlendirilmiştir. Araştırmada fonem tabanlı özellik çıkarımı, renk kuantizasyonu ve sıralama gibi teknikler kullanılarak isimler, kullanıcı adları ve profil renkleri analiz edilmiştir. Fonem tabanlı yaklaşım, isim analizine dayalı olarak %82.5 doğruluk oranı sağlarken, renk tabanlı özelliklerle yapılan analizlerde NB-Tree sınıflandırıcısı %74 doğruluk oranına ulaşmıştır [14]. Bu çalışmada, sosyal ağ metinlerinin duygu analizi için iki yönlü kısa ve uzun vadeli bellek modeli (AT-BiLSTM) önerilmiştir. Bu yöntem, BERT modeliyle metinlerin vektörleştirilmesini sağlamış ve dikkat mekanizması ile entegre edilen BiLSTM ile yön düzeyinde duygu analizi gerçekleştirmiştir. Aynı veri kümesi üzerinde yapılan değerlendirmelerde, önerilen yöntemin %93,72 doğruluk, %93,91 geri çağırma ve %92,38 F1 puanı ile alternatif yöntemlere kıyasla daha yüksek performans sergilediği tespit edilmiştir [15]. Bu çalışmada, metin sınıflandırmasının doğal dil işleme alanında yaygın bir uygulama olduğu ve derin öğrenmenin bu alandaki en sık kullanılan yöntemlerden biri olduğu tespit edilmiştir. Mevcut zorlukları aşmak için BERT-BiGRU modeli önerilmiştir. Bu modelde, kelime vektörleri için geleneksel word2vec yerine BERT kullanılmış ve bağlam bilgisine dayalı kelime temsiline olanak tanınmıştır. Ardından, iki yönlü özellik çıkarımı için BiGRU modeli BERT'e entegre edilmiştir. Deneysel sonuçlara göre, önerilen BERT-BiGRU modeli ile doğruluk, geri çağırma ve F1 puanı 0.9'un üzerinde değerler elde edilmiştir. Modelin, özellikle Çince metin sınıflandırma görevlerinde yüksek performans gösterdiği görülmüştür [16]. Bu çalışmada, bireylerin biyolojik cinsiyetinin adlara dayalı sınıflandırılması üzerine birçok çalışma yapıldığı, ancak Vietnamca adlarla ilgili çalışmaların sınırlı olduğu görülmüştür. Vietnamca adlara dayalı cinsiyet tahmini için 26.000’den fazla tam ad içeren ve cinsiyet etiketleriyle anotasyon yapılmış yeni bir veri seti önerilmiştir. Bu veri seti araştırma amaçlı olarak erişime sunulmuştur. Ayrıca, altı makine öğrenimi algoritması (SVM, Multinomial Naive Bayes, Bernoulli Naive Bayes, Karar Ağacı, Rastgele Orman ve Lojistik Regresyon) ile bir derin öğrenme modeli (LSTM) üzerinde fastText kelime gömme kullanılarak deneyler yapılmıştır. Sonuçlara göre, en yüksek F1 puanı %96 ile LSTM modeliyle elde edilmiştir ve bu model temel alınarak bir web API geliştirilmiştir [17]. Bu çalışmada, Evalita 2018 GxG çapraz tür/alan cinsiyet tahmini yarışmasında İtalyanca metinler üzerine bir araştırma gerçekleştirilmiştir. N-gram özelliklerini kullanan doğrusal bir modelin farklı türler arası dayanıklılığı değerlendirilmiştir. Eğitim ve test türlerinin farklı olduğu durumlarda model performansının belirgin şekilde düştüğü gözlemlenmiştir. Tür bağımsız özellikler yakalamak için soyut özellikler de test edilmiştir. Çalışmada, resmi in-genre test setinde %55 F1 puanı ile birinci sırada yer alınmış, cross-genre test setinde ise %51 F1 puanı elde edilmiştir. Bu sonuçlar, türler arası cinsiyet tahmininde mevcut modellerin sınırlılıklarını göstermiştir [18]. Bu çalışmada, Hint isimlerinin cinsiyet tahmini için Destek Vektör Makinesi (SVM) tabanlı bir sınıflandırma yaklaşımı sunulmuştur. İlk olarak, morfolojik analizlere dayalı olarak bu tür bir sınıflandırma için faydalı olabilecek çeşitli özellikler belirlenmiş ve değerlendirilmiştir. Ardından, bu özelliklerle birlikte n-gram eklerinin kullanımına dayalı yeni bir yaklaşım önerilmiştir. Bu yöntem, temel yaklaşıma kıyasla önemli bir avantaj sağlamaktadır. Tahmin sistemlerinde tüm kelimeler yerine n-gram eklerinin kullanıldığı bu çalışmanın, bu alandaki ilk uygulama olduğuna inanılmaktadır. Sistem, eğitim verisi boyutundaki artışla daha da iyileşmesi beklenen %94,9’luk bir en iyi F1 puanı elde etmiştir [19]. Bu çalışmada, cinsiyeti isme göre tahmin etmek için çeşitli makine öğrenimi algoritmaları incelenmiş ve uygulanmıştır. Yabancı isimler üzerinde yapılan bu tür tahminler, özellikle doğal dil işleme (NLP) alanında önemli bir yer tutmaktadır. Extra Trees, KNN, Naive Bayes, SVM, Random Forest, Gradient Boosting, Light GBM, Logistic Regression, SVC Classifier ve derin sinir ağı modelleri olan MLP, RNN, GRU, CNN ve BiLSTM gibi yöntemler değerlendirilmiştir. Brezilya isimlerinden oluşan bir veri kümesi kullanılarak modeller eğitilmiş ve değerlendirilmiştir. Performans ölçümleri olarak accuracy, recall, precision, F1-score ve confusion matrix analiz edilmiştir. Sonuçlar, isimlere dayalı özellik çıkarımı kullanılarak cinsiyet tahmininin başarılı bir şekilde yapılabildiğini göstermektedir. Bazı modeller, vakaların %95'inden fazlasında doğru tahminler yapabilmiştir ve tekrarlayan modeller, bu ikili sınıflandırma probleminde ileri beslemeli modellere göre üstün performans sergilemiştir [20]. Klinik PDF belgelerindeki gövde metninin denetimsiz olarak çıkarılması, doğal dil işleme (NLP) görevlerinin geliştirilmesi için önemli bir adım olarak kabul edilmiştir. Ancak, bu süreç hala bir zorluk olarak kalmıştır. Yapılan bir çalışmada, büyük veri setleri kullanarak gövde metnini çıkarmak için tasarlanmış gözetimsiz bir algoritma sunulmuştur. DBSCAN kümelemesi kullanılarak, yöntem metin bloklarını içerik ve koordinatlarına göre çıkarmış ve düzenlemiştir. Değerlendirme sonuçları, çeşitli tıbbi uzmanlık kaynakları arasında 0,82 ila 0,98 arasında değişen precision puanları, 0,62 ila 0,94 arasında recall puanları ve 0,71 ila 0,96 arasında F1-score puanları elde edilmiştir [21]. Bu çalışmada, robot destekli müdahalelerde vibro-akustik sinyallerden (VA sinyalleri) özellik çıkarma yöntemleri incelenmiştir. Özellik çıkarma, Cepstrum Dönüşümü (CT), Mel-Frekans Cepstral Katsayıları (MFCC'ler) ve Hızlı Chirplet Dönüşümü (FCT) kullanılarak yapılmış, ardından boyut azaltma için PCA, t-SNE ve UMAP gibi teknikler uygulanmıştır. Sınıflandırma ise en yakın komşular algoritması ile gerçekleştirilmiştir. Sonuçlar, CT ve MFCC'nin boyut azaltma ile birleşiminde yüksek verimlilik göstermiş ve sınıflandırma metriklerinde %99 accuracy, recall ve F1-score değerleri elde edilmiştir. Kümeleme metriği ise 0,61 olarak bulunmuştur [22]. Bu çalışmada, hiyerarşik kümeleme ve yapay sinir ağları (YSA) kullanılarak yol trafik kazalarının ciddiyeti tahmin edilmiştir. Verilerdeki çeşitliliği azaltmak için hiyerarşik kümeleme uygulanmış ve kazalar altı farklı kümeye ayrılmıştır. Ardından, her bir küme için iki YSA modeli oluşturulmuş ve ilk model tüm kazalarla eğitilmiş, ikinci modelde ise kazaların %66’sı eğitim, %34’ü ise test seti olarak kullanılmıştır. Sonuçlar, kümeleme uygulandıktan sonra elde edilen doğruluğun, kümeleme yapılmadan elde edilenden %11-%16 daha yüksek olduğunu, yüzdelik bölünme ile doğruluğun ise %2-%5 arttığını göstermiştir [23]. Bu çalışmada, biyomedikal belgelerin tematik olarak tutarlı gruplarını keşfetmek için BioBERT dil temsilleri ve Gauss Karışım Modeli tabanlı bir kümeleme çerçevesi önerilmiştir. PubMed veritabanından alınan biyomedikal belgeler üzerinde, klasik metin ön işleme teknikleri kullanılmış, ardından BioBERT ile vektör temsiller elde edilmiştir. Kümeleme işlemi için Gauss Karışım Modeli kullanılmıştır. Deneysel sonuçlar, önerilen modelin sırasıyla Fowlkes-Mallows puanı 0,7817, silüet katsayısı 0,3765, ayarlanmış rand endeksi 0,4478 ve Davies-Bouldin puanı 1,6849 elde ederek, kıyaslama modellerinden daha iyi performans gösterdiğini ortaya koymuştur [24]. Bu çalışmada, büyük ölçekli tıbbi metin verisi için verimli bir veri temizleme süreci önerilmiştir. Tekrarlanan yazım hataları ve tutarsızlıklar gibi sorunları gidermek amacıyla metin kümeleme yöntemleri ve değer dönüştürme tekniği kullanılmıştır. Süreç, metin kümeleme ve değer birleştirme aşamalarından oluşmaktadır. Kümeleme aşamasında, anahtar çarpışma ve en yakın komşu yöntemleri karşılıklı olarak kullanılmıştır. Yanlış değerlerin doğru değerlerine dönüştürülmesi için değer dönüştürme aşaması uygulanmıştır. Önerilen süreç, 1995-2015 yılları arasında Samsung Tıp Merkezi'nde parazit analizi için üretilen 574.266 dışkı muayenesi raporuna uygulanmıştır. Yazım hataları ve çoğaltma desenlerinde %98,61 ve %97,78 oranlarında yüksek düzeltme oranları elde edilmiştir. Sonuç olarak, veri doğruluğu toplam kelimeler üzerinden %100 civarında elde edilmiştir [25]. Bu çalışmada, tıbbi metin sınıflandırması için, ERNIE-Health adlı ayırt edici bir önceden eğitimli dil modeli kullanılarak prompt-tuning tekniği ile sınıflandırma yapılması incelenmiştir. Yöntem, metni yeni bir dizeye sararak, kategori etiketini [UNK] token'ı ile değiştirir ve model, aday kategorilerin olasılık dağılımını hesaplamak için eğitilir. Yöntem, KUAKE-Question Intention Classification ve CHiP-Clinical Trial Criterion veri setlerinde test edilmiştir. Deneysel sonuçlar, sırasıyla 0.866 ve 0.861 accuracy değerleri elde edilmiştir. Ayrıca, modelin kayıp (loss) değerlerinin eğitim süresi boyunca fine-tuning yöntemine kıyasla daha hızlı düştüğü gözlemlenmiştir [26]. Bu çalışma, tıbbi metin sınıflandırmasındaki genel performansı etkileyen metin temsili tekniklerini incelemiştir. Bu hedefe ulaşmak için, temel kelime temsili (bag-of-words) ve daha karmaşık tıbbi ifade temsili (bag-of-phrases) destekleyen bir metin sınıflandırma sistemi geliştirilmiştir. Ayrıca, kelime ve ifade temsillerinin birleşimi (hibrit) kullanılarak daha kapsamlı bir analiz yapılmıştır. Sistem, tıbbi bilgi tabanını ve doğal dil işleme tekniklerini entegre ederek metinden tıbbi ifadeler çıkarmaktadır. Farklı temsillerin etkilerini değerlendirmek için OHSUMED veri setindeki MEDLINE belgeleri ile yapılan deneylerde sınıflandırma performansındaki değişim ölçülmüştür. Sınıflandırma performansı, bilgi erişim metrikleriyle ölçülmüş olup, precision (p), recall (r) ve F1-score (F1) değerleri kullanılmıştır. Deneylerde hibrit yaklaşım (p=0.87, r=0.46, F1=0.60) ile bag-of-words yaklaşımından (p=0.85, r=0.44, F1=0.58) ve bag-of-phrases yaklaşımından (p=0.87, r=0.42, F1=0.57) daha iyi sonuçlar elde edilmiştir [27]. Bu çalışmada, deri kanseri görüntülerinin sınıflandırılmasında benzerlik kümelemesi ve dikkat aktarımı yöntemlerinin etkisi araştırılmıştır. Çalışma kapsamında, derin metrik öğrenmeye dayalı denetimsiz kümeleme yöntemi kullanılarak benzer görüntüler bir araya getirilmiş ve bu süreçte elde edilen model ağırlıkları, sınıflandırma modelini eğitmek için öğretici model olarak kullanılmıştır. Dikkat aktarımı yöntemi, modelin hem benzerlik özelliklerini hem de kategori bilgilerini aynı anda öğrenmesini sağlayarak tanı doğruluğunu artırmıştır. Elde edilen sonuçlar, önerilen yöntemin doğruluk oranını %72.53'e yükselterek başlangıç modeline göre %5'lik bir iyileşme sağladığını göstermiştir. Özellikle, üç farklı kötü huylu lezyonun F1-Skorlarında önemli artışlar gözlenmiştir: Bazal hücreli karsinom (BCC) 0.65'ten 0.73'e, Skuamöz hücreli karsinom (SCC) 0.28'den 0.37'ye ve Melanom (MEL) 0.54'ten 0.58'e yükselmiştir. Bu bulgular, benzerlik kümelemesi ve dikkat aktarımı yöntemlerinin, deri lezyonlarının tanısında toplam doğruluğu ve F1-Skorunu artırarak daha başarılı bir sınıflandırma sağladığını ortaya koymuştur [28]. Bir çalışmada, diyabetle ilgili biyomedikal literatürün keşfini ve klinik karar verme sürecini iyileştirmek amacıyla, kullanıcı dostu bir etkileşimli sınıflandırma ve konu keşfi metodolojisi geliştirilmiştir. Bu yaklaşım, sağlık profesyonellerinin sınırlı bilgisayar bilimi bilgisiyle bile metin verilerini keşfetmelerini kolaylaştırmak için tasarlanmıştır. Yöntem, hiyerarşik kümeleme, hedefe yönelik sınıflandırma, kullanıcı etkileşimlerini en aza indiren aktif öğrenme ve görsel bir kullanıcı arayüzü gibi dört ana bileşenden oluşmaktadır. Performans değerlendirmesi, PubMed veritabanından alınan 50.911 diyabetle ilgili özet üzerinde gerçekleştirilmiştir. Hiyerarşik kümeleme için elde edilen F1 skoru 0,73 olarak raporlanmış ve scikit-learn kütüphanesi ile karşılaştırıldığında rekabetçi bir performans sergilemiştir. Aktif öğrenme stratejisinde, ağırlıklı F1 skoru 0,62 olarak elde edilmiş ve bu, alternatif stratejilere benzer bir performans göstermiştir. Yöntem ayrıca, artan veri miktarına rağmen düşük bellek kullanımı sunmuş, bu da büyük veri kümelerinde avantaj sağlamıştır [29]. Bir çalışmada, semantik olarak ilişkili ancak farklı temsillere sahip kelimelerin ayrımını yapmanın, özellikle biyomedikal literatürde zorluk oluşturduğu belirtilmiştir. Bu bağlamda, demans ve deliryum gibi semantik olarak ilişkili kavramların ayrımını yapabilmek için, semantik öğeler arasındaki etkileşim sıklığı verilerinden yararlanılmıştır. Çalışmada, semantik predikasyonlardan elde edilen veri kümeleri iki farklı Expectation Maximization (EM) kümeleme sürecinde kullanılmıştır: biri etiketlenmemiş (unlabeled), diğeri ise kavram etiketleriyle birlikte. Etiketlenmemiş veri kümelerinde, kümelerin %80'inin beklenen kümelenme sayısını ve benzer oranları sağladığı, etiketli veri kümelerinde ise benzeyen veya eşleşen oranlar gözlemlendiği belirtilmiştir. En iyi performans gösteren sınıflandırıcı, %89 doğruluk oranına ulaşmış ve bireysel kavram sınıflandırmaları için F1 skorları 0.69 ile 1 arasında değişiklik göstermiştir. Sonuçlar, klinik metinlerin doğal dil işleme süreçlerinde nasıl kullanılabileceğine dair bir tartışma ile sunulmuştur [30]. Bu çalışmada, radyoloji metinlerini kullanarak proksimal humerus kırıklarını sınıflandırmak için çeşitli makine öğrenmesi (ML) ve derin öğrenme algoritmalarının performansları değerlendirilmiş. 2010-2019 yılları arasında büyük bir travma merkezine kabul edilen hastaların verileri kullanılmış ve bu veriler X-ray ve BT raporları gibi metin kaynaklarıyla birleştirilmiş. Sınıflandırma amacıyla Neer sınıf etiketleri kullanılmış. Modeller, doğruluk, geri çağırma, kesinlik, F1, ve "One-versus-rest" gibi metriklerle değerlendirilmiş. Bir BERT modeli, %61 doğruluk ve "One-versus-rest" metriklerinde 0.8'in üzerinde performans göstermiş. En yüksek kesinlik, geri çağırma ve F1 puanları sırasıyla %50, %39 ve %39 olarak rapor edilmiştir [31].

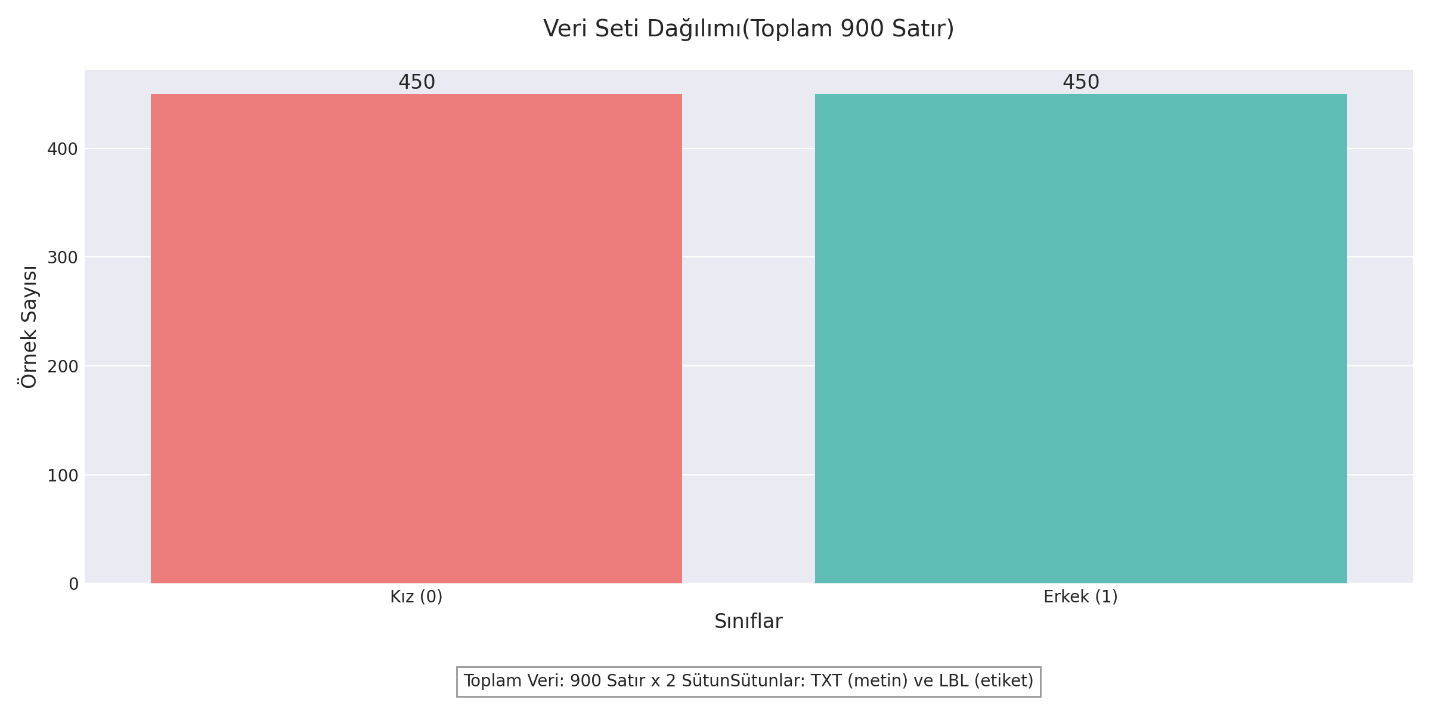
3. Materyal ve Metot

Çalışmada kullanılan material ve yöntemlerin detayları aşağıdaki alt başlıklarda açıklanmıştır. Şekil 3’de kümeleme tabanlı sınıflandırma için önerilen modelin akış şeması gösterilmiştir.

3.1. Veri Kümesi

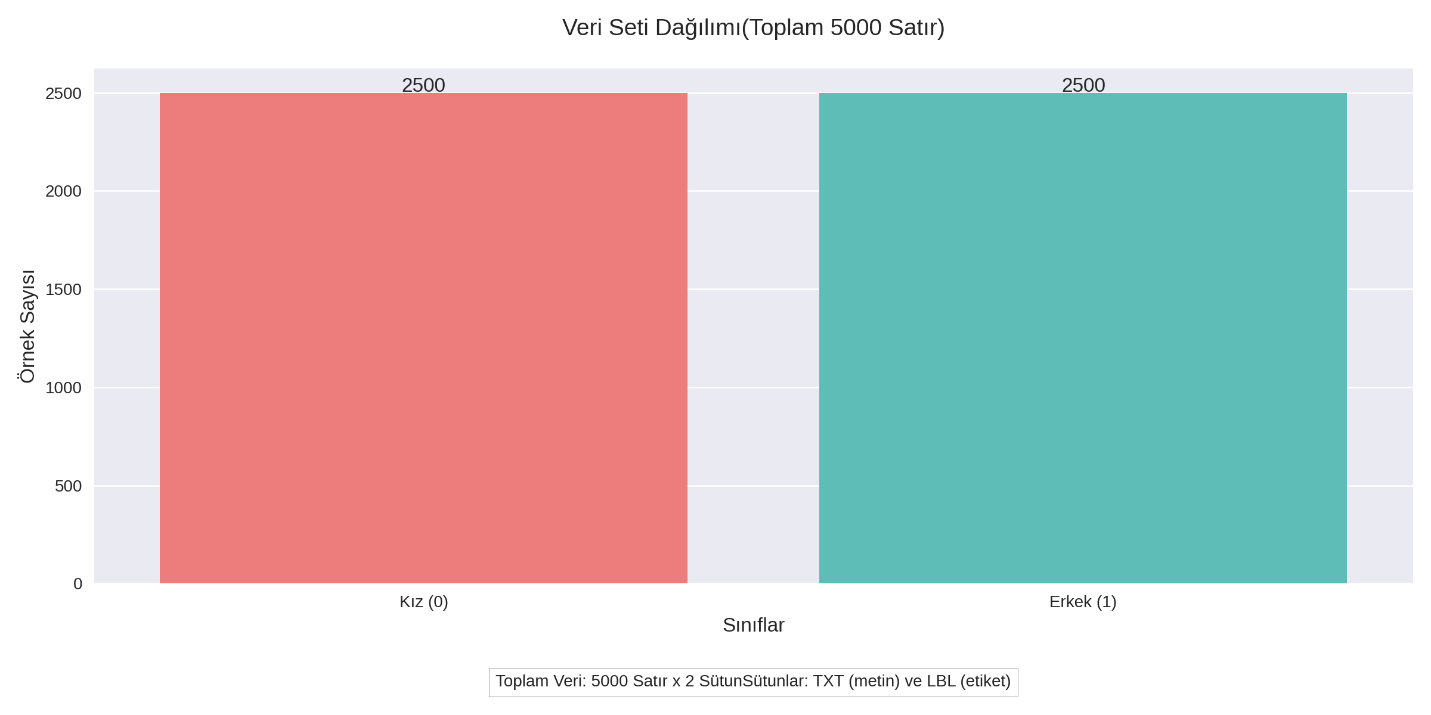
Bu çalışmada kullanılan veri kümesi, Türkçe ve İngilizce metinlerden oluşmaktadır. Türkçe veri kümesi, toplamda 900 adet metinden oluşan, Twitter'dan alınmış mesajları içeren bir koleksiyondur. Her satırda bir metin ve ona ait etiket yer almaktadır. Tweetler, rastgele örneklem yöntemiyle toplanmıştır. Tweetlerin konuları oldukça çeşitli olup, siyaset, spor ve eğlence gibi farklı alanları kapsamaktadır, toplanan metinler (tweetler) cinsiyet çıkarımı yapılabilmesi amacıyla her biri belirli bir cinsiyet etiketi ile ilişkilendirilmiştir.

İngilizce veri kümesi ise 5000 adet metinden (tweet) oluşmaktadır ve Twitter'dan alınmış mesajları içeren bir koleksiyondur. Bu veri kümesinde de her metin bir cinsiyet etiketi ile ilişkilendirilmiştir. Veri kümesindeki metinler, sosyal medya kullanıcılarının yazdığı çeşitli ifadelerden seçilmiş olup, cinsiyetin doğru şekilde çıkarılabilmesi için yeterli çeşitliliğe sahiptir. Her iki veri kümesi de, metin sınıflandırma modelinin eğitilmesi ve doğruluğunun test edilmesi amacıyla kullanılmıştır. Bu çalışmada kullanılan veri kümelerinin temel amacı, Türkçe sosyal medya verileri üzerinde cinsiyet tespiti modelinin geliştirilmesidir.

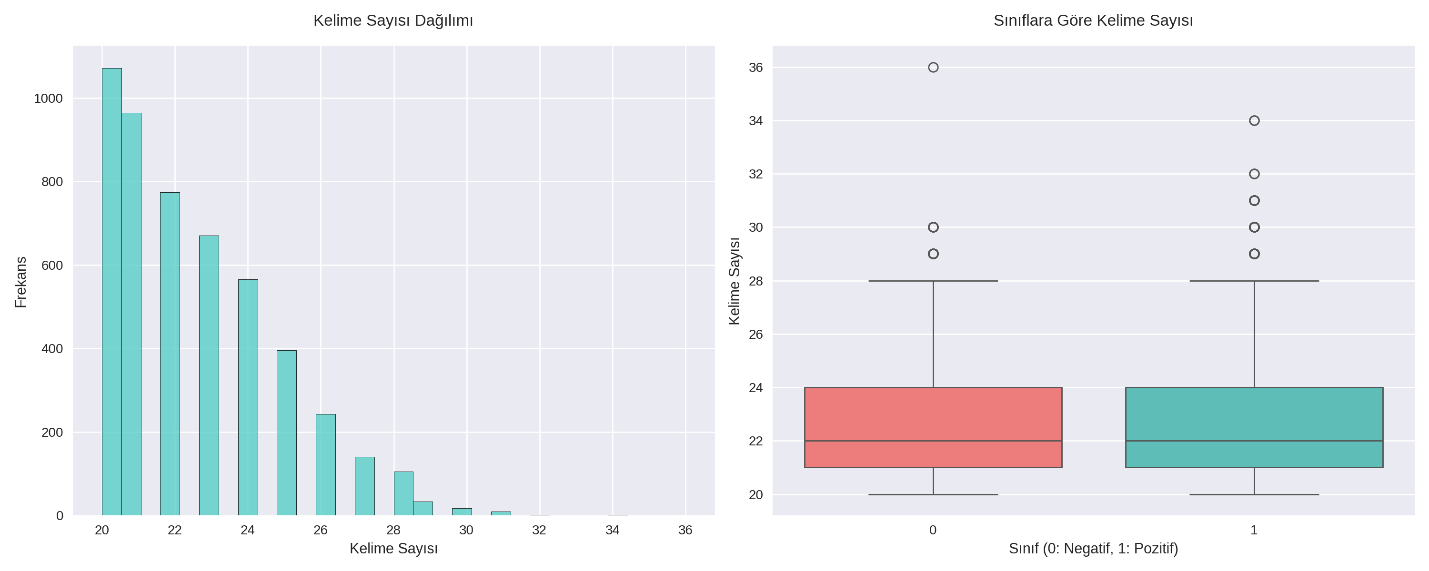


Şekil 1. Türkçe veri kümesindeki kategori değerlerinin dağılımı.

Şekil 2’de ingilizce veri kümesi kategorik olarak dağılımı gösterilmiştir şekil 2’de görüldüğü gibi veri kümesi dengeli veri dağılımı içermektedir. Şekil 3’de veri kümesinde kullanılan kelime sayısının analizi gösterilmiştir.



Şekil 2. İngilizce veri kümesindeki kategori değerlerinin dağılımı.



Şekil 3. kelime sayısı dağılımı.

3.2. Veri Ön İşleme

Veri ön işleme makine öğrenmesi modellerini kurma aşamasında en önemli aşamalardan biridir. Veri ön işlemlerini gerçekleştirmenin iki temel nedeni bulunmaktadır, bunlardan ilki verilerle ilgili sorunlar, ikincisi ise veri analizi için hazırlık. Bu iki aşama güvenilir sonuçlar ve verimli tahmin için önemlidir. Twitter mesajlarının bazı önemli özellikleri vardır. Paylaşılan mesajlar genellikle "@" karakteriyle başlayan kullanıcı adlarını, "#" karakteriyle başlayan hashtag'leri, ifadeleri ve URL'leri içerir. Bu farklı özelliklerin tespit edilebilmesi ve sınıflandırıcıların eğitilebilmesi için veriler çeşitli ön işleme tekniklerinden geçmelidir. Bu nedenle çalışmamızda ön işleme aşamasında gereksiz karakterlerin ayıklanmasının ardından veri seti üzerinde belirli manipülasyonlar yapılmıştır.

Bu çalışmada, metin verilerinin sınıflandırma modeline uygun hale getirilmesi için çeşitli ön işleme adımları gerçekleştirilmiştir. İlk olarak, metinlerdeki gereksiz karakterler, sayılar, URL'ler ve emojiler temizlenmiştir. Emojiler, özel bir düzenli ifade (regex) ile tespit edilip kaldırılmıştır, URL'ler ve sayılar metinlerden çıkarılmıştır. Ayrıca, yalnızca Türkçe ve İngilizce alfabeye ait harfler bırakılarak, diğer tüm karakterler silinmiştir. Metinler küçük harfe dönüştürülerek gereksiz boşluklar kaldırılmış ve noktalama işaretleri temizlenmiştir. Bu adımlar uygulanarak anlamsal olarak önemli olmayan ve analizi zorlaştırabilecek tüm karakterlerden arındırılmıştır. Böylelikle metin, sadece anlamlı kelimelerden oluşur hale gelmiştir.

**Küçük Harfe Çevir:** Metindeki tüm harfleri küçük harfe çevirme işlemi. Bu adım, büyük ve küçük harfler arasındaki farkları ortadan kaldırarak tutarlı bir veri seti oluşturur.

**Fazla Boşlukları Kaldırma:** Bu işlem ise metinde birden fazla bulunan boşlukların tek bir boşluğa indirgenmesiyle gerçekleştirilir. Bu sayede metnin okunabilirliği artırılır ve daha düzenli bir yapıya kavuşturulur.

**Emojileri Kaldır:** Metinde bulunan emojilerin silinmesi. Bu, metnin daha ciddi ve profesyonel görünmesini sağlar.

**URL'leri Kaldır:** Metindeki internet adreslerinin (URL'lerin) kaldırılması. Bu, metni daha okunabilir hale getirir ve dikkat dağınıklığını önler.

**Alfabetik Olmayan Karakterleri Kaldır:** Bu işlem harf dışında kalan tüm karakterler, özel karakterler veya sembollerin (örneğin: @, #, $, vb.) temizlenmesi işlemidir. Bu adım, metnin daha düzenli hale gelmesini sağlar.

**Noktalama İşaretlerini Kaldırma:** Noktalama işaretleri (.,;:!? vb.)virgül, nokta, ünlem işareti gibi noktalama işaretleri genellikle analiz için gerekli değildir. Bu adımda tüm noktalama işaretleri metinden çıkarılır.

**Tekrar Eden Karakterleri Kaldırma:** Bu işlem, metindeki ardışık tekrar eden karakterlerin yalnızca bir kez alınmasını sağlar. Örneğin, " çokkk" kelimesinde "k" harfleri gereksiz şekilde tekrarlanmaktadır. Bu işlem, tekrarları ortadan kaldırarak kelimeyi "çok" haline getirir. Böylece, metnin gereksiz tekrarlarından arındırılmasını sağlar ve veri analizi ile modelleme süreçlerinde daha temiz ve anlamlı bir metin elde edilir.

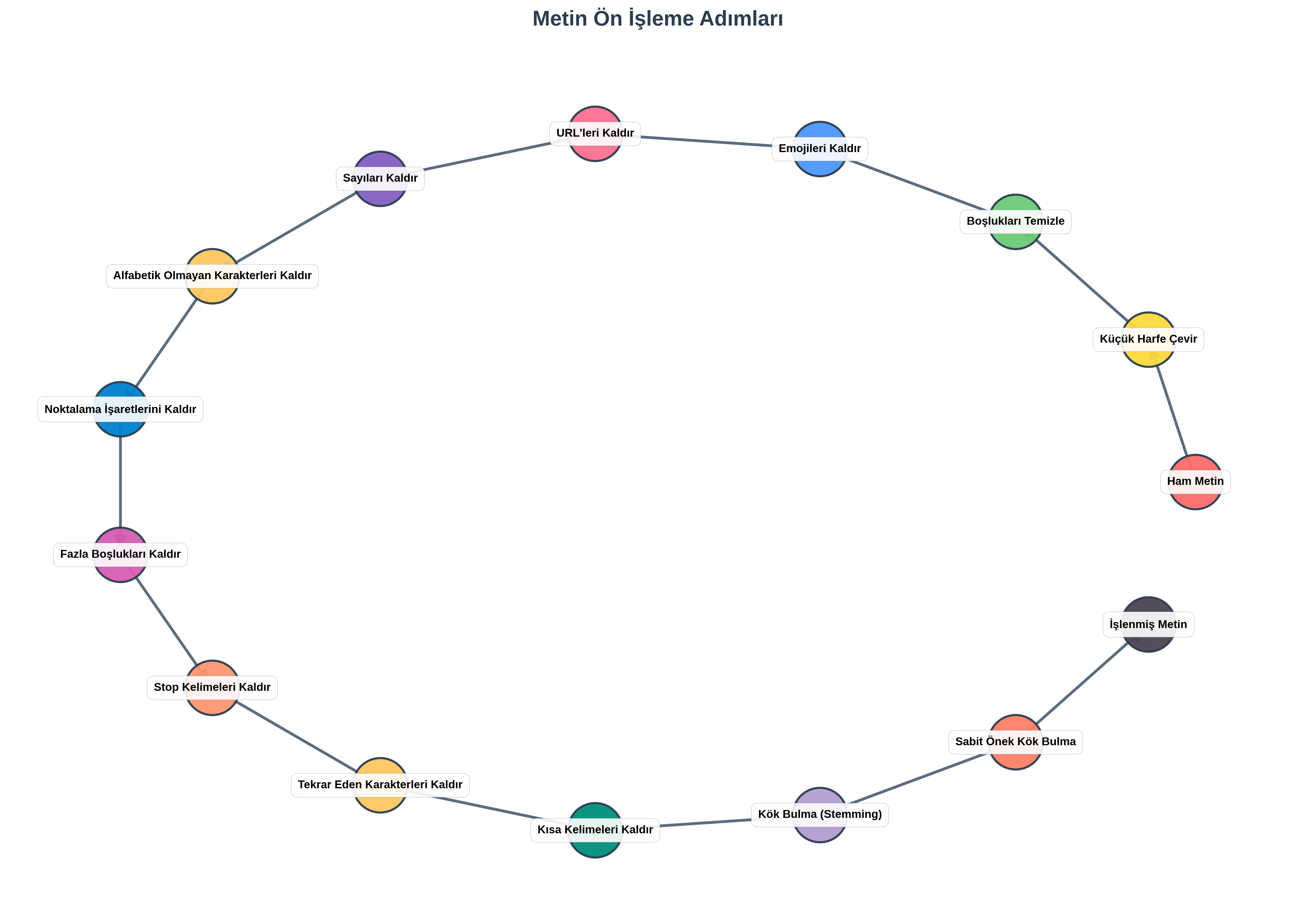
**Kısa Kelimeleri Kaldırma:** Genellikle iki veya üç harften oluşan kısa kelimeler, analizde çok fazla anlam taşımayabilir. Bu nedenle, belirlenen bir uzunluğun altındaki kelimeler çıkarılabilir.

**Stop Kelimeleri Kaldırma**: Durdurma kelimelerini ortadan kaldırma süreci, metnin anlamsal zenginliğine önemli ölçüde katkıda bulunmayan, “ve”, “bir”, “bu”, "ya", "da", "ile", "ama" gibi bağlaçlar ve zamirler ve İngilizce dilinde sıklıkla kullanılabilecek çeşitli diğerleri gibi temel olmayan, yaygın olarak ortaya çıkan terimlerin dikkatli bir şekilde çıkarılmasını gerektirir. Bu durdurma kelimelerinin varlığının analitik çabalara anlamlı bir gelişme sağlamadığını kabul etmek önemlidir; daha ziyade, önemli içeriği gizleme ve böylece analiz edilen metinsel verilerin genel boyutunu ve netliğini azaltma eğilimindedirler. Sonuç olarak, bu yabancı terimlerin kaldırılması, metin analizinin ön işleme aşamasında kritik bir adımdır, çünkü kalan önemli kelimelerin daha odaklanmış bir incelemesini kolaylaştırır ve sonuçta gelişmiş yorumlanabilirlik ve içgörüye yol açar. Bu işlem, hem Türkçe hem de İngilizce için, NLTK kütüphanesinin sağladığı stop kelimeleri listesi kullanılarak gerçekleştirilmiştir. Metin, stop kelimelerden arındırıldıktan sonra, yalnızca anlamlı kelimelerle devam edilmiştir.

**Kök Bulma**: Sözcüksel öğelerin, genellikle kök olarak adlandırılan temel veya temel formlarına sistematik olarak indirgenmesidir başka bir deyişle bu işlemde kelimelerin köklerine inilir. Örneğin, "yürümek", "yürüme" gibi türevler kök olan "yürü"ye dönüştürülür. Türkçe metinler için, SnowballStemmer yerine özel bir sabit önek kök bulma fonksiyonu kullanılmıştır. Bu fonksiyon, kelimenin uzunluğuna bağlı olarak, kelimenin ilk beş harfini alarak kökünü bulmaya çalışır. İngilizce metinler içinse, NLTK kütüphanesinin SnowballStemmer fonksiyonu kullanılarak kelimeler köklerine indirilmiştir. Bu işlemler, metnin analiz için daha tutarlı ve anlamlı hale gelmesini sağlamak amacıyla yapılmıştır.

**Sabit önek kök araması:** Kök arama işlemi sırasında sabit öneklerin de dikkate alınması gerekebilir. Örneğin "yen" kelimesinin "yenilmek" kelimesinin kökü olarak belirlenmesinde ön ekin etkisi dikkate alınmıştır.

Ön işleme aşamasında, Türkçe metinler için durak kelimeler (stopwords) çıkarılmış ve kök bulma işlemi uygulanmıştır. İngilizce metinlerde de benzer şekilde stopwords kelimeleri kaldırılmış ve kelimeler köklerine indirgenmiştir. Ayrıca, metinlerdeki tekrarlayan kelimeler ve fazla uzun kelimeler belirli kurallara göre düzeltilmiştir. Tüm bu adımlar, modelin daha verimli çalışmasını sağlamak ve metinlerden anlamlı özellikler çıkarmak amacıyla yapılmıştır.



Şekil 3. Metin önişleme adımları.

3.3. Geleneksel Yöntemler - Özellik Çıkarma Yöntemleri

Metin verilerinden anlamlı özellikler çıkarmak, doğal dil işleme (NLP) modellerinin başarılı olabilmesi için kritik bir adımdır. Özellik çıkarma, metni sayısal verilere dönüştürerek, makine öğrenmesi algoritmalarının bu veriler üzerinde işlem yapabilmesini sağlar. Çeşitli yöntemler, metinlerden elde edilebilecek özelliklerin türünü ve biçimini belirler. Bu çalışmada kullanılan başlıca özellik çıkarma yöntemleri şunlardır:

* Kelime Çantası (BoW) + TF-IDF
* BoW + TF
* BoW + Binary
* 3-gram + Binary
* 3-gram + TF
* 3-gram + TF-IDF

TF-IDF, bir belgedeki bir kelimenin önemini tüm belge koleksiyonuna (korpus) göre hesaplayan bir istatistiksel metin gösterimi tekniğidir. Bag of Words (BoW), metinleri sayısal olarak temsil etmek için kullanılan bir yöntemdir ve kaynaklarda ayrıntılı olarak açıklanmaktadır. Bu yöntemde, metinlerdeki kelimelerin sırası göz ardı edilir ve yalnızca her kelimenin metinde kaç kez geçtiği önemlidir.

BoW'un Temel Özellikleri:

* Kelime Sırasını Yok Sayar: BoW, metinlerdeki kelimelerin sıralamasını dikkate almaz. Örneğin, "kedi fareyi kovaladı" ve "fareyi kedi kovaladı" cümleleri BoW yöntemiyle aynı şekilde temsil edilir.
* Kelime Frekanslarına Odaklanır: BoW, her kelimenin metinde kaç kez geçtiğini sayar. Bu sayılar, metnin bir vektör olarak temsil edilmesinde kullanılır.
* Basit ve Uygulanması Kolaydır: BoW, kavramsal olarak basittir ve uygulanması kolaydır. Bu nedenle, metin sınıflandırma, bilgi erişim ve konu modelleme gibi çeşitli NLP görevlerinde yaygın olarak kullanılır [32].

N-gram, bir metnin içindeki n sayıda ardışık kelime veya karakter dizisidir. Örneğin, "bu bir örnek" cümlesi için 2 gramlık (bigram) "bu bir", "bir örnek" şeklinde olur. N-gramlar, metnin içindeki kelime veya karakter dizilerini analiz ederek dil modelleri oluşturmak için kullanılır [33].

Kelime Çantası (BoW), bir metindeki kelimelerin sıklığını dikkate alan basit bir metin gösterim yöntemidir. TF-IDF (Terim Sıklığı-Ters Belge Sıklığı) ise, bir kelimenin bir belge için ne kadar önemli olduğunu belirlemek için kullanılan bir istatistiksel ölçüdür [32]. "Kelime Çantası" (BOW) kavramı, metin işleme ve bilgi almada temel bir temsildir. Belgeleri, sözdizimi ve kelime sırasını göz ardı ederek ağırlıklı terimler vektörleri olarak temsil eder [34].

* BoW + Binary: Bu yöntem, belgedeki her bir kelimeye ikili bir değer (varlık veya yokluk) atar. TF veya TF-IDF'den daha az bilgilendiricidir ancak bir kelimenin varlığının önemli olduğu belirli uygulamalarda yararlı olabilir.
* 3-gram + Binary: İkili ağırlıklandırma ile 3-gram kullanımı, üç kelimelik dizileri dikkate alarak metin eşleştirmesinin hassasiyetini artırabilir ve bu da tek kelimelere kıyasla daha fazla bağlam yakalamaya yardımcı olur.
* 3-gram + TF: 3-gramı terim sıklığıyla birleştirmek, tek kelimelerden daha fazla bağlamsal bilgi yakalamaya yardımcı olur ve bu da onu daha ayrıntılı metin analizi görevleri için kullanışlı hale getirir [35].
* 3 gram + TF-IDF: Bu kombinasyon, belgelerdeki kelime dizilerinin hem sıklığını hem de önemini dikkate alarak alma performansını iyileştirmede özellikle etkilidir [36].
* BoW + TF:  BoW ile tek başına terim sıklığını (TF) kullanmak, farklı belgelerdeki kelimelerin önemini hesaba katmadığı için TF-IDF ile karşılaştırıldığında daha az etkilidir. Ancak daha basittir ve hesaplama açısından daha az yoğundur [37].
* BoW + TF-IDF: BoW, TF-IDF ile birleştirildiğinde metin temsili için temel ancak etkili bir yöntemdir. Sınıflandırma ve kümeleme gibi görevlerde iyi performans gösterir ancak düşük temsil verimliliğinden muzdariptir ve kelime düzeyindeki anlamsal benzerliği yakalamada başarısız olur [38].

3.4. Kelime Gömme Modelleri

Kelime gömme modelleri, doğal dil işleme (NLP) ve makine öğrenimi (ML) alanlarında önemli bir rol oynar. Bu modeller, kelimeleri sayısal vektörlere dönüştürerek, benzer anlamlı kelimeleri ve bağlamları yakalarlar. Bu çalışmada, GloVe, FastText ve Word2Vec modelleri kullanılmıştır.

Word2Vec, kelimeleri sayısal vektörlere dönüştüren bir sinir ağı modelidir. Bu model, kelimelerin anlamsal ve sözdizimsel ilişkilerini yakalamaya çalışır, böylece benzer anlama gelen kelimeler vektör uzayında birbirine yakın konumlandırılır.

-CBOW (Continuous Bag of Words): Bağlam kelimelerine bakarak hedef kelimeyi tahmin etmeye çalışır.

-Skip-gram: Hedef kelimeye bakarak bağlam kelimelerini tahmin etmeye çalışır [32].

GloVe (Global Vectors for Word Representation), kelimeleri sayısal vektörlere dönüştürmek için kullanılan bir kelime gömme tekniğidir. GloVe, Word2Vec'ten farklı olarak hem yerel bağlam penceresini hem de küresel matris çarpanlara ayırma yöntemlerini kullanarak kelime gömmeleri oluşturur. FastText, Facebook AI Research tarafından geliştirilen ve kelimelerin iç yapısını (morfolojisini) dikkate alarak kelime gömmeleri oluşturan bir yöntemdir.

FastText'in Çalışma Prensibi:

1. Karakter N-gramlarının Oluşturulması: Her kelime, belirli bir uzunluktaki (n değeri) karakter dizilerine (n-gramlar) ayrılır. Örneğin, "elma" kelimesi için 3-gramlar: "<el", "elm", "lma", "ma>" olur. Başlangıç ve bitiş karakterleri ("<" ve ">") kelime sınırlarını belirtmek için kullanılır.

2. N-gram Vektörlerinin Atanması: Her karakter n-grama bir vektör atanır. Bu vektörler, model eğitimi sırasında öğrenilir.

3. Kelime Vektörünün Hesaplanması: Bir kelimenin vektörü, o kelimeye ait tüm karakter n-gramlarının vektörlerinin toplamı olarak hesaplanır.

FastText, kelimeleri vektörler olarak temsil eden bir modeldir. Bu model, kelimelerin alt birimlerini (subwords) dikkate alarak çalışır. Bu sayede, kelimelerin köklerini ve eklerini analiz ederek daha anlamlı vektörler oluşturur. Bu özellik, özellikle nadir kelimelerin veya yeni kelimelerin temsilinde avantaj sağlar [39].

3.5. Dil Modelleri

Bu çalışmada, çeşitli dil işleme modellerinden faydalanılmıştır. Kullanılan dil modelleri arasında BERT, Megatron, MarianMT, T5, CamemBERT, XLM, CTRL, Reformer, XLNet, ALBERT, ERNIE, FlauBERT, ELECTRA ve RoBERTa yer almaktadır. Bu modeller, doğal dil işleme (NLP) alanında farklı görevleri gerçekleştirmek üzere geliştirilmiş olup çalışmamızın kapsamına uygun çözümler sunmaktadır.

BERT (Transformers'dan Çift Yönlü Kodlayıcı Gösterimleri), Google araştırmacıları tarafından 2018'de geliştirilen ve doğal dil işlemede (NLP) bir dönüm noktası niteliğinde olan bir tekniktir. BERT, ön eğitim için kullanılan temel konfigürasyonda 12 katmanlı, 12 self-attention başlığı ve 768 boyutlu gizli durumlara sahip bir kodlayıcı kullanır. BERT'in başarısının ardında yatan temel nedenlerden biri, her kelimenin bir cümle içindeki tüm tokenlerden bağlamı dikkate almasını sağlayan çift yönlü eğitim yaklaşımıdır. BERT'in önemli özelliklerinden biri de maskeli dil modellemesi (MLM) tekniğini kullanmasıdır. BERT'in ön eğitimi için BookCorpus ve İngilizce Wikipedia metni gibi devasa veri kümeleri kullanılır [40].

MarianMT: MarianMT, C++ dilinde yazılmış hızlı bir sinirsel makine çevirisi araç setidir, Transformer'ların MarianMT modellerinden çıkarım yapmak için bir araç içerdiğini belirtir.  
Megatron: Megatron-LM, NVIDIA tarafından geliştirilen ve büyük dil modellerini eğitmek için kullanılan bir çerçevedir.Megatron-LM'nin MarianMT ve Google BERT modellerinden çıkarım yapmak için bir araç içerdiğini belirtir [41].

T5: T5 (Text-to-Text Transfer Transformer), Google tarafından geliştirilen ve çeşitli NLP görevlerini tek bir metinden metne çerçevesinde birleştirmeyi amaçlayan bir model olarak tanımlanmıştır3. T5, hem kod çözücüye sahip hem de kod çözücüsüz modellerde önemli performans kazanımları sağlayan, önceden eğitilmiş bir dil modelidir. T5, çok dilli bir sürümü olan mT5'i de içeren bir model ailesidir. ERNIE: ERNIE (Enhanced Representation through kNowledge IntEgration), Baidu tarafından geliştirilen ve bilgi grafiği entegrasyonunu vurgulayan bir model ailesidir. ERNIE 3.0, daha sonra 26 kat daha fazla parametre ile daha büyük bir model olan ERNIE 3.0 Titan ile genişletilmiştir. ERNIE 3.0 Titan, NLP görevinde diğer en gelişmiş modelleri geride bıraktı. RoBERTa: RoBERTa (A Robustly Optimized BERT Pretraining Approach), BERT mimarisini temel alan ve ön eğitim sürecini iyileştirerek performansını artıran bir model olarak tanımlanmıştır [42].

FlauBERT, Fransızca diline odaklanan ve çok sayıda Fransızca veriyle eğitilmiş bir dil modelidir. Transformer mimarisine dayanan FlauBERT, özellikle metin sınıflandırma, soru cevaplama ve makine çevirisi gibi görevlerde kullanılır. CamemBERT, tıpkı FlauBERT gibi, Fransızca diline özgü ve büyük miktarda Fransızca veri üzerinde eğitilmiş bir dil modelidir. Transformer mimarisine dayanır ve metin sınıflandırma, soru cevaplama ve makine çevirisi gibi çeşitli Doğal Dil İşleme (NLP) görevlerinde kullanılır.

FlauBERT ve CamemBERT benzer yapılara sahip olsalar da, bazı önemli noktalarda farklılık gösterirler:

* Tokenleştirici: FlauBERT Byte Pair Encoding kullanırken, CamemBERT SentencePiece kullanır.
* Kelime Dağarcığı Boyutu: FlauBERT'in kelime dağarcığı boyutu 50.000 iken, CamemBERT'inki 32.000'dir.
* Parametre Sayısı: FlauBERT 138 milyon parametreye sahipken, CamemBERT 110 milyon parametreye sahiptir.

Eğitim Veri Seti Boyutu: FlauBERT 71 GB'lık bir veri seti üzerinde eğitilirken, CamemBERT 138 GB'lık bir veri seti üzerinde eğitilmiştir [43].

XLM, Facebook AI Research tarafından geliştirilen ve çok dilli doğal dil işleme (NLP) görevleri için tasarlanmış güçlü bir çapraz dilli dil modelidir. XLM, farklı dillerdeki metinlerden anlamsal temsiller öğrenmek için hem denetimli hem de denetimsiz öğrenme tekniklerini kullanır. XLM'in çeşitli çapraz dilli görevlerde, özellikle çapraz dilli sınıflandırma, denetimsiz ve denetimli makine çevirisi ve düşük kaynaklı dil modellemesinde nasıl kullanılabileceğini vurgulamaktadır [44].

CTRL, Salesforce Research tarafından geliştirilen 1.63 milyar parametreli koşullu bir transformatör dil modelidir. Stil, içerik ve göreve özgü davranışı yöneten kontrol kodlarına göre koşullandırılmak üzere eğitilmiştir [45].

Reformer modeli, uzun dizilerle çalışırken Transformer'ların bellek ve hesaplama sınırlamalarını aşmak için Google tarafından geliştirilmiştir. Tek bir hızlandırıcıda yalnızca 16 GB bellek kullanarak 1 milyon kelimeye kadar bağlam pencerelerini işleyebilecek şekilde tasarlanmıştır. Bunu, dikkat hesaplamasının karmaşıklığı ve bellek ayırma sorunlarını çözmek için iki temel teknik kullanarak başarır: Yerellik Duyarlı Karma (LSH) ile dikkat hesaplamalarının karmaşıklığını O(n²)'den O(nlogn)'ye düşürürken, tersine çevrilebilir artık katmanlar sayesinde bellek verimliliğini artırır. Bu özellikler, reformer modellerini uzun dizileri işleyebilen, daha hızlı eğitilebilen ve bellek açısından çok daha verimli hale getirir. Dil modelleme, sohbet robotları ve uzun vadeli bağımlılıkların kritik olduğu diğer NLP görevlerinde etkili bir alternatiftir [46].

ALBERT (A Lite BERT): Google Research ve Toyota Technological Institute tarafından geliştirilen ve BERT'in bellek kullanımını ve eğitim süresini azaltmayı amaçlayan bir modeldir. ALBERT, cross-layer parameter sharing ve sentence ordering objectives tekniklerini kullanır. XLNet: Google Brain ve Carnegie Mellon University tarafından geliştirilen ve BERT'in sınırlamalarını ele alarak performansını artıran bir modeldir. XLNet, permutation language modeling (PLM) tekniğini kullanır [47].

ELECTRA, "Token Değiştirmelerini Doğru Şekilde Sınıflandıran Bir Kodlayıcıyı Verimli Bir Şekilde Öğrenme" anlamına gelir. ELECTRA, doğal dil işleme (NLP) görevlerinde kullanım için metin kodlayıcıları önceden eğitmek için tasarlanmış yeni bir yöntemdir.

* ELECTRA, girişi maskelemek yerine, küçük bir üretici ağdan örneklenen makul alternatiflerle bazı token'ları değiştirerek bozar.
* Ardından, model, bozulan girdideki her bir token'ın bir üretici örneğiyle değiştirilip değiştirilmediğini tahmin eden ayrımcı bir model olarak eğitilir.
* ELECTRA'nın temel avantajı, BERT ve XLNet gibi maskeli dil modelleme (MLM) yaklaşımlarından daha verimli olmasıdır. Bunun nedeni, görevin, yalnızca maskelenmiş küçük alt küme yerine tüm girdi token'ları üzerinde tanımlanmasıdır.
* Sonuç olarak, ELECTRA tarafından öğrenilen bağlamsal temsiller, aynı model boyutu, veri ve hesaplama göz önüne alındığında BERT tarafından öğrenilenlerden önemli ölçüde daha iyi performans gösterir. Kazanımlar özellikle küçük modeller için güçlüdür [48].

3.6. Kümeleme Yöntemleri:

Bu çalışmada, verilerin kümeleme işlemi için çeşitli yöntemler kullanılmıştır. K-Means, MiniBatch K-Means ve Agglomerative Clustering gibi popüler algoritmalar, verinin doğal yapısını keşfetmek ve farklı kümeleri tanımlamak için tercih edilmiştir. Ayrıca, Spectral Clustering ve Hierarchical Clustering gibi daha ileri düzey kümeleme yöntemleri, veri arasındaki benzerlikleri daha derinlemesine incelemek için uygulanmıştır. Birch algoritması ise büyük veri setlerinde daha verimli sonuçlar elde edilmesini sağlamıştır. Bu yöntemler, veri setinin çeşitli özelliklerini keşfetmek ve kümelere ayırmak amacıyla etkin bir şekilde kullanılmıştır.

Kümeleme Yöntemleri:

* K-Means
* Spectral Clustering
* Hierarchical Clustering
* MiniBatch K-Means
* Birch
* Agglomerative Clustering

3.6.1. K-Means

K-Means, en yaygın kullanılan kümeleme algoritmalarından biridir. Temel prensibi, verileri önceden belirlenmiş K sayıda kümeye bölen bir bölümleme yöntemidir. K-Means algoritmasında, başlangıç olarak K tane veri noktası merkez (centroid) olarak rastgele seçilir. Ardından, her bir veri noktası, merkezlere olan mesafeleri (genellikle Öklid uzaklığı) dikkate alınarak en yakındaki merkeze atanır. Bu işlemin ardından, her küme içindeki veri noktalarının ortalaması alınarak merkezler güncellenir. Bu süreç, merkezlerin konumları sabit hale gelene kadar veya belirli bir durma kriteri sağlanana kadar tekrarlanır [49].

3.6.2. Spectral Clustering

Spektral kümeleme, graf teorisi ve lineer cebir kullanarak verileri gruplamak için bir yöntemdir. Kümelemeden önce boyutluluk azaltma yapmak için bir benzerlik matrisinin spektrumunu (özdeğerler ve özvektörler) kullanır. Algoritma, veri noktaları arasındaki benzerlikleri temsil eden bir benzerlik matrisinin özdeğer ve özvektörlerini kullanarak verileri düşük boyutlu bir uzaya projekte edilir. Son olarak, K-Means gibi bir yöntemle bu yeni uzaydaki veriler kümeleme işlemine tabi tutulur. Bu sayede, geleneksel kümeleme algoritmaları daha etkin bir şekilde uygulanabilir [49].

3.6.3. Hierarchical Clustering

Bu yaklaşım, bir küme hiyerarşisi oluşturur; ya aşağıdan yukarıya (agglomerative) ya da yukarıdan aşağıya (divisive) bir yöntemle. Agglomerative kümeleme, her veri noktasının kendi kümesi olarak başladığı ve en yakın küme çiftlerinin birleştirildiği bir süreçtir. Bu yöntem, kümeleme sürecini görsel olarak temsil eden bir dendrogram sağlar. Bu yöntem, önceden belirlenmiş bir küme sayısı gerektirmez.

Nasıl Çalışır:

* Toplayıcı (aşağıdan yukarıya): Her veri noktası başlangıçta kendi başına bir küme olarak başlar ve en yakın çiftler birleştirilir.
* Bölücü (yukarıdan aşağıya): Tüm veri noktaları tek bir kümede başlar ve yinelemeli olarak bölünürler [49].

3.6.4. MiniBatch K-Means

Algoritma, her yinelemede tüm veri seti yerine, her iterasyonda rastgele seçilen küçük bir veri alt kümesi (mini-batch) kullanılır. Bu yaklaşım, hesaplama maliyetini azaltır, büyük veri setlerinde bellek kullanımını azaltır ve bu büyük veri kümeleri için daha hızlı olmasını sağlar [49].

3.6.5. Birch

Büyük veri kümeleri için tasarlanmıştır gelen veri noktalarını artımlı ve dinamik olarak kümeleyen bir hiyerarşik kümeleme yöntemidir ve bir ağaç yapısı oluşturur. Algoritma, veri noktalarını özetleyen bir CF (Clustering Feature) ağacı oluşturur. CF ağacı, bellek kullanımını en aza indirerek büyük veri kümelerinin işlenmesine olanak tanır. Son adımda, CF ağacının yaprak düğümlerinin merkezleri kullanılarak K-Means algoritması çalıştırılır ve nihai kümeleme sonucu elde edilir [49].

3.6.6. Agglomerative Clustering

Agglomerative Clustering, hiyerarşik kümeleme yaklaşımının bir parçasıdır ve aşağıdan yukarıya doğru çalışan bir algoritmadır. Her veri noktası başlangıçta ayrı bir küme olarak kabul edilir. En yakın iki küme birleştirilir ve bu işlem, tüm veriler tek bir küme olana kadar devam eder. Kullanıcı, hangi mesafe ölçütünün (örneğin, Euclidean veya Manhattan) kullanılacağını ve hangi kriterin (örneğin, en kısa mesafe veya en uzun mesafe) temel alınacağını belirleyebilir [49].

3.7. Sınıflandırma için kullanılan algoritmalar

Veri kümesi üzerinde ön işleme adımları uygulantıkdan sonra veri sınıflandırmaya hazır hale geldiğinde sınıflandırma aşamasına geçilmiştir bu aşamada veri kümesini sınıflandırmak için makine öğrenimi sınıflandırma algoritmaları kullanılmıştır. Bu çalışmada kullanılan algoritmalar; Multinomial Naive Bayes (MNB), Lojistik Regresyon (LR), Destek Vektör Makineleri (SVM), Karar Ağacı (DT), K-En Yakın Komşu (KNN), Rastgele Orman (RF) kullanılmıştır.

3.7.1. Multinomial Naive Bayes

Çok terimli Naive Bayes, özellikle metin sınıflandırması problemlerinde kullanılan etkili bir makine öğrenmesi algoritmasıdır. Bir belgenin, o belgedeki kelimelerin sıklığına ve o kelimelerin o kategorideki belgelerde görünme olasılığına bağlı olarak belirli bir kategoriye ait olma olasılığını hesaplayarak çalışır. Çok terimli Naive Bayes algoritması genellikle TF-IDF (Terim Frekansı-Ters Belge Frekansı) adı verilen bir özellik çıkarma tekniği ile birlikte kullanılır [50]. Çok Terimli Naive Bayes, metin sınıflandırması için özellikle uygun olan bir olasılık tabanlı sınıflandırma algoritmasıdır. Bu algoritma, bir metnin belirli bir sınıfa ait olma olasılığını hesaplar ve en yüksek olasılığa sahip sınıfı tahmin olarak seçer [51]. MNB, özellikle özellik sayısının (benzersiz kelimeler) çok büyük olabileceği metin sınıflandırma görevlerinde popülerdir. Her belge bir kelime frekansları vektörü olarak temsil edilir ve MNB belgeleri bu frekanslara göre sınıflandırır [52].

3.7.2. Logistic Regression

Lojistik regresyon (LR), ikili verileri modellemek için kullanılan bir istatistiksel yaklaşımdır. Bu yöntemde, her bir gözlem için bağımlı değişken ​ sadece 0 veya 1 değerlerini alır ve bu durumu Bernoulli olasılık yoğunluk fonksiyonu ile modellemeye çalışır. Lojistik regresyon, genelleştirilmiş lineer modellerin bir parçasıdır ve parametreleri maksimum olabilirlik yöntemi ile tahmin edilir. Lojistik regresyon modeli, bir gözlemin 1 olma olasılığını, bir vektör ​ile ifade edilen açıklayıcı değişkenlere bağlar. Modelin temel denklemi şu şekildedir:

(1)

Burada b vektörü, model parametrelerini içerir ve maksimum olabilirlik yöntemi ile tahmin edilir. Lojistik regresyonun tahminlerini yorumlamak için, belirli bir eşik değeri (genellikle 0.5) üzerinde ise 1, altında ise 0 olarak etiketleme yapılır [53].

3.3.3.. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) sınıflandırma algoritması ilk kez 1967'de Cover ve Hart tarafından önerilmiştir [23]. K-en yakın komşu sınıflandırması, olasılık yoğunluklarının belirgin parametrik yaklaşımlarının bilinmediği veya belirlenmesinin zor olduğu durumlarda karakteristik analiz yapmak amacıyla geliştirildi. KNN algoritması, parametrik olmayan bir sınıflandırma algoritması olup, temel veri kümesi üzerinde herhangi bir varsayımda bulunmaz. Bilinen basitliği ve etkinliği ile öne çıkar. Bu algoritma, denetimli bir öğrenme yöntemidir. Sınıflandırmada, etiketlenmemiş verinin hangi sınıfa ait olduğunu belirleyen çeşitli özellikler kullanılır. KNN genellikle bir sınıflandırıcı olarak kullanılır. Veriyi sınıflandırmak için belirli bir bölgedeki en yakın eğitim örneklerine dayanır. Bu yöntem, uygulanması kolaylığı ve düşük hesaplama süresi nedeniyle tercih edilir. Sürekli veriler için en yakın komşuları hesaplamak için Öklidyen mesafeyi kullanır [54].

3.3.4. Decision Tree

Karar Ağacı, makine öğrenimi ve veri madenciliği alanlarında sıkça kullanılan bir sınıflandırma ve regresyon algoritmasıdır. Karar ağacı algoritması, denetimli bir makine öğrenme algoritmasıdır. Diğer algoritmalarla karşılaştırıldığında, her değişkenin tahmin sürecindeki rolünü net bir şekilde gösterir. DT'ler genellikle yukarıdan-aşağı yaklaşımı ve tekrarlayan böl-ve-hükmet süreci kullanılarak oluşturulur. Bu ağaçlar, kök, dal, düğüm ve yapraklardan oluşan bir yapıya sahiptir. Kök, ağacın başlangıcıdır ve tüm örneği içerir. Her düğüm belirli bir özniteliği temsil eder ve bu öznitelikten yapılan bölünmeler dal oluşturur. Her düğümde, örnekler belirli bir değere göre iki alt kümeye ayrılır. Bu süreç, bir sonuca ulaşana kadar veya veri bölünemeyecek hale gelene kadar devam eder. DT oluşturmanın ilk adımı, kök düğüme yerleştirilecek özniteliğin seçimidir. Ardından, her düğümde mevcut örnekler belirli bir kriterle alt kümelere bölünür. Tüm örnekler aynı sınıflandırmaya sahipse veya veriyi daha fazla bölme avantaj sağlamıyorsa dal oluşturma durur [55].

3.3.5. SVM

Vapnik tarafından geliştirilen Destek Vektör Makineleri (SVM), hem sınıflandırma hem de regresyon problemleri için kullanılan bir denetimli öğrenme algoritmasıdır. n-boyutlu uzay sınıflandırma probleminde, etiketli eğitim verileri (i = 1 ... k) iki pozitif ve negatif sınıfa ait olarak kabul edilir ve sırasıyla = +1 ve = −1 olarak etiketlenir. SVM'nin amacı, sınıf verilerini ayırmak için bir hiper-düzlem bulmaktır, bu şekilde hiper-düzlem ile en yakın veri arasındaki mesafe maksimize edilir. Bu en iyi hiper-düzlem, maksimum marj hiper-düzlemi olarak adlandırılır. Bu şekilde, SVM, özellikle sınıflar arasındaki ilişki doğrusal olmadığında ve diğer sınıflandırma yöntemleri ile karşılaştırıldığında daha az parametre içerdiğinden, pratik uygulamalarda tercih edilen bir yöntemdir [53].

3.3.6. Random Forest

Rastgele Orman sınıflandırıcısı, bağımsız örneklenmiş rastgele vektörlerle oluşturulan ağaç sınıflandırıcılarının bir kombinasyonundan oluşur. Bu yöntem, her bir ağacın giriş vektörünü en popüler sınıfa sınıflandırmak için birim oy verdiği güçlü bir sınıflandırma modelidir. Her düğümde rastgele seçilen özelliklerle ağaçlar büyütülür ve bagging yöntemiyle eğitim veri kümesi oluşturulur. Rastgele orman, Gini İndeksi'ni kullanarak özellik seçimini gerçekleştirir. Ağaçlar, tamamen büyütüldükten sonra budanmaz, bu da aşırı uyuma karşı direnç sağlar. Performansı etkileyen önemli parametreler arasında ağaç sayısı ve her düğümde seçilen özellik sayısı bulunmaktadır [39]. Rastgele Orman modeli, aşırı öğrenme sorunundan kaçındığı için bir karar ağacı modelinden daha iyi performans gösterir. Rastgele orman modelleri, birbirinden tamamen farklı bir dizi karar ağacından oluşur. Rastgele Orman'da bir orman veya enseble oluşturmak için bir dizi ağaç kullanılır ve her ağaç sürekli olarak değerlendirilir [56].

3.4. Değerlendirme Metrikleri

Çalışmada her sınıflandırma algoritması için accuracy, precision, recall, ve F1-score ölçümleri karışıklık matrisi kullanarak ölçülmüştür. Karışıklık matrisi, bir sınıflandırma sistemi tarafından gerçekleştirilen gerçek ve tahmini sınıflandırmalar hakkında bilgi içeren bir makine öğrenimi kavramıdır. Bir karışıklık matrisi iki boyutlu bir yapıya sahiptir; bir boyutu bir nesnenin gerçek sınıfı tarafından indekslenirken, diğer boyutu sınıflandırıcının tahmin ettiği sınıflar tarafından indekslenir [51]. Tüm algoritmaların performans analizi, True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR) ve False Negative Rate (FNR) gibi metrikler aracılığıyla hesaplanır. TPR, test edilen gerçek pozitiflerin pozitif olarak sınıflandırılma oranıdır. TNR, test edilen gerçek negatiflerin negatif olarak sınıflandırılma oranıdır. FPR, negatif örneklerin pozitif olarak sınıflandırılma oranıdır ve bu oran, negatif örneklerin toplam örnek sayısına bölünür. FNR ise pozitif örneklerin negatif olarak sınıflandırılma oranıdır ve bu oran, pozitif örneklerin toplam örnek sayısına bölünür [57].

Confusion matrix üzerinden birkaç temel değerlendirme metriği hesaplanabilir:

Tablo 2. Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Gerçek Değerler | |
| + | - |
| Tahmin Edilen Değerler | + | True Positive (TP) | False Positive (FP) |
| - | False Negative (FN) | True Negative (TN) |

* Doğruluk, doğru tahminlerin oranıdır. Bir sınıflandırıcının bir durumu ne kadar iyi tahmin ettiğini ölçer.

(9)

* Hassasiyet, modelin pozitif durumları tahmin etme yeteneğidir. Gerçekten pozitif tahminler ile tüm pozitif tahminler kümesi arasındaki oran olarak hesaplanır.

(10)

* Kesinlik, geri alınan örnekler arasındaki ilgili durumları değerlendirmek için kullanılan bir ölçüdür. Gerçekten pozitif tahminler ile tüm gerçek pozitif değerler kümesi arasındaki oran olarak hesaplanır.

(11)

* F-Skor, Hassaslık ve Kesinliği aynı şekilde dikkate alan bir ölçümdür. Hassaslık ve kesinlik arasındaki harmonik ortalama olarak hesaplanır [49].

(12)

Siluet Puanı, bir nesnenin diğer kümelere kıyasla kendi kümesine ne kadar benzer olduğunun bir ölçüsüdür. Daha yüksek bir puan, daha iyi tanımlanmış kümeleri gösterir. Davies-Bouldin Endeksi, kümelerin ayrılmasını ve kompaktlığını değerlendirir. Daha düşük bir DBI değeri, daha belirgin ve kompakt kümelerle daha iyi kümeleme performansını gösterir. Calinski-Harabasz Endeksi, kümeler arasındaki dağılımın kümeler içindeki dağılıma oranını ölçer. Daha yüksek bir değer, birbirinden iyi ayrılmış daha iyi tanımlanmış kümeleri gösterir [58].

4. Deneysel Sonuçlar

Bu çalışmada twitter platformundan veri çekilerek oluşturulan türkçe ve ingilizce veri seti kullanılarak geleneksel gömme yöntemleri, kelime gömme modelleri ve BERT tabanlı gömme yöntemleri kullanarak metin sınıflandırma performansını karşılaştırmak amaçlanmıştır. Farklı kümeleme algoritmalarının farklı veri setleri üzerindeki performansları değişebilir. Bu nedenle, uygun kümeleme algoritmasını seçmek için deneysel çalışmalar yapılması gerekmektedir. Cinsiyet tesbitini tahmin etmek için farklı sınıflandırma tekniklerini kullanılmıştır.

5. Sonuç

Cinsiyet çıkarımı, metin analizi ve doğal dil işleme teknikleri kullanılarak gerçekleştirilmektedir. Twitter gibi platformlarda kullanılan dil ve ifadeler, kullanıcıların cinsiyeti hakkında ipuçları sunmakta ve bu ipuçlarının doğru bir şekilde yorumlanması, toplumsal cinsiyet eşitliği konusundaki anlayışımızı derinleştirmektedir. Kümeleme tabanlı metin sınıflandırma, cinsiyet çıkarımı gibi karmaşık sosyal konuları anlamak için güçlü bir araçtır. Hiyerarşik kümeleme, K-means ve yoğunluk bazlı yöntemler gibi teknikler sayesinde metinlerdeki cinsiyet bilgileri etkili bir şekilde analiz edilebilir. Gelecekte bu alandaki çalışmaların artmasıyla birlikte daha kapsamlı ve hassas cinsiyet çıkarım sistemlerinin geliştirilmesi beklenmektedir.

References

[1] Onikoyi, B., Nnamoko, N., & Korkontzelos, I. (2023). Gender prediction with descriptive textual data using a machine learning approach. Natural Language Processing Journal, 4, 100018.

[2] Tantuğ, A. C. (2016). Metin sınıflandırma. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5(2).

[3] YILDIRIM, S., & YILDIZ, T. (2018). Türkçe için karşılaştırmalı metin sınıflandırma analizi. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 24(5), 879-886.

[4] UÇKAN, T., Cengiz, H. A. R. K., Seyyarer, E., & Karci, A. (2019). Ağırlıklandırılmış çizgelerde Tf-Idf ve eigen ayrışımı kullanarak metin sınıflandırma. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi, 8(4), 1349-1362.

[5] Tekin, M. C., & Tunali, V. (2019). Yazılım geliştirme taleplerinin metin madenciliği yöntemleriyle önceliklendirilmesi. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 25(5), 615-620.

[6] Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics, 2015(1), 198363.

[7] Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C. C. J. (2019). Evaluating word embedding models: Methods and experimental results. APSIPA transactions on signal and information processing, 8, e19.

[8] Wolf, T. (2019). Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

[9] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

[10] Mittal, H., Pandey, A. C., Pal, R., & Tripathi, A. (2021). A new clustering method for the diagnosis of CoVID19 using medical images. Applied Intelligence, 51, 2988-3011.

[11] GÜNERKAN, M., ŞAHİNASLAN, E., & ŞAHİNASLAN, Ö. (2022). Gümrük beyannamesi sürecinde öğrenmeye dayalı algoritmaların etkinliğinin incelenmesi. Acta Infologica.

[12] Miller, Z., Dickinson, B., & Hu, W. (2012). Gender prediction on twitter using stream algorithms with n-gram character features.

[13] Onikoyi, B., Nnamoko, N., & Korkontzelos, I. (2023). Gender prediction with descriptive textual data using a machine learning approach. Natural Language Processing Journal, 4, 100018.

[14] Alowibdi, J. S., Buy, U. A., & Yu, P. (2013, December). Empirical evaluation of profile characteristics for gender classification on twitter. In 2013 12th international conference on machine learning and applications (Vol. 1, pp. 365-369). IEEE.

[15] Liu, J. (2023). Sentiment Classification of Social Network Text Based on AT-BiLSTM Model in a Big Data Environment. Int. J. Inf. Technol. Syst. Approach, 16, 1-15.

[16] Yu, Q., Wang, Z., & Jiang, K. (2021). Research on Text Classification Based on BERT-BiGRU Model. Journal of Physics: Conference Series, 1746.

[17] To, H.Q., Nguyen, K.V., Nguyen, N.L., & Nguyen, A.G. (2020). Gender Prediction Based on Vietnamese Names with Machine Learning Techniques. Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval.

[18] Basile, A., Dwyer, G., & Rubagotti, C. (2019). Simple n-gram based models perform well for gender prediction . Sometimes.

[19] Tripathi, A., & Faruqui, M. (2011). Gender prediction of Indian names. *IEEE Technology Students' Symposium*, 137-141.

[20] Rego, R.C., Silva, V.M., & Fernandes, V.M. (2021). Predicting Gender by First Name Using Character-level Machine Learning.

[21] Bensahla, A., Zaghir, J., Gaudet-Blavignac, C., & Lovis, C. (2024). Unsupervised Extraction of Body-Text from Clinical PDF Documents. In Digital Health and Informatics Innovations for Sustainable Health Care Systems (pp. 214-215). IOS Press.

[22] Urrutia, R., Espejo, D., Evens, N., Guerra, M., Sühn, T., Boese, A., ... & Poblete, V. (2023). Clustering Methods for Vibro-Acoustic Sensing Features as a Potential Approach to Tissue Characterisation in Robot-Assisted Interventions. Sensors, 23(23), 9297.

[23] Taamneh, M., Taamneh, S., & Alkheder, S. (2017). Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks. International journal of injury control and safety promotion, 24(3), 388-395.

[24] Davagdorj, K., Wang, L., Li, M., Pham, V. H., Ryu, K. H., & Theera-Umpon, N. (2022). Discovering thematically coherent biomedical documents using contextualized bidirectional encoder representations from transformers-based clustering. International Journal of Environmental Research and Public Health, 19(10), 5893.

[25] Woo, H., Kim, K., Cha, K., Lee, J. Y., Mun, H., Cho, S. J., ... & Kang, M. (2019). Application of efficient data cleaning using text clustering for semistructured medical reports to large-scale stool examination reports: methodology study. Journal of medical Internet research, 21(1), e10013.

[26] Wang, Y., Wang, Y., Peng, Z., Zhang, F., Zhou, L., & Yang, F. (2023). Medical text classification based on the discriminative pre-training model and prompt-tuning. Digital Health, 9, 20552076231193213.

[27] Yetisgen-Yildiz, M., & Pratt, W. (2005). The effect of feature representation on MEDLINE document classification. In AMIA annual symposium proceedings (Vol. 2005, p. 849). American Medical Informatics Association.

[28] Tan, Z., Lin, J., Chen, K., Zhuang, Y., & Han, L. (2023). Skin cancer image recognition based on similarity clustering and attention transfer. Journal of X-Ray Science and Technology, 31(2), 337-355.

[29] Ahne, A., Fagherazzi, G., Tannier, X., Czernichow, T., & Orchard, F. (2022). Improving diabetes-related biomedical literature exploration in the clinical decision-making process via interactive classification and topic discovery: Methodology development study. Journal of medical Internet research, 24(1), e27434.

[30] Workman, T. E., Weir, C., & Rindflesch, T. C. (2017, February). Differentiating sense through semantic interaction data. In AMIA Annual Symposium Proceedings (Vol. 2016, p. 1238).

[31] Dipnall, J. F., Lu, J., Gabbe, B. J., Cosic, F., Edwards, E., Page, R., & Du, L. (2022). Comparison of state-of-the-art machine and deep learning algorithms to classify proximal humeral fractures using radiology text. European journal of radiology, 153, 110366.

[32] Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A survey of text representation and embedding techniques in nlp. IEEE Access, 11, 36120-36146.

[33] Rahman, S., Biplob, K., Rahman, M., Sarker, K. ve Islam, T. (2020). Duygu Sınıflandırmasında N-Gram, TF-IDF ve Topluluk Yöntemlerinin Araştırılması ve Değerlendirilmesi., 391-402. https://doi.org/10.1007/978-3-030-52856-0\_31 .

[34] Wolf, F., Poggio, T.A., & Sinha, P. (2006). Human Document Classification Using Bags of Words.

[35] Zheng, L., Wang, S., & Tian, Q. (2014). Coupled binary embedding for large-scale image retrieval. IEEE transactions on image processing, 23(8), 3368-3380.

[36] Giouvanakis, E., & Kotropoulos, C. (2014, August). Saliency map driven image retrieval combining the bag-of-words model and PLSA. In 2014 19th International Conference on Digital Signal Processing (pp. 280-285). IEEE.

[37] Dias Canedo, E., & Cordeiro Mendes, B. (2020). Software requirements classification using machine learning algorithms. Entropy, 22(9), 1057.

[38] Rui, W., Xing, K., & Jia, Y. (2016). BOWL: Bag of word clusters text representation using word embeddings. In Knowledge Science, Engineering and Management: 9th International Conference, KSEM 2016, Passau, Germany, October 5-7, 2016, Proceedings 9 (pp. 3-14). Springer International Publishing.

[39] Nurdin, A., Aji, B. A. S., Bustamin, A., & Abidin, Z. (2020). Perbandingan Kinerja Word Embedding Word2Vec, Glove, Dan Fasttext Pada Klasifikasi Teks. Jurnal Tekno Kompak, 14(2), 74-79.

[40] Gupta, R. (2024). Bidirectional encoders to state-of-the-art: a review of BERT and its transformative impact on natural language processing. Информатика. Экономика. Управление/Informatics. Economics. Management, 3(1), 0311-0320.

[41] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (pp. 38-45).

[42] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.

[43] Blanc, C., Bailly, A., Francis, É., Guillotin, T., Jamal, F., Wakim, B., & Roy, P. (2022). FlauBERT vs. CamemBERT: Understanding patient's answers by a French medical chatbot. Artificial Intelligence in Medicine, 127, 102264.

[44] Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. Advances in neural information processing systems, 32.

[45] Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858.

[46] Riaz, N., Latif, S., & Latif, R. (2021, May). From transformers to reformers. In 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2) (pp. 1-6). IEEE.

[47] Areshey, A., & Mathkour, H. (2024). Exploring transformer models for sentiment classification: A comparison of bert, roberta, albert, distilbert, and xlnet. Expert Systems, 41(11), e13701.

[48] Clark, K. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

[49] Mussabayev, R., & Mussabayev, R. (2023). Optimizing K-means for Big Data: A Comparative Study. arXiv preprint arXiv:2310.09819.

[50] Riego, N. C. R., & Villarba, D. B. (2023). Utilization of Multinomial Naive Bayes Algorithm and Term Frequency Inverse Document Frequency (TF-IDF Vectorizer) in Checking the Credibility of News Tweet in the Philippines. arXiv preprint arXiv:2306.00018.

[51] Farisi, A. A., Sibaroni, Y., & Al Faraby, S. (2019, March). Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier. In Journal of Physics: Conference Series (Vol. 1192, No. 1, p. 012024). IOP Publishing.

[52] Jiang, L., Wang, S., Li, C., & Zhang, L. (2016). Structure extended multinomial naive Bayes. Information Sciences, 329, 346-356.

[53] Musa, A. B. (2013). Comparative study on classification performance between support vector machine and logistic regression. International Journal of Machine Learning and Cybernetics, 4, 13-24.

[54] Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019, May). A brief review of nearest neighbor algorithm for learning and classification. In 2019 international conference on intelligent computing and control systems (ICCS) (pp. 1255-1260). IEEE.

[55] Colledani, D., Anselmi, P., & Robusto, E. (2023). Machine learning-decision tree classifiers in psychiatric assessment: An application to the diagnosis of major depressive disorder. Psychiatry Research, 322, 115127.

[56] Uddin, K. M. M., Alam, M. J., Uddin, M. A., & Aryal, S. (2023). A Novel Approach Utilizing Machine Learning for the Early Diagnosis of Alzheimer's Disease. Biomedical Materials & Devices, 1-17.

[57] Priya, P. M. (2023, February). Disease Prediction and Classification using Intelligent Algorithms. In 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS) (pp. 494-498). IEEE.

[58] Aryani, D., Kharisma, I. L., Sujjada, A., & Kamdan, K. (2024). Topic Modeling of the 2024 Election Using the BERTopic Method on Detik. com News Articles. Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi, 9(2), 171-180.

[59] Antonakaki, D., Fragopoulou, P., & Ioannidis, S. (2021). A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Syst. Appl.*, 164, 114006. https://doi.org/10.1016/J.ESWA.2020.114006.