KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI

YÜKSEK LİSANS TEZİ

BABELNET İLE ZENGİNLEŞTİRİLEN TWITTER DOKÜMAN UZAYINDA DUYGU ANALİZİ İÇİN İŞBİRLİKÇİ MODEL KURULMASI

SEMİH SEVİM

KOCAELİ 2019

KOCAELİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ

ANABİLİM DALI

YÜKSEK LİSANS TEZİ

BABELNET İLE ZENGİNLEŞTİRİLEN TWITTER DOKÜMAN UZAYINDA DUYGU ANALİZİ İÇİN İŞBİRLİKÇİ MODEL KURULMASI

SEMİH SEVİM

Doç. Dr. Sevinç İLHAN OMURCA
Danışman, Kocaeli Üniversitesi .....

Dr. Öğr. Üyesi Alev MUTLU
Jüri Üyesi, Kocaeli Üniversitesi .....

Dr. Öğr. Üyesi Zeynep Hilal KİLİMCİ
Jüri Üyesi, Doğuş Üniversitesi .....

Tezin Savunulduğu Tarih: 04.07.2019

ÖNSÖZ VE TEŞEKKÜR

Sosyal medya uygulamalarının insan hayatıyla sıkı bir ilişki içerisinde olması günümüzde üretilen veri miktarının en büyük sebeplerinden biri olarak gösterilebilir. Kolay, hızlı ve küresel bir iletişim sağlaması sosyal medya uygulamalarının cazibesini arttıran en önemli etkendir. İnsanların herhangi bir ticari ürün hakkındaki görüşlerinden başlayarak geniş kitleleri ilgilendiren toplumsal konulara kadar geniş bir paylaşım yelpazesine sahip olan sosyal medya uygulamaların barındırdıkları potansiyel birçok kesimin dikkatini çekmiştir. Çeşitli alanlardaki bilimsel disiplinler bu uygulamaların içerdiği kullanışlı bilgileri en verimli şekilde elde edebilmek için çalışmalar gerçekleştirmektedir.

Tez çalışmasının gerçekleştirilmesinde her koşulda desteğini ve tavsiyelerini esirgemeyen danışman hocam Doç. Dr. Sevinç İlhan Omurca'ya, bu süreçteki büyük yardımlarından dolayı Arş. Gör. Ekin Ekinci'ye ve desteğini benden esirgemeyen aileme teşekkür ederim.

Haziran – 2019 Semih SEVİM

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar

NB	:	Naive Bayes
SVM	:	Support Vector Machine (Destek Vektör Makineleri)
DT	:	Decision Tree (Karar Ağacı)
KNN	:	K-Nearest Neighbour (K En Yakın Komşu)
Tf-Idf	:	Term Frequency-Inverse Document Frequency (Terim Frekansı- Ters Doküman Frekansı)
POS	:	Part of Speech

BABELNET İLE ZENGİNLEŞTİRİLEN TWITTER DOKÜMAN UZAYINDA DUYGU ANALİZİ İÇİN İŞBİRLİKÇİ MODEL KURULMASI

ÖZET

Duygu ve düşüncelerin paylaşımında kullanılan sosyal medya uygulamaları insanların birbirleriyle kolay ve hızlı iletişim kurmalarını sağlayan kullanışlı araçlardır. Çeşitli konular ile ilgili görüşlerin paylaşıldığı bu uygulamalar içerdikleri veri miktarı nedeniyle birçok araştırma alanı için kaynak oluşturmaktadır. Örneğin organizasyonların ürünleriyle ilgili pazar analizi yapması, hükümetlerin yaptığı çalışmalar ile ilgili halkın görüşünü belirlemek yapılan araştırmalar arasındadır. Sosyal medya uygulamaları, sağladıkları zengin veri içeriği yanında birçok zorluğu da bünyesinde barındırmaktadır. Üretilen elektronik dokümanların büyük bir bölümü dil bilgisi ve yazım kurallarına uymamaktadır. Ayrıca dokümanların içerdiği ifade simgeleri, kısaltmalar ve argo kelimeler kullanışlı bilginin elde edilmesindeki engellerin arasındadır. Bilgi çıkarımında karşılaşılan en büyük zorluk dokümanların kısa olmasıdır. Sayılan tüm zorluklar doküman analizini daha karmaşık hale getirmektedir. Gerçekleştirilen bu tez çalışmasında bahsedilen zorlukları aşmak için işbirlikçi öğrenme modeli önerilmiştir. Dokümanlardaki yazım hatalarını ve kısaltmaları düzeltmek için bir sözlükten kullanılmıştır. Model için gerekli doküman uzayı BabelNet anlamsal ağı kullanılarak zenginleştirilmiştir. Zenginleştirilen doküman uzayından üç farklı veri kümesi oluşturulmuş, bu veri kümeleri işbirlikçi öğrenme modelinin kurulmasında kullanılmıştır. Tezin sonunda yapılan işlemler ve işbirlikçi öğrenme modelinin sonuçları incelenmektedir.

Anahtar kelimeler: Twitter Duygu Analizi, Kısa Metin sınıflandırma, İşbirlikçi Öğrenme, Zenginleştirilmiş Doküman Uzayı.

AN ENSEMBLE MODEL USING A BABELNET ENRICHED DOCUMENT SPACE FOR TWITTER SENTIMENT CLASSIFICATION

ABSTRACT

Social media applications, which are used to share feelings and thoughts, are useful tools that enable people to communicate with each other easily and quickly. These applications, where opinions about various subjects are shared, provide resources for many research areas due to the amount of data they contain. For example, organizations' market analysis of their product and determining opinion of public about the government's works are among researches are done. Even though social media application provides rich data content, they have a lot of challenges. Most of the electronic documents that were produced don't obey syntactic and spelling rules. Emoticons, abbreviations and slang words which are contained by documents prevent to obtain useful information. The biggest challenge in extracting information is that documents are very short. All challenges that are listed make document analysis more complex. In this thesis ensemble learning model is proposed to overcome the challenges that mentioned. A glossary was used to correct spelling errors and abbreviations in documents. Required document space for proposed model is extended with BabelNet which is a semantic network. Three different datasets are derived from enriched document space were used to establish ensemble learning model. At the end of the thesis, the procedures and the results of the ensemble learning model are examined.

Key words: Twitter Sentiment Analysis, Short Text Classification, Ensemble Learning, Enriched Document Space.

GİRİŞ

Sosyal medya uygulamaları kişilerin duygu ve düşüncelerini herhangi bir zamanda herhangi bir yerden kolaylıkla, genellikle kısa mesaj yoluyla, küresel olarak paylaşabildikleri platformlar olarak tanımlanabilir. En basit haliyle insanlar arasında iletişim kanalı oluşturan bu uygulamalar; internetin de gücünü arkasına alarak her geçen gün kullanıcı sayılarını büyük oranda arttırmış, yaşamın birçok alanında kullanmak üzere yeni işlevler kazanmıştır. Günümüzde büyük bir kesim iletişim hızı ve geniş kullanıcı ağı nedeniyle sosyal veya iş yaşamlarında ve hatta siyasi olaylarda bu uygulamalardan faydalanmaktadır. Bu ve bunun gibi birçok faktör bu uygulamalarının durdurulamaz bir şekilde değerinin artmasına yol açmaktadır.

Günümüzün rekabetçi ortamında bilginin öneminin artmasıyla birlikte kullanışlı bilgilerin elde edilebilmesi için farklı kaynaklardan büyük miktarda veri toplamak ve bu verileri etkili yöntemler ile yorumlamak öncelikli hale gelmektedir. Verileri toplamak için kullanılan gözde kaynaklardan biride sosyal medya uygulamalardır. Kullanıcılar tarafından üretilen büyük miktarda veri yine kullanıcılara, şirketlere veya hükümetlere birçok alanda kullanışlı bilgiler sağlamaktadır. Hükümetlerin yasaları düzenlerken halkın görüşlerinin değerlendirebilmesi, şirketlerin ürünleri ile ilgili pazar analizi yapabilmesi ve elde ettiği bilgiler doğrultusunda iş stratejilerini belirleyebilmesi örnek gösterilebilir [1]. Ayrıca sosyoloji, psikoloji, pazarlama ve bilgisayar bilimleri gibi birçok disiplin çalışmalarında sosyal medya uygulamalarından elde ettikleri verileri kullanmaktadır [2]. Özellikle son yıllarda ortaya çıkan Cambridge Analitica skandalında görüldüğü gibi bu uygulamalardan elde edilecek verilerin taşıdığı potansiyel açıkça görülmektedir. Açığa çıkarılan bilgilere göre ABD seçimlerinde halkı demografik olarak gruplandırabilmek için Facebook üzerinden izinsiz elde edilen veriler analiz edilmiş ve farklı gruplar için ayrı propaganda yöntemleri belirlenmiştir [3]. Bu durumun seçin sonuçlarını büyük oranda etkilediği düşünülmektedir. Twitter, yüksek kullanıcı sayısına sahip olması nedeniyle sosyal, ticari ve politik konularda; felaketlerde ve kazalarda insanları bilgilendirmek için ilk başvurulan uygulamalardan biridir [4]. Yaygın kullanımı ve

içerde büyük miktarda veri nedeniyle de bilgi çıkarımı için önemli bir kaynak haline gelmiştir. İstatistiksel verilere bakıldığında aylık 320 milyonluk aktif kullanıcısı bulunan Twitter'da günde 500 milyona yakın mesaj paylaşılmaktadır [5]. Genellikle Twitter'daki verilerin büyük bir kısmını kısa mesajlar oluşturduğu için analizinde sıklıkla başvurulan yöntemlerden biri de metin madenciliğinde kullanılan duygu analizidir. Duygu analizi en basit haliyle metinlerdeki görüşlerin anlamlarına göre otomatik olarak pozitif ve negatif olarak sınıflandırılması işlemine verilen isimdir. Bu yöntem Twitter verileri e-ticaret, sağlık, eğlence ve politika gibi alanlarda bilgi toplamak ve halkın görüşünü almak için sıklıkla tercih edilmektedir [6-11].

Duygu analizinde araştırmacılara zengin bir kaynak sunan Twitter bazı zorlukları da beraberinde getirmektedir. Kullanıcıların geneli paylaştıkları kısa mesajlarda dil bilgisine önem vermemekte ve kullandıkları kelimelerde imla kurallarına dikkat etmemektedirler. Ayrıca mesaj içeriklerinde günlük yaşamda sıklıkla başvurulan jargonlar, kısaltmalar ve argo kelimeler kullanılmaktadır. Belirtilen nedenlerden dolayı mesajları analiz için uygun hale getiren ön işleme süreci için harcanan zaman uzamaktadır [12-14]. Twitter'ın kullanım şartları gereği mesaj uzunluklarının 280 karakter ile sınırlı olması, mesajların bilgi temsili aşamasında özelliklere atanacak değerlerin kısıtlı kalmasına neden olmakta ve mesajların temsili için kullanılan kelimeler arasında elverişli bir istatistiksel ilişkinin kurulması oldukça zorlaşmaktadır. Bu nedenle analiz için kullanılan sınıflandırma algoritmalarının verimli çalışması, etkili sonuçların elde edilmesi zorlaşmaktadır [15, 16]. Bu sorunun üstesinden gelmek için sık başvurulan çözümlerden biri kısa metinleri belirli yöntemler ile içeriğinin genişletilerek zenginleştirilmesidir [49, 92, 93]. Genişletme işlemi dış kaynakların yardımıyla veya metinlerin içerdiği kelimelerin arasındaki ilişkilerden yararlanılarak gerçekleştirilir.

Gerçekleştirilen tez çalışmasının ilk adımında analiz için kullanılacak olan Twitter mesajları genişletilmiş, daha sonraki adımda genişletilen mesajlardan üç farklı veri kümesi türetilmiştir. Bu kümlerden ilkini üzerinde değişiklik yapılmayan ham veriler oluşturur. İkinci küme mesajların içerdiği terimlere karşılık gelen kavram kümelerinin BabelNet [98] kavram ağı yardımı ile mesajlara eklenerek elde edilir. Üçüncü veri kümesi ise mesajların sadece BabelNet üzerinden genişletilebilen terimleri ile bu terimlere karşılık gelen kavram kümesinin birleştirilmesiyle

oluşturulmuş, genişletilemeyen terimler mesaj içerisinden çıkartılmıştır. Tez çalışmasının ikinci adımında Duygu analizi için etkili bir işbirlikçi model oluşturulmaya çalışılmıştır. Bu nedenle homojen yapıda üç farklı işbirlikçi model oluşturulmuş ve aralarında karşılaştırmalar yapılmıştır. İlk işbirlikçi modelin temel sınıflandırıcılarında Naïve Bayes (NB) algoritması kullanılmış, ikinci modelde Destek Vektör Makinesi (SVM) kullanılmıştır. Üçüncü işbirlikçi modeli Karar Ağacı (DT) algoritması olan C4.5 sınıflandırma algoritması ile oluşturulmuştur. Temel sınıflandırıcı olarak kullanılmasıyla oluşturulmuştur. Temel sınıflandırıcılardaki farklılık eğitim kümeleri ile sağlanmıştır. Temel sınıflandırıcıların sonuçları çoğunluk oylaması yöntemi gerçekleştirilerek birleştirilmiştir.

1. GENEL BİLGİLER

1.1. Tez Çalışmasının Amacı Ve Başlatılma Sebepleri

Sosyal medya uygulamaları sahip oldukları kullanıcı sayıları ve buna paralel olarak üretilen verilerinin miktarı nedeniyle birçok çalışma alanı için zengin bir kaynak haline gelmiştir. Özellikle sosyal medya uygulamaların paylaşılan kısa mesajlar birçok çalışmada çıkarım yapmak için en sık kullanılan verileri oluşturmaktadır. İnsanların sosyal hayattan başlayarak politik konulara kadar pek çok alanda görüşlerini belirtmek ve geniş kitleleri bilgilendirmek için kısa mesajlardan faydalanmaları, bu verilen kullanılma sebebini çok net bir şekilde ortaya koymaktadır. Veri miktarının büyüklüğü ve insan eliyle analiz edilmesinin getireceği yüksek maliyet nedeniyle çoğu araştırmacı çalışmalarında bilgisayarın gücünden yararlanmaktadır. Özelikle bilgisayar bilimlerindeki doğal dil işleme ve sınıflandırma algoritmaları en sık başvurulan yöntemlerdir. Ancak mesajların dil bilgisi açısından zayıf olması ve kısa yapıları nedeniyle sınıflandırma işleminin etkili bir şekilde yapılması oldukça zorlaşmaktadır. Mesajların kısa olması etkili özelliklerin elde edilmesini ve analiz sürecinde mesajlar arasında ilişki kurulmasını zorlaştırmaktadır.

Günümüzde soysal medya mesajlardan en verimli şekilde yararlanabilmek için birçok çalışma yürütülmektedir. Özellikle verilerin doğru şekilde sınıflandırılabilmesi için güçlü sınıflandırma modelleri oluşturmak veya özelliklerin çeşitli yöntemler ile zenginleştirilmeye çalışılması en fazla üzerine durulan yöntemlerdir. Metin analizlerinde özellik kümesinin zenginleştirilmesi için kelime gömme yöntemi sıklıkla kullanılmaktadır. Kelime gömme yöntemiyle vektör haline dönüştürülen kelimelerin matematiksel benzerlikleri hesaplanarak genişletme adımı gerçekleştirilmektedir.

Bu çalışmada Twitter mesajlarından oluşan veri kümesi üzerinde duygu analizinin en etkili şekilde gerçekleştirilebilmesi için güçlü bir işbirlikçi model kurulmaya ve bu modelde kullanmak için zengin bir doküman kümesi oluşturulmaya çalışılmıştır.

Dokümanların zenginleştirilmesi için kullanılan belirli yöntemlerin dışında sözlük temelli genişletme yaklaşımının başarıma olan etkisinin gözlemlenmesi amaçlanmıştır. Ayrıca elde edilen dokümanlarından farklı şekilde temsil edilmesi ve farklı sınıflandırma modellerinin kullanılması ile işbirlikçi modellerin başarımlarının arttırılıp arttırılmayacağını gözlemlenmek istenmiştir.

1.2. Tez Çalışmasının Katkıları

Kısa mesajların duygu analizinde karşılaşılan en büyük zorluk sınıflandırma algoritmaları için yeterli bilginin elde edilememesi, bu nedenle oluşturulan modellerin yeterince verimli çalışmamasıdır. Bu çalışmada kısa mesajların arasındaki ilişkinin belirginleştirilmesi ve özellik çıkarımında daha fazla bilgi elde edebilmek için, literatürde var olan yöntemlerin dışında BabelNet isimli ansiklopedik kaynak kullanılarak doküman genişletme işlemi gerçekleştirilmiştir. Yapılan çalışmaların çoğunda doküman genişletme, kelimelerin vektör temsillerinin oluşturulup aralarındaki ilişkinin araştırılmasına dayanmakta ve benzer kelimler yakalandığında dokümana eklenmektedir. Tez çalışmasında doküman üzerinde gerçekleştirilecek çeşitli metotlara ihtiyaç duyulmadan, önceden internet üzerindeki sözlük ve ansiklopedik kaynak verileri ile hazırlanmış BabelNet'in doküman genişletmedeki avantajları ve dezavantajları gösterilmiştir. Önerilen yöntem kelime gömme yöntemleriyle ile karşılaştırıldığında işlem maliyeti açısında yarar sağlamaktadır.

Çalışma içerisinde yürütülen bir diğer süreç, duygu analizini daha verimli bir şekilde gerçekleştirebilmek için etkili bir sınıflandırma modelinin oluşturulmasıdır. Bu nedenle güçlü bir model elde edebilmek için işbirlikçi öğrenme modelinden yararlanılmıştır. Modelde sınıflandırıcı çeşitliliğini sağlaması için farklı veri kümeleri hazırlanırken BabelNet kullanılmış, veri kümelerinin içerdiği terimlere göre üç farklı veri kümesi oluşturulmuştur. İşbirlikçi modelin asıl yapısını oluşturan zayıf öğrenciler için makine öğrenmesi alanında iyi bilinen üç farklı sınıflandırıcı seçilmiştir. Bu sınıflandırıcılar kullanılarak üç ayrı homojen işbirlikçi model oluşturulmuş, veri kümeleri üzerinde işletildikten sonra çoğunluk oylamasıyla birleştirilmiştir.

1.3. Literatür Taraması

Twitter üzerine yapılan çalışmaların popülerliği ve çalışmalardaki konu zenginliği literatür taraması sonucunda doğrudan göze çarpmaktadır. İnsanların düşüncelerini paylaştıkları bu elektronik sosyal ortamlar, özellikle insan davranışlarının gözlemlendiği istatistiksel araştırmalara büyük miktarda veri sağlamaktadır. Ayrıca Twitter mesajlarının analizinde karşılaşılan birçok zorluk araştırmacıları farklı çözüm yolları aramaya yöneltmiştir. Bu çözüm yollarından genelini özellik çıkarımında denenen farklı metotlar ya da daha güçlü sınıflandırıcı modeller için kullanılan çeşitli yaklaşımlar oluşturmaktadır. Çalışmalar incelendiğinde kullanılan yöntemlerin çoğunu doğal dil işleme algoritmaları oluşturmakta ve özellikle duygu analizi bu yöntemlerin başında gelmektedir.

Jain ve Katkar [17] çalışmalarında Twitter mesajları üzerinde duygu analizi gerçekleştirmek için makine öğrenmesi alanında popüler olan Naive Bayes, Bayesian Network, Random Forest ve KNN (K-Nearest Neighbour) sınıflandırma algoritmalarını kullanmışlar ve bu algoritmalar ile oluşturdukları sınıflandırıcılar arasındaki başarımları karşılaştırmışlardır. Çalışmaları sonucunda KNN algoritması ile oluşturdukları sınıflandırıcın başarımının daha yüksek olduğunu gözlemlemişlerdir.

Duygu analizinde kullanılan en yaygın yöntemlerden birisi sözlük tabanlı yaklaşımlardır. Bu yaklaşımların temelini olumlu veya olumsuz anlamlarına göre göre belirli sayı aralıklarında pozitif veya negatif skorların atandığı kelimelerin oluşturduğu, araştırmacıların hazırlanmış olduğu sözlükler (lexicon) oluşturmaktadır. Kusen ve Strembeck [18] 2016 yılında yapılan Avusturya başkanlık seçimi ile ilgili Twitter'daki mesajların analizini gerçekleştirmek için bu yaklaşıma başvurmuşlar, analiz için kullanacakları mantıksal regresyon modelini kurmak için gerekli özellikleri kullandıkları sözlükten terimlere karşılık gelen polarite skorlarını alarak elde etmişlerdir. Çalışmada SentiStrength[19] isimli polarite sözlüğü yardımıyla mesajlar içerisinden olumlu veya olumsuz anlamlar belirten benzersiz kelimeler belirlenmekte, bu kelimeler kullanılan veri kümesinin özellik parametreleri olarak seçilmektedir. SentiStrenght özellikle duygu analiz için hazırlanmış; kelimelere ve ifadelere belirttikleri anlamlara göre olumlu ise 1 ve 5 arasında, olumsuz ise -1 ve -5

arasında skorların atandığı büyük bir kaynak olmasının yanında duygu analizi için metinlerin sınıflandırılmasında kullanılmaktadır. Örneğin "nefret etmek" ifadesinin anlamı olumsuz olduğu için kaynak içerisinde -4 olarak skoru atanmıştır. Mesajların temsili için kullanılan özelliklere atanacak değerler SentiStrength kullanılarak elde edilmektedir.

Analizlerde özellikleri oluşturmak için kullanılan bir başka yaygın yöntemde kelime çantası ile birlikte kullanılan terim frekansıdır. Bu yöntemde denetimli öğrenme modeli için kullanılacak mesaj kümesindeki tüm benzersiz kelimeler özellikleri, bu kelimelerin mesajlar içerisindeki bulunma frekansı da özelliklere atanacak değerleri oluşturmaktadır. Akhilesh ve arkadaşları [20] Pencap Yasama Meclisi seçimlerini Twitter mesajları üzerinde analiz etmek için bu yönteme başvurmuşlar, analiz için istatistiksel bir model olan Naive Bayes algoritmasını kullanmışlardır.

Twitter üzerinde duygu analizi gerçekleştirilirken karşılaşılan zorlukların arasında mesaj içeriklerinin dil bilgisi kurallarına uymaması da bulunmaktadır. Bu durum özellikle analiz öncesi ön işleme adımında fazladan iş yükü oluşturur. Indrajit ve arkadaşları [21] Twitter gibi kısa metinlerin analizinde özellik çıkarımı için konu modelleme yöntemi kullanımını önermişlerdir. Gerçekleştirdikleri çalışmada konu modellemesi sonucu elde edilen her sanal konu altında oluşan kavramlar, veri kümesini temsil edecek özellikler olarak seçilmektedir. Bu yöntem yapısı gereği kelimelerin birbiri arasındaki sıraya ve dilbilgisi kurallarına bağlı olmadığı için dilbilgisi ihmaline karşı avantaj oluşturmaktadır. Çalışma içerisinde toplanan mesajlar manuel olarak sınıflandırılmış ve MALLET (Machine Learning for Language Toolkit) yardımıyla oluşturulan sınıflar da dikkate alınarak konu modelleme işlemi gerçekleştirilmiştir. Önerilen yöntemin test sürecinde kullanılan sınıflandırma modelleri üzerindeki olumlu etkisi, terim frekansı yöntemiyle karşılaştırıldığında açıkça ortaya konulmuştur. Çalışmanın ileriki aşamalarında verimi arttırmak ve modelin iş yükünü azaltmak için elde edilen konular belirli bir miktara kadar azaltılmaktadır. Ayrıca farklı konuların içerdiği benzer anlamdaki kelimelerde kaldırılarak, başarım düşürülmeden, özellik miktarı minimuma çekilmiştir.

Öztürk ve Ayvaz [22] Twitter mesajları ile Suriyeli mülteci krizi üzerine yaptıkları çalışmada sözlük tabanlı duygu analizi yöntemini kullanmışlardır. Türkçe mesajlar üzerinde kullanılacak kapsamlı bir sözlük olmadığı için çalışmada kullanılmak üzere Türkçe bir sözlük oluşturulmuştur. Sözlük için gerekli kelimeler günlük konuşma dilinde kullanılanlar arasından seçilmiş ve toplam 5405 kelimeden oluşan bir sözlük oluşturulmuştur. Sözlükteki kelimelerden anlamları olumlu olanlar 1 ve 5 arasında, anlamsız olanlar ise -1 ve -5 arasındaki skorlar ile eşleştirilmiştir. Mesajların sınıflandırılması işlemi mesaj skorlarının içerdikleri kelimelerin skorları ile hesaplanması sonucu gerçekleştirilmiştir.

Ghiassi ve arkadaşları [23] yaptıkları çalışmada sözlük tabanlı yaklaşımı n-gram yöntemi ile birleştirerek duygu analizini farklı bir yönden ele almışlar, yapay sinir ağı algoritmasıyla kendi analiz modellerini oluşturmuşlardır. Çalışmada analiz için kendi sözlüklerini oluşturan araştırmacılar, topladıkları Twitter mesajları üzerinden bu işlemi gerçekleştirmişledir. Sözlük için gerekli olan terimler mesajlardan elde edilen kelime grupları içerisinden çıkarılmaktadır. Bunun nedeni olarak kelime gruplarının, tek başına çıkarılan kelimelere göre daha fazla bilgi sağlayacağı düşüncesi öne sürülmektedir. Mesaj kümesi içinden n-gram yöntemiyle çıkarılan kelime grupları arasındaki anlamlı ifadeler araştırmacılar tarafından manuel olarak ayıklanmış, bu ifadelerin içerdiği kelimelerin arasından en sık kullanılanlar sözlük için seçilmiştir ve tekrar araştırmacılar tarafından bu kelimelere skorlar atanmıştır. Analiz için önerilen DAN2 (A Dynamic Architecture for Artificial Neural Networks) modeli normal yapay sinir ağları gibi girdi çıktı ve ara katmanlarından oluşmaktadır. Modelin diğer ağlardan farkı ise ara katmanların sayısının önceden belirlenmemesi, performansta belirli bir sınıra ulaşılana kadar gizli katmanların dinamik olarak üretilmesidir. Araştırmada önerilen modelin başarımı SVM (Support Vector Machine) ile karşılaştırılmış, modelin daha etkili sonuçlar verdiği gözlemlenmiştir.

Saif ve arkadaşları [24] çalışmalarında duygu analizi için sözlük tabanlı yaklaşımın yeterli olmadığını, terimlerin bulunduğu mesajların içeriğine göre olumlu ya da olumsuz olarak farklı anlamlar kazanabileceğini öne sürmüşlerdir. Araştırmacılar bu düşünceden yola çıkarak terimlerin anlamlarını belirlemek için geometrik bir yaklaşım kullanmışlar, terimlerin anlamlarını belirlemek için bir çember modeli geliştirmişlerdir. Yaklaşımda anlamı araştırılan terim bir çemberin merkezine

konumlandırılmış, mesajlar içerisinde bu terim ile birlikte kullanılan kelimeler belirli parametreler yardımıyla bu çembere konumlandırılmıştır. Bu parametrelerin ilkini kelimeler ile terim arasındaki korelasyon oluştururken ikincisini ise sözlük yardımıyla kelimelere atanan polarite skorları oluşturmaktadır. Geometrik yöntemler ile kelimelerin skorlarından yararlanılarak terimlerin anlamları belirlenmiştir.

Filho ve Pardo [25] duygu analizi için kural tabanlı, sözlük tabanlı ve denetimli öğrenme yaklaşımlarını bir araya getirerek hibrit bir model oluşturmuşlarıdır. Hibrit model, alt modeller arasında bir boru hattı görevi görmekte, üst model sınıflandırma sonucunda belirli bir güven oranını sağlayamazsa sınıflandırma işlemi alt model tarafından gerçekleşmektedir. Kural tabanlı yaklaşımın temelini mesajlar içerisindeki ifadelerin (emoticon) sayısı oluşturmakta, mesaj içerisindeki olumlu veya olumsuz ifadelerin sayısı mesajın bütününün olumlu veya olumsuz olduğunu belirlemektedir. Çalışmada denetimli öğrenme yaklaşımı için SVM kullanılmıştır.

Güçlü sınıflandırma modeli oluşturmak için kullanılan yaklaşımlardan biri işbirlikçi öğrenme modelidir. Bu yaklaşımın amacı zayıf sınıflandırıcıların bir araya getirilerek başarımı yüksek, güçlü bir model oluşturmaktır. Yan ve arkadaşları [26] çalışmalarında, etkili bir duygu analizi gerçekleştirebilmek için kendi işbirlikçi modellerini önermişlerdir. İşbirlikçi modeller kurulurken temel sınıflandırıcılar olarak istatistiksel modeller olan Naive Bayes ve Maksimum Entropy yöntemlerinden; ayrıca sözlük tabanlı analizlerde kullanılan SentiStrength'ten ve Pattern'den [27] yararlanılmıştır. Oluşturulan ilk modelde dört temel sınıflandırıcının sonucu oylanırken, ikinci modelde daha kesin sonuçlar elde edebilmek için üç temel sınıflandırıcının sonucu kullanılmakta, mesajlar olumlu veya olumsuz olarak etiketlenmektedir. Çalışma sonunda işbirlikçi modeller bireysel olarak kullanılan temel sınıflandırıcılar ile karşılaştırılmış ve başarımlarının daha yüksek olduğu gösterilmiştir.

Xia ve arkadaşları [28] geliştirdikleri işbirlikçi modelde başarımı arttırmak için eğitim verisi üzerinden farklı özellik kümeleri oluşturmuş, kullanılan temel sınıflandırıcıların sonuçlarını birleştirmek için iki farklı yaklaşım denemiştir. Çalışmada özellik kümeleri oluşturulurken kelimelerin türleri ve birbirleri arasındaki ilişki dikkate alınmaktadır. Kelime türlerine göre oluşturulan üç farklı özellik

kümesinin ilki sadece mesajların içerdiği sıfat ve zarflardan oluşurken; ikinci kümeye yüklemler eklenmiş, üçüncü kümeye isimler de dâhil edilmiştir. Kelimelerin arasındaki ilişkiye göre özellik çıkarımı için n-gram yönteminden yararlanılarak kelime gruplarıyla oluşturulmuştur. Temel sınıflandırıcı olarak NB, Maksimum Entropy, ve SVM yöntemlerinin kullanıldığı çalışmada ilk işbirlikçi model için oylama işlemi yapılmaktadır. İkinci yaklaşım olarak lineer regresyon modeli ile bir üst sınıflandırıcı oluşturulmuş, temel sınıflandırıcıdan gelen sonuçlar bu sınıflandırıcı yardımıyla birleştirilerek genel sonuç elde edilmiştir.

Ankit ve Saleena [29] oluşturdukları işbirlikçi modelde, temel sınıflandırıcıların sonuçlarını birleştirmek için kendi önerdikleri ağırlıklandırma yöntemini kullanmışlardır. SVM, NB, Random Forest ve Logistic Regression öğrenme algoritmaları işbirlikçi modelin temel sınıflandırıcılarını oluşturmaktadır. Bu temel sınıflandırıcıların her biri için, sınıflandırdıkları mesajlara başarımlarıyla doğru orantılı olarak belirli skorlar atanmakta ve bu skorların toplamıyla mesajların sınıfları belirlenmektedir. Skor toplamıyla sınıflandırılma yapılamadığında kosinüs benzerliği kullanılarak bir mesaja kendisine en benzeyen mesajın etiketi atanmaktadır. Çalışmanın sonuçlarında önerilen yöntemin oylamaya göre daha başarılı olduğu gösterilmiştir.

Fersini ve arkadaşları [30] işbirlikçi model oluşturmak için BMA (Bayesian Model Averaging) yöntemine başvurmuşlardır. Çalışmalarında temel sınıflandırıcıların sonucunu birleştirmek için kullanılan geleneksel yöntemlerin sınıflandırıcıların istatistiksel belirsizliğini göz ardı ettiğini öne sürmüşler, bu durumun üstesinden gelmek için BMA modelini kullandıklarını belirtmişleridir. Modelin test edilmesi için yapılan duygu analizi BMA başarımının, sınıflandırıcıların bireysel başarımından ve geleneksel işbirlikçi modellerden daha yüksek olduğunu göstermiştir.

Troussas ve arkadaşları [31] işbirlikçi modellerin Twitter mesajlarının analizin üzerindeki etkisini gözlemlemek için bir çalışma gerçekleştirmişlerdir. Çalışmalarında Bagging, Boosting, Stacking ve oylama modelleri gibi popüler olan işbirlikçi modellerinin sonuçlarını geleneksel denetimli öğrenme modellerinin sonuçları ile karşılaştırmışlardır. Çalışmanın sonucu, işbirlikçi modellerin

sınıflandırma işleminde başarıma olumlu yönde etki ettiğini ve bireysel sınıflandırıcılara göre daha başarılı olduğunu göstermektedir.

Wang ve arkadaşları [1] çalışmalarında işbirlikçi modeller olan Bagging, Boosting ve Random Subspace algoritmaları arasında başarımlarına göre karşılaştırma yapmışlardır. Random Subspace Bagging yöntemine çok benzemekle birlikte aralarındaki en önemli fark veri kümesi yerine özellik kümesinden rastgele seçilen özellikler ile temel sınıflandırıcıların eğitilmesidir. Çalışmanın sonucunda Random Forest modelinin başarım açısından diğer modellere üstün geldiği gösterilmiştir.

Cotelo ve arkadaşları [33] Twitter analizinde üzerinde etkili bir analiz gerçekleştirmek için metinsel özelliklerin yanında yapısal özellikleri de kullanarak kendi iş birlikçi modellerini oluşturmuşlardır. Yapısal özellikler, kullanıcıların benzerliklerine göre önce gruplandırılıp daha sonra bu gruplara ait özelliklerin çıkarılması ile elde edilmiştir. Çalışmanın mimarisini her özellik kümesi için oluşturulan ayrı temel sınıflandırıcı grubu oluşturmaktadır. İki ayrı özellik kümesi kendi temel sınıflandırıcı gruplarında sınıflandırıldıktan sonra elde edilen sonuçlar Stacking yönteminde olduğu gibi bir üst sınıflandırıcı yardımıyla birleştirilmektedir.

Doğal dil işlemede metinlerin bilgi temsilinde kullanılan en yaygın yöntem mesaj kümesi içerisinden elde edilen ve benzersiz kelimeler ile oluşturulmuş terim-doküman matrisidir. Bu matrisin her sütunu metinler kümesinin içerdiği benzersiz kelimelere karşılık gelirken, her bir satır ise ayrı bir mesajı temsil etmektedir ve bu sütunlara mesajların içerdiği kelimelerin sayısıyla doğru orantılı olarak değerler atanmaktadır. Ancak Twitter mesajları gibi kısa mesajların temsili için bu yöntemin kullanılması bazı aksaklıkları da beraberinde getirmektedir. Twitter mesajlarının içerdiği kelime sayısının metin kümesindeki benzersiz kelimelerin sayısına oranla azlığı, bu mesajların temsilinde birçok sütunun boş kalmasına neden olmaktadır. Literatürdeki çalışmalara bakıldığında bu durumla baş etmek için kullanılan yöntemlerden biride özelliklerin zenginleştirilmesidir. Özellik zenginleştirme işlemi için sıklıkla kelime gömme yöntemine başvurulur. Bu yöntem kelimelerin, anlamca benzer kelimelerden oluşan vektörlerle temsil edilmesine dayanmaktadır.

Corrêa Jr ve arkadaşları [34] geliştirdikleri işbirlikçi modeli veri kümesinin farklı temsilleri ile oluşturmuş, modelde temel sınıflandırıcı olarak SVM ve Logistic

Regression modellerini kullanmışlarıdır. Veri kümesinin temsillerinin ilkini doğal dil işleme alanında sıklıkla kullanılan kelime çantası yöntemi ile oluştururken, ikinci temsili oluşturmak için Word2Vec [35,36] kelime gömme yöntemi kullanılmıştır. Word2Vec kelimelerin bulundukları metin içeriğinden yararlanılarak vektör temsillerinin oluşturulmasını sağlayan bir algoritmadır. Bu yöntem aynı zamanda kelimeler arası ilişkilerinde korunmasını sağlar ve bu ilişki yardımıyla vektörler arasındaki benzerlik hesapları daha kullanışlı sonuçlar döndürmektedir. Word2Vec kullanılarak Twitter mesajlarının temsili için iki farklı yöntem kullanılmıştır. İlk yöntem, mesajların içerdiği kelimelerin temsil vektörlerinin ortalaması alınmasıdır. İkinci yöntem ilkine benzemekle beraber Tf-Idf yöntemiyle vektörlerin ağırlıklı ortalaması alınmıştır.

Çoban ve Özyer [37] duygu analizinde performans artışı sağlayabilmek için Word2Vec ve K-Means kümeleme algoritması kullanarak düşük boyutlu özellik kümesi oluşturmayı amaçlamışlardır. Mesajlardan elde edilen benzersiz kelimelerin vektör temsilleri oluşturulduktan sonra önceden belirlenen küme sayısına göre kelimeler kümelere ayrılmıştır. Bu kümeler veri kümesinin özelliklerini temsil ederken, özelliklere atanacak değerler mesajların içerdiği küme üyesi kelimelerin sayısından oluşmaktadır. Çalışmada başarımı kontrol etmek için kelime çantası yöntemiyle farklı bir temsil oluşturulmuş, SVM le yapılan analiz sonucunda hız bakımından olumlu sonuçlar alınmasına rağmen başarım açısından etkili sonuçlar elde edilememiştir.

Hayran ve Sert [38] çalışmalarında kelime gömme yöntemi kullanarak farklı bir özellik çıkarımı yöntemi önermişlerdir. Her mesajın temsili için, tüm mesajların içerdiği benzersiz kelimelerin sütunlara atandığı birer temsil matrisi oluşturulmuştur. Bu temsil matrisi için değerler Word2Vec kullanılarak elde edilmiş, mesajın içerdiği kelimelere temsil vektörleri atanmıştır. Daha sonraki adımda matrisleri vektör haline getirmek için çalışmada önerilen füzyon teknikleri kullanılmış; toplama, aritmetik ortalama ve varyans alma işlemleri ile matris sütunlarından ayrı özellik vektörleri elde edilmiştir.

Rezaeinia ve arkadaşları [39] analiz çalışmalarında görüntü ve sinyal işlemede sıklıkla kullanılan ve başarımı yüksek bir yapay sinir ağ modeli olan CNN

(Convolutional Neural Network) kullanarak sınıflandırma modellerini oluşturmuşlardır. Modelin girdilerini oluşturabilmek için mesajların üzerinde kelime gömme yöntemleri kullanılmış ve mesajların içerdiği kelimeler vektör formatına dönüştürülmüştür. Çalışmada kelime vektörlerinin içerdiği bilgileri arttırabilmek için sözcük türü bilgisi ve sözlüklerden edilen kelime duygu polariteleri araştırmacılar tarafından önerilen yöntemler ile vektör formatına dönüştürülerek kelime vektörüne eklenmiştir. Çalışma sonucundan önerilen yöntemin başarımı arttırdığı gözlenmektedir.

Xiong ve arkadaşları [40] duygu analizinde sadece metinlerin yansıttığı duygular değil, metinleri oluşturan kelimelerin de belirtiği duyguların önemli olduğunu belirtmiş; çalışmalarında bu durumu göz önüne alarak duygu tabanlı kelime gömme işlemi yapabilmek için hibrit bir sinir ağı modeli önermişlerdir. Önerilen model sinir ağı tabanlı kelime gömme modellerine benzemekle beraber, mesajın tümü için iki ayrı alt sinir ağına ayrılmaktadır. Çalışmada CNN ile yapılan analiz sonucunda elde edilen sonuçlar önerilen yöntemin diğer benzer çalışmalara göre daha başarılı olduğunu göstermektedir.

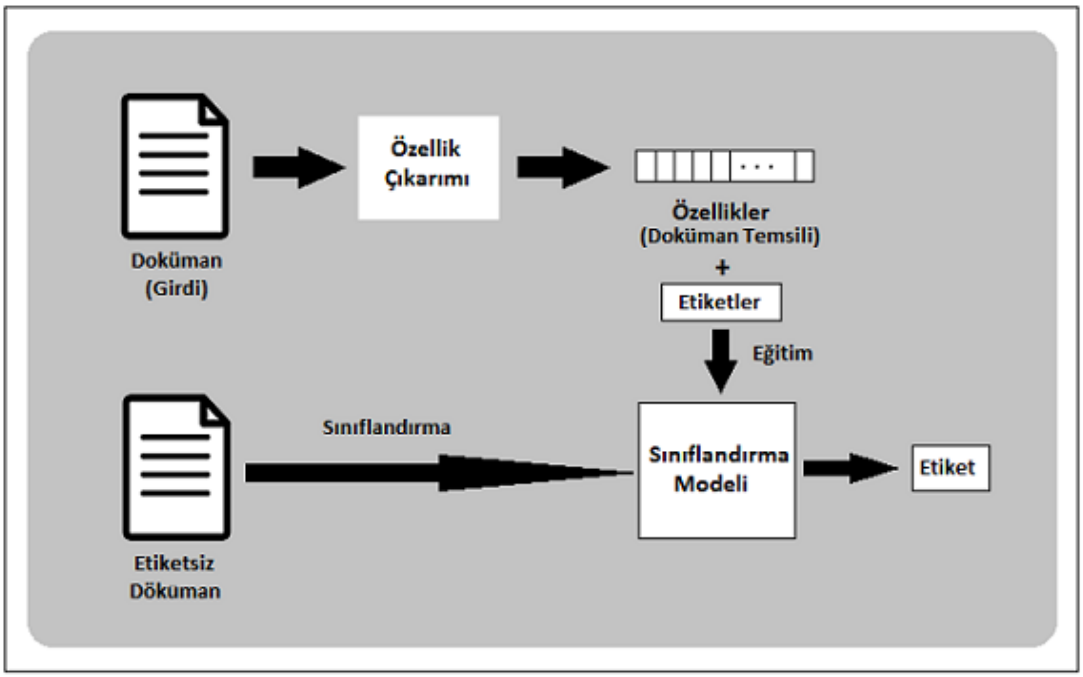
1.4. Tezin Yapısı

Tez, altı farklı bölümden oluşmaktadır. Giriş bölümünde Twitter mesajlarının sınıflandırılması ile ilgili bilgi verilmiş, çalışma alanının öneminde bahsedilmiştir. Genel bilgiler bölümünde yapılan çalışmalarla ilgili literatür taramasına, tezin başlatılma nedenlerine ve sağlayacağı katkılara yer verilmiştir. Doküman sınıflama bölümünde, metinlerin otomatik sınıflandırılmasında kullanılan yöntemler üç farklı konu başlığı altında tanıtılmıştır. Metinlerdeki gürültünün giderilmesi için yapılan işlemler, sayıl değerlerin elde edilmesi için kullanılan yöntem ve sınıflandırma için kullanılan algoritmalar bu üç konu başlığında anlatılmaktadır. Özellik genişletme bölümünde kısa metinlerin sınıflandırmasıyla ilgili problemler anlatılmış, bu problemlerin çözümü için kullanılan yöntemlerden bahsedilmiştir. Özellik genişletilmesiyle ilgili yöntemler bu bölümde anlatılmaktadır. Ayrıca tez çalışmasında da kullanılan BabelNet bu konu başlığında anlatılmıştır. Uygula bölümünde çalışma süreci ile ilgili bilgiler verilmektedir. Çalışmada kullanılan veri kümeleri ve teknolojiler gösterilmekte, sınıflandırma için kullanılan modellerin nasıl

oluşturulduğu anlatılmaktadır. Özellik genişletilmesi ve sınıflandırma sonuçları ilgili çıkarımlar gösterilmiştir.

2. DOKÜMAN SINIFLAMA

Sınıflandırma işlemi en basit şekilde verilen girdilere göre en doğru sınıf etiketinin seçilmesi olarak tanımlanabilir. Doküman sınıflandırma, $D = \{d_1, d_2, \dots, d_n\}$ doküman kümesinin (corpus) elemanlarının daha önceden belirlenmiş olan $S = \{s_1, s_2, \dots, s_m\}$ sınıf etiketleri kümesinin elemanlarıyla eşleştirilmesi, $\langle d_i, s_j \rangle$ ikilisinin elde edilmesi işlemidir. Doküman sınıflandırma duygu analizi, spam e-posta filtrelemesi ve makalelerin belirli konu başlıklarına göre sınıflandırılması gibi birçok çalışmada kullanılan popüler bir metin madenciliği uygulamasıdır. Otomatik doküman sınıflandırma ile ilgili ilk çalışmalar 1960'larda başlamasına rağmen 1990'lardan sonra yapılan çalışmaların sayısı artış göstermeye başlamıştır. Özellikle Web ile ilgili gelişmeler ve internet ortamında üretilen elektronik dokümanların sayısındaki artış araştırmacıları bu alanda çalışmaya teşvik etmiştir [41-44].



Şekil 2.1. Doküman sınıflandırma mimarisi

Doküman sınıflandırma, dokümanın temsili ve sınıflandırma olmak üzere iki önemli aşamadan oluşmaktadır. İlk aşama, dokümanın belirli yöntemler kullanılarak özellik kümesine dönüştürülmesi adımıdır. Özellik kümesinin doküman temsilindeki

başarımı, sınıflandırma modelinin başarımını doğru orantılı olarak etkilemektedir. Özellik kümesinin gösterimi için vektör uzay modeli kullanılmaktadır. Sınıflandırma aşaması ise eğitim verileriyle sınıflandırma modelinin eğitilmesi ve bu modelle yeni dokümanların sınıflandırılması adımlarından oluşur. Sınıflandırma modelleri için makine öğrenmesi alanında sıklıkla kullanılan SVM, karar ağaçları, Naive Bayes ve K-Nearest Neighbor algoritmalar tercih edilmektedir [45-48].

2.1. Doküman Önişleme

Doküman temsili ve sınıflandırma modelin başarımı birbiriyle doğrudan ilişkilidir. Bu nedenle dokümanların içerdiği bilgileri doğru bir şekilde elde edebilmek için yapılan ön işleme adımları büyük önem taşımaktadır. Son yıllarda yaşanan gelişmeler elektronik ortamlarda üretilen dokümanların sayısında büyük artışa sebep olmuş, bu durum doküman sınıflandırma çalışmalarında olumlu ve olumsuz birçok etki doğurmuştur. Yüksek sayıdaki doküman sayısı sınıflandırma çalışmaları için zengin içerik sunmaktadır. Ancak doküman kaynaklı zorluklar dışında bu mesajların oluşumu ile ilgili bazı problemler çalışmalar için artı iş yükü oluşturmaktadır. Özellikle Twitter gibi uygulamalarda üretilen dokümanlarda kullanıcı kaynaklı olarak dilbilgisi ve yazım kurallarının çoğu göz ardı edilmektedir. Bu nedenle bu mesajların içeriği genel ön işleme adımlarına tabi tutulmadan önce gürültüden olabildiğince temizlenmelidir. Daha sonra ise sözcük türü etiketi (Part of Speach), birimlendirme (Tokenization), gövdeleme (Stemming), sözcük birimleştirme (Lemmeziation) ve etkisiz kelimelerin (stop words) kaldırılması gibi işlemler uygulanarak dokümanlar bilgi çıkarımı için uygun hale getirilir.

2.1.1. Twitter önişleme

Twitter mesajları bünyesinde kullanıcı kaynaklı birçok gürültü barındırmaktadır. Mesajlar incelendiğinde çoğunun dilbilgisi ve yazım kurallarına uymadığı görülmektedir. Ayrıca mesajlar; resmi olmayan kısaltmaları, argo kelimeleri ve kullanıcıların türettiği yeni kelimeleri bolca içermektedir. Bu gürültülerin yanı sıra mesajların uygulama yapısı gereği içerebileceği bazı özel karakterler (@, #), duygu bildiren simgeler ve URL (Universal Resource Locator) adresleri bilgi çıkarımını olumsuz olarak etkilemektedir. Doğru bir analiz yapılabilmesi için bu gürültülerin mesaj içeriğinden olabildiğince temizlenmesi gerekir. Twitter üzerine yapılan

çalışmalar incelendiğinde gürültünün kaldırılması için yapılan işlemler aşağıda listede genelleştirilmiştir [12, 49-55].

- Mesaj içeriğinde bulunan URL bağlantıları özel durumlar dışında genellikle mesaj içeriğinden çıkarılmaktadır.
- Kullanıcı adlarını gösteren @ karakteriyle başlayan ifadeler mesajlardan kaldırılmaktadır.
- Mesajların içerdiği *hashtag*'ler bazı durumlarda # karakteri kaldırılıp kullanıldığı gibi bazen mesajlardan tamamen çıkarılmaktadır.
- Duygu belirten simgeler mesajlardan kaldırılırılabileceği gibi bazı durumlarda anlamca eşleşebilecek belirli kelimeler ile değiştirilmektedir.
- Mesajlar, ileriki ön işleme adımlarına kolaylık sağlaması açısından küçük harflere dönüştürülmektedirler.
- Sözlük benzeri kaynaklar yardımıyla mesajlardaki yazım yanlışları ve kısaltmalar düzeltilmektedir
- Mesaj içeriğinden noktalama işaretleri ve rakamlar kaldırılmaktadır.

2.1.2. Birimlendirme ve sözcük türü etiketi

Doküman üzerinden bilgi çıkarımı için yapılan ilk işlem metni birim (token) adı verilen küçük ve kullanışlı parçalara ayırmaktır. Elde edilen birimler metinlerin vektör temsilini oluşturmak için kullanılır ve bu işlemin tamamına birimlendirme (Tokenization) ismi verilmektedir. Birimler kelime çantası (Bag of Words) yönteminde tekil kelimelere karşılık gelirken; n-gram yönteminde kelime veya kelime gruplarına ya da karakter ve karakter gruplarına karşılık gelmektedir. Karakter ve kelime grubunun sayısı gerçekleştirilen çeşitli yaklaşımlara göre farklılık gösterebilir. Parçalama işlemini gerçekleştirmek için genellikle N-gram yöntemine başvurulmaktadır. $N = \{1,2,3 \dots\}$ arasından seçilen değerlere göre metinler n sayıdaki birimlere ayrılmaktadır. n değeri kelime sayısını belirtebilirken, bazı durumlarda karakter sayısını da belirtilebilmektedir. n için en sık kullanılan değerler kümesi $\{1,2,3\}$ olup; bu değerlere göre yöntem unigram, bigram ve trigram olarak adlandırılmaktadır. Örneğin "Kısa mesajların analizi popülerdir" isimli cümle örneğinin üzerine kelime tabanlı bigram yöntemi uygulanmak istenirse $\{(Kısa\ mesaların), (mesajların\ analizi), (analizi\ popülerdir)\}$ listesi elde

edilir. N-gram yöntemi doküman sınıflandırılması, yanlış hecelerin düzeltilmesi, DNA analizi gibi birçok çalışma içerisinde sıklıkla kullanılmaktadır [56, 57].

Kullanılan birçok yazı dilinde kelimeler, metin içerisinde kullanıldığı yere ve diğer kelimeler ile oluşturduğu gruplara bağlı olarak farklı anlamlar kazanabilmektedir. Örneğin Türkçede tükenmek fiili "tükenmez kalem" kelime grubu içerisinde kullanıldığında bir nesneyi belirten isim halini almaktadır. Kelimelerin tekil olarak ele alınması özellikle doküman üzerinden bilgi çıkarımında anlam belirsizliklerine yol açmakta, araştırmacılar için büyük zorluk oluşturmaktadır. Analizlerde belirsizliklerin etkisini en aza indirebilmek için kullanılan yöntemlerden biride sözcük türü etiketi (Part of Speech) yöntemidir [58-60]. Bu teknik, kelimelerin dil bilgisi kurallarına göre metin içerisindeki görevlerine göre parçalanmasıyla gerçekleştirilmektedir. Bu süreçte elde edilen her parça özne, nesne, zarf ve yüklem gibi etiketler ile etiketlenir. Bu işlem yardımıyla kelimeler üzerinde sözcük birimleştirme (Lemmeziation) gibi işlemler gerçekleştirilebilmektedir.

2.1.3. Gövdeleme ve sözcük birimleştirme

Doküman içerisinden özellik çıkarımı için kullanılan işlemlerden biri, dokümanın içerdiği kelimelerin aldığı ekleri kaldırarak kök haline indirgemektir. Yapılan işlemin amacı, aldığı ekler ile yapısal olarak farklılık gösteren kelimelerin çözümlenip kök haline getirilerek aralarındaki anlam ilişkisini ortaya çıkarmaktır. Bu işlem kullanılan iki farklı yöntem gövdeleme (Stemming) ve sözcük birimleştirme (Lemmaziation) olarak adlandırılır. Gövdeleme algoritması benzer köklere sahip olan kelimelerin çekim ve yapım ekleri kaldırılarak ortak bir yapı elde etmeyi amaçlayan bir işlemdir. Gövdeleme algoritmaları, sezgisel olarak belirli ekleri kelimelerden çıkaran kurallar bütününden oluşur. Porter gövdeleme algoritması [62] eski bir yöntem olmasına rağmen sıklıkla kullanılan bir algoritmadır. Kökleri kaldırma işlemini 5 adet kurala göre gerçekleştiren bu algoritma hızlı sonuç ürettiği için tercih edilmektedir. Sonraki süreçte bu algoritmanın geliştirilmesiyle daha iyi sonuçlar üreten Snowball (Porter2) [63] algoritması ortaya çıkmıştır. Lancaster (Paice/Husk) [64] algoritması popüler olan bir başka yöntemdir. Bu algoritma içerdiği 120 adet kural üzerinden ekleri kaldırma işlemini gerçekleştirdiği için sıkı bir algoritma olarak tanınır. Bilgi çıkarımında kullanılan gövdeleme işlemi özünde iki farklı problem barındırmaktadır.

Bunlardan ilki aynı kökü paylaşan ve aralarında anlam ayrımı büyük olan kelimelerdeki bilgiler göz ardı edilmektedir. İkincisi problem ise aldığı eklere göre farklılaşan kelimelerin ek çıkartılması sonucunda anlamlarını yitirmesidir [65].

“Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.”

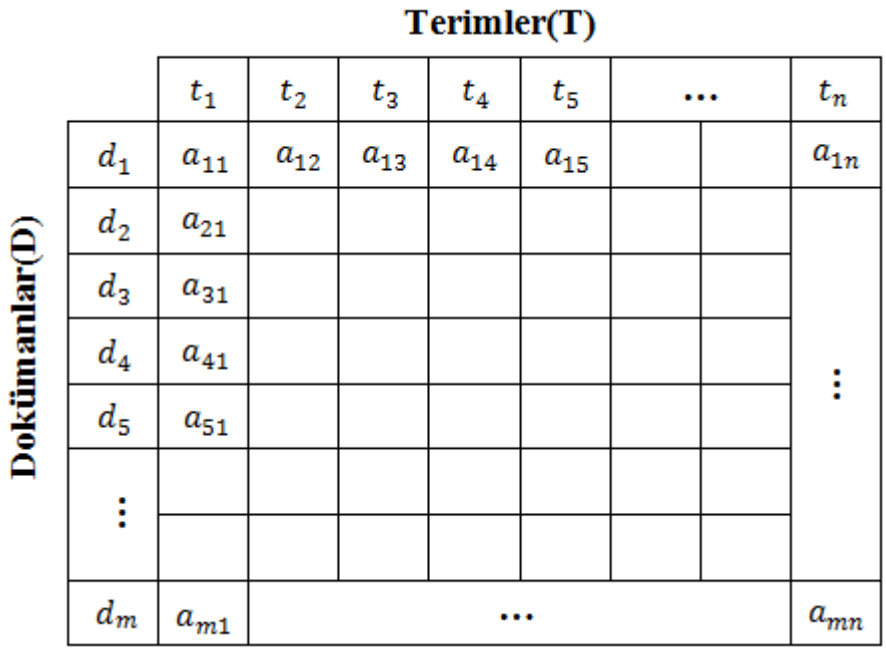
Yukarıdaki İngilizce metin Porter algoritmasıyla işlendiğinde aşağıdaki sonuç elde edilir.

“Stem is the process of reduc a word to it word stem that affix to suffix and prefix or to the root of word known as a lemma”

Sözcük birimleştirme ile kök bulma işleminde kelimelerin anlamları da dikkate alınmakta, bu işlem sözlük yardımıyla gerçekleştirilmektedir. Genellikle yapısı daha karmaşık olan dillerde kullanılırken, gövdeleme algoritmalarına göre daha doğru sonuçlar üretmektedir. Metin içerisindeki kelimeler sözcük türü etiketi yöntemiyle gruplara ayrıldıktan sonra bir sözlük yardımıyla köklerine indirgenirler. WordNET [65] bu işlem için yaygın olarak kullanılan bir kaynaktır.

2.2. Doküman Terim Matrisi

Metin madenciliğinde dokümanların, kullanılacak analiz yöntemine uygun şekilde temsil edilmesi gerekmektedir. Temsilin ve analiz yönteminin başarımı birbiriyle sıkı sıkıya bağlıdır. VUM (Vektör Uzay Modeli) [66] kavramsal basitliği ve kullanışlılığı nedeniyle temsil için yaygın olarak kullanılan bir modeldir. Modeli oluşturmak için ilk adım olarak M elemanlı $D = \{d_1, d_2, \dots, d_m\}$ doküman kümesi ön işleme adımlarından geçirilerek olabildiğince gürültüden temizlenir. Daha sonra bu dokümanlar içerisinden benzersiz kelimeler çıkarılarak N elemanlı $T = \{t_1, t_2, \dots, t_n\}$ terim kümesi oluşturulur. Elde edilen bu terim kümesi ile birlikte her satırı bir doküman ve her sütunu bir terime denk gelen MxN büyüklüğünde $A = \left[a_{ij}\right]_{mxn}$ terim doküman matrisi oluşturulur. Son adım olarak dokümanların içerdiği terimlerin içerikte görünme sayılarına göre sütunlara değerler atanarak terimlerin dokümanlar ile aralarındaki bağlantı gösterilir.



Şekil 2.2. Terim doküman matrisi

Dokümanların içerdiği her kelime analiz işlemine olan etkisi eşit derecede değildir. Bu nedenle genellikle doküman-terim matrisindeki değerleri ağırlıklandırmak için çeşitli işlemler uygulanır. Tf-idf (Term Frequency- Inverse Document Frequency) her kelimenin bireysel olarak önemini ve ayırt ediciliğini göstermek için uygulamalarda sıklıkla kullanılan bir ağırlıklandırma yöntemidir [67, 68].

$$Tf(a_{ij}) = \frac{a_{ij}}{\sum_{j \epsilon N} a_{ij}} \tag{2.1}$$

a_{ij}, t_j teriminin d_i dokümanına görülme sayısını belirtmektedir. Tf (Term Frequency) hesaplamak için a_{ij}, d_i dokümanın içerdiği terimlerin sayılarının toplamına bölünür. Doküman analizinde metinlerin boyutları birbirine göre değişiklik gösterebilir. Sezgisel olarak bakıldığında uzun metinlerin içerdiği kelimelerin görünme sayısı da artmaktadır. Bilgi temsilinin daha doğru sekilde yapılabilmesi için [0,1] aralığındaki Tf değerleri döküman temsilinde kullanılmaktadır.

$$Idf(t_j) = \log \frac{N}{n_i} \tag{2.2}$$

$$TfIdf(a_{ij}) = Tf(a_{ij})\ x\ Idf(t_j) \quad (2.3)$$

n_i , içerisinde t_j terimini bulunduran dokümanların sayısını göstermektedir. Idf (Inverse Document Frequency) işlemi, toplam doküman sayısının n_i sayısına bölünerek elde edilen sonucun logaritması alınmasıyla gerçekleştirilir. Bu işlemin yapılmasının nedeni, terimlerin bulunduğu dokümanların sayısı arttıkça belirleyici özelliğini kaybedeceği düşüncesidir. Bu nedenle terimlerin Idf değerleri cezalandırıcı katsayı olarak Tf değerleri ile çarpılır.

2.3. Sınıflandırma Algoritmaları

Sınıflandırma algoritmaları, makine öğrenmesi alanındaki çalışmalarda tahmin modelleri oluşturmak için kullanılan denetimli öğrenme algoritmalarıdır. Oluşturulan modellerin görevi, verilerin belirli özelliklerinden yararlanarak önceden tanımlı sınıf etiketleri ve veri arasındaki ilişkiyi ortaya çıkararak sınıflandırma işlemini gerçekleştirmektir [69-71]. Modellerin sınıflandırma işlemini gerçekleştirebilmesi için sınıf etiketi bulunan eğitim verileri üzerinden sınıflandırma kurallarını öğrenmesi gerekir. Öğrenme işleminin doğru bir şekilde yapılıp yapılmadığını kontrol etmek için test verileri üzerinden sınıflandırma işlemi gerçekleştirilir ve test verilinin etiketleri ile model çıktıları belirli yöntemler ile karşılaştırılır. Sınıflandırma modelli oluşturmak için kullanılan birçok algoritma mevcuttur. Ancak bu tez içerisinde analiz gerçekleştirmek için de kullanılan SMO (Sıralı Minimal Optimizasyon), Naive Bayes ve karar ağaçları anlatılmaktadır.

2.3.1. Naive bayes

NB (Naive Bayes) sınıflandırıcısı; sınıflandırma ve öğrenme maliyetlerinin az olması, yüksek sayıdaki özellik kümesinde bile hızlı olması, basitliği ve başarımının yüksek olması nedeniyle geniş uygulama alanına sahiptir. NB temeli Bayes teoremine dayanan istatistiksel bir modeldir. Sınıflandırılmak istenen veri kümesinin içerdiği özelliklerin önceden tanımlı sınıflara göre koşullu olasılığı hesaplanır, bu olasılıkların yardımıyla sınıflandırma işlemi gerçekleştirilir. NB sınıflandırıcısının en önemli özelliği, özelliklerin sınıf etiketlerine göre birbirinden istatistiksel olarak bağımsız olduğunun varsayılmasıdır. Ayrıca herhangi bir sınıfın koşullu olasılık dağılımın normal dağılım olduğu varsayılmaktadır [72-74]. Örneğin $T =$

$\{t_1, t_2, \dots, t_n\}$ terim kümesiyle temsil edilen $D = \{d_1, d_2, \dots, d_m\}$ doküman kümesinin NB sınıflandırıcısıyla $E_1, E_2, \dots, E_l$ sınıf etiketlerine göre sınıflandırılmak istendiğinde Denklem (2.4) uygulanır.

$$P(E_i|d_j) = P(E_i|t_1, t_2, \dots, t_n) = \frac{P(t_1, t_2, \dots, t_n|E_i)P(E_i)}{P(t_1, t_2, \dots, t_n)} \quad (2.4)$$

Koşullu olasılık formülü üzerinde zincir kuralı uygulanarak Denklem (2.5) elde edilir.

$$P(t_1, t_2, \dots, t_n|E_i) = P(t_1|E_i)xP(t_2|E_i)x \dots xP(t_n|E_i) = \prod_{k\epsilon N} P(t_k|E_i) \quad (2.5)$$

Özellikler sürekli değerlerden oluşuyorsa koşullu olasılık hesaplanırken normal dağılım formülünden yararlanılır. Bu nedenle normal dağılımda kullanılmak üzere sınıf koşulu alıntında özelliklerin standart sapması(σ) ve ortalaması (μ) hesaplanır[75,76]. Daha sonra bulunan değeler Denklem (2.6)'da işletilir.

$$P(t_k|E_i) = g(t_k, \mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{(t_k-\mu_i)}{2\sigma_i^2}} \quad (2.6)$$

2.3.2. Sıralı minimal optimizasyon

SVM (Support Vector Machine) doğrusal veya doğrusal olmayan sınıflandırma işlemlerinde, regresyon analizinde ve aykırı değerlerin belirlenmesinde kullanılan; güçlü ve farklı çalışmalara kolaylıkla uyum sağlayabilen bir makine öğrenmesi modelidir. Özellikle küçük veya orta ölçekli veri kümeleri üzerinde gerçekleştirilen sınıflandırma işlemleri için uygundur. Doğrusal sınıflandırma için kullanılan SVM modelleri, iki sınıflı ve yüksek boyutlu veri kümesini en iyi şekilde ayırabilecek bir hiperdüzlemin bulunmasıyla oluşturulmaktadır. Hiperdüzlem iki boyutlu uzaydaki doğru denkleminin yüksek boyutlardaki karşılığıdır. Daha sonra modele girdi olarak verilen yeni veriler, hiperdüzleme göre bulundukları grupların sınıf etiketini almaktadırlar [77-79]. $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ etiketli veri kümesinin elemanları $x_i \epsilon R^n$ olup, sınıf etiketleri $y_i \epsilon [-1, 1]$ değerlerini almaktadır. Bu veri kümesinin bölgelere ayıracak hiperdüzlem Denklem (2.7) ile ifade edilir.

$$w.x + b = 0 \quad (2.7)$$

Denklem (2.7) içerisindeki $w \epsilon R^n$ vektörü hiperdüzleme dik olan bir vektördür. Denklemin bir diğer ifadesi olan $b \epsilon R$ ise bias değerini ifade eden skaler bir büyüklüğü tanımlar. Model oluşturmak için bu iki ifadenin bulunması ve oluşturulan hiperdüzlemin ayrılan iki sınıf grubuna maksimum uzaklıkta olması gerekmektedir. Bunun için öncelikle bazı kısıtlamalar tanımlanmalıdır.

$$w.x^+ + b \geq 1 \tag{2.8}$$

$$w.x^- + b \leq -1 \tag{2.9}$$

Sınıf etiketleri için ayrı ayrı tanımlanan yukarıdaki kısıtlamalar, sınıf etiketiyle çarpılarak Denklem (2.10)'a dönüştürülür.

$$y(w.x + b) - 1 \geq 0 \tag{2.10}$$

Hiperdüzlemin sınıf gruplarına olan uzaklığını hesaplayabilmek için veri kümesi dışında w vektöründen yararlanılır.

$$(x^+ - x^-).\frac{w}{\|w\|} \tag{2.11}$$

Hiperdüzleme dik olan w vektörü birim vektöre çevrilerek negatif ve pozitif sınıflı örnekler arasındaki fark ile çarpılır ve uzaklık değeri elde edilir. Yukarıdaki ifade daha önceki tanımlara göre sadeleştirildiğinde Denklem (2.12) elde edilir.

$$\frac{2}{\|w\|} \tag{2.12}$$

Yapılan tüm işlemlerden sonra hiperdüzlemin sınıf gruplarına maksimum uzaklıkta olabilmesi için Denklem (2.13)'teki optimizasyon problemi oluşturulur.

$$\min \frac{1}{2} w^2 \, , Koşulu \;\; Altında \; y(w.x + b) - 1 \geq 0 \tag{2.13}$$

Bu optimizasyon problemi Lagrange Çarpanı yöntemi kullanılarak aşağıdaki gibi tekrar tanımlanır.

$$\max \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i.\alpha_j.y_i.y_j.K(x_i.x_j) \tag{2.14}$$

$$Koşulu\ Altında\ 0 \leq \alpha_i \leq C, \qquad i = 1, \ldots, m \tag{2.14a}$$

$$\sum_{i=1}^{m} \alpha_i . y_i = 0 \tag{2.14b}$$

Denklem (2.14)'te tanımlanan optimizasyon probleminde *C* düzenlileştirme terimini (regularization term), α Lagrange çarpanını ve $K(x_i,x_j)$ kernel fonksiyonunu göstermektedir. Kernel fonksiyonları lineer olarak ayrılamayacak sınıf gruplarını yeni bir uzayda tanımlayarak hiperdüzlemle ayrılabilir hale getirir. SVM modelinin optimizasyonu ikinci dereceden programlama problemidir. Bu nedenle özellikle büyük veri kümeleri üzerinden model oluşturmanın maliyeti yüksektir. SMO (Sıralı Minimal Optimizasyon) bu problemi çözmek için ortaya atılan, ikinci dereceden programlama problemini alt problemlere ayırarak çözen tekrarlamalı bir algoritmadır [80]. Denklem (2.14)'deki gibi bir ikinci dereceden programlama probleminin SMO yöntemiyle optimizasyonunun sağlanması için KKT (Karush-Kuhn-Tucker) koşullarını sağlaması gerekmektedir.

$$\alpha_i = 0 \Rightarrow y_i . f(x_i) \geq 1 \tag{2.15}$$

$$0 < \alpha_i < C \Rightarrow y_i . f(x_i) = 1 \tag{2.15a}$$

$$\alpha_i = C \Rightarrow y_i . f(x_i) \leq 1 \tag{2.15b}$$

Denklem (2.15)'deki $f(x_i)$ ifadesi girdilere göre modelin çıktısını, y_i girdilerin sınıf etiketini, α_i Lagrange çarpanını göstermektedir. SVM üzerinde optimizasyon tek tek Lagrange çarpanları üzerinde gerçekleştirilirken, SMO'da optimizasyon her adımda iki Lagrange çarpanı üzerinde gerçekleştirilir. Kolay uygulanabilirliği ve hızlı olması nedeniyle SMO duygu analizinde özellikle büyük veri kümeleri için sıklıkla kullanılan bir yöntemdir [81].

2.3.3.Karar ağaçları

Karar ağaçları böl ve fethet mantığına dayanan, ağaç veri yapıları şeklinde oluşturulan kural tabanlı makine öğrenmesi algoritmalarıdır. Sınıflandırma işlemi verilerin kökten başlayarak özellik değerlerine göre yaprak düğümlere ulaşmasıyla gerçekleştirilir [82-84]. Ağaç yapısındaki her düğüm veri kümesinin özellikleri ile

temsil edilmektedir. Yapının kök düğümü oluşturulurken veri kümesinin her bir özelliği için bilgi kazanç oranı hesaplanır. Kazanç oranı hesaplama yöntemi algoritmaların yapısına göre farklılık göstermektedir. En yüksek kazanç orana sahip özellik ağacı kök düğümünü temsil eder ve bu özelliğin değerlerine göre dallanmalar oluşmaktadır. Oluşan her dallanma için veri kümesi alt kümelere ayrılır. Daha sonraki düğümler için aynı işlem alt kümeler kullanılarak gerçekleştirilir. ID3, C4.5, J48, AD Tree, BF Tree, CART başta olmak üzere birçok karar ağacı algoritması çeşidi mevcuttur. C4.5, ID3 algoritmasının gelişmiş bir sürümü olarak kazanç oranı olarak entropi hesabını kullanmaktadır. Entropi sonucuna göre bilgi kazancı yüksek olan özellikler düğüm olarak seçilmektedir. Ağaç oluşturulurken ilk adım olarak $E = \{e_1, e_2, \dots, e_n\}$ sınıf etiketi kümesinin entropi değeri Denklem (2.15)'teki gibi hesaplanır.

$$Bilgi(E) = -\sum_{i=1}^{n} \left(\frac{|e_i|}{|E|}\right) \cdot \log_2 \left(\frac{|e_i|}{|E|}\right) \tag{2.15}$$

Denklemde gösterilen $|e_i|$ değeri e_i etiket değerine sahip sınıf sayısını gösterirken, $|E|$ etiket kümesinin eleman sayısını göstermektedir. Daha sonraki adımda düğümlere atanacak özelliklerin karar verilmesi için $T = \{t_1, t_2, \dots, t_m\}$ özellik kümesinin sınıf etiketi kümesini kullanarak Denklem (2.16)'da gösterilen formüle göre parçalı entropi değerleri hesaplanır.

$$Bilgi_T(E) = -\sum_{i=1}^{k} \left(\frac{|t_i|}{|T|}\right) \cdot Bilgi(E) \tag{2.16}$$

Denklemde gösterilen $|t_i|$ değeri t_i etiket değerine sahip sınıf sayısını gösterirken, $|T|$ etiket kümesinin eleman sayısını göstermektedir. Bilgi kazancı elde edilirken $Bilgi(E)$ entropi değeri $Bilgi_T(E)$ entropi değerinden çıkarılır. J48, C4.5 algoritmasının açık kaynak kodlu bir yazılım olan WEKA [85,86] üzerindeki temsilidir. Sayısal, nominal ve metinsel girdi verileri üzerinde işlem gerçekleştirebilen J48 karar ağacı duygu analizinde yüksek performans sağlamaktadır.

3. İŞBİRLİKÇİ ÖĞRENME MODELİ

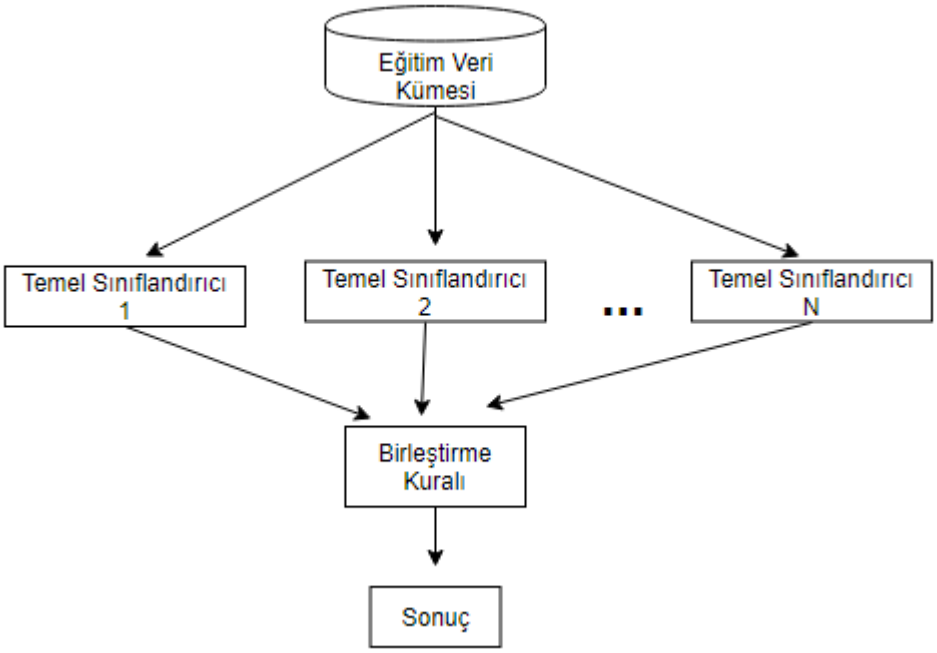
İşbirlikçi öğrenme, belirli bir sınıflandırma probleminin çözümü için birden fazla sınıflandırıcının eğitildiği yöntemin adıdır. Yöntemin temel mantığı, birden fazla zayıf sınıflandırıcının sınıflandırma sonuçlarının bir araya getirilerek başarımı yüksek bir sınıflandırma modeli oluşturmasına dayanmaktadır. Bireysel sınıflandırıcıların bir arada çalışması nedeniyle komite tabanlı öğrenme yöntemi olarak da adlandırılır. Özellikle 90'lı yıllarda popülaritesini arttıran işbirlikçi öğrenme yöntemi zamanla makine öğrenmesinde önemli bir araştırma alanı haline gelmiştir [106,108]. İşbirlikçi modeller eğitim kümesi, temel sınıflandırıcılar ve sınıflandırıcıların sonuçlarını birleştirme yöntemlerine göre oluşturulur [32]. Temel sınıflandırıcıların doğruluğu ve çeşitliliği işbirlikçi modelin başarımını etkileyen büyük etkenler arasındadır. Modellerin çeşitliliği, aralarındaki korelasyonu düşürerek birbirlerine daha az bağımlı hale gelmelerini sağlamaktadır [103].

İşbirlikçi öğrenme modellerinin sınıflandırma işlemlerinde kullanılması birçok fayda sağlamaktadır. Bunlardan bazıları şu şekilde sıralanabilir [109];

- Birden fazla sınıflandırıcı eğitim verisi üzerinde benzer sonuçları üretirken, test veri kümesi üzerinde farklı sonuçları üretebilmektedir. Sezgisel olarak bakıldığında bu sınıflandırıcıların sonuçlarının birleştirilmesi verilerin doğru sınıflandırılması şansını arttırmaktadır. Her temel sınıflandırıcının sonucu en iyi modelden düşük olsa bile sonuçların birleştirilmesi yanlış sınıflandırma riskini azaltmaktadır.
- Bazı durumlarda büyük miktarda veri ile sınıflandırıcı modelin eğitilmesi gerekebilir. Ancak bu durum pratik bir çözüm yöntemi değildir. Verilerin alt kümelere ayrılarak birden fazla sınıflandırma modelinin eğitiminde kullanılması ve sonuçların birleştirilmesi daha verimli bir çözüm sağlamaktadır.
- Güçlü bir sınıflandırıcı için veri uzayının tamamının en iyi şekilde temsil edilmesi gerekmektedir. Veri kümelerinin uzay temsili için yeterli olmadığı durumlarda her sınıflandırıcı için veri kümesinden yeniden örnekleme

yöntemleriyle elde edilen veri alt kümeleri sınıflandırıcıları eğitmek için kullanılabilir. Bu sınıflandırıcıların sonuçlarının birleştirilmesiyle elde edilen başarımın yüksek olduğu gözlemlenmiştir.

- Sınıflandırma işleminde kullanılacak veri kümeleri içerdiği gruplar karmaşık sınırlara sahip olabilirler. Bu tür veri kümeleri doğrusal sınıflandırıcılar ile sınıflandırılamaz. Ancak birçok doğrusal sınıflandırıcıyı bir araya getiren işbirlikçi modellerin karmaşık sınırları öğrenebildiği görülmüştür.
- Sınıflandırma işlemlerinde farklı kaynaklardan elde edilen, farklı özelliklere sahip veri kümelerinin kullanılması gerekebilir. Bu veri kümeleri ile tek bir sınıflandırıcının oluşturulması mümkün değildir. Her veri kümesi ile ayrı bir sınıflandırma modelinin oluşturulması, sınıflandırıcılardan elde edilen sonuçların birleştirilmesi bu tür problemlerin çözümü için etkili bir yöntemdir.



Şekil 3.1. İşbirlikçi öğrenme modeli

İşbirlikçi modeller temel sınıflandırıcılarının oluşturulmasına göre heterojen ve homojen işbirlikçi modeller olmak üzere ikiye ayrılırlar. Heterojen modeller aynı eğitim veri kümesi üzerinde farklı temel sınıflandırıcıların kurulmasıyla veya aynı sınıflandırıcıların parametreleri değiştirilerek oluşturulur. Homojen işbirlikçi modeller farklı eğitim kümeleri ile oluşturulan aynı sınıflandırma modelleri ile kurulmaktadır. Farklı eğitim kümeleri çeşitli yöntemler ile aynı veri kümesinden

türetilirler. Heterojen modellerdeki başarım farklı sınıflandırma modellerinin birbirlerini tamamlamasına dayanırken, homojen sistemlerde sınıfların farklı eğitim kümelerine göre optimizasyonuna dayanmaktadır [107]. Bagging ve Boosting homojen işbirlikçi modellere örnek olarak gösterilebilirken, Stacking bir Heterojen işbirlikçi modeldir.

3.1. Bagging

Bagging, Breiman [110] tarafından önerilen, birden fazla zayıf sınıflandırıcının bir araya getirilerek güçlü bir sınıflandırıcı modelin oluşturulduğu işbirlikçi öğrenme modelidir. Homojen temel sınıflandırıcılardan oluşan Bagging modeli, sınıflandırıcı çeşitliliğini sağlamak için farklı veri kümeleri kullanmaktadır. V veri kümesinden elde edilen $(V_1, V_2, \dots, V_N)$ veri alt kümeleri ayrı ayrı $(S_1, S_2, \dots, S_N)$ sınıflandırıcılarının eğitiminde kullanılmaktadır. Veri alt kümelerinin elde edilmesinde Bootsrap örnekleme yöntemine başvurulur. Bootstrap örnekleme veri uzayı ile ilgili istatistiksel çıkarımlarda sıklıkla başvurulan bir yöntemdir. Örneğin V_1 veri alt kümesinin örnekleri V veri kümesinden rastgele seçilir ve seçilen her örnek tekrar kullanılmak üzere yerine yerleştirir. Bu durumda veri kümesinde aynı örneğin birden fazla kez görülmesi mümkündür.

$$e = \sum_{n=1}^{N} I(\arg\max(S_n = e)) \tag{3.1}$$

Bagging modelinde temel sınıflandırıcılar eğitildikten sonra sonuçları Denklem (3.1)'deki gibi çoğunluk oylaması ile veya sınıflandırıcıların sonuçlarının ortalaması alınarak birleştirilmektedir. Model test aşamasında $(S_1, S_2, \dots, S_N)$ temel sınıflandırıcılardan elde edilen e sınıf etiketi değeleri birbiriyle karşılaştırıldığında çoğunluktaki değer verinin sınıfı olarak kabul edilmektedir. Bagging yöntemi sınıf etiketlerinin sayısından bağımsız oluşturulabilen bir yöntemdir [106]. Yapay sinir ağları ve karar ağaçları gibi çıktılarının girdilerdeki küçük değişiklikler ile etkilenebildiği öğrenme algoritmaları ile kullanıldığında en etkili sonuçlar elde edilmektedir [109]. Uygulaması kolay bir yöntem olması nedeniyle ve az sayıda örneğe sahip küçük veri kümelerinde iyi çalıştığı için çalışmalarda sıklıkla kullanılmaktadır.

3.2. Boosting

Zayıf sınıflandırıcılar rastgele tahminlerden biraz daha iyi sonuç verirken, güçlü sınıflandırıcılar sınıflandırma işlemini mükemmele yakın bir şekilde gerçekleştirirler. Araştırmacılar zayıf ve güçlü sınıflandırma problemlerinin birbirine benzer olup olmadığını merak etmişler, eğer bu sınıflandırıcılar arasında benzerlik var ise zayıf sınıflandırıcıların performansının artırılıp arttırılmayacağını araştırmışlardır [106]. Schapire [111] önerdiği Boosting modeli ile zayıf sınıflandırıcıların güçlendirilebileceğini kanıtlamıştır. İlk Boosting modeli yaklaşımı Adaboost algoritmasıdır [112]. Adaboost modelinin kurulumunda ilk adım olarak bir S zayıf sınıflandırıcısı V eğitim kümesi ile oluşturulur. S zayıf bir sınıflandırıcı olduğu için V veri kümesi içerisinde yanlış sınıflandırdığı eğitim örnekleri bulunacaktır. İkinci adımda S' in yanlış sınıflandırdığı örnekler belirli değeler ile ağırlıklandırılarak V_1 veri kümesi oluşturulur. Oluşturulan bu yeni küme ile S sınıflandırıcısı ile aynı algoritmadan türetilen S_1 sınıflandırıcısının eğitimi gerçekleştirilir. Bu adımlar tekrarlanarak isteğe bağlı olarak $(S, S_1, \dots)$ temel sınıflandırıcıları elde edilir ve her bir sınıflandırıcıya başarımıyla orantılı olarak ağırlık değerleri atanır. Atanan ağırlık değerleri temel sınıflandırıcıların birleştirilmesi aşamasında ağırlıklı oylama için kullanılmaktadır.

4. DOKÜMAN UZAYINI GENİŞLETME

4.1. Özellik Mühendisliği

Veri madenciliği uygulamaları veriler üzerinden kullanışlı bilgiler elde edebilmek için uygulama alanını anlama, veri toplama, veri önişleme, makine öğrenmesi modellerin uygulanması ve elde edilen sonuçların değerlendirilmesi gibi genel adımlardan oluşmaktadır. Veri önişleme adımı genel süreç içerisinde geçirilen zaman açısından maliyeti en yüksek olan adımdır. Veri önişleme adımı kendi içerisinde ham verilerin gürültülerden arındırılması, veri kaynaklarının bir araya toplanması, özelliklerin türetilmesi, ilgili özelliklerin seçilmesi, normalizasyon gibi birçok işlem barındırmaktadır [88].

Veri madenciliğinde kullanılan modeller doğrudan ham veriler kullanılarak işletilemez. Bu nedenle ham veriler üzerinde çeşitli işlemler gerçekleştirilerek kullanışlı özellikler türetilir. Özellikler, işlenmemiş ham verinin sayısal veya sembolik olarak temsilleridir. Ham verilerin yapısı çeşitlilik gösterdiğinden özellik elde etmek için kullanılan işlemlerde farklılık göstermektedir. Ayrıca özellik elde etme yöntemleri kullanılacak makine öğrenmesi modeline göre de farklılık gösterebilir, bir model için uygun olan özellikler başka bir model üzerinde kullanılamayabilir [89]. Güçlü modelleri oluşturmada kullanılmak üzere ham veriyi en iyi şekilde temsil eden özellikleri elde etmek için yapılan işlemlerin bütünü özellik mühendisliği (feature engineering) olarak adlandırılmaktadır.

Doğru türetilen özellikler güçlü ve esnek modellerin oluşturulmasına yardımcı olmakta, özelliklerin ayırt edici karakteristiği sayesinde problemler analiz edilip çözüme kavuşturulabilmektedir [90]. Modellerde kullanılacak özelliklerin sayısı da önemli bir konudur. Yetersiz sayıda özellik modellerin tam kapasitesiyle çalışmasına engel olabilmektedir. Ayrıca fazla sayıda ilgisiz verilerin kullanılması modellerin eğitimi süresini arttırmakta, performansını ve başarımını olumsuz olarak etkilemektedir [89].

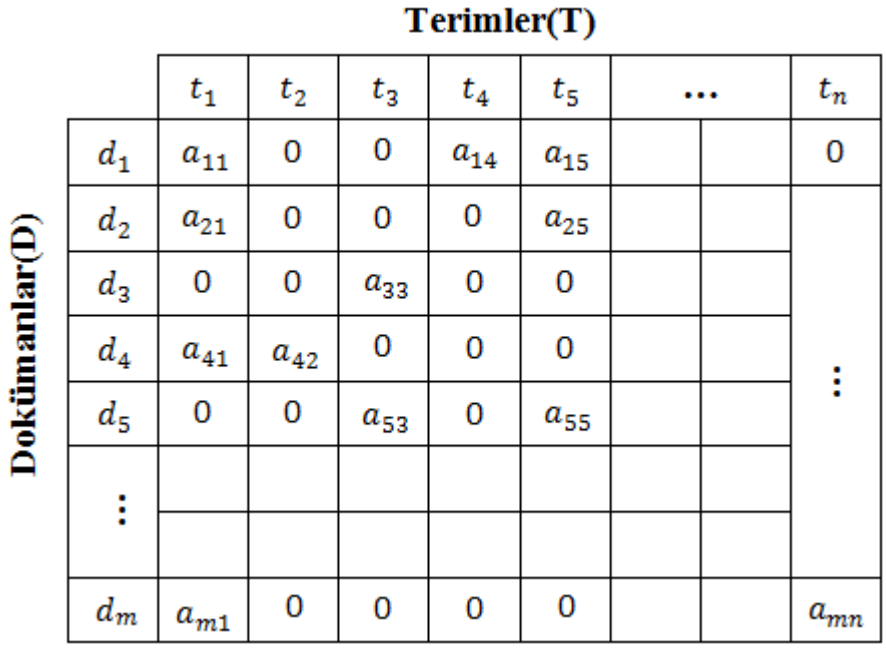
4.2.Kısa Metin Sınıflandırma

Gün geçtikçe insanların internet ile olan etkileşimi hızlı bir şekilde artmakta, buna bağlı olarak üretilen elektronik metinlerin miktarı devasa boyutlara ulaşmaktadır. Özellikle kısa metinler; e-ticaret siteleri, sosyal medya ve anlık mesajlaşma uygulamaları, makale ve haber başlıkları gibi birçok ortam karşımıza çıkmaktadır [91]. Bu tür veriler, yoğunluğu ve insanların görüşlerini yansıtması sebebiyle birçok çalışmada kaynak olarak kullanılmaktadır. Örneğin sosyal medya uygulamalarından elde edilen kısa mesajlar; sosyoloji, psikoloji, pazarlama ve bilgisayar bilimleri gibi birçok disiplinin çalıma konusu haline gelmiştir [2]. Bilgisayar bilimlerinde kısa mesajlar üzerinde en sık yapılan çalışma metinlerin sınıflandırılmasıdır. Kısa metinlerin sınıflandırılması birçok zorluğu içerisinde barındırmaktadır. İnsanlar genellikle internetteki paylaşımlarında dilbilgisi ve imla kurallarını göz ardı etmekte; argo kelimeleri, kısaltmaları, internet ortamında türetilen belirli jargonları sıklıkla kullanmaktadır [12-14]. Özellik çıkarımı yapılmadan önce bu gürültülerin giderilmesi gerekmektedir.

Metin sınıflandırmada, özellik temsili için kullanılan yöntemlerden biri vektör uzay modelidir [45]. Modelde her metin hayrı bir vektör ile temsil edilmektedir. Bu vektörlerin bir araya getirilmesiyle terim-doküman matrisi gibi matrisler oluşturulur. Sınıflandırma işlemi gerçekleştirilen metin kümesinin içerdiği benzersiz karakterler, kelimler vektörün bileşenlerini oluşturmaktadır. Belirli yöntemler kullanılarak kelime ve karakterlerin veya bunların oluşturduğu grupların metin içerisinde görünme sayısına göre vektöre değerler atanmaktadır. Kısa metinlerin sınıflandırılmasında karşılaşılan en büyük problem, temsil vektörünün bileşenlerini oluşturan özelliklerin sayısının metinlerin içerdiği kelimelere oranla çok yüksek olmasıdır [15,16]. Bu sebeple vektör içerisindeki bileşenlerin çoğu, özellikler metin içerisinde görülmediği için boş bırakılmakta ve bu durum gereksiz bir işlem maliyeti oluşmaktadır.

Metinlerin özellikler ile temsilinde oluşan boşlukları giderebilmek için yapılan işlemlerden birisi metinlerin içeriğinin genişletilmesidir. Genişletme işlemi iki farklı yöntem ile gerçekleştirilmektedir. Yöntemlerin ilki herhangi bir dış kaynaktan elde edilen bilgilerin metin içeriğine eklenmesiyle gerçekleştirilir. İkinci yöntemde hiçbir

dış kaynağa ihtiyaç duyulamadan, metinlerin içerdiği kelimeler arasındaki ilişkiden yararlanarak içeriğin boyutu arttırılmaktadır [49, 92, 93]. Kelimelerin arasındaki ilişkinin elde edilmesi için kullanılan yöntemlerden biride kelime gömme işlemidir.



Şekile 4.1. Kısa metinlerin terim döküman matrisi örneği

4.3. Kelime Gömme

Kelime gömme, kelimelerin boyutları önceden belirlenen bir uzaydaki vektör temsillerinin oluşturulmasıyla gerçekleştirilir. Dağılım yapısına (Distributional Structure) göre benzer içerikler içinde bulunan kelimeler yakın anlamlara sahiptir [96]. Bu sebeple vektör temsili oluşturulurken metinin içeriği ve söz dizimi göz önünde bulundurulmaktadır. Kelime gömme yöntemi ile birlikte kelimelerin benzerlikleri vektörler yardımıyla hesaplanmakta, benzer kelimeler kullanılarak metin içeriği zenginleştirilmektedir [37, 95]. Vektöreler arasındaki benzerlikler genellikle Öklid mesafesi ve kosinüs benzerliği aracılığıyla hesaplanmaktadır.

4.3.1. Coals

Coals, Rohde ve arkadaşları [97] tarafından önerilen, kelime vektör temsili için yapılan ilk çalışmalardan biri olup, eş-oluşum matrisi kullanılarak oluşturulan bir kelime temsil modelidir. Eş-oluşum matrisleri her satırı ve sütunu belirli bir kelimeye karşılık gelen, karşılıklı kelimelerin metin içeriklerinde beraber bulunma sayısını

gösteren bir simetrik kare matristir. COALS modelinde eş-oluşum matrisinin oluşturulması için gerekli bilgi pencere yapısı ile metinlerin içerisinden toplanmaktadır. Pencere yapısı, hedef kelimenin aynı içerikte beraber bulunduğu, aralarında belirli bir ilişki bulunabilecek kelimeleri belirlemek için kullanılır. Coals modeli uzunluğu 4 birim olan bir pencere yapısı kullanarak hedef olarak belirlediği kelimenin sağında ve solunda maksimum 2 birim uzaklıkta bulunan kelimeleri belirler. Belirlenen kelimeler hedef kelimeye yakınlıklarına göre ağırlıklandırılmaktadır. Yakın olan kelimeler 4 ve uzak olan kelimeler 3 değerini almak üzere ağırlık değerleri atanır. Atanan bu ağırlık değerleri kelimeler ile her karşılaşıldığında eş oluşum matrisindeki yerlerine eklenir.

Coals modelinde, başlangıçta oluşturulan eş-oluşum matrisi doğrudan vektör temsilinde kullanılmamaktadır. Vektöre temsili oluşturulmadan önce eş-oluşum matrisi belirli normalizasyon işlemlerinden geçer. Bazı kelimler anlam bakımında fazla bilgi sağlamamalarına rağmen içeriklerde sık olarak bulunabilmektedir. Bu durumun üstesinden gelebilmek için eş-oluşum matrisindeki değerlerin korelasyonları hesaplanmaktadır. $W = [w_{ab}]_{ixj}$ eş-oluşum matrisinin korelasyon değerleri Denklem (4.1)'e göre hesaplanmaktadır.

$$w'_{ab} = \frac{T w_{ab} - \sum_j w_{aj} \cdot \sum_i w_{ib}}{\left(\sum_j w_{aj} \cdot \left(T - \sum_j w_{aj}\right) \cdot \sum_i w_{ib} \left(T - \sum_i w_{ib}\right)\right)^{1/2}} \tag{4.1}$$

$$T = \sum_i \sum_j w_{ij} \tag{4.1a}$$

Korelasyon değerler -1 ve 1 arasında değişmektedir. Korelasyon 0'a eşitse iki kelimenin arasındaki ilişkinin rastgele bir kelime ile aynı olduğunu göstermektedir. Coals modelinde pozitif korelasyon değerlerinin anlam temsilinde daha güvenilir olduğu düşünüldüğü için matris içerisindeki negatif değerler yerine 0 değeri atanır.

Eş-oluşum matrisinin boyutu kelime temsili oluşturmak için çok büyüktür. Coals modelinde kelimelerin daha küçük bir uzayda daha verimli bir şekilde temsil edilebilmesi için eş-oluşum matrisinin boyutu küçültülür. Matrisin boyutlarını küçültmek için SVD (Singular Value Decomposition) yöntemine başvurulmaktadır.

Yapılan tüm işlemler sonunda oluşan matris kelimeleri temsil etmek için kullanılır. Matrisin her satırı karşılık geldiği kelime için bir vektör olarak kullanılmaktadır.

	a	as	chuck	could	how	if	much	wood	woodch.	would	,	.	?
a	0	5	9	6	1	10	4	8	18	9	10	0	0
as	5	4	2	1	0	0	7	10	3	2	1	0	5
chuck	9	2	0	8	0	5	1	9	11	2	4	3	3
could	6	1	8	0	0	4	0	6	8	0	2	2	2
how	1	0	0	0	0	0	4	3	0	2	0	0	0
if	10	0	5	4	0	0	0	0	10	3	8	0	0
much	4	7	1	0	4	0	0	10	2	3	0	0	3
wood	8	10	9	6	3	0	10	2	8	5	0	4	6
woodch.	18	3	11	8	0	10	2	8	0	8	10	1	1
would	9	2	2	0	2	3	3	5	8	0	5	0	0
,	10	1	4	2	0	8	0	0	10	5	0	0	0
.	0	0	3	2	0	0	0	4	1	0	0	0	0
?	0	5	3	2	0	0	3	6	1	0	0	0	0

Şekil 4.2. Başlangıçtaki oluşturulan Coals eş-oluşum matrisi [97]

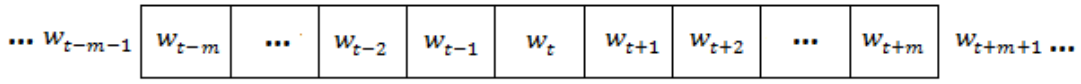
	a	as	chuck	could	how	if	much	wood	woodch.	would	,	.	?
a	0	0	0.120	0.093	0	0.291	0	0	0.310	0.262	0.291	0	0
as	0	0.175	0	0	0	0	0.364	0.320	0	0	0	0	0.365
chuck	0.120	0	0	0.306	0	0.146	0	0.177	0.220	0	0	0.297	0.175
could	0.093	0	0.306	0	0	0.182	0	0.149	0.221	0	0	0.263	0.151
how	0	0	0	0	0	0	0.438	0.265	0	0.263	0	0	0
if	0.291	0	0.146	0.182	0	0	0	0	0.291	0.076	0.372	0	0
much	0	0.364	0	0	0.438	0	0	0.358	0	0.136	0	0	0.268
wood	0	0.320	0.177	0.149	0.265	0	0.358	0	0	0.034	0	0.333	0.317
woodch.	0.310	0	0.220	0.221	0	0.291	0	0	0	0.221	0.291	0	0
would	0.262	0	0	0	0.263	0.076	0.136	0.034	0.221	0	0.246	0	0
,	0.291	0	0	0	0	0.372	0	0	0.291	0.246	0	0	0
.	0	0	0.297	0.263	0	0	0	0.333	0	0	0	0	0
?	0	0.365	0.175	0.151	0	0	0.268	0.317	0	0	0	0	0

Şekil 4.3. Normalizayon sonucu oluşturulan COALS eş-oluşum matrisi [97]

4.3.2. Word2Vec

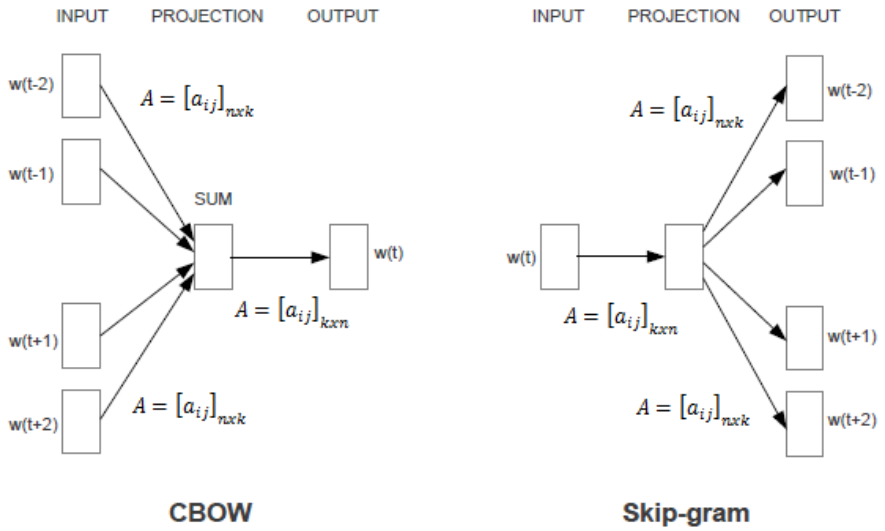
Word2Vec [35, 36] , Mikolov ve arkadaşları tarafından önerilen, içerik tabanlı bir kelime gömme yöntemidir. Temsil vektörleri oluşturulurken içerikteki hedef kelime ve hedef kelimeyi çevreleyen diğer kelimeler kullanılmaktadır. Vektör temsili oluşturulacak hedef kelimeyi ve içerik olarak kullanılacak çevre kelimeleri belirlemek için boyutu önceden belirlenen pencere benzeri bir yapı metin içerisinde dolaştırılır. Vektöre dönüştürülecek kelime pencerenin merkezinde bulunurken, pencerenin boyutuna göre çevre kelimelerin oluşturduğu içerik belirlenmektedir. Örneğin 2m boyutlarında bir pencere kullanılarak $C = \{\dots, w_{t-1}, w_t, w_{t+1}, \dots\}$ metni

içerisinde w_t kelimesini çevreleyen içerik belirlenmeye çalışıldığında en sağdaki kelimenin w_{t-m} en soldaki kelimenin ise w_{t+m} olduğu görülür.

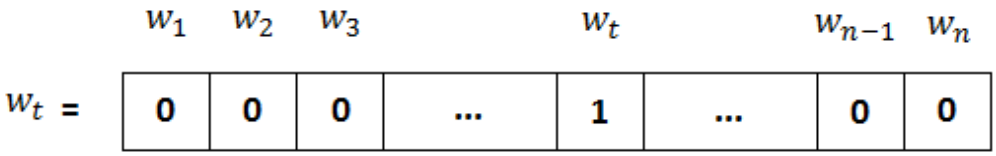


Şekil 4.4. 2m boyutlarına sahip pencere modeli

Word2Vec, CBOW (Continuous Bag Of Word) ve Skip-Gram olmak üzere iki farklı uygulama şekline sahiptir. CBOW ve Skip-Gram farklı şekillerde tasarlanmış girdi, çıktı ve bir gizli katmandan oluşan iki yapay sinir ağı modelidir. Gizli katmanın çıktısı vektör temsilini vermektedir. Bu nedenle gizli katmanın düğüm sayısı ve vektör temsilinin boyutu birbirine eşittir. Bunun dışında one-hot vektörünün ve pencerenin boyutu modellerin parametre sayısını etkiler. CBOW bir kelimeyi metin içerisinde kendisini çevreleyen kelimeler yardımıyla tahmin etmeye çalışırken; Skip-Gram, tam tersi şekilde hedef kelime ile çevresindeki kelimeleri tahmin etmeye çalışmaktadır. Modellerin oluşturulmasında kullanılacak kelimelerin işlemden geçirilebilmesi için one-hot gösterimi kullanılmaktadır. One-hot gösterimi, her boyutu metinlerin içerdiği $W = \{w_1, w_2, \dots, w_n\}$ benzersiz kelimelerine karşılık gelen n boyutlu bir vektör gösterimidir. Temsil edilecek kelimeye karşılık gelen boyut 1 değerini alırken, diğer boyutlara 0 değeri atanmaktadır [14].



Şekil 4.5. Word2Vev CBOW ve Skip-Gram mimarisi [35]



Şekil 4.6. w_t kelimesinin one-hot gösterimi

CBOW modelin girdi ve çıktıları, *n* boyutlu one-hot vektörleridir. Gizli katman içerisinde *k* adet düğüm olduğu varsayılırsa, gizli katman ve girdi arasında her girdi düğümü için $A = [a_{ij}]_{nxk}$ ağırlık matrisi oluşturulur. CBOW' da gizli katmanın değerleri elde edilirken her girdi vektörü kendi *A* ağırlık matrisi ile çarpılır ve *k* boyutlu *h* vektörü elde edilir. Daha sonra *h* vektörü içerik sayısı olan *2m* değerine bölünür. Gizli katman ile çıktıyı bağlayan ağırlık değerleri $B = [b_{ij}]_{kxn}$ ağırlık matrisi ile temsil edilirse, *h* vektörü ile *B* matrisi çarpımı Softmax fonksiyonundan geçirilerek çıktı oluşturulur.

$$y_i = \frac{\exp(u_i)}{\sum_{j=1}^{n} \exp(u_j)} \qquad (Softmax) \tag{4.2}$$

Denklem (4.2)'de gösterilen u_i, w_t çıktı vektörünün *i*. boyutundaki değere karşılık gelmektedir ve *h* vektörü ve *B* matrisi çarpımında elde edilir. Modelin optimizasyonunu sağlamak için Denklem (4.3)'te gösterilen Log-Likehood maliyet fonksiyonu kullanılmaktadır.

$$y_i - \log \sum_{j=1}^{n} \exp(y_j) \tag{4.3}$$

Skip-gram modelinin CBOW'a göre farkı girdisinin w_t hedef kelimesiyken, çıktısının $\{w_{t-m}, \ldots, w_{t+m}\}$ içeriği olmasıdır. Ayrıca gizli katmanın çıktısı tek vektör girdisi olması nedeniyle içerik sayısına bölünmez. Bunun dışında Skip-Gram modelinin çıktı ve maliyet fonksiyonları CBOW ile aynıdır.

4.3.3. GloVe (global vector)

GloVe, Pennington ve arkadaşları [94] tarafından önerilen vektör temsil modelidir. Word2Vec vektör temsili, yapay sinir ağı modeli yardımıyla oluşturulduğu için

eğitim maliyeti oldukça yüksektir. Modelin eğitimi sırasında metin kümesinin birden fazla kez taranması ve bilgi temsilleri işlem yükünü ciddi oranda arttırmaktadır. GloVe bu problemin üzerinden gelebilmek bilgi temsilinde eş-oluşum matrisinden (co-occurrence matrix) yararlanmaktadır.

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Şekil 4.7.Glove eş-oluşum matrisi [94]

Eş oluşum matrisi; satır ve sütunların metin kümesi içerisinden çıkarılan benzersiz kelimelerle gösterildiği, kelimeler arasındaki ilişkiyi belirtmek için kullanılan bir matristir. Eş oluşum matrisinin en büyük avantajı metin kümesinin bir kere taranmasının yeterli olmasıdır. GloVe için kullanılan eş oluşum matrisi oluşturulurken Word2Vec' te kullanılan pencere yönteminden yararlanılmaktadır. Metin üzerinde dolaştırılan pencere sayesinde aynı içerikte bulunan kelimelerin sayıları belirlenir. $X = [x_{ij}]_{lxk}$ matrisinin x_{ij} elemanı i kelimesi ile j kelimesinin aynı içerik içerisinde kaç kez kullanıldığını belirtmektedir. Bu matirisin sütunları içerik kelimelerini temsil etmektedir. Bir sonraki adımda x_{ij} , $\sum_k x_{ik}$ değerine bölünerek $P(j|i)$ koşullu olasılıkları elde edilir ve $P = [p_{ij}]_{lxk}$ olasılık matrisi oluşturulur. GloVe'da kelimelerin arasındaki anlam ilişkileri olasılıkların oranlarına göre gözlemlenmektedir. Örneğin benzersiz kelimeler içerisinde *i* ve *j* kelimelerinin *k* içeriğine göre aralarındaki ilişki incelenmek istendiğinde p_{ik}/p_{jk} olasılık oranı kullanılmaktadır. Kelime vektörleri oluşturmak için bu olasılık oranları bir başlangıç noktası olarak kullanılmaktadır.

$$F(w_i, w_j, \bar{w}_k) = \frac{p_{ik}}{p_{jk}} \tag{4.4}$$

Denklem (4.4)'teki w_i ve w_j, i ve j kelimelerinin vektör temsillerini belirtirken; $\bar{w}_k$ ise içerik kelime vektörlerini temsil etmektedir. Bu vektörlerin bulunabilmesi için F tanımlanması gerekmektedir. Kelimeler arasındaki benzerlik incelendiği için $(w_i - w_j)$ vektöre farkı kullanılabilir. F'in sinir ağlarında olduğu gibi karmaşık bir

yapıda olması istenmediğinden $(w_i - w_j)^T . \bar{w}_k$ noktasal çarpımıyla tanım basitleştirilir.

$$F\left((w_i - w_j)^T . \bar{w}_k\right) = \frac{p_{ik}}{p_{jk}} \tag{4.5}$$

Eş-oluşum matrisinde kelime ve içerik kelimesi arasındaki fark isteğe bağlı olarak oluşturulmaktadır ve bu kelimelerin rolleri değiştirebilir. Ancak Denklem (4.5) bu duruma uygun değildir. Bu nedenle denklem sağ ve solundaki parametrelerin aynı parametreler kullanılacak şekilde tekrar tanımlanır.

$$F\left((w_i - w_j)^T . \bar{w}_k\right) = \frac{F({w_i}^T . \bar{w}_k)}{F\left({w_j}^T . \bar{w}_k\right)} \tag{4.6}$$

$$F({w_i}^T . \bar{w}_k) = p_{ik} = \frac{X_{ik}}{X_i} \tag{4.7}$$

Denklem (4.7)'deki $F = exp$, $(R, +)$ ve $(R_{>0}, +)$ grupları arasında homomorfizma olduğu varsayılarak Denklem (4.9) elde edilir.

$${w_i}^T . \bar{w}_k = \log(p_{ik}) = \log(X_{ik}) - \log(X_i) \tag{4.8}$$

Denklem (4.8)'in sağında bulunan $\log(X_i)$ değeri w_i'nin b_i bias değeri olarak yazılıp, denklemin simetrik olmasını sağlamak için b_k bias değeri de eklenirse Denklem (4.9) elde edilir.

$${w_i}^T . \bar{w}_k + b_i + b_k = \log(X_{ik}) \tag{4.9}$$

Kelime vektörlerini belirlemek için özel olarak tanımlanan Denklem (4.10) ağırlıklı maliyet fonksiyonu kullanılmaktadır. Bu maliyet fonksiyonun içerdiği ağırlık değerleri yine model için özel olarak tanımlanan Denklem (4.10a) tarafından üretilir.

$$J = \sum_{i,j} f(X_{ik})({w_i}^T . \bar{w}_k + b_i + b_k - \log(X_{ik}))^2 \tag{4.10}$$

$$f(x) = \begin{cases} (x/x_{max})^{\alpha}, & x < x_{max} \\ 1 \quad\quad\quad , & diğer \end{cases} \tag{4.10a}$$

Denklemde (4.10a)'da x_{max} değeri 100'e eşitken, α değeri $3/4$'e eşittir.

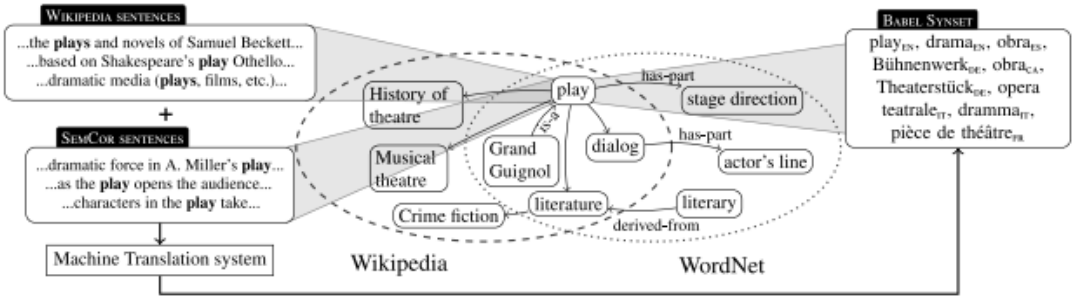
4.4. BabelNet

İnsanlar yeni bir dil öğrenmeye başladıklarında bu dil ile iletişim kurabilmek için belirli miktarda sözlük bilgisine ihtiyaç duymaktadır. İletişimin doğru bir şekilde gerçekleşmesi, kelimelerin anlamlarının açıkça anlaşılmasıyla gerçekleştirilir. Bu durum doğal dil işleme çalışmalarında da aynıdır. Yapılan bir çalışmanın başarımının yüksek olabilmesi kelimelerin anlam ayrımlarının doğru bir şekilde yapılabilmesine bağlıdır. Bu sebeple çalışmaların büyük çoğunluğu anlam ayrımını yapabilmek için anlamsal olarak iyi tanımlanmış güçlü bir sözlüğe ihtiyaç duymaktadır. BabelNet, WordNet [65,99] kavramlarının ve Wikipedia sayfalarının bir araya getirilmesiyle oluşturulan ansiklopedik bir sözlük ve bir kavram ağıdır [98]. BabelNet, WordNet'in içerdiği kavramlar ve Wikipedia sayfalarının belirli yöntemler ile eşleştirilmesiyle elde edilen, synset adı verilen birbiriyle ilişkili kavram kümelerinden oluşmaktadır.

WordNet; dil bilgisi kurallarına göre oluşturulmuş, kavramların kendisine anlamsal olarak benzeyen kavramları içeren synset isimli kümeler ile temsil edildiği, doğal dil işleme uygulamalarında sıklıkla kullanılan zengin bir bilgi kaynağıdır. Örneğin "paper" kavramının WordNet üzerindeki küme temsillerinden biri $\{sheet_n^2, peace\ of\ paper_n^1,\ sheet\ of\ paper_n^1\}$ kümesidir. Küme içerisinde üst simgeler kelimenin kaç farklı anlama geldiğini gösterirken, alt simgeler metin içerisindeki görevini belirtir. Ayrıca WordNet her synset için bir açıklama metnini ve synsetlerin birbirleriyle olan anlamsal veya yapısal ilişikisini içermektedir. Wikipedia; gönüllülük esasına dayanan, birçok dilde bilgilerin paylaşıldığı, web tabanlı bir ansiklopedidir Adlandırılmış varlıkları veya kavramları açıklayan makaleler sayfalarla temsil edilmektedir. Kavramlar arasında anlam ayrımlarını yapabilmek için makale başlıklarında parantez içinde anlam etiketleri bulunabilir. Makale içerisinde özet niteliğinde, kavramla ilgili kısa bilgi veren tablolar olabilir. Wikipedia içerisinde anlam ayrımı için sayfalar, benzer anlamlara gelen kavramlar için yönlendirme sayfaları, sayfaların kategorilerini ve sayfalar içerisinde diğer kavramlara yönlendirme yapan dahili linkleri barındırmaktadır.

BabelNet üzerinde bilgi $G = (E, V)$ graf yapısıyla temsil edilmektedir. V kavram veya isimlendirilmiş varlıkları temsil eden düğümlerin kümesidir. Düğümler WordNet synsetlerinin ve Wikipedia sayfalarından elde edilen kavramlar ile

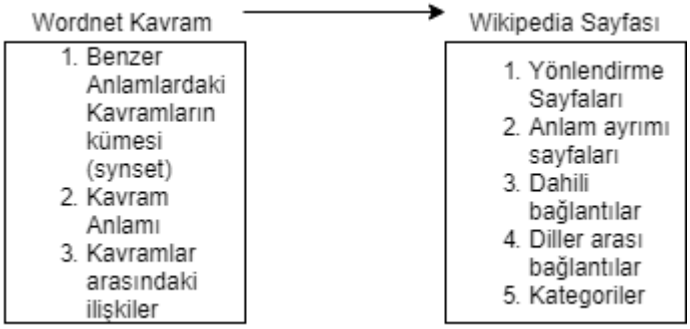
genişletilmiş halidir. $E \subseteq V\ x\ R\ x\ V$ kenarlar kümesi ise kavramlar arasındaki anlam ilişkisini göstermektedir. $R = \{is\ a,\ part\ of, \ldots, \epsilon\}$ kavramlar arasındaki anlam ilişkisi türleridir ve ϵ tanımlanmamış anlam ilişkisini gösterir.



Şekil 4.8. BabelNet graf temsili [98]

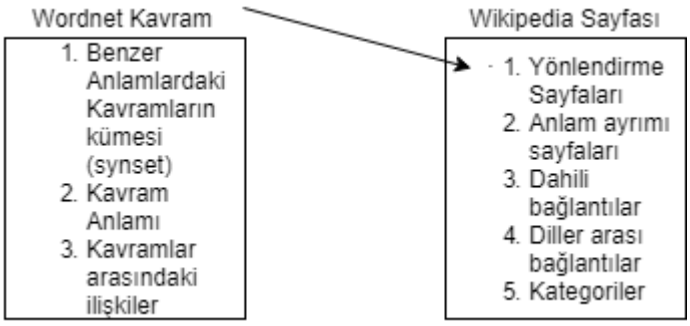
BabelNet synset kümelerinin oluşturulması süreci üç adımda gerçekleştirilmektedir. İlk adımda WordNet synsetleri Wikipedia'dan elde edilen kavramlar ile genişletilerek zengin içerikli BabelNet synsetler elde edilir. İkinci adımda elde edilen synsetin içerdiği kavramlar farklı dillere çevrilerek tekrar synset içerisine ilave edilir. Son adımda elde edilen tüm BabeNet synset kümeleri arasındaki ilişkiler Wordnet'ten içerisinden belirlenerek sysnsetler birbirine bağlanır.

WordNet synset kümesi ve Wikipedia sayfalarının kümesinin elemanlarını birbirleriyle eşleştirebilmek için özel bir yöntem geliştirilmiştir. Yöntemin ilk aşamasında Wikipedia sayfası olan w'lerin başlıklaları ile aynı anlamı taşıyan WordNet synsetleri doğrudan eşleştirilir.



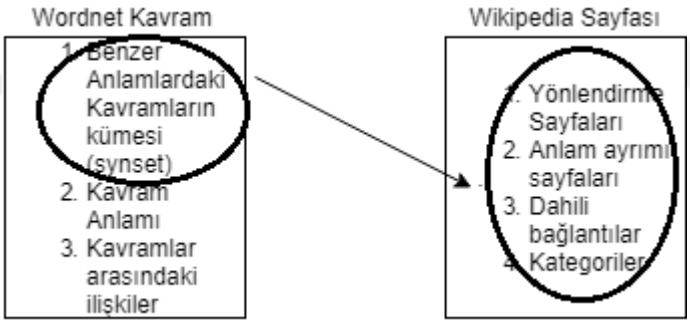
Şekil 4.9. Babelnet WordNet eşleştirme birinci adımı

İlk işlem sonucunda eşleştirilemeyen *w* sayfalarını için *w* sayfalarına yönlendirme yapan *d* Wikipedia sayfalarının başlıkları kullanılmakta, d ile eşleştirilen synset *w* sayfası ile de eşleştirilmektedir.



Şekil 4.10. Babelnet WordNet eşleştirme ikinci adımı

Birinci ve ikinci işlem sonucunda bir *w* Wikipedia sayfası ile birden fazla Wordnet synseti eşleşirse, anlam ayrımı yapılabilmesi için bu iki kaynağın içerikleri kullanılarka koşullu Denklem (4.13)' teki gibi bir koşullu olasılık değeri hesaplanır.



Şekil 4.11. Babelnet WordNet eşleştirme üçüncü adımı

$$p(s|w) = argmax_s \ \frac{p(s,w)}{p(w)} \tag{4.13}$$

$p(w)$ maksimum değerin bulunmasında etkili olmadığı için formülden kaldırılır. Bileşik olasılığın hesaplanabilmesi için kelime çantası metodu ve graf tabanlı bir metot önerilmektedir. Bu yöntemler için iki kaynaktan anlam ayrımında kullanılacak özellikler toplanmaktadır. Wikipedia için anlam ayrımında kullanılan özellikler anlam etiketleri, kavram linkleri, yönlendirme linkleri ve kategorilerdir. WordNet üzerinden elde edilen özellikleri synset içerisindeki kavramlar, benzer anlamdaki

synset kümeleri içerisindeki kavramlar ve synset kümelerinin karşılık geldiği kavramların anlamlarını açıklayan metinler oluşturmaktadır.

Babelnet synset kümeleri çok dilli bir yapıya sahip oldukları için içlerinde kavramların farklı dillerdeki karşılıklarını da barındırmaktadır. Bu nedenle eşleştirilme sırasında kurulan synset farklı dillere çevrilmektedir. Kavramların çevirisinde ilk kullanılan kaynak Wikipedia'dır. Wikipedia herhangi bir sayfanın yabancı dillerdeki sayfalarına yönlendirme linkini bünyesinde barındırmaktadır. Bu sayede birçok kavramın farklı dillerdeki karşılığı doğrudan elde edilebilmektedir. Wikipedia dışında kalan kavramlar için sözlük kullanılmaktadır.

Elde edilen synset kümeleri ile graf oluşturabilmek için kümeler arasındaki ilişkinin belirlenmesi ve bu ilişkinin gücünün hesaplanması gerekmektedir. Wordnet'ten elde edilen ilişki çeşitleri ve Wikipedia linkleri bu amaç doğrultusunda kullanılır. İlişkinin gücünün belirlenmesi için ise dice katsayısı kullanılmaktadır. Dice katsayısı, iki s synset kümesinin kesişiminin iki katının synset kümelerin eleman sayılarının toplamına bölünmesiyle elde edilmektedir.

$$\frac{2|s \cap s'|}{|s| + |s'|} \tag{4.14}$$

5. DENEYSEL TASARIM VE SONUÇLARI

5.1.Veri Kümeleri

Tez çalışmasında kullanılan veri kümelerinin tamamı benzer çalışmalarda önerilen, sınıflandırıcı modellerin performansını değerlendirmek için kullanılan yaygın veri kümeleridir. Veri kümelerin tamamı Twitter üzerinden toplanan, erişime açık, İngilizce kısa metinlerden oluşmaktadır. Veriler çeşitli etiketleme araçlarıyla pozitif ve negatif olarak etiketlenmiştir. Sınıfların dengeli bir şekilde dağıldığı görülmektedir. Veriler ilgili özet bilgiler Tablo 5.1'de gösterilmektedir.

Tablo 5.1. Veri Kümeleri

Veri Kümesi	Pozitif	Negatif	Konu
Iphone6	371	161	Akıllı Telefon
Hobbit	354	168	Film
UMICH	3995	3091	Film
Archeage	724	994	Oyun
Ststest	183	177	Genel
StsGold	596	739	Genel

5.2. Kullanılan Teknolojiler

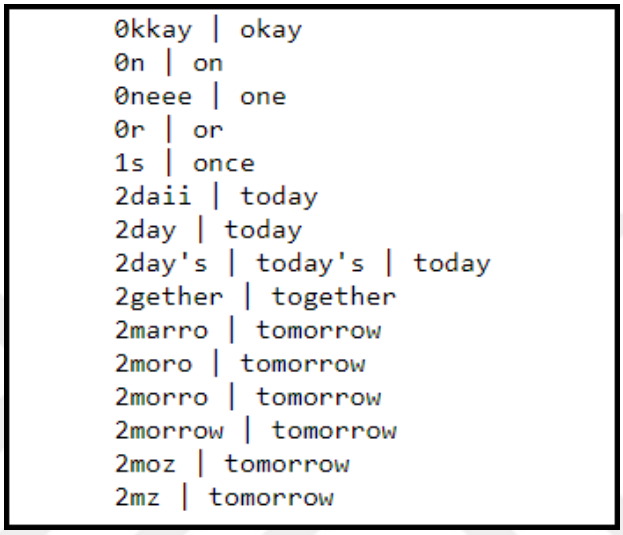
Tez çalışması gerçekleştirilirken Java nesneye yönelik programlama dili kullanılmıştır. Doğal dil işleme alanında yapılan işlemler ile ilgili birçok Java kütüphanesinin bulunması bu dilin kullanılmasında önemli rol oynamıştır. Metin önişleme adımında LanguageTool ve Stanford CoreNLP [100] kütüphaneleri kullanılmıştır. LanguageTool [101] kütüphanesi kelimelerin yazım kurallarına uygun olup olmadığını kontrol eden, belirli ölçüde yanlış yazılan kelimeler düzelten bir Java kütüphanesidir. CoreNLP kütüphanesi; birimlendirme, sözcük birimleştirme, gövdeleme ve sözcük türü etiketi gibi birçok temel metin önişleme adımını

bünyesinde barındıran güçlü bir doğal dil işleme kütüphanesidir. Çalışmada özellik uzayı genişletilirken BabelNet API ve Babelfy API kullanılmıştır. BabelNet synset adı verilen kavram kümelerinde oluşan bir kavramlar grafıdır. Grafta her düğüm bir synsete, her synsette bir kavrama karşılık gelmektedir ve her synsetin kendine özgü bir kimlik numarası mevcuttur. BabelNet API üzerin yapılan sorgulama ile bir kavrama anlamca yakın olan kavramların listesi elde edilmektedir. Anlam ayrımı için kelimenin metin içerisindeki görevinden yararlanılmaktadır. Babelfy API, bir kavramın anlamını açık hale getirerek en doğru synsetin kimlik numarasının elde edilmesini sağlar. Sorgular sonucu elde edilen kimlik numarası BabelNet API üzerinde kullanılarak synsetler elde edilir. Her API için günlük sorgu sayısı 2000 ile sınırlıdır. İşbirlikçi sınıflandırma modelinin gerçekleştirilmesinde WEKA kütüphanesinden [85,86] yararlanılmaktadır. WEKA birçok denetimli ve denetimsiz öğrenme algoritmasını sahip, makine öğrenmesi alanında sıklıkla kullanılan bir araçtır. Sağladığı Java kütüphanesi ile öğrenme algoritmaları rahatlıkla projelere entegre edilebilmektedir.

5.3. Metin Önişleme

Metin madenciliğinde modellerde kullanılacak gerekli bilgilerin elde edilmesi için veriler belirli önişleme adımlarından geçirilmektedir. Kısa metinlerin önişleme adımlarının büyük bir kısmı normal metinler ile aynı olmasına rağmen yapısı gereği kısa metinlerin fazladan işlemlerden geçmesi gerekmektedir. Twitter mesajları gibi kısa metinlerin büyük bir çoğunluğu elektronik ortamlarda üretildiği için dilbilgisi ve yazım hataları içermektedirler. Twitter mesajları içeriklerinde emojiler, semboller, hashtagler, linkler, karakter tekrarları ve internet ortamında kullanılan çeşitli kısaltmaları ve jargonları barındırmaktadır. Çalışmada kullanılan gürültülü İngilizce metinlerin standart bir metin haline getirilmesi için Texas Üniversitesi Bilgisayar Bilimleri Bölümü tarafından geliştirilen bir sözlük ile normalizasyon işlemi gerçekleştirilmiştir [102]. Bu sözlük özellikle Twitter'da yaygın olarak bulunan yazım yanlışlarını, kısaltmaları ve bu ifadelerin doğru yazımlarını içermektedir. Mesajlardaki yazım yanlışları sözlük yardımıyla olabildiğince düzletildikten sonra LanguageTool [101] kütüphanesi kullanılarak var olan imla hataları olabildiğince mesaj içerisinden temizlenmeye çalışılmıştır.

Yazım yanlışlarını düzeltmenin dışında mesajların içerisinden emojiler, kullanıcı adları, URL'ler, retweet işaretçileri, kullanıcı mentionları kaldırılmıştır. Bunların dışında mesaj içeriğindeki noktalama işaretleri ve rakamlar kaldırılmış, geri kalan tüm karakterler küçük harflere dönüştürülmüştür. Önişleminin son adımında Stanford CoreNLP [100] kütüphanesi kullanılarak mesaj içerikleri üzerinde sözcük birimleştirme işlemi gerçekleştirilerek kelimeler gövde hallerine indirgenmiştir.



Şekil 5.1. Normalizasyon sözlüğü örnek içerik [102]

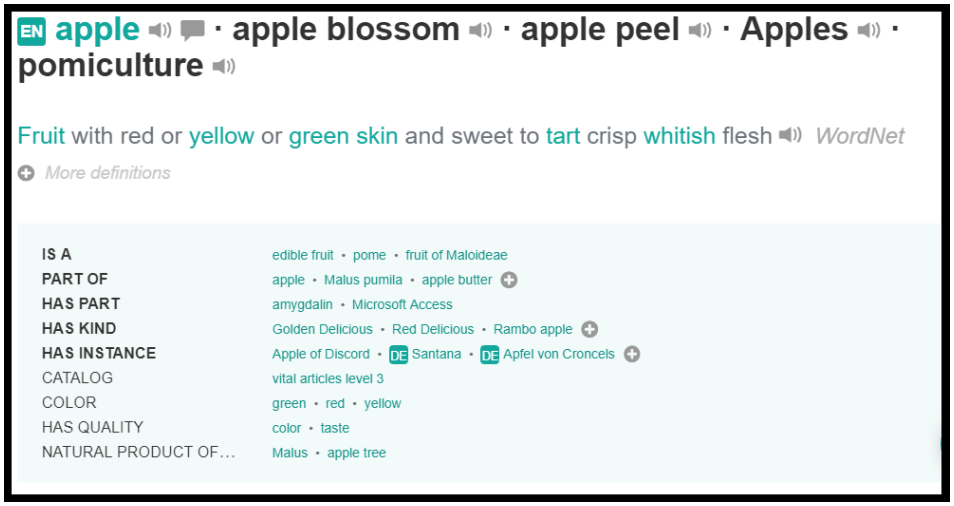
Tablo 5.2. Metin önişleme örnekleri

The Da Vinci Code book is just awesome.	the the vinci code book be just awesome
the hobbit was sooooo good omfg :D:D:D	the hobbit be so good
be sure to check out my video on how i edit #instagram photos! http://t.co/umrsdddjhd #youtuber #iphone6	be sure to check out my video on how edit instagram photo youtube iphone
@archeage pls revise fishing; u cant make the line; wool almost impossible to obtain this early and we want to basic fish at least.....	please revise fishing you cant make the line wool almost impossible to obtain this early and we want to basic fish at least
Fine-tuning part of a song Maddy and I r making. Sounding good. I feel pro. Not. Hahaha	fine tuning pa of song maddy and be make sound good feel pro not havana
@MissSydneyJ Im good, lol... I feel awake	in good lol feel awake

5.4.Doküman Uzayının Genişletilmesi

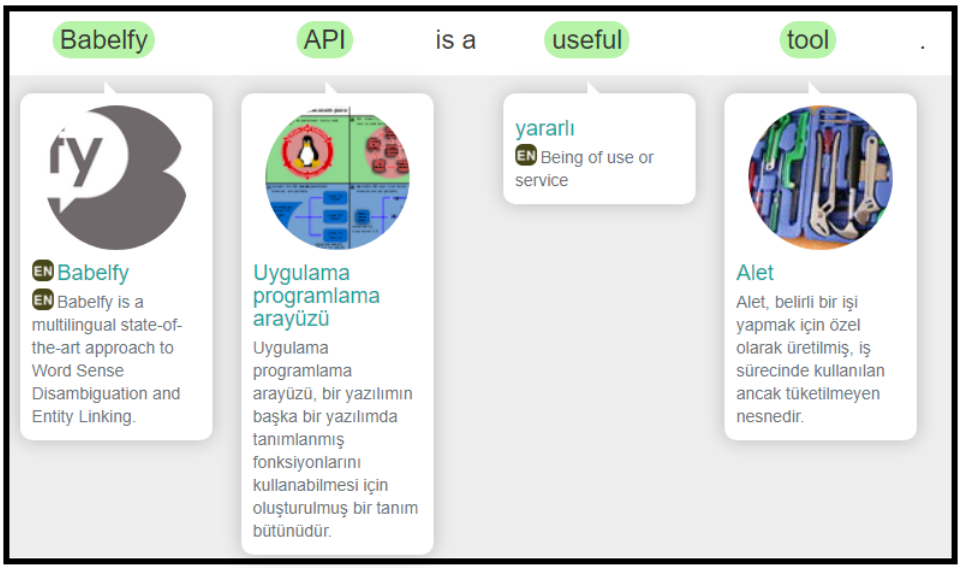
Metin madenciliğinde metinlerin sınıflandırma modellerinde işletilebilmeleri için karakteristiklerini yansıtan bilgi temsillerinin oluşturulması gerekmektedir. Bu nedenle birçok çalışmada metin temsilleri için vektör uzay modelinden yararlanılır [66]. Vektör uzay modelinde her boyut metin kümesindeki benzersiz kelimelere denk gelmektedir ve bu boyutlara genellikle kelimelerin metin içerisinde görünme sayısı ile orantılı değerler atanır. Ancak kısa metinlerin sınıflandırılmasında bu durum dezavantaj oluşturmaktadır. Boyutun laneti olarak adlandırılan bu durumda metinlerin kısalığıyla ilişkili olarak kelimelerin içerik içinde görünme sayısı azalmakta, hatta hiç görünmemesi durumu ile karşılaşılmaktadır. Bu sebeple vektörün büyük bir kısmına değer atanamamaktadır. Bu durumun üstesinden gelmek için yapılan işlemlerden biri metin içerersindeki kelimelere benzer anlamlar taşıyan kelimelerin metin içeriğine eklenmesidir. Bu çalışmada metin içeriğine zenginleştirmek için BabelNet ansiklopedik sözlüğü kullanılmıştır. BabelNet her düğüm, bir kavram ve bu kavrama anlamca benzer olan kavramlardan oluşan bir synset isimli kümelerden oluşmaktadır. Ayrıca BabelNet kavramlar arasında anlam ilişkinin hangi türden olduğunu da göstermektedir. Graf içerisinde her düğüm BabelNet Id isimli kendi kimlik parametreleri ile temsil edilmektedir. Örneğin “bn:00289737n” parametresindeki “bn” ifadesi düğüm kaynağının BabelNet olduğunu, “n” ifadesi ise düğümü temsil eden kavramın isim olduğunu belirtmektedir.

BabelNet üzerinden bilgiye ulaşmak için BabelNet API kullanılmaktadır. BabelNet API sorgu tabanlı olarak çalışmakta, herhangi bir terim ile benzer anlam taşıyan kavramlar sorgu yardımıyla elde edilmektedir. API ile sorgu işlemi günlük 2000 ile sınırlandırılmıştır. İncelenmek istenen terim doğrudan sorgu içerisinde kullanılabileceği gibi terimin BabelNet Id değerleri biliniyorsa sorgu için kullanılabilmektedir. Kelimeler ile yapılan sorgular için birden fazla aday synset kümesi döndürülürken, BabalNet Id ile yapılan sorgular doğrudan aranan kelimeyle ilgili sonucu döndürmektedir.



Şekil 5.2. BabelNet web uygulaması sorgu soncu

BabelNet Id değerlerinin sorgu içinde kullanılması daha kesin sonuçlar elde edilmesini yardımcı olmaktadır. Çalışmada bir terimin BabelNet Id değerini elde etmek için Babelfy API kullanılmıştır. Babelfy API; metinlerin sorgu olarak kullanıldığı, kelimelerin metin içerisindeki görevlerine göre anlam ayrımlarını yaparak BabelNet Id değerlerini döndüren kullanışlı bir araçtır. BabelNet kaynak olarak kullanan Babelfy API günlük 2000 sorguya kadar izin verirken bir metin içerisindeki kavramların BabelNet Id değerlerini döndürmektedir.



Şekil 5.3. Babelfy web uygulaması sorgu sonucu

Çalışmada önişleme adımından geçirilen metinler Babelfy API ile sorguya sokularak, içerdikleri kavramların BabelNet Id değerleri elde edilmiştir. Elde edilen bu değerler BabelNet API' de sorguya sokularak kelimeler karşılık gelen synset kümeleri elde

edilmiştir. Elde edilen kümelerdeki kavramlar metin içeriklerine doğrudan eklenerek doküman uzayı zenginleştirilmiştir.

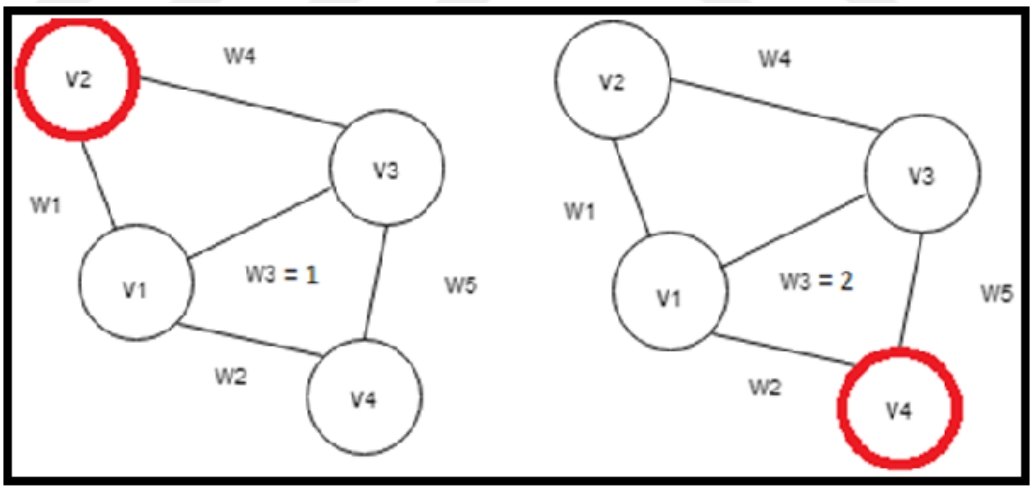
Tablo 5.3. Genişletme örnekleri

the the vinci code book be just awesome	the the vinci code book be just awesome "laws" "codes" "pages" "written" "written work" "wonder" "admiration"
the hobbit be so good	the hobbit be so good manufacture tolkien species smaller skeleton primate prehistoric pleistocene person man mammal island indonesian indonesia imaginary human homo hominid hobbit hairy floriensis floresiensis floresiens floresian floresensis florensis flore extinction early asia quality good especially desirable unite the technology social science revolution product politics microeconomic manufacturing manufacture labor kingdom japan industry industrial horology history factory economy economic craft artisan and
be sure to check out my video on how edit instagram photo youtube iphone	be sure to check out my video on how edit instagram photo youtube iphone "secure" "Physically" "dependable" "accuracy" "departure" "Announce" "hotel" "audible" "movie" "television" "television program" "Twitter" "Tumblr" "social networking service" "Flickr" "Facebook" "social networking platforms" "service" "transparent" "light-sensitive" "person" "website" "San Bruno
please revise fishing you cant make the line wool almost impossible to obtain this early and we want to basic fish at least	please revise fishing you cant make the line wool almost impossible to obtain this early and we want to basic fish at least "indirectly" "Seek" "formation" "hair" "fabric" "sheep" "capable" "possession" "expected" "period" "events" "usual" "look for" "reason" "Pertaining" "group" "jawless vertebrates" "vertebrates" "bony fishes" "vertebrates" "fishes"
fine tuning pa of song maddy and be make sound good feel pro not havana	fine tuning pa of song maddy and be make sound good feel pro not havana "smooth" "relatively" "textures" "pitches" "musical instruments" "instruments" "amplification" "electronic" "communication system" "public" "short" "words" "musical composition" "Financially" "desirable" "especially" "qualities" "intuitive" "awareness" "proposal" "action"
in good lol feel awake	in good lol feel awake "desirable" "especially" "qualities" "Internet slang" "Internet" "laugh" "slang" "acronym" "Undergo"

5.4.1.Babelfy

Babelfy, kavramların ve isimlendirilmiş varlıkların üzerinde anlam ayrımı yapılabilmesi için geliştirilen bir yöntemdir. Bu işlem, BabelNet anlamsal ağını kaynak olarak kullanır ve üç farklı adım gerçekleştirmektedir. İlk adımda BabelNet içindeki her düğüm anlamsal bir imza ile eşleştirilir. Anlamsal imza bir düğümün ilişkili olduğu diğer düğümlerin kümesidir. Bu adımda anlam ayrımı yapılacak metin kullanılmaz. İkinci adımda ise yönteme girdi olarak verilen metin anlamlı parçalara ayrılır. Her parça için olası aday kavram listesi BabelNet yardımıyla elde edilir. Son adımda ise önceden elde edilen anlamsal imzalar ile aday anlamlar bir graf yapısında toplanır ve metnin graf temsili oluşturulur. Bu temsilden elde edilen yoğun alt graf sayesinde metin parçaları ile eşleşen en uygun aday anlamlar belirlenir.

Babelfy yönteminde kullanılan anlamsal imzaların elde edilmesi için öncelik BabelNet grafındaki her kenar için bir ağırlık değeri belirlenir. (v, v') düğümlerini bağlayan bir kenarının ağırlık değeri bu iki düğümü komşu olan diğer düğümlerin sayısına eşittir.



Şekil 5.4. Babelnet bağlantıların ağırlıklandırılması

$$ağırlık(v, v') \coloneqq |\{(v, v', v'') : (v, v'), (v', v''), (v, v'') \in E\}| + 1 \qquad (5.1)$$

Anlamsal imzaların belirlenmesi için Denklem (5.1) ile elde edilen ağırlık değerlerinden yararlanılır. Bir düğümün kendisiyle ilişkili olan düğümleri belirlenirken BabelNet graf ağı üzerinde rastgele yürüyüş gerçekleştirilir. Yüyüyüşün başlangıç noktası anlamsal imzası aranan düğüm olurken, gidilecek yön ağırlık

değerleri ile elde edilen bir $P(v'|v)$ olasılık değeri ile belirlenir. Olasılığı yüksek olan kenarlar üzerinde ilerleriken karşılaşılan her düğüm bir liste içerisinde tutulur.

$$P(v'|v) = \frac{ağırlık(v, v')}{\sum_{v'' \epsilon V} ağırlık(v, v'')} \tag{5.2}$$

Metnin graf temsili oluşturulurken F metni f parçalarına ayrılmaktadır. Elde edilen her f parçası için BabelNet kullanılarak v aday kavramlar belirlenir. Belirlenen bu adaya anlamlar $G = \{E, V\}$ grafının düğümlerini oluşturmaktadır.

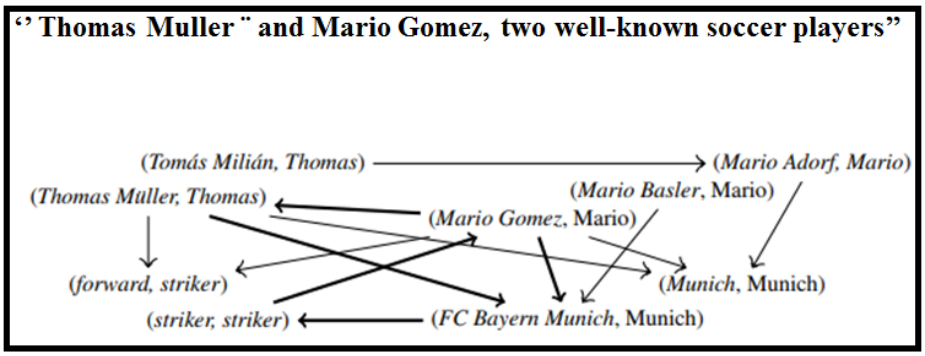
$$V = \{(v, f)|v \epsilon aday(f), f \in F\} \tag{5.3}$$

Eğer v aday kavramlardan biri başka bir aday kavramın anlamsal imza listesinde bulunuyorsa bu iki düğüm arasına bir kenar çizilir. Metnin graf temsili belirtilen şekilde elde edilmektedir. Elde edilen graf içerisinden metin parçaları için en uygun aday anlamaların belirlenebilmesi için indirgeme işlemi gerçekleştirilmektedir. İndirgeme işlemi Denklem (5.4)' de belirtilen skorlar ile gerçekleştirilir. Zayıf skorlara sahip aday anlamlar graftan kaldırılarak yoğun bir graf oluşturulur.

$$skor\big((v, f)\big) = \frac{w_{(v,f)} \deg((v, f))}{\sum_{v' \in aday(f)} w(v', f) \deg((v', f))} \tag{5.4}$$

$$w_{(v,f)} = \frac{\left|\{f' \in F : \exists v'\ k.a\ \big((v, f), (v', f')\big) veya \big((v', f'), (v, f)\big) \in E\}\right|}{|F| - 1} \tag{5.4a}$$

Denkleme (5.4)'de $\deg((v, f))$ değeri v aday düğümünün bağlantı sayısını belirtirken, $w_{(v,f)}$ değeri bir v aday düğümün diğer f' metin parçalarının v' aday düğümlerine olan bağlantı sayısın F metin parça sayısına oranını göstermektedir.



Şekil 5.5. Babelfy anlam ayrımı graf örneği [113]

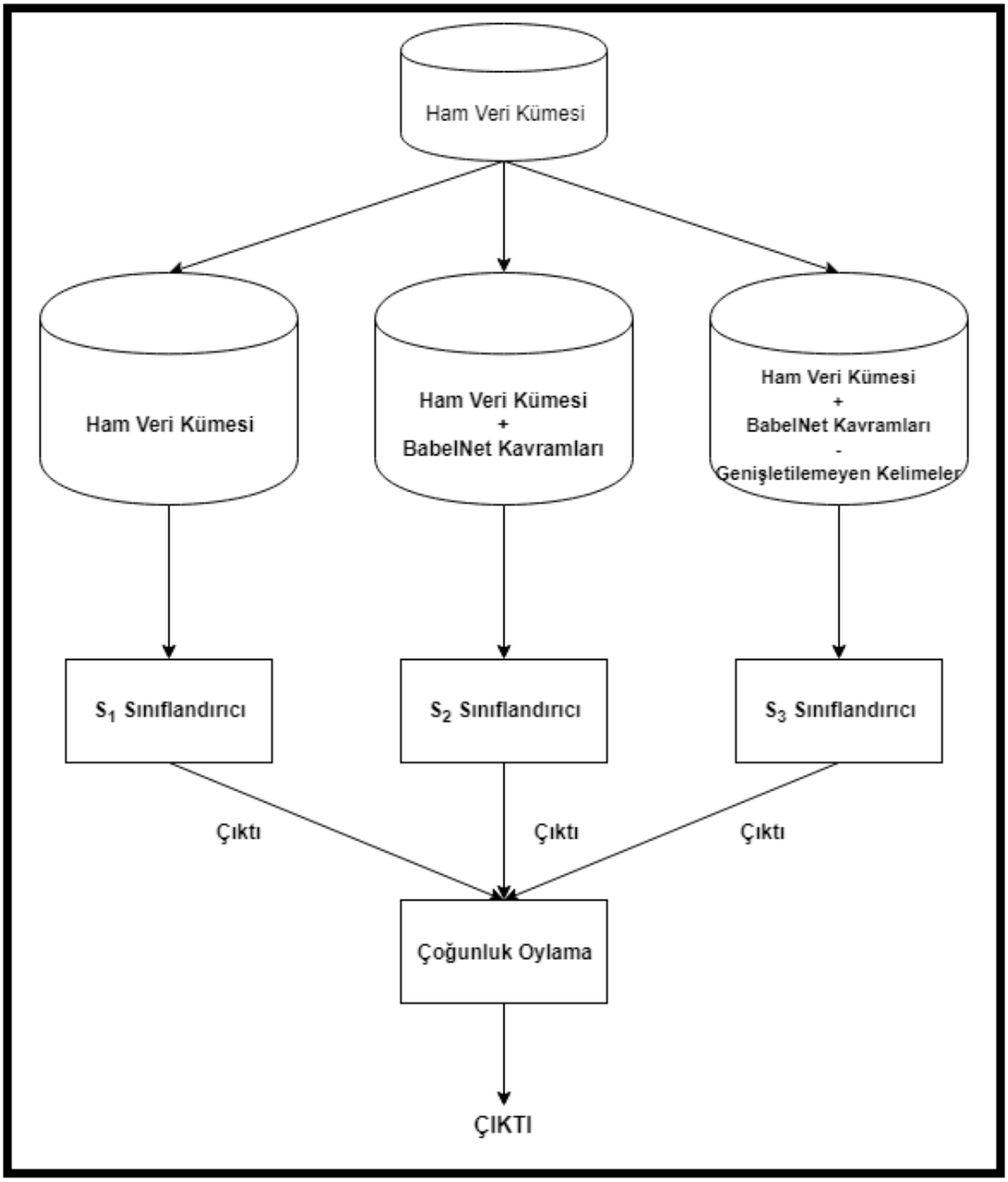
5.5.İşbirlikçi Öğrenme Modeli

İşbirlikçi öğrenme birden fazla temel sınıflandırma modelinin bir arada kullanılarak yüksek başarımlı bir sınıflandırma modelinin oluşturulmasıdır [1, 31]. Oluşturulan modelin başarımı, kendisini meydana getiren temel sınıflandırıcıların başarımından yüksektir. Sınıf etiketi bulunan veri kümeleri, temel sınıflandırıcılar ve sınıflandırıcıların sonuçlarını bir araya getirecek yöntemler işbirlikçi modelleri oluşturan temel parçalardır [32]. Başarımı yüksek bir işbirlikçi model doğruluk ve çeşitliliğe ihtiyaç duymaktadır [103]. Temel sınıflandırıcıların çeşitliliği kendi aralarındaki korelasyonu azaltarak işbirlikçi modelin kazancını arttırmaktadır.

İşbirlikçi modeller öğrenmenin gerçekleştirilme şekline ve temel sınıflandırıcıların sonuçlarının nasıl birleştirildiğine göre farklılık göstermektedir [104]. Sonuçların birleştirilmesi için yapılan işlemlerden biri temel sınıflandırıcılara atanan ağırlık değerlerine göre sonuçların birleştirilmesidir. Ağırlık değerleri sabit olabileceği gibi farklı yöntemlere göre güncellenebilir. Sonuçların birleştirilmesinde kullanılan bir diğer yöntem çoğunluk oylamasıdır [32, 105]. Çoğunluk oylamasında, temel sınıflandırıcıların sonuçlarının çoğunlukta olanı işbirlikçi modelin sonucu olarak kabul edilmektedir. Çoğunluk oylama birleştirme kuralının sabit ve basit olması nedeniyle temel sınıflandırıcıların birleştirilmesinde yaygın olarak kullanılan yöntemlerden biridir. Sonuçların birleştirilmesinde kullanılan bir diğer yöntem üst sınıflandırıcılardır. Bu yöntemde temel sınıflandırıcılar üzerinde oluşturulan bir üst sınıflandırıcı temel sınıflandırıcılardan elde ettiği sonuçlara göre değerler üretmekte, üretilen değerler işbirlikçi modelin sonucunu oluşturmaktadır [31]. Eğitim verilerini öğrenme şekilleri, veri kümesinin ya da özellik kümesinin belirli kurallar ile bölünmesine göre farklılık göstermektedir. Bagging ve Boosting modelleri veri kümesinin bölünmesiyle oluşturulurken, Random Subspace özellik kümesinin bölünmesiyle oluşturulur [106].

Tez çalışmasında gerçekleştirilen işbirlikçi model homojen temel sınıflandırıcılardan oluşurken, modele girdi olarak verilen veri setlerinin farklı temsilleri sınıflandırıcıların çeşitliliği sağlanmaktadır. Aynı öğrenme algoritmasından türetilen temel sınıflandırıcının her biri için farklı bir veri kümesi temsi eğitim verisi olarak kullanılmıştır. Eğitim verisi olarak kullanılan temsillerden ilki ham veri kümesinden

oluşturulurken, ikincisi ise ham veri kümesinin BabelNet kavramları ile genişletilmiş halinden oluşturulmuştur. Üçüncü temel sınıflandırıcıda kullanılacak eğitim veri kümesi ise genişletilen veri kümesi içerisinden, BabelNet ile genişletilemeyen terimlerin çıkartılmasıyla oluşturulmuştur. Üç temel öğreniciden elde edilen değerler çoğunluk oylamasına göre birleştirilmiştir.



Şekil 5.6. Türetilen işbirlikçi model mimarisi

5.6.Deneysel Sonuçlar

Kısa metinler ile yapılan çalışmada karşılaşılan en büyük sorun boyutun lanetidir. Özellik uzayının çok büyük olması verilerin birbirine göre daha ayrık olmasına sebep olmaktadır. Ayrıca verilerin temsili sırasında büyük miktarda özelliğe değer atanamamaktadır. Çalışmada verilerin arasındaki ilişkinin daha güçlü hale gelmesi için özellik uzayı genişletilmiştir. Genişletilme sonucunda özellik kümelerinin boyutu iki katından daha fazla artmıştır. Ancak veri kümelerinin farklı şekillerde genişletilmesi model kurulumundaki çeşitlilik açısından önemli yönde etki etmiştir.

Tablo 5.4. Genişletilmiş Uzay

Özellik Uzayı			
	Ham Veri	Bütün Terimler + BabelNet Kavramları	Genişletilen Terimler + BabelNet Kavramları
Iphone 6	1085	2616	2414
Hobbit	954	2182	1992
UMICH	1650	3615	3387
Archeage	1963	4120	3848
Ststest	1027	2516	2372
StsGold	2522	5346	4975

Tablo 5.4'teki değerler incelendiğinde her veri kümesinin özellik uzayının iki katından daha fazla arttığı gözlemlenmektedir. Tablo 5.1' deki veri kümelerinin eleman sayıları dikkate alındığında, UMICH veri kümesinin özellik uzayı diğer kümele oranla daha düşüktür. Bu durumun sebebinin UMICH veri kümesinin gürültüsünün azlığı ile ilişkili olduğu düşünülmektedir. Twitter verilerin içerdiği gürültü ve mesajların birbirinden ayrık olması, doküman uzayının doğru şekilde genişletilmesine engel olmuştur. Özellikle terimlerin yanlış yazımları Babelfy API'nin çalışmasını olumsuz yönde etkilemiş, terimlerin anlam ayrımı yeterince iyi yapılamamıştır. Bu durum genişletilen uzayın özelliklerinin artmasına sebep olmuştur. Belirtilen sebeplerden yola çıkıldığında Twitter mesajlarındaki yazım yanlışlarını düzeltmek için güçlü bir kaynağa veya yönteme ihtiyacın olduğu rahatlıkla söylenebilir.

Tez çalışmasında veri kümelerinin oluşturulması ve mimarinin belirlenmesinden sonraki son adım verilerin modeller üzerinde işletilmesidir. Oluşturulan üç farklı işbirlikçi modeller için temel sınıflandırıcılar sırasıyla NB, SMO ve C4.5 olarak seçilmiştir. Veri kümelerinin %20 si test için ayrılırken, %80'i modellerin eğitiminde kullanılmıştır. Modellerin başarımın ölçülmesinde F1-Score yöntemi kullanılmıştır.

Tablo 5.5. Iphone veri kümesi sınıflandırma sonuçları

İphone	NB	SVM	C4.5
Ham Veri	0.772	0.774	0.689
Tüm Terimler + Kavramlar	0.742	0.725	0.696
Genişletilen Terim + Kavramlar	0.759	0.733	0.685
İşbirlikçi Model	0.795	0.813	**0.839**

Tablo 5.5'te Iphone hakkında yorumları içeren veri kümesi ile oluşturulan sınıflandırıcıların sonuçları görülmektedir. NB ve SVM sınıflandırıcılarının genişletilmiş veri kümesi üzerindeki başarımları azalırken, C4.5 sınıflandırıcısının başarımı sadece genişletilen kelimelerin ve kavramların bulunduğu veri kümesinde azalmıştır. Tüm işbirlikçi modellerin başarımı temel sınıflandırıcılardan yüksek olmakla beraber, en yüksek başarıma sahip işbirlikçi model C4.5 algoritmasıyla kurulmuştur. Sınıflandırma sonucundaki başarım yaklaşık olarak yüzde 15 artmıştır. Bu artış herhangi bir sınıflandırma modeli için ciddi bir başarıyı ifade etmektedir. Bu yüksek artışın sebebinin veri kümesine ya da işbirlikçi modele bağlı olup olmadığı eldeki veriler ile belirlenememektedir.

Tablo 5.6. Hobbit veri kümesi sınıflandırma sonuçları

Hobbit	NB	SVM	C4.5
Ham Veri	0.724	0.832	0.811
Ham Veri + Kavramlar	0.549	0.834	0.830
Genişletilen Terim + Kavramlar	0.549	0.832	0.822
İşbirlikçi Model	0.809	0.898	**0,916**

Tablo 5.6'da Hobbit film yorumlarından oluşan veri kümesi ile ilgili analiz sonullarını göstermektedir. Sonuçları incelediğinde NB' in diğer sınıflandırıcılara göre daha zayıf kaldığı gözlemlenmektedir. Ancak işbirlikçi model sonucundan başarımı en çok yükselen NB temel sınıflandırıcılarıdır. İşbirlikçi modellerin tamamı başarımı arttırmıştır. Tüm iş birlikçi modeller arasında başarımı en yüksek olan C4.5 temel sınıflandırıcılarla kurulan modeldir. Özellik kümesinin genişletilmesi C4.5 sınıflandırıcısı dışındaki sınıflandırıcılarda başarıma etki etmemiştir. Hobbit veri kümesindeki elde edilen başarım, Iphone veri kümesine kıyasla biraz daha az olmasına rağmen ciddi bir artış göstermiştir.

Tablo 5.7'de sonuçları gösterilen UMICH veri kümesi, diğer veri kümeleri içerisinde en az gürültülü ve en fazla örnek sayısına sahip olanıdır. Film yorumları oluşturulan bu veri kümesini sınıflandırma kümeleri üzerindeki sonuçları aşırı yüksektir. Başarımların arasındaki farklar dikkate değer boyutlarda değildir. En yüksek sınıflandırma başarımları SVM ve C4.5 temel sınıflandırıcılarından elde edilmiştir. Tablo 5.7'de görüldüğü gibi temel sınıflandırıcılar bireysel olarak güçlü modeller oluşturmaktadır.

Tablo 5.7. UMICH veri kümesi sınıflandırma sonuçları

UMICH	NB	SVM	C4.5
Ham Veri	0.959	**0.991**	**0.991**
Ham Veri + Kavramlar	0.960	0.986	0.987
Genişletilen Terim + Kavramlar	0.957	0.978	0.983
İşbirlikçi Model	0.977	0.989	0.990

Tablo 5.8. Archeage veri kümesi sınıflandırma sonuçları

Archeage	NB	SVM	C4.5
Ham Veri	0.751	0.825	0.773
Ham Veri + Kavramlar	0.749	0.814	0.775
Genişletilen Terim + Kavramlar	0.750	0.796	0.770
İşbirlikçi Model	0.824	**0.837**	0.821

Archeage oyun yorumları ile oluşturulan veri kümesinin sınıflandırma algoritmalarındaki başarımları Tablo 5.8'de görülmektedir. Temel sınıflandırıcılar arasında başarımı en yüksek olan SVM sınıflandırıcısıdır. Ayrıca işbirlikçi modeller arasında başarımlar birbirine yakın olmasına rağmen en yüksek başarıma sahip olan model SVM temel sınıflandırıcıları ile kurulandır. Genişletilen veri kümelerinin C4.5 sınıflandırıcıları dışında başarıma büyük bir etkisi olmamıştır.

Tablo 5.9. Ststest veri kümesi sınıflandırma sonuçları

Ststest	NB	SVM	C4.5
Ham Veri	0.709	0.709	0.617
Ham Veri + Kavramlar	0.690	0.712	0.635
Genişletilen Terim + Kavramlar	0.655	0.689	0.608
İşbirlikçi Model	0.711	**0.717**	0.677

Mesajları arasında konu ilişkisi bulunmayan Ststest veri kümesi ile kurulan sınıflandırıcıların başarımları Tablo 5.9' da görülmektedir. Diğer kümeleri ile karşılaştırıldığında sınıflandırıcılarının başarımlarının daha düşük olduğu görülmektedir. C4.5 temel sınıflandırıcılarının başarımları NB ve SVM ile karşılaştırıldığında en düşük değere sahip olandır. Genişletilmiş veri kümesi ile oluşturulan SVM ve C4.5 sınıflandırıcılarının başarımlarının arttığı görülmektedir. İşbirlikçi modeller içerisinde en yüksek başarıma SVM sınıflandırıcıları ile kurulan model sahiptir.

Tablo 5.10'da görüldüğü gibi StsGold veri kümesi ile kurulan sınıflandırıcıların başarımları Ststest'ten yüksek olmasına rağmen diğer veri kümelerine göre düşük kalmıştır. Genişletilmiş veri kümelerinin başarıma olumlu yönde etkisi gözlemlenmemektedir. İşbirlikçi modeller içerisinde başarımı en yükse olan SVM algoritmasıyla kurulan modeldir.

Tablo 5.10. StsGold veri kümesi sınıflandırma sonuçları

StsGold	NB	SVM	C4.5
Ham Veri	0.710	0.785	0.714
Ham Veri + Kavramlar	0.698	0.754	0.702
Genişletilen Terim + Kavramlar	0.695	0.709	0.695
İşbirlikçi Model	0.764	**0.791**	0.773

Tablo 5.11 tüm temel sınıflandırıcıların ve işbirlikçi modellerin en yüksek başarıma sahip olanlarının sonuçlarına göre oluşturulmuştur. Koyu renkle yazılan değerler, her veri kümesi için en yüksek başarıma sahip modelleri göstermektedir. UMICH dışındaki tüm veri kümelerinde işbirlikçi modelin başarımı temel sınıflandırıcıların başarımını geçmiştir. UMICH veri kümesinin temel sınıflandırıcıları hali hazırda çok yüksek değerler sonuçlar verdiği için işbirlikçi modelin zayıf veya güçlü olmasıyla ilgili kesin bir yargıda bulunulamamaktadır.

Tablo 5.11. Sınıflandırma sonuçları

	Sınıflandırma Modelleri			
	NB	SVM	C4.5	İşbirlikçi
Iphone6	0,772	0,774	0,696	**0,839**
Hobbit	0,824	0,834	0,830	**0,916**
UMICH	0,960	**0,991**	**0,991**	0,990
Archeage	0,751	0,825	0,775	**0,837**
Ststest	0,690	0.709	0,617	**0,717**
StsGold	0,710	0,785	0,714	**0,791**

Veri kümelerine göre başarımlar karşılaştırıldığında StsGold ve Ststest veri kümelerimin başarımlarının daha düşük olduğu gözlemlenmektedir. StsGold ve Ststest konulardan bağımsız olarak oluşturulan veri kümeleridir. Bu nedenle mesajların içerdiği duygu temelli kelimelerin herhangi bir konudan bağımsız olması

sınıflandırmadaki başarımı olumsuz yönde etkilemektedir. Bir kelime kullanıldığı içeriğe göre farklı anlamlar kazanabilmektedir. Örneğin "uzun" kelimesi "Cevap vermeleri uzun sürdü" ifadesinde olumsuz anlam taşırken, "uzun batarya ömrü" ifadesinde olumlu olarak kullanılmaktadır. Bu nedenle Tablo 5.2'de de görülmektedir ki duygu tabanlı kelimelerin konudan bağımsız kullanılması sınıflandırmanın başarımını düşürmektedir.

Tez çalışmasında sınıflandırıcı modellerin eğitimi ve değerlendirilmesi dört çekirdekli 2.50 GHz hızındaki işlemci ve 8 GB ram bileşenlerine sahip olan bilgisayar üzerinde gerçekleştirilmiştir. Tablo 5.12' de modellerin kurulması için geçen süreler milisaniye cinsinden gösterilmiştir.

Tablo 5.12. Modellerin kurulum süresi (milisaniye)

		NB	**SVM**	**C4.5**
Iphone 6	Ham	8529	5780	6549
	Ham + Kavramlar	11428	3088	9365
	Genişletilen + Kav.	10953	3035	9451
Hobbit	Ham	8624	5559	4179
	Ham + Kavramlar	13887	2796	7082
	Genişletilen + Kav.	12476	2990	4865
UMICH	Ham	130089	90450	298895
	Ham + Kavramlar	329946	223045	522847
	Genişletilen + Kav.	255222	248647	568285
Archeage	Ham	24286	31223	64088
	Ham + Kavramlar	48660	51534	140382
	Genişletilen + Kav.	44999	49655	149301
Ststest	Ham	4504	5125	5224
	Ham + Kavramlar	5938	1828	7675
	Genişletilen + Kav.	5453	1672	6704
StsGold	Ham	24492	22903	47407
	Ham + Kavramlar	49802	33882	113687
	Genişletilen + Kav.	45354	30839	112402

Sonuçlar göz önüne alındığında işbirlikçi modellerin başarımının tek başına kullanılan sınıflandırıcılara göre yüksek olduğu görülür. Ancak veri kümesinin ve modellerin seçilmesi, modellerin eğitilmesi daha fazla zaman almaktadır. Zamanın önemli olmadığı durumlarda işbirlikçi modellerin kullanılması tavsiye edilebilir.

Ağırlıklı oylama işlem süresi tüm modellerde 5-10 milisaniye arasında değiştiği için Tablo 5.12'de gösterilmemiştir.

6. SONUÇLAR VE ÖNERİLER

Sosyal medya uygulamaları hayatın birçok alanında kendisine yer edinmiş iletişim araçlarıdır. İnsanların düşüncelerini kolay ve hızlı bir şekilde geniş kitlelere yayabilmelerini sağlayan bu araçlar üzerinde günümüzde birçok farklı işlem gerçekleştirilebilmektedir. Bir ürünün pazarlanması ya da afet anında insanların bilgilendirilmesi bu işlemlerin arasında sayılabilir. Taşıdığı potansiyel nedeniyle sosyal medya uygulamaları birçok bilim dalının araştırma konusu haline gelmiştir. Ancak bu uygulamaların sağladığı veriler üzerinden kullanışlı bilgilerin elde edilmesi çalışmalarda karşılaşılan en büyük zorluklardan birisidir. Sosyal medya uygulamalarının karakteristliğinden kaynaklı olarak paylaşılan mesajlar arasındaki ilişiklerin belirlenmesi oldukça zorlaşır.

Bu tez çalışmasında popüler bir sosyal medya uygulaması olan Twitter'dan elde edilen mesajlar üzerinde analiz işlemi gerçekleştirilmiştir. Mesajlarında elde edilecek bilginin olabildiğince arttırılması için mesaj içerikleri zenginleştirilmiştir. Zenginleştirme işlemi gerçekleştirilirken BabelNet isimli ansiklopedik sözlük kullanılmış, mesajların içerdiği kelimelere anlamca en yakın olanlar mesaj içerisine yerleştirilmiştir. Ayrca analiz işlemini en verimli şekilde gerçekleştirebilmek için çalışma içerisinde işbirlikçi öğrenme modeli önerilmiştir. Bu model Bagging öğrenme modelinin bir türevi olup, temel sınıflandırıcıları oluşturmak için kullanılan veri kümeleri BabelNet yardımıyla türetilmiştir. Öğrenme modeli üç popüler sınıflandırma modeli olan NB, SMO ve C4.5 ayrı ayrı kurulmuş ve bu modellerin sonuçları çoğunluk oylaması ile birleştirilmiştir.

Yapılan çalışmada karşılaşılan en büyük zorluk mesajların gürültüden temizlenmesi aşamasıdır. Bu aşamanın sonucu doğrudan genişletme aşamasını da etkileyeceği için dikkatle yapılmaya çalışılmış ancak eldeki yardımcı kaynaklar ile yeterli verim sağlanamamıştır. Metin içerisinden temizleyen kısaltmalar ve yanlış yazılan kelimeler Babelfy tarafından adlandırılmış varlık olarak algılanmış, bu doğrultuda mesaj içeriklerine ilgisiz kavramlar eklenmiştir.

(ha - "Cyrillic script" "letter" sun "planets" "heat" "solar system" "light" "star")

(ps - "inorganic phosphate" "allotropic" "nitrogen family" "element" "nitrogen" "commonly" "living cells" "phosphate rocks" "phosphates" "nonmetallic" "phosphate" "rocks" "family" "cells" try "test" see "Perceive" "power" "perceive" town "urban area" "city" "smaller" broadcast internationally brazilian "characteristic" "Brazil" "Brazil" great "Relatively" fan "current of air" "air" "creating" "device" "current" "surface" "surfaces")

Bu sebeple sosyal medya verileri ile çalışmak isteyenlere, verilerin içerdiği gürültünün giderilmesi için kullanılmak üzere güçlü bir kaynak ya da yöntem bulunması önerilmektedir.

Çalışmada BabelNet ile gerçekleştirin içerik zenginleştirme adımı sonucunda veri kümelerin boyutları büyük miktarda artmıştır. Bu sorunun nedenlerinden biri daha önce bahsedilen önişleme adımı eksiklikler gösterilebilir. Özellik boyutunun genişletilmesi her ne kadar işlem yükünü arttırdıysa da temel sınıflandırıcıların farklı bilgileri yakalayabilmesini sağlamış, bireysel olarak başarılı düşük olsa bile sonuçları birleştirildiğin işbirlikçi modelin başarımı artmıştır. Verikümelerine göre işbirlikçi modellerin başarımı farklılık göstermesine rağmen temel sınıflandırıcılar göz önüne alındığında başarımların daha yüksek olduğu görülmektedir. Analiz için kullanılan bazı veri kümelerinde başarımın yüzden 10'dan daha fazla arttığı görülmüştür. Bu nedenle özellikle kısa metinlerin özelliklerin genişletilmesinde BabelNet'in başarılı bir kaynak olduğu çalışmada açıkça ortaya konulmaktadır. Ayrıca anlam ayrımında kullanılan Babelfy'ın metinlerin içeriğini dikkate alması özellik genişletmenin başarımını arttırmış, veri kümesi içerisindeki farklı ilişkilerin göz ardı edilmemesini sağlamıştır. İleriki çalışmalarında özellik genişletmenin ne tür katkıları olduğu incelenebilir. Özellikle kurulan işbirlikçi modele farklı temel sınıflandırıcılar ekleyerek başarımın ne oranda etkilendiği gözlemlenebilir.

KAYNAKLAR

[1] Wang G., Sun J., Ma J., Xu K., Gu ., Sentiment Classification: The Contribution of ensemble Learning, *Decision Sport System*, 2014, **ISSN 0167-9236** (57), 77-93.

[2] Ravi K., Ravi V., A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications, *Knowledge-Based Systems*, 2015, **ISSN 0950-7051**(89), 14-46.

[3] https://www.channel4.com/news/topic/cambridge-analytica, (Ziyaret Tarihi : 17 Haziran 2019).

[4] Yoo S., Song J., Jeong O., Social Media Contents Based Sentiment Analysis and Prediction System, *Expert Systems with Applications*, 2018, **ISSN 0957-4174**(105), 102-111

[5] Crannell W. C., Clark E., Jones C. James T. A., Moore J., A Pattern-matched Twitter Analysis of US Cancer-patient Sentiments, *Journal of Surgical Research*, 2016, **ISSN 0022-4804**(206), Issue 2, 536-542.

[6] Zhang L., Hall M., Bastola D., Utilizing Twitter Data for Analysis of Chemotherapy, *International Journal of Medical Informatic*, 2018, **ISSN 1386-5056**(120), 92-100.

[7] Goldberg Y., *Modeling with Recurrent Networks*, Hirst G., *Neural Network Methods for Natural Language Processing*, 1st edition, Morgan and Claypool Publishers, San Rafael, 185-193, 2017.

[8] Hussein D. M., A Survey on Sentiment Analysis Challenges, *Journal of King Saud University - Engineering Sciences*, 2018, **ISSN 1018-3639**(30), Issue 4, 330-338.

[9] Daniel M., Neves R. F., Horta N., Company Event Popularity for Financial Markets Using Twitter and Sentiment Analysis, *Expert Systems with Applications*, 2017, **ISSN 0957-4174**(71), 111-124.

[10] Z. Doshi, S. Nadkarni, K. Ajmera and N. Shah, TweerAnalyzer: Twitter Trend Detection and Visualization, 2017, *International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 1-6.

[11] Keib K., Himelboim I., Han J., Important Tweets matter: Predicting Retweets in the #BlackLivesMatter talk on Twitter, *Computers in Human Behavior*, 2018, **ISSN 0747-5632**(85), 106-115.

[12] Singh T., Kumari M., Role of Text Pre-processing in Twitter Sentiment Analysis, *Procedia Computer Science*, 2016, **ISSN 1877-0509**(89), 549-554.

[13] Prassad A. G., Sanjana S., Bhat S. M., Harish B. S., Sentiment Analysis for Sarcasm Detection on Streaming Short Text Data, *Conference: 2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)*, 2017.

[14] Smailovic J., Grcar M., Lavrac N., Znidarsic M., Stream-based Active Learning for Sentiment Analysis in the Financial Domain, *Information Science*, 2014, **ISSN 0020-0255**(285), 181-203.

[15] I. Taksa, S. Zelikovitz and A. Spink, Using Web Search Logs to Identify Query Classification Terms, *Fourth International Conference on Information Technology (ITNG'07)*, Las Vegas, NV, 2007.

[16] Z. Faguo, Z. Fan, Y. Bingru and Y. Xingang, Research on Short Text Classification Algorithm Based on Statistics and Rules, *2010 Third International Symposium on Electronic Commerce and Security*, Guangzhou, 2010.

[17] Jain A. P., Katkar V., Sentiments Analysis of Twitter Data Using Data Mining, *2015 International Conference on Information Processing (ICIP)*, December 2015.

[18] Kusen E., Strembeck M., Politics, Sentiments, and Misinformation: An Analysis of the Twitter Discussion on the 2016 Austrian Presidential Elections, *Online Social Networks and Media*, 2018, **ISSN 2468-6964**(5), 37-50.

[19] Thelwall M., Buckley K., Paltoglou G., Cai D., Kappas A., Sentiment Strength Detection in Short Informal Text, *American Society for Information Science and Technology*, December 2010, (61), 2544-2558.

[20] Singh A. K., Gupta D., Singh R. M., Sentiment Analysis of Twitter User Data on Punjab Legislative Assembly Election, 2017, *International Journal of Modern Education and Computer Science*, 2017, 60-68.

[21] Mukherjee I., Sahana S. K., Mahanti P. K., An Improved Information Retrieval Approach to Short Text Classification, *International Journal of Information Engineering and Electronic Business*, July 2017, 31-37.

[22] Öztürk N., Ayvaz S., Sentiment Analysis on Twitter: A Text Mining Approach to the Syrian Refugee, *Telematics and Informatics*, 2018, **ISSN 0736-5853**(35), Issue 1, 136-147.

[23] Ghiassi M., Skinner J., Zimbra D., Twitter Brand Sentiment Analysis: A Hybrid System Using n-gram Analysis and Dynamic Artificial Neural Network, *Expert Systems with Applications*,2013, **ISSN 0957-4174**(40), Issue 16, 6266-6282.

[24] Saif H., He Y., Fernandez M., Alani H., Contextual Semantics for Sentiment Analysis of Twitter, *Information Processing & Management*, 2016, **ISSN 0306-4573**(52), Issue 1, 5-19.

[25] Pedro B. F., Thiago P., NILC_USP: A Hybrid System for Sentiment Analysis in Twitter Messages, *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, Atlanta, Georgia, USA, Jun 2013.

[26] Yan Y., Yang H., Wang H., Two simple and effective ensemble classifiers for Twitter sentiment analysis, *2017 Computing Conference*, London, 2017.

[27] https://www.clips.uantwerpen.be/pattern , (Ziyaret Tarihi: 17 Haziran 2019).

[28] Xia R., Zong C., Li S., Ensemble of Feature Sets and Classification Algorithms for Sentiment Classification, *Information Sciences*, 2011, **ISSN 0020-0255**(181), Issue 6, 1138-1152.

[29] Ankit, Saleena N., An Ensemble Classification System for Twitter Sentiment Analysis, Procedia Computer Science, 2018, ISSN 1877-0509(132), 937-946.

[30] Fersini E., Messina E., Pozzi F. A., Sentiment analysis: Bayesian Ensemble Learning, *Decision Support Systems*, 2014, **ISSN 0167-9236**(68), 26-38.

[31] Troussas C., Krouska A., Virvou M., Evaluation of ensemble-based sentiment classifiers for Twitter data, *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Chalkidiki, 2016.

[32] L. Rokach, Ensemble-based classifiers, *Artificial Intelligence Review*, 2010, (33), 1-39.

[33] Cotelo J. M., Cruz F. L., Enriquez F., Troyano J. A., Tweet Categorization by Combining Content and Structural Knowledge, *Information Fusion*, 2016, **ISSN 1566-2535**(31), 54-64.

[34] Corrêa E. A. Jr., Marinho V. Q., Santos L. B., NILC-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis, *Proceedings of the 11th International Workshop on Semantic Evaluation*, Vancouver, Canada, 2017.

[35] Mikolov T., Chen K., Corrado G. S., Dean J., Efficient Estimation of Word Representations in Vector Space, *CoRR*, 2013, **abs/1301.3781**.

[36] Mikolov T., Chen K., Corrado G. S., Dean J., Distributed Representations of Words and Phrases and Their Compositionality, *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada, 2013.

[37] Çoban Ö., Ozyer G. T., Word2vec and Clustering based Twitter Sentiment Analysis, *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, September 2018.

[38] A. Hayran and M. Sert, Sentiment Analysis on Microblog Data Based on Word Embedding and Fusion Techniques, *25th Signal Processing and Communications Applications Conference (SIU)*, Antalya, 2017.

[39] Rezaeinia S. M., Rahmani R., Ghodsi A., Veisi H., Sentiment Analysis Based on Improved Pre-trained Word Embeddings, *Expert Systems with Application*, 2019, **ISSN 0957-4174**(117), 139-147.

[40] Xion S., Lv H., Zhao W., Ji D., Towards Twitter Sentiment Classification by Multi-level Sentiment-enriched Word Embeddings, *Neurocomputing*, **ISSN 0925-2312**(275), 2459-2466.

[41] Sebastiani F., Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, 2001, (34), 1-37.

[42] Wu Z., Zhu H., Li G., Cui Z., Hui H., Li J., Chen E., Xu G., An Efficient Wikipedia Semantic Matching Approach to Text Document Classification, *Information Sciences*, 2017, **ISSN 0020-0255**(393), 15-28.

[43] Stein R. A., Jaues P. A., Valiati J. F., An Analysis of Hierarchical Text Classification Using Word Embeddings, *Information Sciences*, 2019, **ISSN 0020-0255**(471), 216-232.

[44] Kim D., Seo D., Cho S., Kang P., Multi-co-training for Document Classification Using Various Document Representations: TF–IDF, LDA, and Doc2Vec, *Information Sciences*, 2019, **ISSN 0020-0255**(477), 15-29.

[45] Chagheri S., Roussey C., Calabretto S., Dumoulin C., Technical documents classification, *13th International Symposium on the Management of Industrial and Corporate Knowledge*, Lausanne, Switzerland, Jun 2011.

[46] D'cunha A., SenA. K., Hierarchical Approach for Scientific Document Classification, *International Conference on Computing, Communication & Automation*, Noida, 2015.

[47] Sun X., Zhang Q., Wang Z., An Efficient Document Classification Algorithm Based on Kernel LDE, *2009 International Conference on Industrial Mechatronics and Automation*, Chengdu, 2009.

[48] Bird S., Klein E., Loper E., *Natural Language Processing with Python*, 1st edition, O'Reilly, USA, June 2009.

[49] Zhang L., Ghosh R., Dekhil M., Hsu M., Liu B., Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis, January 2011.

[50] Gull R., Shoaib U., Rasheed S., Abid W., Zahoor B., Pre Processing of Twitter's Data for Opinion Mining in Political Context, *Procedia Computer Science*, 2016, **ISSN 1877-0509**(96), 1560-1570.

[51] Desai M., Mehta M. A., Techniques for Sentiment Analysis of Twitter Data: A Comprehensive Survey, *2016 International Conference on Computing, Communication and Automation (ICCCa)*, Noida, 2016,

[52] Jianqiang Z., Xiaolin G., Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis, *IEEE Access*, 2017, (5), 2870-2879.

[53] Celikyilmaz A., Hakkani-Tür D., Feng J., Probabilistic Model-based Sentiment Analysis of Twitter Messages, *2010 IEEE Spoken Language Technology Workshop*, Berkeley, CA, 2010.

[54] Wagh R., Punde P., Survey on Sentiment Analysis using Twitter Dataset, *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*¸ March 2018.

[55] Symeonidis S., Effrosynidis D., Arampatzis A., A Comparative Evaluation of Pre-processing Techniques and Their Interactions For Twitter Sentiment Analysis, *Expert Systems with Applications*, **2018, ISSN 0957-4174**(110), 298-310.

[56] Tomovic A., Janicic P., Keselj V., n-Gram-based Classification and Unsupervised Hierarchical Clustering of Genome Sequences, *Computer Methods and Programs in Biomedicine*, 2006, **ISSN 0169-2607**(81), 137-153.

[57] Huang H. H., Yu C., Clustering DNA Sequences Using the Out-of-place Measure with Reduced n-grams, *Journal of Theoretical Biology*, 2016, **ISSN 0022-5193**(406), 51-72.

[58] Sun S., Luo C., Chen J., A Review of Natural Language Processing Techniques for Opinion Mining Systems, *Information Fusion*, 2017, **ISSN 1566-253**5 (36), 10-25.

[59] Das B. R., Sahoo S., Panda C. S., Patnaik S., Part of Speech Tagging in Odia Using Support Vector Machine, *Procedia Computer Science*, 2015, **ISSN 1877-0509**(48), 507-512.

[60] Ingersoll G. S., Morton T. S., Farris A. L., *Taming Text: How to Find, Organize, and Manipulate It*, 1st edition, Manning, December 2012.

[61] Porter M. F., An algorithm for suffix stripping, *Program*, 1980, (14), 130-137.

[62] Porter M. F., Snowball: A Language for Stemming Algorithms, January 2001.

[63] Paice C. D., Another Stemmer, *ACM SIGIR*, 1990, (24), 56-61.

[64] Paice C.D. An Evaluation Method for Stemming Algorithms, *In: Croft B.W., van Rijsbergen C.J. (eds) SIGIR '94*, 1994.

[65] Fellbaum C., A Semantic Network of English: The Mother of All WordNets, *Computers and the Humanities*, 1998, (32), 209-220.

[66] Salton G., Yang C.S., On the Specification of Term Values in Automatic Indexing, *Journal of Documentation*, 1973, (29), 351-372.

[67] Trstenjak B., Mikac S., Donko D., KNN with TF-IDF based Framework for Text Categorization, *Procedia Engineering*, 2014, **ISSN 1877-7058**(69), 1356-1364.

[68] Omurca S. İ., Baş S., Ekinci E., An Efficient Document Categorization Approach for Turkish Based Texts , *International Journal Of Intelligent Systems And Applications In Engineering*, 2015, (3), 2015.

[69] Kotsiantis S., Supervised Machine Learning: A Review of Classification Techniques, *Informatica*, 2007, (31), 249-268.

[70] Omurca S. İ., Ekinci E., An alternative evaluation of post traumatic stress disorder with machine learning methods, *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, Madrid, 2015.

[71] Nielsen J. D., Rumi R., Salmeron A., Supervised Classification Using Probabilistic Decision Graphs, *Computational Statistics & Data Analysis*, 2009, **ISSN 0167-9473**(53), 1299-1311.

[72] Deisy C., Baskar S., Ramraj N., Koori J. S., Jeevanandam P., A Novel Information Theoretic-interact Algorithm (IT-IN) for Feature Selection Using Three Machine Learning Algorithms, *Expert Systems with Applications*, 2010, **ISSN 0957-4174**(37), 7589-7597.

[73] Lee L. H., Isa D., Automatically Computed Document Dependent Weighting Factor Facility for Naïve Bayes Classification,

[74] Parveen H., Pandey S., Sentiment analysis on Twitter Data-set using Naive Bayes algorithm, *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, Bangalore, 2016.

[75] Saleh A., Menai M. E. B., Naïve Bayes Classifiers for Authorship Attribution of Arabic Texts, December 2014.

[76] Farid D., Md., Zhang L., Rahman C. M., Hossain M. A., Strachan R., Hybrid Decision Tree and Naïve Bayes Classifiers for Multi-class Classification Tasks, *Expert Systems with Applications*, 2014, **ISSN 0957-4174**(41), 1937-1946.

[77] Geron A., *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, 1st edition, O'Reilly, USA, March 2017.

[78] Şimşek N. Ö. Ö., Gurgen F., Fuzzy Support Vector Machines for ECG Arrhythmia Detection, *Pattern Recognition (ICPR)*, September 2010.

[79] Melgani F., Bruzzone L., Classification of Hyperspectral Remote Sensing Images with Support Vector Machines, *IEEE Transactions on Geoscience and Remote Sensing*, 2004, (42), 1778-1790.

[80] Platt J. C., Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, July 1998.

[81] Gamon M., Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis, *Proceedings of the 20th International Conference on Computational Linguistics*, 2004.

[82] Witten I. H., Frank E., *Data Mining Practical Machine Learning Tools and Techniques*, 2nd edition, Elsevier, San Francisco, June 2005.

[83] Ekinci E., Takçı H., Elektronik Postaların Adli Analizinde Yazar Analizi Tekniklerinin Kullanılması, Yüksek Lisans Tezi, Gebze Yüksek Teknoloji Mühendislik ve Fen Bilimleri Enstitüsü, 2012, 334457

[84] Mitchell R. S., Sherlock R. A., Smith L. A., An Investigation Into the Use of Machine Learning for Determining Oestrus in Cows, *Computers and Electronics in Agriculture*, 1996, **ISSN 0168-1699**(15), 195-213.

[85] Witten I. H., Frank E., Hall M. A., Pal C., *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques*, 4th edition, Morgan Kaufmann, 2016.

[86] Zhang L., Ghosh R., Dekhil M., Hsu M., Liu B., Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis, January 2011.

[87] Njølstad P. C. S., Høysæter L. S., Wei W., Gulla J. A., Evaluating Feature Sets and Classifiers for Sentiment Analysis of Financial News , *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Warsaw, 2014.

[88] Davis J. J., Foo E., Automated Feature Engineering for HTTP Tunnel Detection, *Computers & Security*, 2016, **ISSN 0167-4048**(59), 166-185.

[89] Zheng A., Casari A., *Feature Engineering for Machine Learning*, 1st edition, O'Reilly, USA, April 2018.

[90] Jalal A. A., Feature Engineering in Hybrid Recommender Systems, *The Third International Conference on Data Mining, Internet Computing, and Big Data*, Konya, Turkey 2016.

[91] Song G., Ye Y., Du X., Huang X., Bie S., Short Text Classification: A Survey, *Journal of multimedia*, May 2014.

[92] Yang L., Li C., Ding Q., Li L., Combining Lexical and Semantic Features for Short Text, *Procedia Computer Science*, 2013, **ISSN 1877-0509**(22), 78-86.

[93] Man Y. Feature Extension for Short Text Categorization Using Frequent Term Sets, *Procedia Computer Science*¸ 2014, **ISSN 1877-0509**(31), 663-670.

[94] Pennington J., Socher R., Manning C. D., GloVe: Global Vectors for Word Representation, 2014.

[95] Amer N. O., Mulhme P., Gery M., Abdulahhad K.., Sign in Word Embedding for Social Book Suggestion, *CLEF*, 2016.

[96] Harris Z. S. Distributional Structure, *WORD*, 1954, (10), 146-162.

[97] Rohde D. L. T., Gonnerman L. M., Plaut D. C. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence, November 7, 2005

[98] Navigli R., Ponzetto S. P., BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network, *Artificial Intelligence*, 2012, **ISSN 0004-3702**(193), 217-250.

[99] Kilgarriff A., Fellbaum C., WordNet: An Electronic Lexical Database, *Language*, September 2000.

[100] Manning C. D., Surdeanu M., Bauer J., Finkel J., Bethard J. S., McClosky D., The Stanford CoreNLP Natural Language Processing Toolkit, *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014.

[101] https://www.languagetool.org/ (Ziyaret Tarihi: 17 Haziran 2019).

[102] http://www.hlt.utdallas.edu/~yangl/data/Text_Norm_Data_Release_Fei_Liu/ (Ziyaret Tarihi: 17 Haziran 2019).

[103] Windeatt T., Ardeshir G., Decision Tree Simplification For Classifier Ensembles, *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 2004.

[104] Tümer K.., Ghosh J., Error Correlation and Error Reduction in Ensemble Classifiers, *Connection Science*, 1996, (8), 385-404.

[105] Amasyali M. F., Ersoy O., Performance based pruning and weighted voting with classification ensembles, *2011 IEEE 19th Signal Processing and Communications Applications Conference (SIU)*, Antalya, 2011.

[106] Zhou Z. H., *Ensemble Methods: Foundations and Algorithms*, Chapman & Hall, 2012.

[107] Wang Z., Wang Y., Srinivasan R. S., A Novel Ensemble Learning Approach to Support Building Energy Use Prediction, *Energy and Buildings*, 2018, **ISSN 0378-7788**(159), 109-122.

[108] Li W., Ding P., Zhang X., Duan C., Qiu R., Lin J., Shi X., Ensemble Learning Methodologies to Improve Core Power Distribution Abnormal Detectability, *Nuclear Engineering and Design*, **ISSN 0029-5493**(2019), 160-166.

[109] Polikar R., Ensemble based systems in decision making, *IEEE Circuits and Systems Magazine*, Third Quarter 2006 , (6), 21-45.

[110] Breiman L., Bagging Predictors, *Machine Learning*, August 1996, (24), 123-140.

[111] Schapire R. E., The Strength of Weak Learnabilirty, *Machine Learning*, 1990, 1997-227.

[112] Freund Y., Schapire R. E., A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, *Journal of Computer and System Sciences,* 1997, **ISSN 0022-0000**(55), 119-139.

[113] Moro A., Raganato A. Navigli R., Entity Linking meets Word Sense Disambiguation: a Unified Approch, *Transactions of the Association for Computational Linguistics (TACL),* 2014, 231-244.

[114] Moro A., Cecconi R., Navigli R., Multilingual Word Sense Disambiguation and Entity Linking for Everybody, *Proc. of the 13th International Semantic Web Conference, Posters and Demonstrations (ISWC 2014)*, Riva del Garda, Italy, 19-23 October 2014.

KİŞİSEL YAYIN VE ESERLER

Sevim S., Omurca S. İ., Ekinci E., An Ensemble Model using a BabelNet Enriched Document Space for Twitter Sentiment Classification, *IJ. Information Technology and Computer Science*, DOI:10.5815/ijitcs.2018.01.03

ÖZGEÇMİŞ

Semih Sevim 1993'de Kocaeli'de doğdu. Lise öğreninmini İzmit Atılım Anadolu Lisesi'nde tamamladı. 2011 yılında Trakya Üniversitesi Bilgisyar Mühendisliği bölümünde okumaya başladıktan sonra 2012 yılında Kocaeli Üniversitesi'nin aynı lisans programına yatay geiş yaptı. 2015 yılında lisans hayatını sonlandırdıktan sonra 2016' Kocaeli Ünivertesi Bilgisayar Fen Bilimleri Enstitüsü Bilgisayar Mühendsiliği Anabilim dalında yüksek lisans yapmaya başladı.