METİNLERDE DUYGU ANALİZİ VE SINIFLANDIRMA İÇİN YENİ YÖNTEMLER

Muhammet Yasin PAK

Yüksek Lisans Tezi

Bilgisayar Mühendisliği Anabilim Dalı

Haziran, 2015

JÜRİ VE ENSTİTÜ ONAYI

Muhammet Yasin PAK'ın **"Metinlerde Duygu Analizi ve Sınıflandırma İçin Yeni Yöntemler"** başlıklı **Bilgisayar Mühendisliği** Anabilim Dalındaki, Yüksek Lisans Tezi 25.06.2015 tarihinde, aşağıdaki jüri tarafından Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca değerlendirilerek kabul edilmiştir.

		Adı Soyadı	**İmza**
Üye (Tez Danışmanı) :		**Doç. Dr. Serkan GÜNAL**	**……..**
Üye	**:**	**Doç. Dr. Hüseyin POLAT**	**……..**
Üye	**:**	**Yrd. Doç. Dr. Semih ERGİN**	**……..**

Anadolu Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun …….. tarih ve …….. sayılı kararıyla onaylanmıştır.

Enstitü Müdürü

ÖZET

Yüksek Lisans Tezi

METİNLERDE DUYGU ANALİZİ VE SINIFLANDIRMA İÇİN YENİ YÖNTEMLER

Muhammet Yasin PAK

Anadolu Üniversitesi

Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Serkan GÜNAL

2015, 53 sayfa

Duygu analizi, yazılı bir metin içerisinde, yazarın belirli bir konu hakkındaki duygu ve düşüncesinin analiz edilmesini, duygu sınıflandırma ise duygunun pozitif ve negatif olarak sınıflandırılmasını amaçlar. Duygu sınıflandırma için kullanılan gözetimli öğrenme metotları, sınıflandırıcının eğitimi için çok sayıda etiketli veriye ihtiyaç duymaktadır. Bununla birlikte, alan bağımlı bir problem olan duygu sınıflandırma probleminde başarılı bir sınıflandırma yapılabilmesi için her alana özgü eğitim verisinin toplanması gerekmektedir. Ancak bu işlem zaman alıcı ve maliyetli bir işlem olmaktadır. Alanı bilinmeyen bir metin sınıflandırılmak istendiğinde ise alana ait etiketli veya etiketsiz veri toplamak mümkün olmamaktadır. Problemin çözümü için bu tez çalışmasında, alan sınıflandırma temelli duygu sınıflandırma yaklaşımı önerilmiştir. İki aşamadan oluşan yaklaşımın ilk aşamasında, alanı bilinmeyen metnin, etiketli verisi olan mevcut alanlardan hangisine dâhil edilebileceği bulunmakta, ikinci aşamada ise metin, dâhil edildiği alana ait duygu sınıflandırıcı kullanılarak sınıflandırılmaktadır. Deneysel çalışmalarda, metnin alanının mevcut alanlar içerisinde bulunması ve bulunmaması durumu analiz edilmiştir. Bu amaçla, Türkçe ve İngilizce metinler içeren iki ayrı veri kümesi kullanılmıştır. Önerilen yaklaşım sayesinde, beklenen sonuçlara yakın ve bazen de üzerinde bir sınıflandırma başarımı elde edilmiştir. Ayrıca, yaklaşımın her iki dil için de kullanılabilir olduğu doğrulanmıştır.

Anahtar Kelimeler: Makine Öğrenimi, Metin Madenciliği, Duygu Analizi, Duygu Sınıflandırma, Alan Sınıflandırma

ABSTRACT

Master of Science Thesis

NEW TECHNIQUES FOR SENTIMENT ANALYSIS AND CLASSIFICATION IN TEXTS

Muhammet Yasin PAK

Anadolu University

Graduate School of Sciences

Computer Engineering Program

Supervisor: Assoc. Prof. Dr. Serkan GUNAL

2015, 53 pages

Sentiment analysis aims to analyze the author's thoughts and feelings in a written text and sentiment classification focuses on classifying feelings as positive and negative. Supervised learning methods used for sentiment classification need several labeled data to train classifier. Moreover, since the sentiment classification is domain-dependent problem, it is necessary to collect labeled data for each domain in order to obtain high performance; however, this process is time consuming and costly. When a text whose domain is unknown is to be classified, it is not possible to collect labeled and unlabeled data. In this dissertation, a new approach called as domain classification-based sentiment classification is proposed to solve this problem. In the first phase of the proposed approach consisting of two phases, the domain with labeled data, which a text of an unknown domain belongs to, is found out. In the second phase, the text is classified using the sentiment classifier within the domain it belongs to. In the experiments, the cases that the domain of text exist and does not exist within the available domains were analyzed. For this purpose, two datasets including Turkish and English texts were employed. Thanks to the proposed method, classification performances, which are close to or even better than the ones expected, were attained. Also, it was verified that the proposed method is applicable for both languages as well.

Keywords: Machine Learning, Text Mining, Sentiment Analysis, Sentiment Classification, Domain Classification

TEŞEKKÜR

Yüksek lisans tezim boyunca benden yardımlarını esirgemeyen, karşılaştığım sorunlarda bilgi ve deneyimlerini benimle paylaşan değerli hocam ve tez danışmanım Doç. Dr. Serkan GÜNAL'a teşekkürü bir borç bilirim.

Tez süresince desteklerini aldığım Araş. Gör. Rasım ÇEKİK ve Araş. Gör. Ahmet ARSLAN başta olmak üzere Anadolu Üniversitesi Bilgisayar Mühendisliği Bölümü'nde görev yapmakta olan tüm hocalarıma ve çalışma arkadaşlarıma teşekkürlerimi sunarım.

Çalışma hayatım boyunca her zaman yanımda olan ve destek veren sevgili eşime ve aileme teşekkür ederim.

Muhammet Yasin PAK

Haziran, 2015

İÇİNDEKİLER

Sayfa

ÖZET ..... **ii**

ABSTRACT ..... **iii**

TEŞEKKÜR ..... **iv**

ŞEKİLLER DİZİNİ ..... **vii**

ÇİZELGELER DİZİNİ ..... **viii**

KISALTMALAR DİZİNİ ..... **x**

1. GİRİŞ **1**

2. İLGİLİ ÇALIŞMALAR **4**

2.1. Türkçe İçin Yapılmış Duygu Analizi Çalışmaları 4

2.2. İngilizce İçin Yapılmış Duygu Analizi Çalışmaları 7

3. PROBLEM ANALİZİ VE ÖNERİLEN METOT **10**

3.1. Gözetimli Öğrenme Metodu İçin Etiketli Eğitim Verisinin Elde Edilmesi 10

3.2. Duygu Sınıflandırma Probleminin Alan Bağımlılığı 11

3.3. Alan Bilgisine Bağlı Olarak Duygu Sınıflandırma Yaklaşımları 12

3.3.1. Alanı Bilinen Bir Verinin Duygu Sınıflandırması 12

3.3.2. Alanı Bilinmeyen Bir Verinin Duygu Sınıflandırması 12

3.4. Önerilen Metot: Alan Sınıflandırıcı Temelli Duygu Sınıflandırma Yaklaşımı 14

4. DENEYSEL ÇALIŞMALAR **16**

4.1. Veri Seti 16

4.2. Ön işleme 16

4.3. Öznitelik seçimi ve çıkartımı 17

4.4. Sınıflandırma Algoritmaları 20

4.5. Deney Sonuçları ve Analizi 21

4.5.1. Alan içi ve alanlar arası duygu sınıflandırma 22

4.5.2. Hedef verinin mevcut kaynak alanlardan birine ait olma durumunda duygu sınıflandırma 26

4.5.3. Hedef verinin mevcut kaynak alanlardan farklı bir alana ait olma durumunda duygu sınıflandırma 36

5. SONUÇ VE DEĞERLENDİRME 49

KAYNAKLAR 50

ŞEKİLLER DİZİNİ

3.1. Alan sınıflandırıcı temelli duygu sınıflandırma yaklaşımı..... 15
4.1. Türkçe veriler için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırma sonuçları (%): NB sınıflandırıcı..... 30
4.2. Türkçe veriler için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırma sonuçları (%): SVM sınıflandırıcı..... 31
4.3. İngilizce veriler için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırma sonuçları (%): NB sınıflandırıcı..... 32
4.4. İngilizce veriler için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırma sonuçları (%): SVM sınıflandırıcı..... 32
4.5. Türkçe veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma sonuçları (%): NB sınıflandırıcı..... 39
4.6. Türkçe veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma sonuçları (%): SVM sınıflandırıcı..... 41
4.7. İngilizce veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma sonuçları (%): NB sınıflandırıcı..... 42
4.8. İngilizce veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma sonuçları (%): SVM sınıflandırıcı..... 43

ÇİZELGELER DİZİNİ

4.1. Türkçe veriler için her bir alana ait çıkarılan en ayırt edici 15 öznitelik 19
4.2. İngilizce veriler için her bir alana ait çıkarılan en ayırt edici 15 öznitelik 19
4.3. Türkçe veriler için alan içi ve alanlar arası duygu sınıflandırma sonuçları (%): NB sınıflandırıcı 23
4.4. Türkçe veriler için alan içi ve alanlar arası duygu sınıflandırma sonuçları (%): SVM sınıflandırıcı 23
4.5. Türkçe veriler için en iyi sonuçlar için kullanılan ağırlıklandırma yöntemleri ve öznitelik sayıları 23
4.6. İngilizce veriler için en iyi sonuçlar için kullanılan ağırlıklandırma yöntemleri ve öznitelik sayıları 24
4.7. İngilizce veriler için alan içi ve alanlar arası duygu sınıflandırma sonuçları (%): NB sınıflandırıcı 25
4.8. İngilizce veriler için alan içi ve alanlar arası duygu sınıflandırma sonuçları (%): SVM sınıflandırıcı 25
4.9. Türkçe için her bir verinin kendi alanına ait sınıflandırıcıda sınıflandırılması durumunda elde edilecek sonuçlar (%) 26
4.10. İngilizce için her bir verinin kendi alanına ait sınıflandırıcıda sınıflandırılması durumunda elde edilecek sonuçlar (%) 26
4.11. Türkçe veriler için alan bağımsız sınıflandırıcı sonuçları (%) 27
4.12. Türkçe veriler için alan bağımsız sınıflandırıcı sonuçlarının en iyi sonuçlarla karşılaştırılması (%) 28
4.13. İngilizce veriler için alan bağımsız sınıflandırıcı sonuçları (%) 28
4.14. İngilizce veriler için alan bağımsız sınıflandırıcı sonuçlarının en iyi sonuçlarla karşılaştırılması (%) 29
4.15. Türkçe için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırıcı temelli duygu sınıflandırma sonuçları (%) 33
4.16. Türkçe için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırıcı temelli duygu sınıflandırma ile en iyi sonuçlar arasındaki fark (%) 34
4.17. Türkçe verileri için uygulanan yöntemlerin karşılaştırılması (%) 34

4.18. İngilizce için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırıcı temelli duygu sınıflandırma sonuçları (%) 35
4.19. İngilizce için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırıcı temelli duygu sınıflandırma ile en iyi sonuçlar arasındaki fark (%) 35
4.20. İngilizce verileri için uygulanan yöntemlerin karşılaştırılması (%)..... 35
4.21. Türkçe verilerin diğer kaynak alanları kullanılarak sınıflandırma sonuçları (%)..... 37
4.22. İngilizce verilerin diğer kaynak alanları kullanılarak sınıflandırma sonuçları (%)..... 38
4.23. Türkçe veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma temelli yaklaşım ile diğer sonuçların karşılaştırılması (%): NB sınıflandırıcı 44
4.24. Türkçe veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma temelli yaklaşım ile diğer sonuçların karşılaştırılması (%): SVM sınıflandırıcı 46
4.25. İngilizce veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma temelli yaklaşım ile diğer sonuçların karşılaştırılması (%): NB sınıflandırıcı 46
4.26. İngilizce veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma temelli yaklaşım ile diğer sonuçların karşılaştırılması (%): SVM sınıflandırıcı 48

KISALTMALAR DİZİNİ

ME	: Maximum Entropy
SVM	: Support Vector Machine
TF-IDF	: Term Frequency-Inverse Document Frequency
NB	: Naïve Bayes

1. GİRİŞ

Çevresiyle her an iletişim içinde olan insanlar, sahip olduğu duygu ve düşüncelerini diğer insanlarla paylaşma isteği duymakta ve ayrıca başkalarının da düşüncelerine başvurarak herhangi bir konu hakkında fikir edinebilmektedirler. Benzer şekilde şirketler, sundukları hizmetler ve ürünleri geliştirmek ve aynı zamanda ihtiyaçların tespiti amacıyla müşterilerin fikirlerine ihtiyaç duymaktadır. Günümüzde duygu ve düşünce paylaşımı, İnternet'in aktif olarak kullanılmasıyla birlikte farklı bir boyut kazanmıştır. Örneğin; İnternet kullanımı yaygınlaşmadan önce insanlar, herhangi bir hizmet veya ürün hakkında bilgi almak istediklerinde, çevresinden aldığı olumlu ya da olumsuz fikirlere başvurmaktaydılar. Şirketler ise müşterilerin fikir ve düşüncelerine ulaşmak için anket benzeri yöntemler uygulamaktaydı. Günümüzde ise hem insanlar hem de şirketler, bu müşteri yorumlarına forumlar, bloglar ve sosyal medya gibi ortamlardan daha kolay ve kapsamlı bir şekilde ulaşabilmektedirler. Paylaşılan bu bilgiler özellikle şirketlere, insanların bir konu hakkındaki duygu ve düşüncelerini daha kolay bir biçimde değerlendirme ve analiz etme fırsatı vermiştir. Ancak analiz edilmesi gereken verinin çok büyük miktarlara ulaşması, değerlendirmelerin manuel olarak yapılmasını imkânsız hale getirmiştir. Bilgisayarların yüksek işlem gücünün kullanılması ve çeşitli veri işleme yaklaşımları ile bu problem en aza indirgenmeye çalışılmaktadır. Bu ihtiyaçların sonucu olarak ortaya çıkan duygu analizi, doğal dil işleme ve makine öğrenimi alanlarının ilgilendiği bir konu olmuştur.

Duygu analizi çalışma alanının alt problemlerinden biri olarak ele alınan duygu sınıflandırma, bir konu hakkında yazılmış ifadelerin analiz edilmesiyle yazarın sahip olduğu duyguyu genellikle olumlu, olumsuz veya nötr gibi kategorilere sınıflandırmayı amaçlar. Şu ana kadar bu konuda birçok çalışma yapılmış olup bu çalışmaların çoğunda, dokümanda ifade edilen duygunun tek bir nesne hakkında olduğunun varsayıldığı, doküman düzeyinde duygu sınıflandırma problemi üzerine yoğunlaşılmıştır (Liu, 2010). Doküman düzeyinde sınıflandırma için kullanılan tekniklerden bazıları gözetimsiz (unsupervised) öğrenme metotları iken, çoğunlukla gözetimli (supervised) öğrenme metotları kullanılmıştır.

Gözetimli öğrenme metodunda, sınıflandırıcının eğitilmesi için çok miktarda etiketli veri elde edilmesi gerekmektedir. Ayrıca, yapılan çalışmalarda duygu sınıflandırmanın alan bağımlı (domain dependent) bir problem olduğu gözlemlenmiştir (Aue ve Gamon, 2005; Blitzer ve ark., 2007). Bundan dolayı bir alandaki verilerle eğitilmiş sınıflandırıcı başka bir alandaki verilerin sınıflandırılmasında kullanıldığında, çoğu zaman iyi bir başarı gösterememektedir. Problemin kesin çözümü için, sınıflandırma yapmak istenilen her bir alanla ilgili etiketli veri toplamak gerekir ki, bu da birçok alanın var olduğu düşünüldüğünde, yüksek maliyetli ve zaman alıcı bir işlem olmaktadır.

Alan bağımlı bir problem olan duygu sınıflandırma işlemi için etiketli veri toplama problemini çözmek üzere yapılan çalışmalar, sınıflandırma için ihtiyaç duyulacak veri miktarını mümkün olduğu kadar azaltmayı hedeflemektedirler. Duygu sınıflandırma yapılacak alana ait yeteri kadar etiketli verinin olmaması durumunda, farklı bir alana ait etiketli veri kullanılabilmektedir. Ancak farklı alana ait etiketli veriler tek başına yeterli olmamakta, bunun yanında sınıflandırılacak alana ait etiketsiz veriler de kullanılmaktadır. Bazı durumlarda ise etiketsiz verilerle birlikte sınıflandırılacak alana ait az miktarda etiketli veri gerekebilmektedir. Ayrıca, yeterli etiketli veriye sahip olmayan verinin sınıflandırılmasında, tek bir farklı kaynaktan yararlanılabilirken, birden fazla alandan kaynak veriler de kullanılabilmektedir.

Duygu sınıflandırma işlemi yapılacak alana ait yeteri kadar etiketli veri toplamak zor olurken, etiketsiz veriye ise kolay bir şekilde ulaşılabilmektedir. Ancak sınıflandırılacak verinin alanı hakkında bir bilgiye sahip olunmadığında, alana ait etiketli veya etiketsiz veri toplamak mümkün olmayacaktır. Bu çalışmada hangi alana ait olduğu bilinmeyen bir veri için, alanıyla ilgili etiketli veya etiketsiz veri kullanmadan, gözetimli öğrenme metodu ile duygu sınıflandırma işleminin nasıl yapılacağı problemi ele alınmıştır. Çalışmada, birden fazla alana ait etiketli veriye sahip olunduğu varsayılmış, alanı bilinmeyen verinin sınıflandırılması için alan sınıflandırıcı temelli duygu sınıflandırma yaklaşımı önerilmiştir.

Alan sınıflandırıcı temelli duygu sınıflandırma yaklaşımı iki aşamadan oluşmaktadır. İlk aşama olan “alan sınıflandırma” aşamasında, alanı bilinmeyen

verinin alanı tespit edilmekte, ikinci aşama olan "duygu sınıflandırma" aşamasında ise tespit edilen alana ait duygu sınıflandırıcı ile sınıflandırma işlemi yapılmıştır. Yapılan deneyler hem Türkçe hem de İngilizce dili için gerçekleştirilmiş, Naïve Bayes (NB) ve Support Vector Machine (SVM) (Destek Vektör Makinesi) sınıflandırma algoritmaları kullanılmıştır. Deneyler sonucunda, önerilen alan sınıflandırma temelli yaklaşımın alanı bilinmeyen veriler için uygulanabilir olduğu görülmüştür.

Tezin sonraki bölümleri şu şekilde organize edilmiştir: Literatürde Türkçe ve İngilizce dilindeki metinler üzerine yapılan duygu analizi çalışmaları Bölüm 2'de anlatılmıştır. Tez kapsamında ele alınan problemler ve önerilen yaklaşımlar Bölüm 3'te açıklanmıştır. Gerçekleştirilen deneysel çalışmalar ve analiz kısmı Bölüm 4'te yer almıştır. Son bölümde ise sonuçlar yorumlanmış ve gelecek çalışmalar için ele alınabilecek konular ifade edilmiştir.

2. İLGİLİ ÇALIŞMALAR

2.1. Türkçe İçin Yapılmış Duygu Analizi Çalışmaları

Türkçe duygu analizi çalışmalarının ilki sayılan çalışmada (Eroğul, 2009), İngilizce için kullanılan yöntemlerin Türkçe verilerde uygulanabilirliği incelenmiş ve aynı zamanda Türkçe için yeni yöntemler önerilmiştir. Çeşitli öznitelikler kullanarak Türkçe ve İngilizce için sonuçlar karşılaştırılmış, Türkçe veriler için en yüksek sonuç %86 (F1-ölçümü) olarak elde edilmiştir.

Taner (2011) yaptığı çalışmada, doğal dil işleme tekniklerini ve kelimeler arasındaki bağlantıları kullanarak özellik tabanlı duygu analizi için bir altyapı oluşturmak, aynı zamanda ontoloji yapısını kullanarak alanları, kutupluluk (polarity) bilgisini ve sonuçları ayrı ayrı modellemeyi amaçlamıştır.

Albayrak (2011), çalışmasında Türkçe ifadeler ile psikolojik durum arasındaki ilişkiyi araştırmıştır. Depresyonlu, depresyonsuz, anksiyeteli, anksiyetesiz kişilerden toplanan yazılar, morfolojik analizler sonucunda elde edilen öznitelikler kullanılarak incelenmiştir. Test sonuçları, Türkçe ifadelerdeki kullanılan kelimelerin psikolojik durum hakkında önemli bilgiler verdiğini göstermiştir.

Akbaş (2012), yaptığı çalışmada konu (aspect) temelli duygu çıkarımı yapan bir sistem geliştirmiştir. Türkçe Twitter verileri kullanılarak yapılan çalışmada, veriler konulara göre gruplanmış, manuel olarak oluşturulan Türkçe duygu kelime listesiyle birlikte kelime seçme algoritması kullanılarak, duygu gücü belirlenmiş kelimelerin otomatik üretilmesi önerilmiştir.

Diğer bir çalışmada (Kaya ve ark., 2012), politik haberlerin duygu analizi üzerinde çalışılmış, farklı öznitelikler ve ağırlıklandırma yöntemleri kullanılarak kapsamlı bir çalışma yapılmış, dört farklı sınıflandırma algoritmasına göre elde edilen sonuçlar karşılaştırılmıştır. En yüksek doğruluk oranı %76 civarı elde edilirken, literatürde ürün yorumları için elde edilen sonuçlara göre yeterince iyi olmadığı görülmüştür. Maximum Entropy (ME) ve N-gram temelli karakter dil modeli (N-gram based character Language Model) sınıflandırıcılarının gösterdiği performans, NB ve SVM sınıflandırıcılarına göre daha yüksek olmuştur.

Vural ve ark. (2013) yaptıkları çalışmada, İngilizce için sıklıkla kullanılan SentiStrength kütüphanesini Türkçe'ye çevirmiş ve film verileri için gözetimsiz öğrenme metodunu kullanarak duygu analizi yapmışlardır. Üç farklı skorlama tekniği kullanılan çalışmada, gözetimli makine öğrenimi yaklaşımlarıyla karşılaştırıldığında umut verici sonuçlara ulaşmıştır.

Çetin ve Amasyalı (2013a) çalışmalarında, çok miktarda etiketlenmiş veri elde etmenin zorluğuna bir çözüm olarak ortaya çıkan ve kullanılan eğitim verisini azaltarak, aynı veya mümkünse daha iyi başarı elde etmeye çalışan aktif öğrenme konusunu ele almışlardır. Türkçe Twitter verileri kullanılarak yapılan çalışmada, eğitim verisinin miktarı yarı yarıya azaltılarak, hem doğruluk hem de performans açısından daha iyi bir sonuç elde edilmiştir.

Çetin ve Amasyalı (2013b), Türkçe Twitter verileri üzerinde yapmış oldukları diğer bir çalışmada, eğiticili ve eğiticisiz terim ağırlıklandırma yöntemlerini, çeşitli sınıflandırma algoritmaları kullanarak analiz etmişlerdir. İngilizce için yapılan çalışmalarla paralellik gösteren sonuçlarda terim ağırlıklandırmada eğiticili yöntemlerin eğiticisiz yöntemlere göre daha başarılı olduğu görülmüştür.

Başka bir çalışmada (Demirtaş ve Pechenizkiy, 2013), diller arası (cross-lingual) duygu analizi için makine çevirisinin yaptığı katkılar ve getirdiği sınırlamaların araştırılması amaçlanmıştır. Bunun için belirli bir alan ile ilgili İngilizce ve Türkçe dillerinden etiketli veriler kullanılmıştır. Farklı bir dile ait verinin sınıflandırılmasının alanlar arası duygu sınıflandırma problemine benzediği ve alan adaptasyonu için var olan yaklaşımların kullanabileceği ifade edilmiştir.

Boynukalın ve Karagöz (2013) çalışmalarında, Türkçe metinlerde duygu analizini sevinç, üzüntü, öfke ve korku olmak üzere dört sınıfı ele alarak gerçekleştirmiştir. Mevcut Türkçe veri seti olmadığından İngilizce anket cevaplarından oluşan veriler manuel olarak Türkçe'ye çevrilmiştir. Çeşitli sınıflandırma algoritmaları ile yapılan deneylerde Türkçe'ye özgü metotlar eklenmiş ve başarılı sonuçlar elde edilmiştir.

Tutar (2013) çalışmasında, bir alanın ontoloji bilgilerini kullanarak ontoloji tabanlı bir duygu analiz motoru tasarlamayı amaçlamıştır. Otomotiv

verileri üzerinde yapılan çalışmada, alanla ilgili haritalanmış kavramlar (ürün, özellik) kullanılarak, sözlük temelli yaklaşımlara göre daha başarılı sonuçlar elde edilmiştir.

Türkçe köşe yazıları üzerinde duygu sınıflandırma yapılması amaçlanan diğer bir çalışmada (Kaya ve ark., 2013), sınıflandırma performansını artırmak amacıyla köşe yazarlarının etiketsiz Twitter verilerinden yararlanarak etiketli köşe yazıları için bilgi transferi yapılmıştır (transfer learning). Önemli bilgi taşıyan yüksek frekanslı özniteliklerin kullanımının başarıyı ciddi anlamda arttırdığı ancak transfer edilen öznitelik miktarının artmasının (%10 ve üzeri) performansı düşürebildiği görülmüştür.

Sevindi (2013) yaptığı çalışmasında, film yorumları üzerinde makine öğrenimi ve sözlük tabanlı olmak üzere iki yaklaşımla duygu analizini gerçekleştirmiştir. Elde edilen sonuçlara göre, makine öğrenimi yaklaşımda SVM sınıflandırıcısıyla 0,8258 F-skor değeri, sözlük tabanlı yaklaşımda ise 0,5969 F-skor değeri elde edilmiştir.

Özsert ve Özgür (2013) yaptıkları çalışmada, kelime kutbunun belirlenmesi için çoklu dil kullanımını önermiştir. İngilizce çekirdek kelimeleri kullanarak diğer diller için pozitif ve negatif çekirdek kelimeleri üretecek yarı otomatik bir sistem oluşturmuşlardır. İngilizce ve Türkçe için yapılan değerlendirmelerde her iki dil için performans artışı gözlemlenmiştir.

Akba ve ark. (2014) yaptıkları çalışmada, Türkçe film yorumlarını kullanarak öznitelik seçme metotlarını değerlendirmişlerdir. SVM ve NB sınıflandırıcıları ile yaptıkları deneylerde iki sınıflı problem için %83,9, üç sınıf için %63,3 doğruluk oranıyla en yüksek performansı SVM sınıflandırıcısıyla elde etmişlerdir.

Türkçe Twitter verileri üzerinde çalışılan diğer bir çalışmada (Taşlıoğlu, 2014) duygu sınıflandırma doğruluğunu arttırmak amacıyla ironi içeren ifadelerin otomatik olarak çıkarılması ve ele alınması amaçlanmıştır.

Diğer bir çalışmada (Uçan, 2014), duygu ifadelerinin tüm dillerde ortak olabileceğini düşünerek, var olan bir İngilizce duygu sözlüğünün Türkçe'ye çevrilmesi ile Türkçe duygu sözlüğü oluşturulmuştur. Elde edilen sözlük

kullanılarak deneyler gerçekleştirilmiş ve makine öğrenme yöntemleri ile karşılaştırılarak geliştirilen yöntemin başarılı olduğu görülmüştür.

Vural (2013) çalışmasında, duygu analizi kullanarak, düşünce içeren web sayfalarının daha hızlı bir şekilde keşfedilmesini sağlayan düşünce odaklı bir web tarayıcı altyapısı önermiş ve bununla ilgili deneyler gerçekleştirmiştir.

Nizam ve Akın (2014) çalışmalarında, Türkçe Twitter verileri üzerinde gözetimsiz makine öğrenmesi yöntemleri kullanarak duygu analizi çalışması gerçekleştirmişlerdir. Çalışmada pozitif, negatif ve nötr sınıftan oluşan verinin farklı dağılımlar göstermesinin, sınıflandırmadaki başarım sonuçlarına etkisi incelenmiştir. Yapılan deney sonucunda, eşit dağılımlı veri kümesi ile yapılan sınıflandırmanın dengesiz veri kümesi ile yapılan sınıflandırmadan daha iyi sonuç verdiği görülmüştür.

2.2. İngilizce İçin Yapılmış Duygu Analizi Çalışmaları

Alan bağımlı bir problem olan duygu sınıflandırma işlemi için sınıflandırma yapılacak her bir alan için etiketli veri toplama işleminin zor olmasından dolayı bu problemi çözmek üzere çeşitli çalışmalar yapılmıştır. Duygu sınıflandırma işleminde, bir alana ait etiketli verilerin başka bir alana ait verinin duygu sınıflandırma işleminde kullanılması, alan adaptasyon problemi veya alanlar arası duygu sınıflandırma problemi adı altında ele alınmıştır. Bu konuda literatürde, İngilizce veriler için birçok çalışma yapılmış ve problemin çözümü için bazı çalışmalarda tek bir kaynak alandan faydalanırken, bir kısmında ise çoklu kaynak alanlar kullanarak farklı yaklaşımlar öne sürülmüştür. Çalışmalarda adaptasyon problemi ele alınırken, hedef alana ait sadece etiketsiz veriler kullanılabildiği gibi, etiketli az miktarda hedef alan verisiyle birlikte de kullanılabilmektedir.

Duygu analizinde alan adaptasyonu ile ilgili temel bir çalışmada (Aue ve Gamon, 2005), büyük miktarda etiketlenmiş verisi olmayan hedef alan için yapılan sınıflandırma dört farklı yaklaşıma göre aşağıdaki gibi incelenmiştir.

Bahsedilen yaklaşımlar şunlardır:

i) Etiketli veriye sahip olan diğer alanların toplam verilerinde eğiterek test etmek

ii) İlk yaklaşıma benzemekle birlikte hedef alanda gözlemlenen öznitelik kümesini sınırlandırarak eğitmek ve test etmek

iii) Mevcut etiketli veriye sahip alanların sınıflandırıcılarını birleştirerek kullanmak

iv) Hedef alanda etiketlenmemiş büyük miktarda veri ile etiketlenmiş az miktarda veriyi birleştirmek

İlk iki yaklaşımda hedef alanla ilgili etiketsiz veri kullanılmamış, en başarılı sonuç dördüncü yaklaşımda elde edilmiştir.

Tek kaynak alana dayalı bir çalışmada (Blitzer ve ark., 2007), yapısal uyumluluk öğrenme modeli kullanılarak, kaynak ve hedef alanlar arasındaki pivot öznitelikler seçilmiş ve ayrıca az miktarda etiketlenmiş hedef alan verisi yardımıyla yapısal uyumluluk problemlerinin nasıl düzeltileceği gösterilmiştir.

İki aşamadan oluşan diğer bir yaklaşımda (Jiang ve Zhai, 2007), "genelleştirme" aşaması ile alanlar arasında ortak kullanılabilen öznitelikler çıkarılmakta, "adaptasyon" aşamasında ise yarı-gözetimli öğrenme kullanılarak hedef alana özgü öznitelikler elde edilmiştir. Diğer bir çalışmada (Lin ve ark., 2012), alanlar arası duygu analizinde genellikle başarısız olan gözetimli öğrenme yöntemlerine karşın zayıf gözetimli bir yapıda olan JST (Joint Sentiment-Topic) modelini tasarlamışlardır. LDA (Latent Dirichlet allocation) (Blei ve ark., 2003) tabanlı olan JST ile duygu ve konuyu eş zamanlı olarak belirlenebilirken duygu analizi için alanlar arası taşınabilir bir model oluşturulmuştur.

Adaptasyon problemini ele alan diğer bir çalışmada (Read, 2005), alan, konu ve zamana bağlı olmadığı düşünülen duygu ikonları kullanılmış ve oluşturulan eğitim verisiyle bağımsız bir sınıflandırıcı gerçekleştirilmiştir. Alan bağımsız ve alan bağımlı kelimeler arasındaki ilişkiyi bütünüyle ortaya çıkaran diğer bir yaklaşım (Pan ve ark., 2010), alan adaptasyon problemine sebep olan farklı alanlardaki alana özgü kelimeleri, alan bağımsız kelimeler yardımıyla eşleştiren SFA (Spectral Feature Alignment) algoritması önerilmiştir. Başka bir çalışmada ise (Bollegala ve ark., 2011), birden fazla kaynak alandan etiketli ve etiketsiz veri, hedef alandan ise sadece etiketsiz veri kullanarak duygu duyarlı bir

sözlük oluşturulmuş ve farklı alanlardaki benzer duygu belirten kelimeler arasındaki ilişki çıkarılmıştır. Oluşturulan bu sözlük ile öznitelik vektörlerini genişletmek için kullanılmıştır.

Alan adaptasyonu dışında çoklu alan duygu sınıflandırmanın ele alındığı ilk çalışmada (Li ve Zong, 2008), farklı alanlardaki eğitim verileri birleştirilerek performansın artırılması hedeflenmiştir. Öznitelik düzeyinde, bütün alanların öznitelikleri birleştirilerek sınıflandırıcı bütün verilerle eğitilmiş, sınıflandırıcı düzeyinde ise, her bir alanın kendi verileriyle eğitilmesiyle oluşturulan taban sınıflandırıcılar birleştirilerek sınıflandırma yapılmıştır. Literatürde çoklu kaynak alan adaptasyonu için önerilen yaklaşımlar; öznitelik gösterim yaklaşımları ve eğitilmiş sınıflandırıcıların kombinasyonu yaklaşımı olarak iki kategoride ele alınabilir (Sun ve ark., 2015). Öznitelik gösterim yaklaşımında, hedef ve kaynak alan dağılımları arasındaki farklılıkları azaltmak için öznitelik gösterimi değiştirilmektedir. Alanlar arasında değişiklik gösteren özniteliklerin ağırlığını azaltmak veya kaldırmak ve benzer öznitelikleri ön plana çıkararak hedef alan ve kaynak alan benzetilmeye çalışılır. Eğitilmiş sınıflandırıcıların kombinasyonunda ise çoklu kaynak alanların her biri için eğitilmiş olan sınıflandırıcılar, hedef alanın sınıflandırılması için birlikte kullanılırlar. Buradaki temel nokta ise birleştirme işleminde sınıflandırıcıların ağırlıklandırılmasının nasıl yapılacağı konusudur.

Hedef alanla ilgili etiketli verinin kullanılmadığı çalışmada (Whitehead ve Yaeger, 2009), hedef alanın etiketsiz verilerinden yararlanarak sözlüksel benzerlik hesaplanmış ve en benzer alana ait sınıflandırıcı kullanılmıştır. Ayrıca kaynak alan sınıflandırıcılarının birleştirilmesi için basit veya ağırlıklı skor yöntemi uygulanmıştır. Basit yöntemde sınıflandırıcıların çoğunluğunun kararına bakılırken, ağırlıklı yöntemde bu karar sözlüksel benzerlikle elde edilen ağırlıklar hesaba katılarak verilmiştir.

3. PROBLEM ANALİZİ VE ÖNERİLEN METOT

Duygu sınıflandırma için kullanılan teknikler; makine öğrenimi yaklaşımı ve veri sözlüğü temelli yaklaşım (lexicon-based approach) olmak üzere iki ana başlık altında ele alınabilir (Medhat ve ark., 2014). Makine öğrenimi yaklaşımında, çeşitli makine öğrenimi algoritmaları ve dilsel öznitelikler kullanılırken, veri sözlüğü temelli yaklaşımda ise derlenmiş duygu ifadelerinden oluşan duygu sözlükleri kullanılmaktadır. Makine öğrenimi yaklaşımı altında uygulanan gözetimsiz öğrenme metotlarında, etiketli veri kullanmadan sınıflandırma yapılırken, gözetimli öğrenme metotlarında çok sayıda etiketli eğitim verisinden yararlanılır. Literatürdeki çalışmaların birçoğu duygu sınıflandırma problemini doküman düzeyinde ele almaktadır. Doküman düzeyinde yapılan sınıflandırma işleminde, sınıflandırılacak dokümanın tek bir konu hakkında duygu içerdiği varsayılır. Bu düzeyde yapılan çalışmaların birçoğunda gözetimli öğrenme yaklaşımına dayalı yöntemler uygulanmış ve geliştirilmiştir (Liu, 2010)

3.1. Gözetimli Öğrenme Metodu İçin Etiketli Eğitim Verisinin Elde Edilmesi

Gözetimli öğrenme yaklaşımında, sınıflandırma performansı yeterli sayıda etiketli eğitim verisinin var olmasına bağlıdır. Örneğin, film yorumlarının pozitif veya negatif olarak sınıflandırılabilmesi için sınıflandırıcının eğitiminde kullanılmak üzere çok sayıda pozitif ve negatif olarak etiketlenmiş etiketli film verisine ihtiyaç duyulmaktadır. Verilerin etiketlenmesi, manuel veya otomatik olarak yapılabilmektedir. Etiketleme işleminin manuel olarak yapılması, özellikle duygu analizi söz konusu olduğunda çok maliyetli bir iş olabilmektedir. Çünkü yorumların tek tek değerlendirilip pozitif, negatif veya nötr sınıflarından hangisine dâhil edileceğini belirlemek, örneğin konu sınıflandırma (topic classification) problemine kıyasla daha zor olmaktadır. Verilerin manuel olarak etiketlenmesinin zorluğundan dolayı, bu işlem çoğu zaman otomatik olarak yapılmaya çalışılmaktadır. Bu amaçla, kullanıcıların ürün hakkındaki yorumları toplanırken bu yorumların yanında verdikleri oylar da elen alınmaktadır. Oluşturulan değer

aralıklarıyla, (örneğin, oyların 5 üzerinden verildiği durumda; 4-5: pozitif, 1-2: negatif, 3: nötr) yorumların etiketlenme işlemi otomatik olarak yapılmaktadır. Etiketleme işleminin kullanıcının verdiği oylar yardımıyla yapılabilmesi, etiketli veri toplama işlemini kolaylaştırırken bu yöntem ile sınıflandırılacak her alan için etiketli veri toplamak, çoğu zaman uygulanabilir bir çözüm olmamaktadır.

3.2. Duygu Sınıflandırma Probleminin Alan Bağımlılığı

Yapılan çalışmalarda duygu sınıflandırma probleminin alan bağımlı bir problem olduğu görülmüştür (Aue ve Gamon, 2005; Blitzer ve ark., 2007). Sınıflandırma yapılmak istenen verinin ait olduğu alan, sınıflandırıcının eğitimi için kullanılan verinin alanından farklı ise sınıflandırma işlemi yeterli başarıyı gösterememektedir. Örneğin, hedef alan olarak otomotiv alanındaki yorumlar duygu sınıflandırma yapılmak istendiğinde, kaynak alan olarak etiketli film yorumları ile eğitilmiş bir sınıflandırıcıyı kullanmak, sınıflandırma başarısını ciddi anlamda düşürmektedir. Bu performans düşüklüğü kaynak ve hedef alanlar arasındaki öznitelik dağılımlarının farklı olmasından kaynaklanmaktadır (Ben-David ve ark., 2007). Kaynak ve hedef alanlar arasında, duygu ifade eden bazı ortak kelimeler olsa da, alana özgü duygu ifadelerinin de genellikle kullanılmakta olduğu görülmektedir. Örneğin “iyi”, “kötü” ve “güzel” gibi kelimeler tüm alanlar için aynı duyguyu belirten alan bağımsız kelimelerdir. Bu kelimeler alan bağımsız olduğundan, her alan için ortak öznitelik olarak kullanılabilmektedir. Diğer taraftan, her alana özgü duygu ifade eden kelimeler vardır, örneğin; “kırık” ve “hızlı” gibi kelimeler çoğunlukla otomotiv alanında bir duygu ifade edip film alanında kullanılmazken, “sıkıcı” ve “başyapıt” gibi kelimeler ise genellikle film alanında kullanılan duygu ifadeleridir. Bu kelimelerin her alanda farklı olmasından dolayı ortak öznitelikler olarak kullanılamayacağından, bu durum alan adaptasyonu problemine neden olmaktadır. Alan adaptasyonu problemi için akla gelen ilk çözüm, sınıflandırılacak hedef veriyle eğitim için kullanılacak kaynak verinin aynı alana sahip olmasını sağlamaktır. Ancak her bir alan için etiketli veri toplamak, birçok farklı alan söz konusu olduğundan daha önce bahsedildiği üzere ele alınması çok zor bir süreci ifade etmektedir.

3.3. Alan Bilgisine Bağlı Olarak Duygu Sınıflandırma Yaklaşımları

3.3.1. Alanı bilinen bir verinin duygu sınıflandırması

Duygu sınıflandırma işleminin gözetimli öğrenme metotları kullanılarak yapılabilmesi için etiketli eğitim verisine ihtiyaç duyulmaktadır. Sınıflandırma yapılacak verilerin hangi alana ait olduğu biliniyorsa, o alanla ilgili etiketli eğitim verisi toplanmakta ve bu verilerle eğitilen sınıflandırıcı, gelen veriyi pozitif veya negatif olarak sınıflandırmaktadır. Verinin alan bilgisi bilinmesine rağmen her alan için veri toplamanın zorluğundan dolayı sınıflandırma için mevcut eğitim verilerinden yararlanılmak istenmektedir. Ancak, duygu sınıflandırma probleminin alan bağımlı olmasından dolayı, sınıflandırılmak istenen hedef veri için farklı alanlardaki verilerin doğrudan kullanılması, tatmin edici sonuçlar vermemektedir. Genellikle alan adaptasyonu olarak ele alınan bu problemin çözümü için duygu sınıflandırma konusunda birçok çalışma yapılmıştır. Çalışmaların genel amacı, adaptasyon problemine neden olan hedef alan ve kaynak alan arasındaki dağılım farkını olabildiğince azaltmak ve böylece sınıflandırma başarısını artırmaktır.

3.3.2. Alanı bilinmeyen bir verinin duygu sınıflandırması

Bazı durumlarda duygu sınıflandırma yapılmak istenen verinin hangi alana ait olduğu bilinmeyebilir. Böyle bir durumda, sınıflandırıcının eğitilmesi için gerekli olan etiketli alan verisi toplanamamakta ve dolayısıyla gözetimli öğrenme metotlarının kullanımı konusunda ele alınması gereken bir problem ortaya çıkarmaktadır. Alanı bilinmeyen hedef alan verisinin sınıflandırma problemi için elimizde tek bir kaynak alana ait etiketli eğitim verisi olduğu varsayıldığında, sınıflandırıcı, sadece bu etiketli veriler kullanılarak eğitilmek durumundadır.

Etiketli eğitim verisi olarak birden fazla kaynak alan mevcut olduğunda ise farklı alternatif yaklaşımlar ortaya konabilir. Bu tez kapsamında, elimizde birden fazla kaynak alana ait etiketli eğitim verisi olduğu ve sınıflandırılacak hedef alan verisinin hangi alana ait olduğunun bilinmediği varsayılmaktadır. Alan bağımlı bir

problem olduğu bilinen duygu sınıflandırma probleminde, en yüksek sınıflandırma doğruluğunun elde edilmesi için hedef ve kaynak alanların aynı olması gerektiği ifade edilmişti. Bunun yanında, eğer mevcut kaynak alanlara ait veriler kullanılarak alan bağımsız bir duygu sınıflandırıcı oluşturulabilseydi, alanlar arasındaki farklılıklardan kaynaklanan sınıflandırma başarısının düşüklüğü engellenmiş olurdu. Bahsedilen varsayımlar ve ifade edilen bilgiler ışığında karşılaşılabilecek olası durumları ve ifade edilen her bir durum için bu çalışma kapsamında ele alınan yaklaşımları şu şekilde ifade edebiliriz:

- İlk olarak, sınıflandırma yapılacak hedef verinin sahip olduğu alan, etiketli verisi olan mevcut kaynak alanlar içerisinden birisiyle aynı olabilir. Örneğin; elimizde film, kitap ve bilgisayar alanlarına ait etiketli veriler olması durumunda, gelen veri bu alanlardan biri olan film alanına ait olabilir. Böyle bir durumda ortaya konulabilecek birçok yaklaşım olabileceği gibi bu çalışmada ele alınan yaklaşımlar şu şekildedir:
    - Film, kitap ve bilgisayar alanlarına ait tüm eğitim verileri birleştirilerek duygu sınıflandırıcıyı eğitmek ve gelen hedef veriyi bu genel sınıflandırıcı ile sınıflandırmak.
    - Alanı bilinmeyen hedef verinin mevcut olan kaynak alanlar içerisinden hangisine ait olduğunu belirlemek ve bunun sonucunda, hedef veriyi sınıflandırabilmek için ait olduğu belirlenen alanın etiketli verileriyle eğitilmiş sınıflandırıcıyı kullanmak.
- İkinci olarak, sınıflandırma yapılacak hedef verinin sahip olduğu alan, etiketli verisi olan mevcut kaynak alanlardan farklı bir alan olabilir. Elimizde film, kitap ve bilgisayar alanlarına ait etiketli kaynak veriler olduğunu varsayıldığında; gelen verinin alanı mevcut alanlardan farklı olarak, örneğin; otomotiv alanına ait olabilir. Bu durumla ilgili farklı yaklaşımlar öne sürülebilir, ancak bu çalışma için şu yaklaşımlar ele alınmıştır:
    - İlk duruma benzer şekilde, film, kitap ve bilgisayar alanlarına ait etiketli verileri kullanarak genel bir sınıflandırıcı oluşturmak ve gelen hedef veriyi bu genel sınıflandırıcı ile sınıflandırmak.

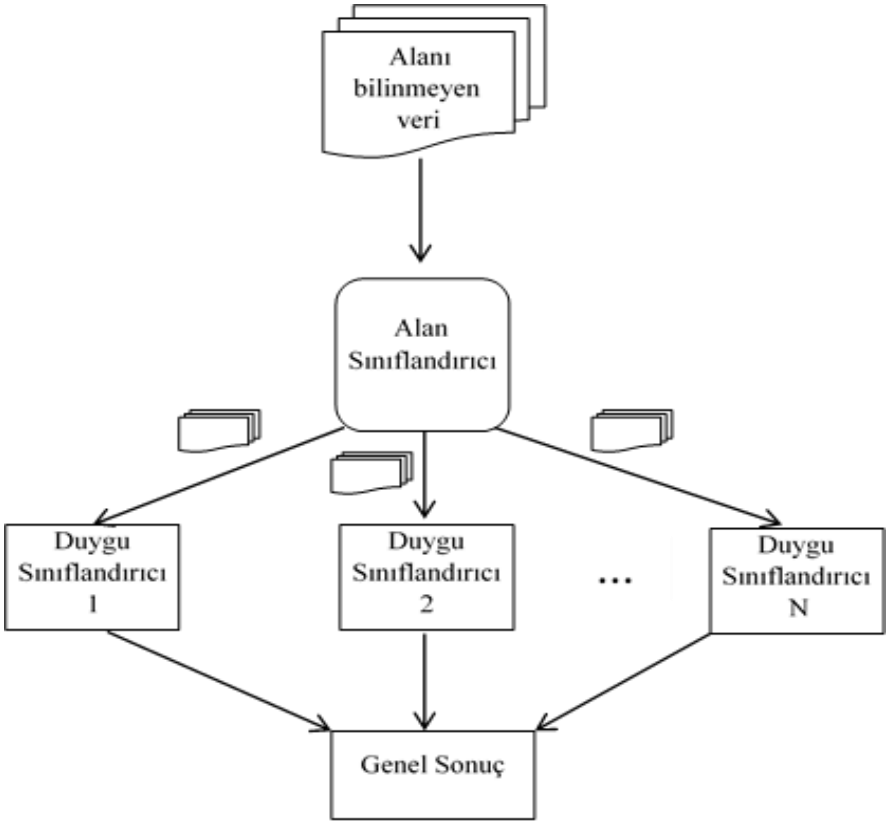
- Alanı bilinmeyen verinin mevcut alanlardan hiçbirisine ait olmadığı durumda, verinin var olan alanlar içerisinde hangisine daha benzer olduğunu bulmak ve gelen veriyi benzer olduğu alana ait sınıflandırıcıda sınıflandırmak.

Duygu sınıflandırılması yapılacak olan verinin hangi alana ait olduğunun bilinmediği ve dolayısıyla hedef alanla ilgili etiketli veya etiketsiz herhangi bir veriye sahip olunmadığı durumda gözetimli öğrenme metodu ile sınıflandırma yapılabilmesi için nasıl bir yöntem izlenmesi gerektiği ile ilgili bir çalışma araştırdığımız kadarıyla bulunamamıştır. Bu çalışmada, yukarıda belirtilen durumların ve ortaya konulan yaklaşımların uygulanabilirliği yapılan deneysel çalışmalarla incelenmiştir.

3.4. Önerilen Metot: Alan Sınıflandırıcı Temelli Duygu Sınıflandırma Yaklaşımı

Alan bilgisi olmayan bir veri sınıflandırılırken hedef verinin mevcut kaynak alanlardan birine veya bu alanlardan farklı bir alana ait olabileceği belirtilmişti. Hedef veri, mevcut kaynak alanlardan birine aitse, bu alanlardan hangisine ait olduğunun, hedef alanın mevcut alanlardan farklı olması durumunda ise var olan alanlardan hangisine dâhil edilebileceğinin bulunması için bir mekanizmaya ihtiyaç duyulmaktadır. Böyle bir işlemin sonucunda alan bilgisi elde edilen hedef veri, atanmış olduğu alanın etiketli eğitim verileri kullanılarak sınıflandırılabilecektir. Bu tezde öne sürülen, alan sınıflandırıcı temelli duygu sınıflandırma yaklaşımı bahsedilen gereksinimleri karşılamak amacıyla öne sürülmüştür. Alan sınıflandırıcı temelli duygu sınıflandırma yaklaşımı, iki temel aşamadan oluşmaktadır. Önerilen yaklaşımın aşamaları ve genel mimarisi Şekil 3.1'de gösterilmiştir. İlk aşama olan "alan sınıflandırma" aşamasında, alanı bilinmeyen hedef verinin mevcut alanlar içerisinden hangisine ait olduğu veya hangi alana dâhil edilebileceği alan sınıflandırıcı yardımıyla tespit edilmektedir. Örneğin; N adet kaynak alana ait etiketli veriye sahip olunması durumunda, alan sınıflandırma aşaması için bu etiketli verilerle eğitilmiş bir alan sınıflandırıcı oluşturulur. Bu alan sınıflandırıcı N alanlı bir durum için, N-sınıflı sınıflandırma problemi olarak düşünülebilir. Alanı bilinmeyen hedef veri, alan sınıflandırıcı

kullanılarak N kaynak alandan birine dâhil edilmektedir. Daha önce bahsedildiği gibi alanı bilinmeyen veri için iki durum söz konusu olabilmektedir. İlk durumda hedef verinin alanı mevcut N kaynak alandan biri olabilmektedir ve bu durumda alan sınıflandırıcıdan beklenilen sınıflandırma işlemini doğru bir şekilde yaparak bu hedef veriyi kendi alanına sınıflandırmaktır. Hedef verinin alanının mevcut N kaynak alandan farklı bir alan olması durumunda ise alan sınıflandırıcının doğru sınıflandırma yapması doğal olarak mümkün olmamaktadır. Böyle bir durumda ise sınıflandırıcıdan, mevcut N kaynak alan içerisinden hedef verinin alanına en benzer alanı bulması beklenmektedir. "Duygu sınıflandırma" aşamasında ise alan sınıflandırma aşaması sonucunda alan bilgisi elde edilen hedef veri, dâhil edildiği alana ait duygu sınıflandırıcı kullanılarak sınıflandırılmaktadır. Bunu gerçekleştirmek üzere, gözetimli öğrenme metodunda uygulandığı üzere her bir kaynak alan, sahip olduğu etiketli eğitim verileri kullanılarak eğitilmiş ve N adet kaynak alan için N adet duygu sınıflandırıcısı oluşturulmuştur. Kendi alanına ait pozitif ve negatif yorumlardan oluşan etiketli verilerle eğitilen duygu sınıflandırıcıları, alan sınıflandırıcı sonucunda alanı belirlenen verinin pozitif ve negatif sınıflarından hangisine ait olduğunu belirlemektedir.



Şekil 3.1. Alan sınıflandırıcı temelli duygu sınıflandırma yaklaşımı

4. DENEYSEL ÇALIŞMALAR

4.1. Veri Seti

Önerilen yaklaşım farklı yapılardaki iki dil için test edilmek istenmiş, bu nedenle çekimli bir dil olan İngilizce ve eklemeli bir dil olan Türkçe kullanılarak deneyler gerçekleştirilmiştir. İngilizce için birçok çalışmada da kullanılan bir veri seti (Blitzer ve ark., 2007) kullanılmış, Türkçe veri setinin elde edilmesi için ise bir e-ticaret sitesi (www.hepsiburada.com) ve film sitesi (www.beyazperde.com) kullanılmıştır. Türkçe veri seti oluşturulurken ürün yorumları otomatik olarak etiketlenmiş ve bu işlem için kullanıcıların yorumları yazarken yanında vermiş oldukları oylar kullanılmıştır. Örneğin, hepsiburada.com internet sitesindeki yorumlar için, kullanıcıların 5 üzerinden vermiş oldukları puanlar dikkate alınarak, 4 ve 5 yıldızlı yorumlar pozitif, 1 ve 2 yıldızlı yorumlar negatif olarak etiketlenmiş, 3 yıldızlı yorumlar ise ihmal edilmiştir. Benzer şekilde beyazperde.com sitesindeki yorumlar için 5 üzerinden 3,5 ve üzeri oy verilen yorumlar pozitif, 2 ve altındaki puanlara sahip yorumlar ise negatif olarak etiketlenmiştir. Türkçe ve İngilizce veri setleri 5 farklı alana ait verilerden oluşmaktadır. Türkçe veri seti içerisinde; bilgisayar, kozmetik, oto aksesuar, telefon ve film alanlarından, her bir alan için 350 negatif ve 350 pozitif yorum olmak üzere toplam 3.500 yorum bulunmaktadır. İngilizce verilerde ise; kitap, DVD, elektronik, sağlık ve mutfak alanlarından her bir alan için 1.000 negatif, 1.000 pozitif yorum olmak üzere toplam 10.000 yorum bulunmaktadır. Duygu sınıflandırma ve alan sınıflandırma aşamalarının her ikisinde de aynı veri setleri kullanılmıştır. Her sınıf için eşit miktarda yorum ele alınmış ve oluşturulan veri setinin dengeli dağılımda bulunması sağlanmıştır. Böylece, sonuçların daha doğru bir şekilde karşılaştırılabilmesi amaçlanmıştır. Test ve eğitim verilerini ayırmak ve verilerin tamamını hem test hem de eğitim verisi olarak kullanabilmek amacıyla 5-katlamalı çapraz doğrulama (5-fold cross validation) yöntemi kullanılmıştır.

4.2. Ön İşleme

Duygu sınıflandırma probleminde de diğer metin sınıflandırma problemlerinde olduğu gibi öznitelik çıkarımı (feature extraction) işleminden

önce birtakım ön işlemler yapılmaktadır. Bu çalışmada da öncelikle genel ön işlemler olan, kelimelerin tüm harflerinin küçük harfe dönüştürülmesi, sayı ve noktalama işaretlerinin kaldırılması ve birden fazla olan boşlukların kaldırılması gibi işlemler yapılmış, öznitelik olarak ayırt edici bir özeliği olmayan "ve" ve "ise" gibi gereksiz kelimeler (stopwords) çıkarılmıştır. Ayrıca, eklemeli bir dil olan Türkçe için özellikle gerekli olan kök bulma işlemi (stemming) gerçekleştirilmiş. Türkçe dili için Zemberek (Akın ve Akın, 2007) adı verilen doğal dil işleme kütüphanesi kullanılarak, kelimeler yalın hale getirilmiş ve bunun yanında olumsuzluk eki almış kelimeler de ayrıca ele alınmıştır. Örneğin; "beğenmedim" kelimesi "neg_beğen" şeklinde ele alınmıştır. Bu işlemlerin dışında, özellikle duygu sınıflandırma probleminde daha sık karşılaşılan duygu ikonları da ön işlemeye dâhil edilmiştir. Kullanıcıların yorum içerisinde kullandıkları ":)" ve ":D" gibi işaretler pozitif duygu ifade ederken, ":(" ve ":-[" gibi ifadeler negatif duygu ifade etmektedir. Duygu ifade eden bu duygu ikonları, noktalama işaretlerinin kaldırma işlemi yapılmadan önce tespit edilmiş, pozitif ve negatif duygu belirten iki öznitelik olarak ele alınmıştır. Aynı ön işlemler İngilizce dili için de yapılmış, sadece kök bulma ve negatif durumların ele alınması işlemleri bu çalışmada gerçekleştirilmemiştir.

4.3. Öznitelik Seçimi ve Çıkartımı

Öznitelik çıkartımı işlemi için, metin sınıflandırmada da sıklıkla kullanılan kelime çantası (the bag of words) yaklaşımı kullanılmıştır. *N*-gram modeli olarak ifade edilen yöntemde, her kelime (unigram) birer öznitelik olabileceği gibi, kelime çiftleri (bigram) veya genel olarak ifadeyle, yan yana olan *n* adet kelime birleşerek öznitelik oluşturabilmektedir. Bu çalışmada her bir kelime birer öznitelik olarak alınmıştır. Özniteliklerin çıkartımı işleminden sonra, elde edilen özniteliklerden ayırt ediciliği yüksek olanların seçilmesi ve öznitelik sayısının azaltılması amacıyla öznitelik seçimi (feature selection) tekniklerinden biri olan bilgi kazanımı (information gain) tekniği kullanılmıştır.

Bilgi kazanımı, bir terimin varlığı ya da yokluğu bilgisinin, herhangi bir sınıfa doğru sınıflandırma kararı verme işleminde ne kadar katkıda bulunduğunu ölçer (Yang ve Pedersen, 2007). Bir terim için elde edilen bilgi kazanımı

değerinin maksimum değere ulaşması, sınıflar arasında ideal bir ayırt ediciliğe sahip olduğunu gösterir. Aşağıda belirtilen formüle göre, t terimi için bilgi kazanımı değeri hesaplanmaktadır.

$$IG(t) = -\sum_{i=1}^{M} P(C_i)\log P(C_i) + P(t)\sum_{i=1}^{M} P(C_i \mid t)\log P(C_i \mid t) + P(\bar{t})\sum_{i=1}^{M} P(C_i \mid \bar{t})\log P(C_i \mid \bar{t}) \quad (4.1)$$

Formülde, M sınıf sayısını, $P(C_i)$, C_i sınıfının olasılığını, $P(t)$ ve $P(\bar{t})$, sırasıyla t teriminin varlığının ya da yokluğunun olasılıklarını, $P(C_i \mid t)$ ve $P(C_i \mid \bar{t})$, sırasıyla t teriminin varlığına ya da yokluğuna bağlı olarak C_i sınıfının koşullu olasılıklarını ifade eder.

Bu teknikle her bir özniteliğin bilgi kazanım değeri hesaplanmış ve öznitelikler bu değerlere göre sıralanmıştır. Ayırt ediciliği yüksek olandan düşük olana doğru sıralanan öznitelikler belirli oranlarda seçilerek deneyler yapılmış ve sonuçlar farklı öznitelik sayılarına göre değerlendirilmiştir. Çizelge 4.1 ve Çizelge 4.2'de ürün yorumlarının öznitelik çıkartımı ve seçimi sonucunda elde edilen duygu sınıflandırmada kullanılacak en ayırt edici 15 öznitelik her bir alan için sıralanmıştır. Henüz sınıflandırma aşamasına geçilmese bile çıkarılan özniteliklerden bazı çıkarımlar yapabilmek mümkün olabilmektedir.

Örneğin; Çizelge 4.1'deki Türkçe alan verilerinden çıkarılan öznitelikler incelendiğinde alanlar arasındaki ilişkiler görülmektedir. Öznitelikler, her alanda ortak olarak görülebilenler en yukarıda, alana özgü olarak ortaya çıkanlar aşağı kısımlarda olacak şekilde gösterilmiştir. İlk 10 özniteliğin her alan için hemen hemen aynı olduğu görülmekte, daha sonra ise alanlar için farklı öznitelikler göze çarpmaktadır. Örnek vermek gerekirse, bilgisayar alanı için "hız, bellek"; film alanı için "sıkıcı, berbat"; kozmetik alanı için "fiyat, yorum"; oto aksesuar alanı için "fena, pırıl"; ve telefon alanı için "kırık, süper" kelimelerinin sadece kendi alanlarında öne çıkmış kelimeler olduğu görülmektedir. Daha öncede bahsedildiği gibi alana özgü bu kelimeler alanlar arası duygu sınıflandırma yapıldığında sınıflandırma başarısı düşürmektedir. Çizelge 4.2'de ise İngilizce diline ait her bir alan için çıkarılan öznitelikler gösterilmiştir. Türkçe'ye benzer olarak, İngilizce'de de her alan için ortak duygu ifade eden kelimelerin olduğu görülmektedir. Ancak alana özgü kelimelerin daha çeşitli olduğu bilgisi çıkarılabilir.

Çizelge 4.1. Türkçe veriler için her bir alana ait çıkarılan en ayırt edici 15 öznitelik

Bilgisayar	Film	Kozmetik	Oto Aksesuar	Telefon
tavsiye	harika	tavsiye	tavsiye	tavsiye
teşekkür	mükemmel	teşekkür	teşekkür	teşekkür
değil	güzel	harika	harika	harika
ama	kötü	değil	değil	değil
gayet	süper	ama	ama	ama
herkes	sıradan	eder	gayet	gayet
fakat	iğrenç	neg_et	eder	eder
fena	berbat	çok	mükemmel	mükemmel
neg_kaçır	sıkıl	neg_kaçır	herkes	fakat
kötü	sıkıcı	neg_beğen	güzel	çok
ancak	iyi	fırça	neg_et	süper
hız	yazık	fiyat	fena	kırık
mikrofon	gerçek	yorum	hediye	karşıla
bellek	para	pek	jant	idare
ses	yok	taş	pırıl	koşul

Çizelge 4.2. İngilizce veriler için her bir alana ait çıkarılan en ayırt edici 15 öznitelik

Kitap	DVD	Elektronik	Sağlık	Mutfak
waste	waste	waste	waste	waste
great	great	great	great	great
higly	not	highly	highly	highly
not	nothing	not	not	return
excellent	best	return	return	excellent
nothing	bad	excellent	best	price
bad	money	price	money	perfect
wonderful	boring	perfect	love	love
boring	was	little	easy	wonderful
poorly	horrible	service	was	easy
war	worst	plenty	flaxseed	after
life	ridiculous	room	returned	back
don	season	call	loves	send
author	pluto	support	manufacurer	warranty
also	commander	customer	aspirin	refund

Son olarak, her bir dokümanın öznitelik vektörüne dönüştürülme işlemi gerçekleştirilmiştir. Öznitelik vektörünün oluşturulma aşamasında, en sık kullanılan terim ağırlıklandırma yöntemlerinden (Lan ve ark., 2009); özniteliğin varlığına ya da yokluğuna bakılma yöntemi (binary) ve TF-IDF (Term Frequency-Inverse Document Frequency) temelli yöntem uygulanmış ve sonuçları karşılaştırılmıştır.

4.4. Sınıflandırma Algoritmaları

Gözetimli öğrenme metotları içerisinde birçok sınıflandırma algoritması kullanılmaktadır. Bu algoritmalar içerisinden duygu sınıflandırma için sıklıkla kullanılan birçok sınıflandırma yöntemi olmasına rağmen bu çalışmada diğer metotlara göre daha başarılı sonuçlar verdiği görülen NB ve SVM sınıflandırıcıları tercih edilmiştir (Pang ve ark., 2002). Çalışmanın aşamalarında kullanılan alan ve duygu sınıflandırıcıların her ikisi için de bu sınıflandırıcılar kullanılmış ve performansları karşılaştırılmıştır.

NB sınıflandırma algoritması, koşullu olasılıkları kullanan Bayes teoremine dayanmaktadır. Öznitelik vektörünü oluşturan özniteliklerinin istatistiksel olarak bağımsız olduğu kabul edilir (Theodoridis ve Koutroumbas, 2008). Metin sınıflandırma problemi için sıklıkla kullanılmaktadır. *d*-boyutlu *x* öznitelik vektörü, (4.2)'ye göre yapılan hesaplar sonucunda c_i i=1,2,.,M sınıflarından birine atanır:

$$Sınıf = enbüyükle \prod_{k=1}^{d} p(x_k|c_i) \tag{4.2}$$

(4.2)'de $p(x_k|c_i)$ ifadesi, c_i sınıfı için *x* öznitelik vektörünün *k'*nıncı özniteliğin olasılığını ifade eder. Normal bir dağılım gösterdiği düşünüldüğünde, olasılık yoğunluk fonksiyonunun değeri şu şekilde hesaplanmaktadır:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{(x-m)^2}{2\sigma^2}) \tag{4.3}$$

(4.3)'te m ve σ, değerleri sırasıyla x özniteliğinin ortalamasını ve standart sapmasını göstermektedir.

SVM, literatürde sıklıkla kullanılan en etkin sınıflandırma algoritmalarından biridir. SVM modelinde, örneklerin her biri, uzaydaki noktalar olarak gösterilmektedir. Sınıfları ayırmak için hiperdüzlemlerden (hyperplanes) yararlanılır (Theodoridis ve Koutroumbas, 2008). Her bir hiper düzlem, yönü (w) ve uzaydaki gerçek konumu (w_0) ile nitelendirilir. Doğrusal sınıflandırıcı, basitçe şu şekilde tanımlanabilir:

$$w^T x + w_0 = 0 \tag{4.4}$$

İki sınıfı ayıran $w^T x + w_0 = 1$ ve $w^T x + w_0 = -1$ hiperdüzlemler arasındaki bölge marjin (margin) olarak adlandırılmaktadır. Marjin genişliği, $2/\|w\|$ ile hesaplanır. SVM algoritmasının temelinin altında yatan fikir, mümkün olan maksimum marjinin elde edilmesidir. Marjinin en yüksek değeri için şu değerin minimum olması gerekmektedir:

$$J(w, w_0, \varepsilon) = \frac{1}{2}\|w\|^2 + K\sum_{i=1}^{N}\varepsilon_i \tag{4.5}$$

Burada *K*, kullanıcı tanımlı sabit bir değeri, ε, marjin hatasını (margin error) ifade etmektedir. Marjin hatası, bir sınıfa ait verinin hiperdüzlemin yanlış tarafında kalmasıyla oluşur. Maliyeti minumuma indirmek için, marjin büyüklüğü ve hata arasında optimum çözüm elde edilmeye çalışılır.

4.5. Deney Sonuçları ve Analizi

Alanı bilinmeyen bir verinin duygu sınıflandırma işlemi için olası durumlar ve bu durumlar için önerilen yaklaşımlar daha önceki bölümde açıklanmıştı. Tezin bu kısmında ise öncelikle mevcut alanlara ait etiketli eğitim verileriyle her bir alan için duygu sınıflandırıcılar oluşturulmuştur. Daha sonra bu duygu sınıflandırıcılar ile mevcut alanlara ait test verileri sınıflandırılmıştır. Sonraki deneylerde ise alanı bilinmeyen veri için ifade edilen her bir durum ve yaklaşımlar ile ilgili deneyler gerçekleştirilmiştir. Deneyler hem Türkçe hem de İngilizce veriler için gerçekleştirilirken, verilerin sınıflandırılması için NB ve SVM sınıflandırma algoritmaları kullanılmıştır. Ayrıca farklı öznitelik sayıları ve ağırlıklandırma yöntemleri kullanılarak sonuçlar karşılaştırılmıştır.

4.5.1. Alan içi ve alanlar arası duygu sınıflandırma

Deneylerin ilk kısmında, duygu sınıflandırma işleminde alan adaptasyonu problemini görebilmek amacıyla her bir alana ait duygu sınıflandırıcı, hem kendi alanına ait verilerle hem de diğer alanların verileriyle test edilmiştir. Aynı zamanda deney sonucunda elde edilen değerler, daha sonraki aşamalarda yapılacak deneylerin sonuçlarının değerlendirmesi için temel oluşturmaktadır.

Deneyler için, Türkçe ve İngilizce olmak üzere iki farklı dile ait veriler kullanılmıştır. Türkçe için her bir alana ait 350 pozitif ve 350 negatif olmak üzere 3500 yorum kullanılarak deneyler gerçekleştirilmiştir. Bir alan kendi alanına ait verilerle sınıflandırılırken, 280 pozitif ve 280 negatif yorum ile eğitilmiş, 70 pozitif ve 70 negatif yorum ile test edilmiştir. Bu işlem 5-çapraz doğrulama ile tekrarlanmıştır. Daha sonra, her bir alana ait veriler, diğer alanlara ait duygu sınıflandırıcılar ile sınıflandırılmış ve bunun için her bir alan için 280 pozitif ve 280 negatif yorum eğitim için kullanılmıştır. Her bir alana ait veriler, diğer alanlara ait sınıflandırıcıların her birinde sınıflandırılmış ve deneyler 5 kez tekrarlanmıştır. Deney tekrarlarının nedeni, yapılan alan içi ve alanlar arası duygu sınıflandırma deneyleri için eğitimde kullanılan veri sayılarını eşit tutmak ve böylece daha doğru bir karşılaştırma yapabilmektir. Çizelge 4.3 ve Çizelge 4.4'te Türkçe yorumlar için duygu sınıflandırma sonuçları gösterilmiştir. Bu sonuçlar, farklı öznitelik sayıları ve ağırlıklandırma yöntemleri kullanılarak, NB ve SVM sınıflandırma algoritmaları ile elde edilen en iyi sonuçları ifade etmektedir. En iyi sonuçları veren öznitelik sayıları ve ağırlıklandırma yöntemleri Çizelge 4.5'te gösterilmiştir. Daha önceki çalışmaları da destekler şekilde (Pang ve ark., 2002), ağırlıklandırma yöntemleri içerisinde, Binary ağırlıklandırma yönteminin TF-IDF hesabına dayalı ağırlıklandırma yönteminden genellikle daha başarılı sonuçlar verdiği görülmüştür.

Sınıflandırma algoritmaları açısından bakıldığında, Türkçe veriler için yapılan bu deneyde, NB ve SVM algoritmalarının hemen hemen aynı derecede başarı gösterdiği görülmüştür. Duygu sınıflandırma amacıyla, her bir alana ait etiketli veriler kullanılarak eğitilen sınıflandırıcılar farklı alanlardaki test verileri ile sınıflandırılmışlardır. Çizelge 4.3 ve Çizelge 4.4'te kalın olarak yazılmış

değerler, her bir alan için test edilen verilerin sınıflandırıcılar arasında aldığı en iyi sonuçları ifade etmektedir. Sonuçlardan da görüleceği üzere, test verisi genellikle kendi alana ait duygu sınıflandırıcı ile sınıflandırıldığında en iyi sonuçları vermektedir. Bu sonuç duygu sınıflandırma probleminin alan bağımlı bir problem olduğunu göstermektedir. Ancak test sonuçlarına göre otomotiv ve telefon alanları için bu genellemenin dışına çıkıldığı görülmektedir.

Çizelge 4.3. Türkçe veriler için alan içi ve alanlar arası duygu sınıflandırma sonuçları (%): NB sınıflandırıcı

Test verisi / Sınıflandırıcı	Bilgisayar	Kozmetik	Oto Aksesuar	Telefon	Film
Bilgisayar	**78,14**	78,00	74,91	**76,43**	67,51
Kozmetik	75,34	**82,00**	**79,17**	75,66	70,17
Oto Aksesuar	74,31	79,54	77,43	74,54	65,26
Telefon	76,80	74,46	74,34	72,14	69,29
Film	63,49	65,69	64,97	66,46	**80,00**

Çizelge 4.4. Türkçe veriler için alan içi ve alanlar arası duygu sınıflandırma sonuçları (%): SVM sınıflandırıcı

Test verisi / Sınıflandırıcı	Bilgisayar	Kozmetik	Oto Aksesuar	Telefon	Film
Bilgisayar	**78,71**	77,40	77,03	**75,31**	64,37
Kozmetik	74,06	**80,86**	**78,14**	74,31	68,86
Oto Aksesuar	76,17	78,40	78,00	75,14	66,09
Telefon	77,14	76,14	76,80	73,43	66,51
Film	63,14	65,74	64,74	62,74	**78,71**

Çizelge 4.5. Türkçe veriler için en iyi sonuçlar için kullanılan ağırlıklandırma yöntemleri ve öznitelik sayıları

	Bilgisayar	Kozmetik	Oto Aksesuar	Telefon	Film
NB	Bin. N/5	Bin. N/4	Bin. N	Bin. N/2	Bin. N/4
SVM	Bin. N/5	Bin. N/5	Bin. N/4	TF-IDF N/5	TF-IDF N/4

Örneğin, otomotiv alanına ait verilerin kozmetik alanına ait sınıflandırıcıda, telefon verilerinin ise bilgisayar alanına ait sınıflandırıcıda sınıflandırıldığında daha iyi sonuç verdiği görülmüştür. Bu durumun, bazı alanların birbirine yakın olmasından veya eğitim verilerinin bazı durumlarda tutarsız olmasından kaynaklandığı düşünülmektedir. Eğitim verileri için toplanan pozitif ve negatif yorumlar arasında, özellikle negatif yorumlar için, negatif duygudan ziyade, nötr duygu daha baskın olabilmekte ve bu da yanlış sınıflandırmalara yol açabilmektedir.

Aynı deneyler, İngilizce veriler üzerinde de yapılmış ve İngilizce için her bir alana ait 1.000 pozitif ve 1.000 negatif olmak üzere 10.000 yorum kullanılarak deneyler gerçekleştirilmiştir. Her bir alana ait duygu sınıflandırıcıları eğitilirken, eğitim ve test için kullanılacak veri miktarları Türkçe verilerle benzer şekilde belirlenmiştir. En iyi sonuçlar için kullanılan öznitelik sayıları ve ağırlıklandırma yöntemleri Çizelge 4.6'da ifade edilmiştir. Türkçe deney sonuçlarına benzer şekilde, öznitelik vektörü oluşturulurken Binary yöntemi kullanıldığında daha iyi sonuç verdiği gözlemlenmiştir. Türkçede olduğu gibi İngilizce veriler için yapılan deneylerde de iki farklı sınıflandırma algoritması kullanılmış, Çizelge 4.7'de NB, Çizelge 4.8.'de ise SVM algoritmaları kullanılarak elde edilen sonuçlar gösterilmiştir. İngilizce veriler için, SVM algoritması ile elde edilen sonuçların NB algoritmasına göre genellikle daha iyi olduğu görülmüştür.

Çizelge 4.6. İngilizce veriler için en iyi sonuçlar için kullanılan ağırlıklandırma yöntemleri ve öznitelik sayıları

	Kitap	DVD	Elektronik	Sağlık	Mutfak
NB	Bin. N/3	Bin. N/5	Bin. N/5	Bin. N/5	Bin. N/5
SVM	TF-IDF N/2	Bin. N/4	Bin. N	Bin. N/3	TF-IDF N/2

Kalın olarak yazılmış olan değerlere bakıldığında, alanların hepsinde, verilerin kendi alanına ait sınıflandırıcı kullanılarak sınıflandırıldığında en iyi sonuçları verdiği görülebilmektedir.

Çizelge 4.7. İngilizce veriler için alan içi ve alanlar arası duygu sınıflandırma sonuçları (%): NB sınıflandırıcı

Sınıflandırıcı \ Test verisi	Kitap	DVD	Elektronik	Sağlık	Mutfak
Kitap	**80,15**	76,38	64,91	64,97	66,8
DVD	76,71	**80,30**	67,6	66,41	68,47
Elektronik	67,15	70,60	**80,65**	76,13	79,26
Sağlık	70,10	73,61	76,27	**81,45**	81,03
Mutfak	67,84	70,83	78,27	78,04	**83,70**

İngilizce verilerle yapılan deney sonucunun, Türkçe verilerle elde edilen sonuçlarla karşılaştırıldığında, daha tutarlı olduğu söylenilebilir. Türkçe yorumlar için elde edilen sonuçların bir kısmında, bazı verilerin kendi alanları yerine farklı alanlarda sınıflandırıldığında daha başarılı sonuçlar ortaya koyduğu görülmüştü. İngilizce'de ise her veri kendi alanında en iyi doğruluğa ulaşmıştır. Bunun yanında, sınıflandırma doğrulukları incelendiğinde birbirine yakın alanların sonuçlarının da birbirine yakın olduğu göze çarpmaktadır. Örneğin; sonuçlara göre kitap ve DVD alanları birbirine daha yakınken, elektronik, sağlık ve mutfak alanlarının da birbirlerine daha benzer alanlar olduğu ifade edilebilir.

Çizelge 4.8. İngilizce veriler için alan içi ve alanlar arası duygu sınıflandırma sonuçları (%): SVM sınıflandırıcı

Sınıflandırıcı \ Test verisi	Kitap	DVD	Elektronik	Sağlık	Mutfak
Kitap	**79,35**	77,78	68,17	70,51	70,74
DVD	76,10	**80,6**	71,82	72,55	71,60
Elektronik	66,67	72,32	**82,05**	77,42	81,33
Sağlık	67,75	73,34	78,43	**82,90**	81,60
Mutfak	67,11	70,15	79,95	79,16	**84,55**

Alan bağımlı bir problem olan duygu sınıflandırma işleminde, en iyi sınıflandırma başarısı için, verinin normal şartlarda kendi alanına ait sınıflandırıcıda sınıflandırılması gerektiği ifade edilmişti. Çizelge 4.9 ve Çizelge 4.10'da sırasıyla Türkçe ve İngilizce veriler için her bir verinin kendi alanına ait duygu sınıflandırıcıda sınıflandırılması durumunda elde edilecek doğruluk

değerleri ifade edilmektedir. Bu değerler, her bir alan verisinin kendi alanında sınıflandırılması sonucu elde edilen değerlerin ortalaması alınarak hesaplanmıştır.

Çizelge 4.9. Türkçe için her bir verinin kendi alanına ait sınıflandırıcıda sınıflandırılması durumunda elde edilecek sonuçlar (%)

Öznitelik sayısı	NB/Binary	NB/TF-IDF	SVM/Binary	SVM/TF-IDF
N	77,34	75,00	76,26	76,46
N/2	77,37	75,40	76,91	76,17
N/3	77,31	75,00	76,66	76,26
N/4	77,14	75,49	77,03	77,00
N/5	76,97	75,29	77,63	76,46
Ortalama	**77,23**	**75,24**	**76,90**	**76,47**

Çizelge 4.10. İngilizce için her bir verinin kendi alanına ait sınıflandırıcıda sınıflandırılması durumunda elde edilecek sonuçlar (%)

Öznitelik sayısı	NB/Binary	NB/TF-IDF	SVM/Binary	SVM/TF-IDF
N	81,04	77,47	81,12	80,79
N/2	80,80	75,60	81,41	81,59
N/3	80,79	77,67	81,49	81,50
N/4	80,82	77,77	81,43	81,20
N/5	81,07	77,12	81,03	81,05
Ortalama	**80,90**	**77,13**	**81,30**	**81,23**

4.5.2. Hedef verinin mevcut kaynak alanlardan birine ait olma durumunda duygu sınıflandırma

Duygu sınıflandırma işleminde, hedef verinin alanının bilinmediği bir durumda, ele alınacak ilk senaryo, hedef verinin alanının mevcut kaynak alanlar içerisinde olabilme durumudur. Böyle bir durumda uygulanabilecek iki yaklaşımdan bahsedilmişti. Bu yaklaşımlardan ilki, mevcut alanlara ait etiketli verilerin tamamının, duygu sınıflandırıcının eğitilmesinde kullanılması ve oluşturulan bu alan bağımsız duygu sınıflandırıcı ile hedef verinin sınıflandırılmasıydı. Bu çalışmaya özgü olan ikinci yaklaşımda ise alan sınıflandırma ve duygu sınıflandırma olmak üzere iki aşamadan oluşan bir metot önerilmişti. Tez kapsamında önerilen bu yaklaşımın uygulanabilirliğini görmek

için, öncelikle ilk yaklaşım olan alan bağımsız bir duygu sınıflandırıcı oluşturmak amacıyla deneyler gerçekleştirilmiştir. Daha sonra ise bu deney sonuçları, alan sınıflandırıcı temelli yaklaşım kullanılarak elde edilen sonuçların başarısını ölçmek için kullanılmıştır. Her iki deney Türkçe ve İngilizce olmak üzere iki farklı dilde gerçekleştirilmiş ve 5 farklı alana ait veriler kullanılmıştır. Deneyler farklı sınıflandırma algoritmaları ve ağırlıklandırma yöntemleri ile tekrarlanmış ve karşılaştırılmıştır.

Alan bağımsız bir duygu sınıflandırıcı oluşturmak amacıyla, Türkçe için her bir alandan 700'er yorum olmak üzere toplam 3.500 yorum kullanılmış ve 700'ü test 2.800'ü eğitim verisi olmak üzere ayrılmıştır. Alan bağımsız sınıflandırıcı, 1.400 pozitif ve 1.400 negatif yorumla eğitilmiş ve 350 pozitif ve 350 negatif olmak üzere 700 adet yorumla test edilmiştir. Elde edilen sonuçlar Çizelge 4.11'de gösterilmiştir. NB ve SVM sınıflandırıcılarıyla, farklı ağırlıklandırma yöntemleri ve öznitelik sayıları ile deneyler tekrarlanmıştır. En iyi sonuçların SVM sınıflandırma algoritması ve Binary ağırlık yöntemi ile elde edildiği görülmüştür.

Çizelge 4.11. Türkçe veriler için alan bağımsız sınıflandırıcı sonuçları (%)

Öznitelik sayısı	NB/Binary	NB/TF-IDF	SVM/Binary	SVM/TF-IDF
N	75,66	74,17	76,06	75,31
N/2	75,94	73,57	76,57	76,03
N/3	75,86	74,28	76,83	75,71
N/4	75,63	73,83	76,51	75,23
N/5	74,03	74,43	75,91	74,83
Ortalama	**75,42**	**74,06**	**76,38**	**75,42**

Çizelge 4.12'de ise Çizelge 4.9 ve Çizelge 4.11'deki sonuçlar karşılaştırılmıştır. Tablodaki değerler sonuçlar arasındaki farkları göstermektedir. Sonuçlar arasında ortalama %1'lik bir farkın oluştuğu görülmektedir. Türkçe için elde edilen bu sonuç, alanı bilinmeyen bir verinin, kaynak alanlardan birine dâhil olması durumunda, alan bağımsız bir duygu sınıflandırıcının kullanımının sadece %1'lik bir performans kaybına yol açtığı, yani en iyiye yakın bir sonuç elde edilebildiğini göstermektedir

Çizelge 4.12. Türkçe veriler için alan bağımsız sınıflandırıcı sonuçlarının en iyi sonuçlarla karşılaştırılması (%)

Öznitelik sayısı	NB/Binary	NB/TF-IDF	SVM/Binary	SVM/TF-IDF
N	-1,68	-0,83	-0,20	-1,15
N/2	-1,43	-1,83	-0,34	-0,14
N/3	-1,45	-0,72	0,17	-0,55
N/4	-1,51	-1,66	-0,52	-1,77
N/5	-2,94	-0,86	-1,72	-1,63
Ortalama	**-1,80**	**-1,18**	**-0,52**	**-1,05**

Aynı deneyler İngilizce veriler için de yapılmış ve deney için her bir alandan 2.000'er yorum olmak üzere 10.000 yorum kullanılmıştır. Çapraz doğrulama yöntemiyle tekrarlanan deneyde, alan bağımsız duygu sınıflandırıcı, 4.000 pozitif ve 4.000 negatif yorumla eğitilmiş, 2.000 adet yorumla test edilmiştir. Deney sonuçları Çizelge 4.13'te belirtilmiştir. İngilizce veriler için en iyi sonuçlar SVM sınıflandırıcısıyla ve TF-IDF ağırlıklandırma yöntemiyle elde edilmiştir.

Çizelge 4.13. İngilizce veriler için alan bağımsız sınıflandırıcı sonuçları (%)

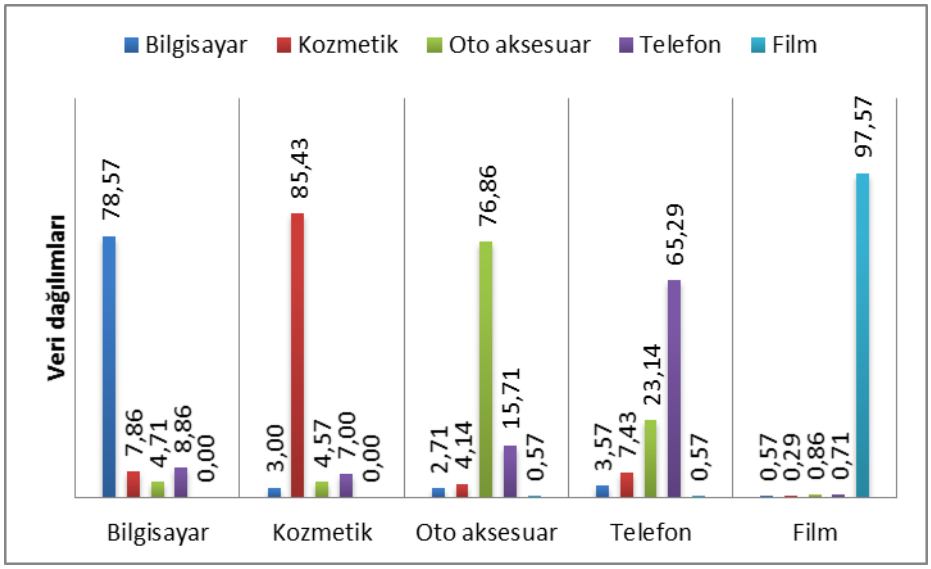
Öznitelik sayısı	NB/Binary	NB/TF-IDF	SVM/Binary	SVM/TF-IDF
N	78,00	73,88	77,57	78,19
N/2	77,50	71,63	78,09	78,39
N/3	78,28	74,56	78,04	78,26
N/4	78,26	74,33	78,00	78,21
N/5	78,56	73,60	78,07	78,03
Ortalama	**78,12**	**73,60**	**77,95**	**78,22**

Çizelge 4.14'te ise Çizelge 4.10 ve Çizelge 4.13'teki sonuçlar karşılaştırılmıştır. Sonuçlar arasında yaklaşık %3'lük bir farkın olduğu görülmektedir. İngilizce için elde edilen bu sonucun Türkçe'ye göre biraz daha farklı olduğu görülmektedir. İngilizce için alan bağımsız sınıflandırıcı sonuçları elde edilebilecek en iyi sonuçlara yakınlaşsa da Türkçe'ye göre daha az başarılı olduğu ifade edilebilir.

Çizelge 4.14. İngilizce veriler için alan bağımsız sınıflandırıcı sonuçlarının en iyi sonuçlarla karşılaştırılması (%)

Öznitelik sayısı	NB/Binary	NB/TF-IDF	SVM/Binary	SVM/TF-IDF
N	-3,04	-3,59	-3,55	-2,60
N/2	-3,30	-3,97	-3,32	-3,20
N/3	-2,51	-3,11	-3,45	-3,24
N/4	-2,56	-3,44	-3,43	-2,99
N/5	-2,51	-3,52	-2,96	-3,02
Ortalama	**-2,78**	**-3,53**	**-3,34**	**-3,01**

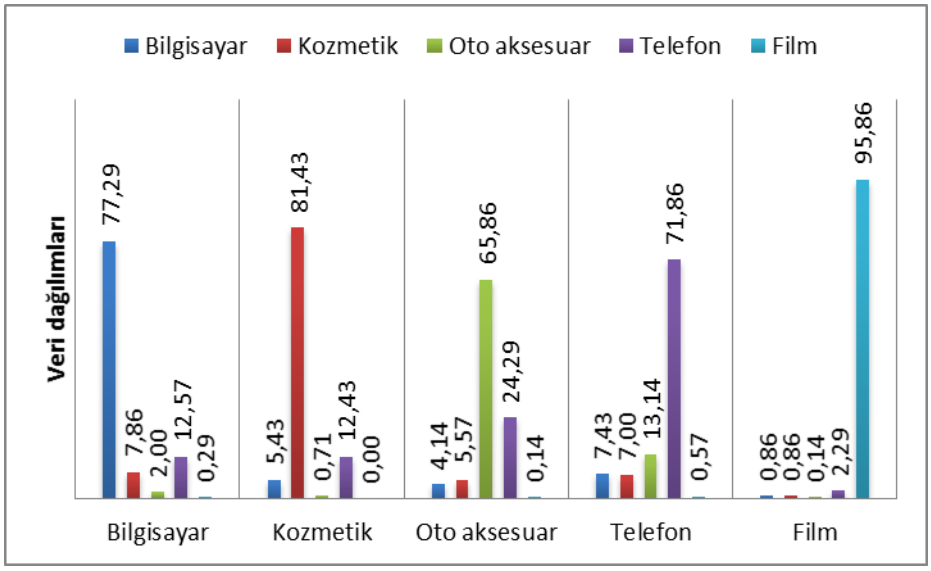
Mevcut kaynak alan verilerin tamamının kullanılmasıyla oluşturulan alan bağımsız bir sınıflandırıcının, alanı bilinmeyen bir verinin alanının mevcut kaynaklar içerisinde olması durumunda, duygu sınıflandırma için kullanılabilir olduğu görülmüştür. Ancak, sınıflandırılacak verinin alanı bilinseydi, veri kendi alanına ait sınıflandırıcıda sınıflandırılacak ve Türkçe veriler için %1'lik, İngilizce verilerde ise %3'lük fark kapanmış olacaktı. Bu amaca yönelik olarak, çalışma kapsamında önerilen alan sınıflandırıcı temelli yaklaşımda, ilk olarak alanı bilinmeyen verinin alanı tespit edilmeye çalışılmıştır. Alan sınıflandırma işleminde, veri 5 farklı alandan birine sınıflandırılacağı için bu işlem, 5 sınıflı sınıflandırma problemi olarak düşünülebilir. Türkçe ve İngilizce için yapılan deneyde, Türkçe veriler için alan sınıflandırıcısının eğitilmesi amacıyla her bir alana ait 560'ar veri kullanılmış ve her bir alandan 140'ar veriyle test edilmiştir. Alan sınıflandırma için hem NB hem de SVM algoritmaları kullanılmış ve çapraz doğrulama ile tekrarlanan deneyler ile elde edilen sonuçlar Şekil 4.1 ve Şekil 4.2'de gösterilmiştir. Test edilen her bir alan için alan sınıflandırma sonuçları gösterilmekte olup, test verilerinin kendi alanına ve diğer alanlara dağılımları belirtilmiştir.



Şekil 4.1. Türkçe veriler için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırma sonuçları (%): NB sınıflandırıcı

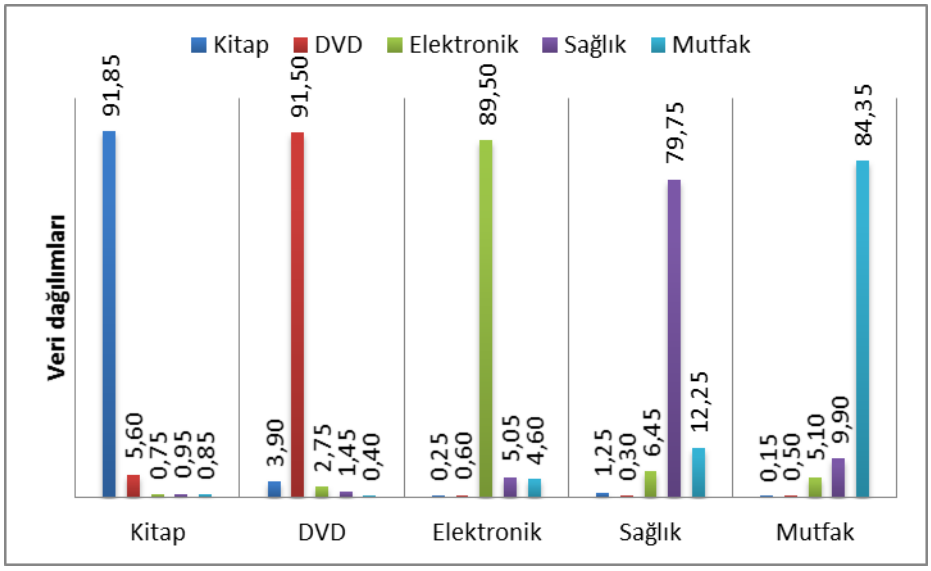
Örneğin, Şekil 4.1'de; alan sınıflandırma sonucunda, bilgisayar alanına ait verilerin %78,57'sinin doğru olarak sınıflandırıldığı, %8,86'sının telefon alanına, % 4,71'inin ise oto aksesuar alanına sınıflandırıldığı görülmüştür.

Başka bir örnek vermek gerekirse, Şekil 4.1 ve Şekil 4.2'de film alanına ait sonuçlarda, film verilerinin yüksek bir sınıflandırma doğruluğuna sahip olduğu görülmektedir. Ancak bazı durumlarda alan sınıflandırıcının yeterince iyi bir performans gösteremediği görülmüştür. Örneğin, oto aksesuar ve telefon alanına ait sonuçlarda verilerin büyük bir kısmının diğer alanlara dağılmış olduğu gözlemlenmiştir. Çizelge 4.3 ve Çizelge 4.4'te ifade edilen sonuçlar yorumlanırken, oto aksesuar ve telefona ait sonuçların diğer alanlara göre beklenilen performansı gösteremediği belirtilmişti. Dolayısıyla bu deney sonucu, yapılan bu değerlendirmeleri destekler mahiyette olmuştur. Son olarak, Şekil 4.1 ve Şekil 4.2'de sınıflandırma algoritmalarının performansları karşılaştırıldığında, Türkçe veriler için, NB ile yapılan sınıflandırmanın SVM algoritmasına göre daha iyi sonuçlar verdiği görülmektedir.

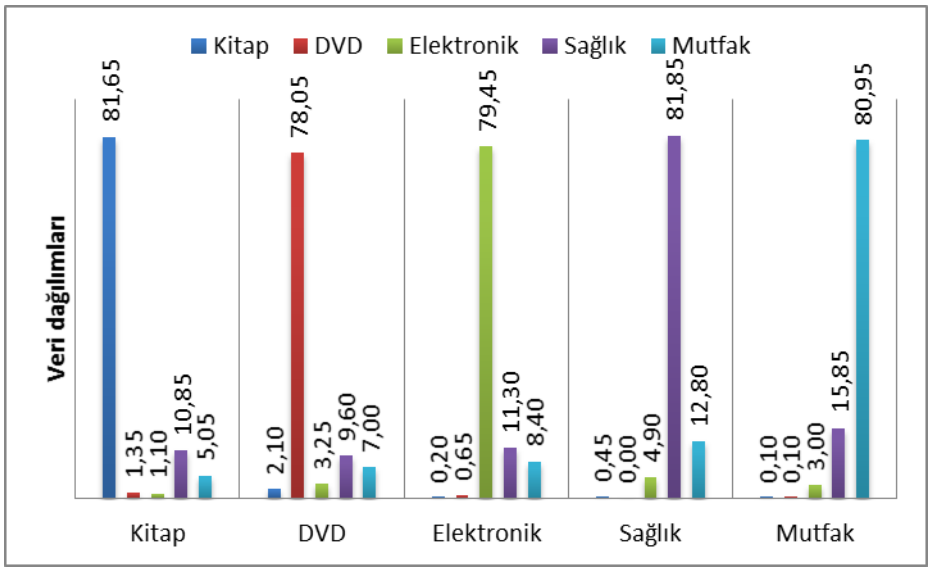


Şekil 4.2. Türkçe veriler için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırma sonuçları (%): SVM sınıflandırıcı

Alan sınıflandırma için yapılan deneyler İngilizce veriler için de tekrarlanmıştır. Alan sınıflandırıcısı, her bir alandan 1.600'er eğitim verisiyle eğitilmiş ve her alan için 400'er veriyle test edilmiştir. NB ve SVM algoritmaları ile ayrı ayrı gerçekleştirilen deneyde elde edilen sonuçlar, sırasıyla Şekil 4.3 ve Şekil 4.4'te gösterilmiştir. Örneğin; Şekil 4.3'te alan sınıflandırma sonucunda, kitap verilerinin %91,85 doğrulukla sınıflandırıldığı görülmektedir. Kitap verilerinin %5,60'ı DVD alanına, geri kalan veriler ise diğer üç alana dağılmıştır. Sağlık alanına ait sonuçlar incelendiğinde ise %79,75'lik bir doğruluk oranı elde edildiği, verilerin %12,25'inin mutfak, %6,45'inin ise elektronik alanına dâhil edildiği görülmektedir.



Şekil 4.3. İngilizce veriler için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırma sonuçları (%): NB sınıflandırıcı



Şekil 4.4. İngilizce veriler için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırma sonuçları (%): SVM sınıflandırıcı

Elde edilen sonuçlar ışığında, alan sınıflandırma için NB ve SVM algoritmalarını karşılaştıracak olursak, İngilizce veriler için, sağlık alanı haricinde NB ile yapılan sınıflandırmanın SVM'e göre çok daha iyi sonuçlar verdiği ifade edilebilir. Ayrıca, Şekil 4.3 ve Şekil 4.4 incelendiğinde alanların yakınlıklarına göre gruplandığı görülmektedir. Çizelge 4.7'de de belirtildiği gibi kitap ve DVD

alanları, bir grup, elektronik, sağlık ve mutfak alanlarının ise, ikinci bir grup olarak ortaya çıktığı gözlemlenmektedir.

Alan sınıflandırıcı temelli duygu sınıflandırma yaklaşımında, alan sınıflandırıcı aşamasından sonra duygu sınıflandırma aşaması gelmektedir. Alan sınıflandırma sonrasında alanı bilinmeyen verinin alanı tespit edildikten sonra, tespit edilen alana ait duygu sınıflandırıcı ile sınıflandırma yapılmıştır. Türkçe veriler için deneyler yapılırken, her bir alana ait 140'ar test verisi, alan sınıflandırıcı sonucunda Şekil 4.1 ve Şekil 4.2'deki elde edilen yüzdeliklere göre alanlara dağıtılmış ve ayrılan bu veriler, alanlara ait her bir duygu sınıflandırıcı ile sınıflandırılmıştır. Sonuç olarak, iki aşamadan oluşan alan sınıflandırma temelli duygu sınıflandırma sonuçları, Türkçe veriler için Çizelge 4.15'te gösterilmiştir. Alan bağımsız sınıflandırıcı ile yapılan duygu sınıflandırma sonuçlarının belirtildiği Çizelge 4.11 ile karşılaştırıldığında önerilen yaklaşımın uygulanması sonucu sınıflandırma performansının arttığı gözlemlenmiştir.

Çizelge 4.15. Türkçe için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırıcı temelli duygu sınıflandırma sonuçları (%)

Öznitelik Sayısı	NB/Binary	NB/TF-IDF	SVM/Binary	SVM/TF-IDF
N	77,63	75,60	76,77	76,89
N/2	77,60	75,29	77,71	76,34
N/3	77,37	75,31	77,29	76,43
N/4	77,26	75,86	77,80	77,17
N/5	77,23	75,08	78,00	76,60
Ortalama	**77,42**	**75,43**	**77,51**	**76,69**

Çizelge 4.16'da ise sonuçlar Çizelge 4.9'da ifade edilen en iyi sonuçlarla karşılaştırılmıştır. Çizelge 4.9'daki sonuçlar alan sınıflandırma sonucunun her alan için %100 olması durumunda elde edilecek sonuçları göstermektedir. Dolayısıyla, Çizelge 4.16'ya bakıldığında Türkçe veriler için, önerilen yaklaşımın kullanılmasıyla bu üst sınırla aradaki farkın kapandığı, hatta çoğu durumda en iyi sonuçların bile üzerine çıkıldığı görülmektedir. Elde edilen sonuçların en iyi sonuçların üzerinde olması, örneğin, Türkçe veriler içerisinde otomotiv ve telefon alanına ait verilerin diğer alanlarda daha başarılı olmasından kaynaklanmaktadır.

Çizelge 4.16. Türkçe için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırıcı temelli duygu sınıflandırma ile en iyi sonuçlar arasındaki fark (%)

Öznitelik Sayısı	NB/Binary	NB/TF-IDF	SVM/Binary	SVM/TF-IDF
N	0,29	0,60	0,51	0,43
N/2	0,23	-0,11	0,80	0,17
N/3	0,06	0,31	0,63	0,17
N/4	0,12	0,37	0,77	0,17
N/5	0,26	-0,21	0,37	0,14
Ortalama	**0,19**	**0,19**	**0,62**	**0,22**

Çizelge 4.17'de ise genel bir karşılaştırma yapılabilmesi amacıyla, Türkçe veriler için, alan bağımsız sınıflandırıcı sonuçları, alan sınıflandırıcı temelli duygu sınıflandırma sonuçları ve alan sınıflandırıcının %100 sınıflandırma yapabilmesi durumunda elde edilecek doğruluk değerleri verilmiştir.

Çizelge 4.17. Türkçe verileri için uygulanan yöntemlerin karşılaştırılması (%)

Yöntemler	NB/Binary	NB/TF-IDF	SVM/Binary	SVM/TF-IDF
Üst sınır	77,23	75,24	76,9	76,47
Alan bağımsız	75,42	74,06	76,38	75,42
Önerilen Yaklaşım	77,42	75,43	77,51	76,69

Alan sınıflandırıcı temelli duygu sınıflandırma için yapılan deneyler İngilizce veriler için de tekrarlanmıştır. Her bir alana ait 400'er veri test olarak kullanılmış, alan sınıflandırma sonucu alanlara dağılan veriler, alana özgü her bir sınıflandırıcı ile sınıflandırılmıştır. Deneyler, farklı öznitelik sayıları, ağırlıklandırma yöntemleri ve sınıflandırma algoritmaları ile gerçekleştirilmiş, sonuçlar Çizelge 4.18'de verilmiştir. Sonuçlar, Çizelge 4.13'teki değerler ile karşılaştırıldığında önerilen yaklaşımın kullanılmasıyla sınıflandırma doğruluklarının artmış olduğu görülmektedir.

Çizelge 4.18. İngilizce için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırıcı temelli duygu sınıflandırma sonuçları (%)

Öznitelik sayısı	NB/Binary	NB/TF-IDF	SVM/Binary	SVM/TF-IDF
N	81,00	77,53	80,59	80,62
N/2	80,73	75,65	80,74	81,02
N/3	80,84	77,71	80,99	80,77
N/4	80,83	77,71	80,82	80,80
N/5	81,08	77,16	80,60	80,66
Ortalama	**80,90**	**77,15**	**80,75**	**80,77**

Artış miktarını daha rahat görebilmek amacıyla, Çizelge 4.19'da İngilizce veriler için alan sınıflandırıcı temelli yaklaşım ile en iyi sonuçlar arasındaki farklar gösterilmiştir. Görüleceği üzere önerilen yaklaşım ile duygu sınıflandırma için hedeflenen en iyi sonuçlara hemen hemen ulaşılmıştır. Çizelge 4.20'de ise yapılan 3 farklı deney sonuçlarının ortalama değerleri karşılaştırma amacıyla gösterilmiştir.

Çizelge 4.19. İngilizce için hedef verinin mevcut alanlar içerisinde olması durumunda alan sınıflandırıcı temelli duygu sınıflandırma ile en iyi sonuçlar arasındaki fark (%)

Öznitelik Sayısı	NB/Binary	NB/TF-IDF	SVM/Binary	SVM/TF-IDF
N	-0,04	0,06	-0,53	-0,17
N/2	-0,07	0,05	-0,67	-0,57
N/3	0,05	0,04	-0,5	-0,73
N/4	0,01	-0,06	-0,61	-0,4
N/5	0,01	0,04	-0,43	-0,39
Ortalama	**-0,01**	**0,03**	**-0,55**	**-0,45**

Çizelge 4.20. İngilizce verileri için uygulanan yöntemlerin karşılaştırılması (%)

Yöntemler	NB/Binary	NB/TF-IDF	SVM/Binary	SVM/TF-IDF
Üst sınır	80,90	77,13	81,30	81,23
Alan bağımsız	78,12	73,60	77,95	78,22
Önerilen Yaklaşım	80,90	77,15	80,75	80,77

Sonuç olarak, alanı bilinmeyen bir verinin duygu sınıflandırma işleminde, verinin alanı mevcut kaynak alanlar arasında bulunuyorsa, bu çalışmada önerilen

alan sınıflandırıcı temelli yaklaşım ile yapılan duygu sınıflandırma işlemi, Türkçe veriler için hiç bir kayıp olmadan gerçekleştirilmekte, hatta bazı durumlarda normalden daha iyi performans göstermektedir. İngilizce verilerde ise alan sınıflandırma temelli yaklaşım kullanılarak duygu sınıflandırma işlemi çok az bir kayıpla yapılabilmektedir.

4.5.3. Hedef verinin mevcut kaynak alanlardan farklı bir alana ait olma durumunda duygu sınıflandırma

Duygu sınıflandırma işleminde, hedef verinin alanının bilinmediği bir durumda, ele alınacak ikinci senaryo, hedef verinin alanının mevcut kaynak alanlardan farklı olması durumudur. Böyle bir durumda ise uygulanabilecek iki yaklaşım belirtilmişti. İlk olarak, mevcut alanlara ait etiketli verilerin tamamı kullanılarak eğitilen alan bağımsız bir duygu sınıflandırıcı oluşturulabileceği ve hedef verinin bu sınıflandırıcı ile sınıflandırılabileceği ifade edilmişti. İkinci bir çözüm olarak, alanı bilinmeyen hedef verinin, alan sınıflandırıcı kullanarak kendi alanından farklı olan mevcut alanlardan içerisinden hangisine dâhil edilebileceğinin bulunabileceği ve bunun sonucunda hedef verinin dâhil edilen alana ait duygu sınıflandırıcıda sınıflandırılabileceği belirtilmişti.

Öncelikle ilk yaklaşım olan mevcut kaynak alan verilerini kullanarak alan bağımsız bir duygu sınıflandırıcı oluşturmak amacıyla deneyler gerçekleştirilmiştir. Bu deneyleri sonucu, daha sonra alan sınıflandırıcı temelli yaklaşım sonuçlarının performans değerlendirilmesi için ölçüt olacaktır. Her iki deney Türkçe ve İngilizce olmak üzere iki farklı dilde gerçekleştirilmiş ve 5 farklı alana ait veriler kullanılmıştır. Deneyler farklı sınıflandırma algoritmaları ve ağırlıklandırma yöntemleri ile tekrarlanmış ve karşılaştırılmıştır.

Alan bağımsız bir duygu sınıflandırıcı oluşturmak amacıyla, Türkçe verilerle yapılan deneyde her bir alan, diğer alanlara etiketli verilerle eğitilmişlerdir. Her bir alan için, diğer alanlardan 70’er pozitif 70’er negatif olmak üzere, dört alandan toplam 560 eğitim verisi ile sınıflandırıcı eğitilmiş ve her bir alana ait test verileri ile bu deney 5 kez tekrarlanmıştır. Eğitim verisinin parçalara ayrılmasının sebebi, eğitim için kullanılan veri miktarının, Çizelge 4.3’te sonucu

belirtilen deneylerdeki eğitim veri miktarına eşit tutulmaya çalışılarak, sonuçların daha doğru yapılabilmesidir. Deneyler, NB ve SVM sınıflandırıcılarıyla, farklı ağırlıklandırma yöntemleri ve öznitelik sayıları ile tekrarlanmıştır. Türkçe veriler için sonuçlar Çizelge 4.21'de gösterilmiştir. İfade edilen bu sonuçlar, Çizelge 4.3'teki en iyi sonuçlara karşılık gelen ağırlıklandırma yöntemleri ve öznitelik sayıları kullanılarak elde edilen sonuçları belirtmektedir. SVM kullanılarak elde edilen sonuçların NB ile edilen sonuçlara göre daha iyi olduğu görülmektedir.

Çizelge 4.21'de her bir hedef alanın sınıflandırılması için kullanılan eğitim verilerine ait kaynaklar gösterilmektedir. Örneğin; bilgisayar alanına ait veriler, kozmetik, oto aksesuar, telefon ve film alanına ait veriler kullanılarak sınıflandırılmıştır.

Çizelge 4.21. Türkçe verilerin diğer kaynak alanları kullanılarak sınıflandırma sonuçları (%)

Hedef Alan	Kaynak Alanlar	NB	SVM
Bilgisayar	Kozmetik, Oto Aksesuar, Telefon, Film	73,63	74,74
Kozmetik	Bilgisayar, Oto Aksesuar, Telefon, Film	75,29	76,37
Oto Aksesuar	Bilgisayar, Kozmetik, Telefon, Film	74,11	77,89
Telefon	Bilgisayar, Kozmetik, Oto Aksesuar, Film	74,26	74,91
Film	Bilgisayar, Kozmetik, Oto Aksesuar, Telefon	66,69	64,57

Bu sonuçlar, Çizelge 4.3'teki sonuçlarla karşılaştırılacak olursa, örneğin; Çizelge 4.3'te, bilgisayar alanına ait veriler kendi alanında sınıflandırıldığında, %78,14 doğruluk oranı elde edilirken, Çizelge 4.21'de ise alan bağımsız sınıflandırıcı ile %73,63 doğruluk oranı elde edilmiştir. Film verilerinde ise kendi alanında %80 doğrulukla sınıflandırma yapılırken, diğer alanların birleşimi ile oluşturulan alan bağımsız sınıflandırıcı ile %66,69 doğruluk elde edilmiştir. Birbirlerine benzemeleri açısından diğer alanlardan farklı olan film alanında, alan bağımsız kullanarak elde edilen doğruluğun düşük olduğu görülmektedir. Diğer yandan, telefon verileri ele alındığında, Çizelge 4.3'te kendi alanında %72,14'lük doğruluk oranı elde edilirken, alan bağımsız sınıflandırıcı sonucunda ise %74,11 doğrulukla sınıflandırıldığı görülmüştür.

Aynı deneyler İngilizce veriler için de gerçekleştirilmiştir. Her bir alan için, sınıflandırıcı, diğer alanlardan 200'er pozitif ve 200'er negatif olmak üzere,

dört alandan toplam 1.600 eğitim verisi ile eğitilmiş ve her bir alana ait veriler ile test edilmiştir. NB ve SVM sınıflandırıcılarıyla, farklı ağırlıklandırma yöntemleri ve öznitelik sayıları ile deneyler tekrarlanmış, sonuçlar Çizelge 4.22'de gösterilmiştir. Sonuçlar, Çizelge 4.6'daki en iyi sonuçlara karşılık gelen ağırlıklandırma yöntemleri ve öznitelik sayıları kullanılarak elde edilen sonuçları göstermektedir.

Sonuçlar incelendiğinde, örneğin; kitap alanına ait veriler, oluşturulan alan bağımsız sınıflandırıcıda sınıflandırılmış ve Çizelge 4.22'de görüldüğü gibi NB sınıflandırıcı ile %72,65 doğrulukla sınıflandırılmıştır. Mutfak alanına ait veriler ise %77,82 doğrulukla diğer alanlara göre en iyi performansı göstermiştir. Çizelge 4.7'de verilerin kendi alanlarına ait sınıflandırma sonuçlarında ise kitap alanı için %80,15 doğruluk elde edilirken, mutfak alanında ise %83,70 doğruluk sağlanmıştır. Bu sonuçlar Türkçe için yapılan deneylerle karşılaştırıldığında, telefon ve film alanında olduğu gibi çok farklı sonuçlarla karşılaşılmamıştır.

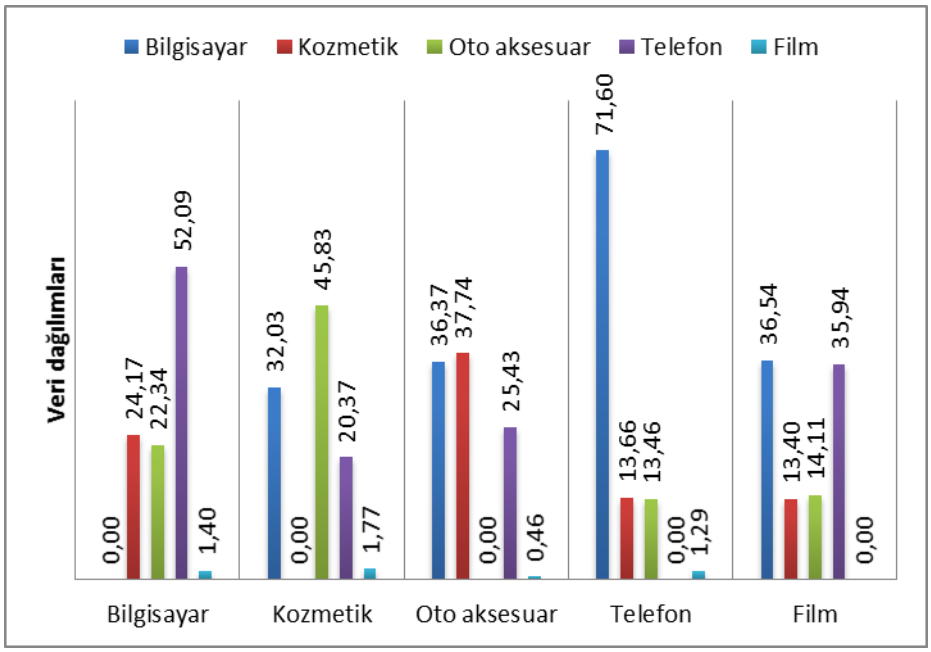
Çizelge 4.22. İngilizce verilerin diğer kaynak alanları kullanılarak sınıflandırma sonuçları (%)

Hedef Alan	Kaynak Alanlar	NB	SVM
Kitap	DVD, Elektronik, Sağlık, Mutfak	72,65	71,59
DVD	Kitap, Elektronik, Sağlık, Mutfak	74,82	75,46
Elektronik	Kitap, DVD, Sağlık, Mutfak	75,12	77,53
Sağlık	Kitap, DVD, Elektronik, Mutfak	76,13	77,5
Mutfak	Kitap, DVD, Elektronik, Sağlık	77,82	78,75

Sonuç olarak, alanı bilinmeyen bir verinin alanının mevcut kaynaklar alanlardan farklı olması durumunda, alan bağımsız bir sınıflandırıcı oluşturmak, uygulanabilir bir yaklaşım olarak görülmektedir. Alan bağımsız sınıflandırıcı ile elde edilen sonuçlar ile en iyi sonuçlar arasında Türkçe için ortalama %5, İngilizce için ise %6'lık bir fark oluştuğu görülmektedir. Ancak, hedef verinin sınıflandırılması için mevcut alanların tamamı kullanılarak oluşturulan alan bağımsız sınıflandırıcı yerine mevcut kaynak alanların içerisinde hedef alana en yakın alanı bulmak ve bulunan alana ait sınıflandırıcıyı kullanmanın daha iyi bir başarı göstereceği düşünülmüştür. Bu amaçla, önerilen alan sınıflandırıcı temelli

yaklaşımda, ilk olarak alanı bilinmeyen veriye en yakın alan tespit edilmeye çalışılmıştır. Alan sınıflandırıcı, beş farklı alana ait verilerin her birini diğer dört alandan birine sınıflandırmaktadır.

Türkçe ve İngilizce için yapılan deneylerde, Türkçe veriler için alan sınıflandırıcısının eğitilmesi amacıyla her bir alana ait 140'er yorum olmak üzere toplamda 560 yorum kullanılmıştır. Alan sınıflandırma için hem NB hem de SVM algoritmaları kullanılmış ve deney her bir alana ait 700 yorumu eğitim için kullandığından beş defa tekrarlanmış, sonuçlar Şekil 4.5 ve Şekil 4.6'da gösterilmiştir. Test edilen her bir alan için alan sınıflandırma sonuçları gösterilmekte olup, test verilerinin kendi alanına ve diğer alanlara dağılımları belirtilmiştir.



Şekil 4.5. Türkçe veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma sonuçları (%): NB sınıflandırıcı

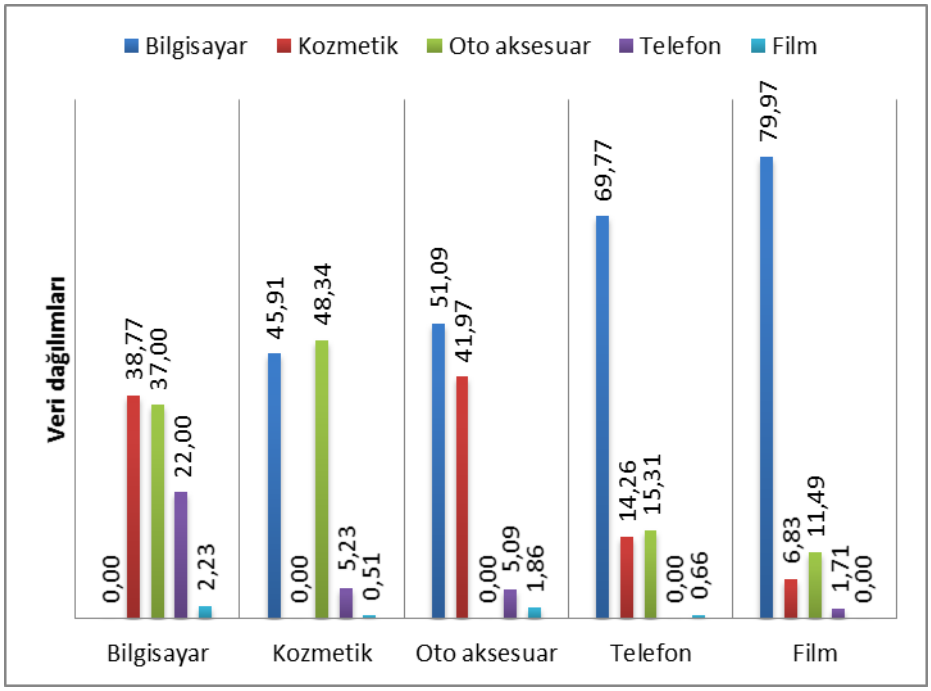
Örneğin, Şekil 4.5'te; alan sınıflandırma sonucunda, bilgisayar alanına ait verilerin %52,09'u telefon alanına ait olarak sınıflandırılırken, %24,17'si kozmetik alanına, %22,34'ü oto aksesuar ve %1,40'ı film alanına sınıflandırılmıştır. Her bir alan için sonuçlar incelendiğinde, alan sınıflandırma sonucunda belirli alanlara dağılımın daha fazla olduğu görülmektedir. Örneğin,

kozmetik alanına ait veriler; oto aksesuar ve bilgisayar alanlarına, oto aksesuar verileri; bilgisayar ve kozmetik alanlarına, telefon verileri; bilgisayar alanına, film verileri ise bilgisayar ve telefon alanlarına sınıflandırma eğiliminde oldukları görülmektedir.

Çizelge 4.3'te her bir alanının kendi alanı ve diğer alanlara ait duygu sınıflandırıcılarındaki sınıflandırma sonuçları gösterilmişti. Bu sonuçlarda bazı alanların birbirlerine daha yakın sonuçlar verdiği ifade edilmişti. Şekil 4.5'te sonuçları incelendiğinde de Çizelge 4.3'teki sonuçları destekleyen sonuçlar olduğu görülmektedir. Sadece film alanına ait alan sınıflandırma sonucu tam olarak eşleşmekte, bu sonuç da film alanının diğer alanlara göre tamamen farklı bir yapıda olmasından kaynaklanmaktadır.

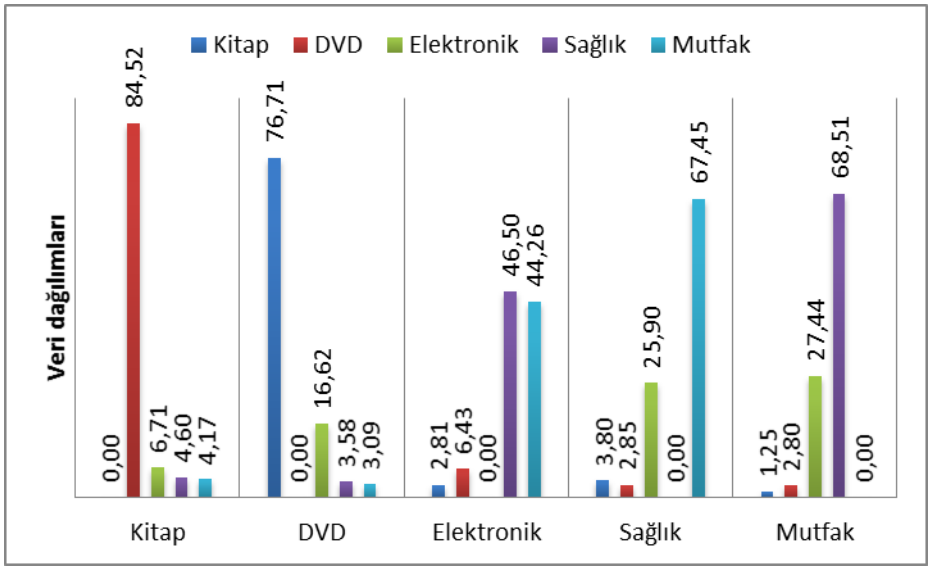
Türkçe veriler için alan sınıflandırma işlemi, SVM algoritması kullanılarak tekrarlanmıştır. Şekil 4.6'da bu deneyin sonuçları gösterilmektedir. NB ile elde edilen sonuçlarla karşılaştırıldığında bu sonuçların yeterince başarılı olmadığı görülmüştür. Örneğin, Çizelge 4.4'teki sonuçlarla karşılaştırılacak olursa, Çizelge 4.4'te bilgisayar alanına ait veriler, kendi alanı dışındaki alanlarda sırasıyla, telefon, oto aksesuar ve kozmetik alanlarında daha başarılı olmuştur. Şekil 4.6'ya bakıldığında ise aynı dağılımın elde edilemediği görülmüştür. Diğer alanlar için de benzer durumların söz konusu olduğu görülmüştür. Ancak bazı durumlarda daha iyi bir sonuç elde edildiği de söylenebilir. Örneğin, Çizelge 4.4'te telefon verilerinin kendi alanından ziyade bilgisayar alanında daha başarılı olduğu görülmektedir. Alan sınıflandırma sonucunda ise telefon verilerinin %69,77'sinin bilgisayar alanına sınıflandırıldığı görülmektedir.

Genel bir değerlendirme yapılacak olursa, hedef verinin alanının mevcut kaynak alanlar arasında olmaması durumunda, Türkçe veriler için gerçekleştirilen alan sınıflandırma sonucunda NB algoritması iyi bir sınıflandırma gerçekleştirirken, SVM sınıflandırma algoritmasının tutarlı bir ayırım yapamadığı görülmüştür.



Şekil 4.6. Türkçe veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma sonuçları (%): SVM sınıflandırıcı

Alan sınıflandırma işlemi İngilizce veriler içinde tekrarlanmış ve alan sınıflandırıcısının eğitilmesi amacıyla her bir alana ait 400'er yorum olmak üzere toplamda 1.600 yorum kullanılmıştır. Alan sınıflandırma için hem NB hem de SVM algoritmaları kullanılmış ve deneyde her bir alana ait 2.000 yorum eğitim için kullandığından, deney 5 defa tekrarlanmış, sonuçlar Şekil 4.7 ve Şekil 4.8'de gösterilmiştir. Test edilen her bir alan için alan sınıflandırma sonuçları gösterilmekte olup, test verilerinin kendi alanına ve diğer alanlara dağılımları belirtilmiştir.



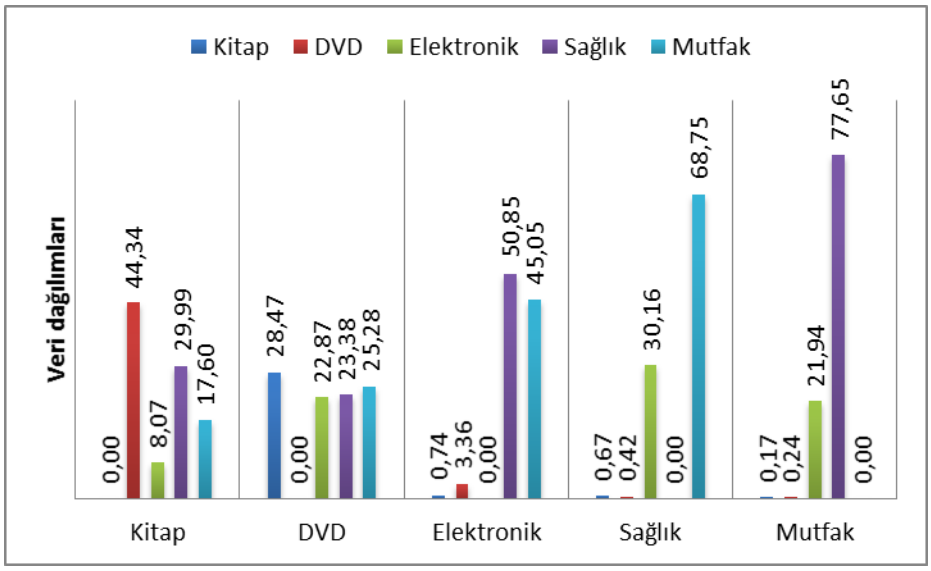
Şekil 4.7. İngilizce veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma sonuçları (%): NB sınıflandırıcı

Örneğin, Şekil 4.7'de, kitap alanına ait verilerin %84,52'si DVD alanına ait olarak sınıflandırılırken, %6,71'i elektronik alanına, %4,60'ı sağlık ve %4,17'si mutfak alanına sınıflandırılmıştır. Sonuçlar incelendiğinde, alan sınıflandırma sonucunda her bir alan için, belirli alanlara dağılımın daha yüksek olduğu görülmektedir. Örneğin, DVD alanına ait verilerin; çoğunlukla kitap alanına, elektronik verilerinin; sağlık ve mutfak alanlarına, sağlık verilerinin; mutfak alanına, mutfak verilerinin ise sağlık alanına sınıflandırma eğilimlerinin yüksek oldukları görülmektedir.

Çizelge 4.7'de İngilizce veriler için NB algoritması kullanılarak her bir alanının kendi alanı ve diğer alanlara ait duygu sınıflandırıcılarındaki sınıflandırma sonuçları gösterilmişti. Bu sonuçlarda bazı alanların birbirlerine daha yakın sonuçlar verdiği ifade edilmişti. Şekil 4.7'deki sonuçlar incelendiğinde de Çizelge 4.7'deki sonuçları desteklediği görülmüştür.

Alan sınıflandırma işlemi, SVM algoritması kullanılarak tekrarlanmış ve deney sonuçları Şekil 4.8'de gösterilmiştir. Elde edilen sonuçlar, Şekil 4.7'de NB için elde edilen sonuçlarla karşılaştırıldığında daha az başarılı olduğu görülmektedir. Örneğin, Şekil 4.8'de, kitap alanına ait verilerin %44,34'si DVD alanına ait olarak sınıflandırılırken, %29,99'u sağlık alanına, %17,60'ı mutfak ve %8,07'si elektronik alanına sınıflandırılmıştır. Çizelge 4.8'deki sonuçlara

bakıldığında ise DVD alanında iyi bir doğruluk elde edilmişken, hemen hemen yakın değerlere sahip diğer alanlarda ise iyi bir başarı gösteremediği görülmüştür. Dolayısıyla, Şekil 4.8'de kitap verilerinin DVD alanına daha yüksek oranla sınıflandırılması beklenmekteydi. Bunun yanında elektronik, sağlık ve mutfak alanları için elde edilen sonuçların NB ve SVM algoritmalarının her ikisi için de benzer sonuçlar ürettiği görülmüştür. Ancak genel olarak karşılaştırıldığında İngilizce veriler için gerçekleştirilen alan sınıflandırma sonucunda NB algoritması iyi bir sınıflandırma gerçekleştirirken, SVM bazı alanlarda yeterli başarıyı gösterememiştir.



Şekil 4.8. İngilizce veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma sonuçları (%): SVM sınıflandırıcı

Alanı bilinmeyen ve mevcut kaynak alanlardan farklı olan hedef verinin alan sınıflandırıcı sonucunda hangi alana dâhil edileceği tespit edildikten sonra duygu sınıflandırma işlemi gerçekleştirilmektedir. Türkçe için her bir alan verisi Şekil 4.5 ve Şekil 4.6'daki sonuçlara göre alan sınıflandırma sonucunda belirli alanlara dağılmıştı. Hedef verinin en yakın olduğu alan tespit edildikten sonra bir sonraki aşamada, hedef veri, atanmış olduğu alana ait duygu sınıflandırıcıda sınıflandırılmaktadır. Türkçe veriler için her bir alana ait alan sınıflandırma temelli yaklaşım sonuçları NB ve SVM sınıflandırma algoritmaları için sırasıyla Çizelge 4.23 ve Çizelge 4.24'te gösterilmektedir. Çizelge 4.24'te alan

sınıflandırma temelli yaklaşıma ait sonuçların yanı sıra, karşılaştırma yapabilmek amacıyla daha önceki bölümlerde elde edilen sonuçlar da ayrıca belirtilmiştir.

Çizelge 4.23'te belirtilen diğerlerinin ortalama sonuçları, Çizelge 4.3'te her bir alan için diğer alanlara ait sınıflandırıcılardaki sınıflandırma sonuçlarının ortalamasını ifade etmektedir. Çizelge 4.21'deki alan bağımsız sınıflandırıcı sonuçları da bu tabloda tekrar belirtilmiştir. Çizelge 4.23'de her bir alan için diğer alanlara ait sonuçlardan elde edilen en iyi sonuçlar, "diğerlerinin en iyisi" başlığı altığında gösterilmiştir. Çizelge 4.23'de her bir alan için kendi alanında sınıflandırılma durumunda elde edilen sonuçlar ise aynı alana ait sonuçlar olarak ifade edilmiştir.

Çizelge 4.23. Türkçe veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma temelli yaklaşım ile diğer sonuçların karşılaştırılması (%): NB sınıflandırıcı

Hedef Alan	Diğerlerinin Ortalaması	Alan Bağımsız Sınıflandırıcı	Diğerlerinin En İyisi	Aynı Alana Ait Sonuç	Alan Sınıflandırma Temelli Yaklaşım
Bilgisayar	72,48	73,63	76,80	78,14	76,34
Kozmetik	74,42	75,29	79,54	82,00	77,23
Oto Aksesuar	73,35	74,11	79,17	77,43	77,09
Telefon	73,27	74,26	76,43	72,14	76,06
Film	68,06	66,69	70,17	80,00	67,86

Çizelge 4.23'teki sonuçlar incelenecek olursa, hedef verinin alan bağımsız sınıflandırıcı ile sınıflandırılmasında sonucunda elde edilen sonuçlar, hedef verinin diğer kaynak alanların her birinde ayrı ayrı sınıflandırıldığında elde edilen ortalama doğruluk değerinden çoğunlukla yüksek olduğu görülmektedir. Aynı alana ait sonuçlar, daha önceden de bahsedildiği gibi ulaşılması hedeflenen en üst değerleri belirtmektedir. Ancak oto aksesuar ve telefon verileri için elde edilen deney sonucunda bu kuralın dışına çıkıldığı görülmektedir. Çizelge 4.23'te özellikle telefon alanına ait sonuçların tamamına bakıldığında kendi alanına ait sonuçların diğer tüm sonuçlardan düşük olduğu görülmektedir. Hedef verinin kendinden farklı kaynaklar arasında elde edilen en iyi sonuçlar ile alan bağımsız

sınıflandırıcı sonuçları karşılaştırılacak olursa, alan bağımsız sınıflandırma sonuçları, en iyi sonuçlar arasındaki farkın %3,5 olduğu görülmektedir.

Hedef verinin alanının kaynak alanlardan farklı olması durumunda, hedef verinin, diğer alanlar içerisinde en iyi sonucun elde edildiği alanda sınıflandırma durumunda, en iyi sonucun elde edilebileceği ifade edilebilir. Örneğin, Çizelge 4.23'te diğerlerinin en iyi sonuçlarına Çizelge 4.3'teki sonuçlar ile birlikte bakıldığında; bilgisayar, kozmetik, oto aksesuar, telefon ve film verileri için sırasıyla; telefon, oto aksesuar, kozmetik, bilgisayar ve kozmetik alanlarının duygu sınıflandırma için en iyi sonuçları verdikleri görülmektedir. Çizelge 4.4'teki alan sınıflandırma sonuçlarına bakıldığında, her bir alan için bahsedilen en iyi alanlara yapılan sınıflandırma ne kadar yüksekse, alan sınıflandırıcı temelli yaklaşım sonuçları, en iyi sonuçlara o kadar yaklaşmış olur. Alan sınıflandırma temelli yaklaşım sonuçları ile diğer alanların en iyi sonuçları karşılaştırıldığında, bilgisayar ve telefon alanlarına ait sonuçların en iyi sonuçlara oldukça yaklaşıldığı, kozmetik oto aksesuar ve film alanlarında ise alan bağımsız sınıflandırıcı sonuçlarına göre daha iyi doğruluk oranlarının elde edildiği görülmektedir. Alan bağımsız sınıflandırma sonucu ile en iyi sonuçlar arasındaki %3,5'lik fark, alan sınıflandırma temelli yaklaşım ile %1,5'e düşürülmüştür.

Çizelge 4.23'te Türkçe veriler için NB sınıflandırma algoritması ile gerçekleştirilen deneyler, Çizelge 4.24'te SVM sınıflandırma algoritması ile tekrarlanmıştır. Sonuçlar Çizelge 4.23'teki sonuçlarla çoğunlukla benzerlik gösterirken bazı durumlarda farklı sonuçlar elde edilmiştir. Diğer alanların ortalama sonuçları ve alan bağımsız sınıflandırma sonucunda SVM sınıflandırıcının genellikle daha iyi sonuç verdiği görülmektedir. Hedef verinin kendinden farklı kaynaklar arasında elde edilen en iyi sonuçlar ile alan bağımsız sınıflandırıcı sonuçları karşılaştırılacak olursa, alan bağımsız sınıflandırma sonuçları, en iyi sonuçlar arasındaki farkın %1,9 olduğu görülmektedir. Alan sınıflandırma temelli yaklaşım sonuçlarında ise kozmetik ve oto aksesuar alanları için NB sınıflandırıcıya göre daha yüksek sonuçlar elde edilmiştir. Alan bağımsız sınıflandırma sonucu ile en iyi sonuçlar arasındaki %1,9'luk fark, alan sınıflandırma temelli yaklaşım ile %1,3'e düşürülmüştür.

Çizelge 4.24. Türkçe veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma temelli yaklaşım ile diğer sonuçların karşılaştırılması (%): SVM sınıflandırıcı

Hedef Alan	Diğerlerinin Ortalaması	Alan Bağımsız Sınıflandırıcı	Diğerlerinin En iyisi	Aynı Alana Ait Sonuç	Alan Sınıflandırma Temelli Yaklaşım
Bilgisayar	72,63	74,74	77,14	78,71	75,60
Kozmetik	74,42	76,37	78,40	80,86	78,11
Oto Aksesuar	74,18	77,89	78,14	78,00	77,37
Telefon	71,87	74,91	75,31	73,43	75,31
Film	66,46	64,57	68,86	78,71	64,91

Türkçe veriler için yapılan deneyler, İngilizce veriler için de tekrarlanmış, NB ve SVM algoritmaları kullanılarak gerçekleştirilen deneylerin sonuçları sırasıyla Çizelge 4.25 ve Çizelge 4.26'da gösterilmiştir.

Çizelge 4.25. İngilizce veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma temelli yaklaşım ile diğer sonuçların karşılaştırılması (%): NB sınıflandırıcı

Hedef Alan	Diğerlerinin ortalaması	Alan Bağımsız Sınıflandırıcı	Diğerlerinin En iyisi	Aynı Alana Ait Sonuç	Alan Sınıflandırma Temelli Yaklaşım
Kitap	70,45	72,65	76,71	80,15	75,80
DVD	72,85	74,82	76,38	80,30	76,00
Elektronik	71,76	75,12	78,27	80,65	77,43
Sağlık	71,39	76,13	78,04	81,45	76,97
Mutfak	73,89	77,82	81,03	83,70	80,27

Çizelge 4.25'te Türkçe deney sonuçlarında olduğu gibi, alan sınıflandırma temelli yaklaşıma ait sonuçların yanı sıra, karşılaştırma yapabilmek amacıyla daha önceki bölümlerde elde edilen sonuçlar da ayrıca belirtilmiştir. Çizelge 4.25'deki sonuçlar incelendiğinde, alan bağımsız sınıflandırıcı ile elde edilen sonuçların, diğer kaynak alanların her biri için elde edilen sonuçların ortalama doğruluk değeri arasında yaklaşık %2,8'lik fark olduğu görülmektedir. Aynı alana ait sonuçlar, İngilizce veriler için her bir alana ait en yüksek sonucu göstermektedir.

Hedef verinin alanının kaynak alanlardan farklı olması durumunda, duygu sınıflandırma işleminin yüksek doğrulukla yapılabilmesi için diğer alanlar içerisinde en iyi sonucun elde edildiği alanda sınıflandırma yapılması gerekmektedir. Örneğin, Çizelge 4.25 ve Çizelge 4.7'deki sonuçlar birlikte ele alındığında; kitap, DVD, elektronik, sağlık ve mutfak alanına ait veriler için sırasıyla; DVD, kitap, mutfak, mutfak ve sağlık alanlarının duygu sınıflandırma için en iyi sonuçları verdikleri görülmektedir. Alan sınıflandırma sonuçlarına bakıldığında, her bir alan için en iyi sonuçların elde edildiği alanlara yapılan sınıflandırma ne kadar yüksek olursa, alan sınıflandırıcı temelli yaklaşım sonuçları, en iyi sonuçlara o kadar yaklaşmış olur.

Alan sınıflandırma temelli yaklaşım sonuçları ile diğer alanların en iyi sonuçları karşılaştırıldığında, alan sınıflandırma temelli yaklaşım sonucunda her bir alan için en iyi sonuçlara oldukça yaklaşıldığı görülmektedir. Alan bağımsız sınıflandırma sonucu ile en iyi sonuçlar arasındaki %2,8'lik fark, alan sınıflandırma temelli yaklaşım ile %0,8'e düşürülmüştür.

İngilizce veriler için NB sınıflandırma algoritması ile gerçekleştirilen deneyler, SVM sınıflandırma algoritması ile tekrarlanmış ve sonuçları Çizelge 4.26'da gösterilmiştir. Çizelge 4.25'teki sonuçlarla karşılaştırıldığında SVM ile gerçekleştirilen deney sonuçlarının daha başarılı olduğu görülmektedir. Hedef verinin kendinden farklı kaynaklar arasında elde edilen en iyi sonuçlar ile alan bağımsız sınıflandırıcı sonuçları karşılaştırılacak olursa, alan bağımsız sınıflandırma sonuçları, en iyi sonuçlar arasındaki farkın yaklaşık %2,7 olduğu görülmektedir. Alan sınıflandırma temelli yaklaşım ile elde edilen sonuçlara bakıldığında alan sınıflandırma temelli yaklaşım ile en iyi sonuçlar arasındaki farkın yaklaşık %1,2 olduğu görülmektedir.

Çizelge 4.26. İngilizce veriler için hedef verinin mevcut alanlardan farklı olması durumunda alan sınıflandırma temelli yaklaşım ile diğer sonuçların karşılaştırılması (%): SVM sınıflandırıcı

Hedef Alan	Diğerlerinin Ortalaması	Alan Bağımsız Sınıflandırıcı	Diğerlerinin En iyisi	Aynı Alana Ait Sonuç	Alan Sınıflandırma Temelli Yaklaşım
Kitap	69,41	71,59	76,10	79,35	73,30
DVD	73,40	75,46	77,78	80,60	74,73
Elektronik	74,59	77,53	79,95	82,05	79,05
Sağlık	74,91	77,50	79,16	82,90	79,10
Mutfak	76,32	78,75	81,60	84,55	**82,19**

Çizelge 4.26'da kitap ve DVD alanları için diğer alanların en iyi sonuçlarıyla alan sınıflandırma temelli yaklaşım sonuçları karşılaştırıldığında, alan sınıflandırıcı temelli yaklaşım ile sınıflandırma başarısının düşmüş olduğu gözlemlenmektedir. Bu durum Şekil 4.8'de görüldüğü gibi alan sınıflandırma işleminde SVM sınıflandırma algoritmasının kitap ve DVD alanında başarısız olmasından kaynaklanmaktadır. Bu iki alan dışında alan sınıflandırma temelli yaklaşımı ile en iyi sonuçlara oldukça yaklaşıldığı görülmektedir. Mutfak alanına ait verilerin sınıflandırılmasında, alan sınıflandırma temelli yaklaşım sonucunun diğer alanların en iyi sonucunu geçtiği görülmüş ve çizelgede kalın olarak ifade edilmiştir. Mutfak alanı için en iyi sonuç Çizelge 4.8'de görüleceği üzere sağlık alanında elde edilmişti. Şekil 4.8'de alan sınıflandırılma işleminde, mutfak alanına ait veriler %100 olarak sağlık alanına sınıflandırılmış olsaydı %81,6 doğruluk oranı elde edilecekti. Şekil 4.8'de mutfak verilerinin %21,94'ünün elektronik alanına sınıflandırılması ile birlikte, bazı verilerin sağlık alanı yerine elektronik alanında sınıflandırılmasının başarıyı arttırdığı görülmüştür.

5. SONUÇLAR VE DEĞERLENDİRME

Bu tezde, duygu sınıflandırma işleminde karşılaşılabilecek bazı durumlar ele alınmış ve bu duruma çözüm olarak yeni bir metot geliştirilmiş ve geliştirilen alan sınıflandırıcı temelli duygu sınıflandırma metodunun uygulanabilirliği, yapılan birçok deney ile ölçülmüştür. Türkçe ve İngilizce diline ait 5 farklı alandan veriler üzerinde yapılan çalışmada, NB ve SVM sınıflandırıcıları kullanılmış, alanı bilinmeyen bir verinin duygu sınıflandırma işlemi gerçekleştirilmiştir.

Alanı bilinmeyen bir veri için yapılan duygu sınıflandırma işleminde, hedef alanın mevcut kaynak alanlar arasında bulunması durumunda, önerilen metodun Türkçe veriler için kullanılması ile verinin alanı bilinmesi durumunda elde edilecek en iyi sonuçlara ulaşıldığı ve hatta az da olsa sonuçları geçtiği görülmüştür. İngilizce verilerde ise hemen hemen en iyi sonuçlara ulaşan sonuçlar elde edilmiştir. Alanı bilinmeyen verinin alanının mevcut alanlardan farklı olduğu durumda gerçekleştirilen deneylerde ise, alan sınıflandırma temelli duygu sınıflandırma metodu ile Türkçe ve İngilizce veriler için mevcut kaynak alanlar arasında elde edilen en iyi sonuçlara çok yakın sonuçlar elde edilmiştir.

Önerilen metodun ilk aşaması olan alan sınıflandırma aşamasında, mevcut kaynak alanlar içerisinde olan hedef veri, alan sınıflandırıcı ile yüksek doğrulukla kendi alanına sınıflandırılabilmiştir. Mevcut alanlardan farklı olan hedef verinin en yakın olduğu alan ise, alan sınıflandırıcı ile tespit edilebilmiş ve böylelikle duygu sınıflandırma başarısı arttırılmıştır. Alan sınıflandırma konusunda NB sınıflandırma algoritmasının SVM'e göre daha başarılı olduğu görülmüştür.

Ayrıca, yapılan deneyler sonucunda Türkçe verilerin bazı alanlarına ait etiketli verilerinde tutarsızlık olduğu saptanmış ve bu tutarsızlığın neden olduğu yanlış sınıflandırmaların, alan sınıflandırıcı temelli yaklaşım ile çözülebildiği görülmüştür.

Önerilen bu yaklaşımın sonraki çalışmalara ışık tutacağına inanılmaktadır. Özellikle, Türkçe dili için bu konuda yapılabilecek birçok çalışmadan bahsedilebilir. Alanı bilinmeyen hedef veri için yapılacak sınıflandırmada, mevcut alanlara ait sınıflandırıcıların sonuçları, hedef alanına benzerlikleri ölçüsünde ağırlıklandırılarak genel sonucun elde edilmesi sağlanabilir.

KAYNAKLAR

Akba F., Uçan, A., Akçapınar Sezer, E. ve Sever, H. (2014), "Assessment of Feature Selection Metrics for Sentiment Analyses: Turkish Movie Reviews", *In Proceedings of the 8th European Conference on Data Mining*, Lisbon, Portugal, 180-184.

Akbaş, E. (2012), *Aspect Based Opinion Mining on Turkish Tweets*, Yüksek Lisans Tezi, Bilkent Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Akın, A.A. ve Akın, M.D. (2007), "*Zemberek, An Open Source NLP Framework for Turkic Languages*", http://zemberek.googlecode.com/.

Albayrak, N.B. (2011), *Opinion and Sentiment Analysis Using Natural Language Processing Techniques*, Yüksek Lisans Tezi, Fatih Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Aue, A. ve Gamon, M. (2005), "Customizing Sentiment Classifiers to New Domains: A Case Study", *In Proceedings of the Recent Advances in Natural Language Processing*, Borovets, Bulgaria, 21-23.

Ben-David, S., Blitzer, J., Crammer, K. ve Pereira, F. (2007), "Analysis Of Representations for Domain Adaptation", *In Proceedings of the 9th Annual Conference on Neural Information Processing Systems* 19, British Columbia, Canada, 137-144.

Blei, D. M., Ng, A. Y., ve Jordan, M. I. (2003), "Latent dirichlet allocation", *The Journal of Machine Learning Research*, **3**, 993-1022.

Blitzer, J., Dredze, M. ve Pereira, F. (2007), "Biographies, Bollywood, Boom-Boxes and Blenders: Domain Adaptation for Sentiment Classification", *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 440-447.

Bollegala, D., Weir, D., ve Carroll, J. (2011), "Using Multiple Sources To Construct A Sentiment Sensitive Thesaurus for Cross-Domain Sentiment Classification", *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, **1**, 132-141.

Boynukalın, Z. ve Karagöz, P. (2013), "Emotion analysis on Turkish texts", *Information Sciences and Systems*, 159-168.

Çetin, M. ve Amasyalı, M.F. (2013a), "Active Learning for Turkish Sentiment Analysis", *In Innovations in Intelligent Systems and Applications*, Albena, Bulgaria, 1-4.

Çetin, M. ve Amasyalı, M.F. (2013b), "Eğiticili ve Geleneksel Terim Ağırlıklandırma Yöntemleriyle Duygu Analizi", *In Signal Processing and Communications Applications Conference*, Girne.

Demirtaş, E. ve Pechenizkiy, M. (2013), "Cross-Lingual Polarity Detection With Machine Translation", *In Proceedings of the 2nd International Workshop on Issues of Sentiment Discovery and Opinion Mining*, Chicago, USA, 9/1-8

Eroğul, U. (2009), *Sentiment analysis in Turkish*, Yüksek Lisans Tezi, Orta Doğu Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Jiang, J. ve Zhai, C. (2007), "A Two-Stage Approach to Domain Adaptation for Statistical Classifiers", *In Proceedings of the ACM 16th Conference on Information and Knowledge Management*, Lisbon, Portugal, 401-410.

Kaya, M., Fidan, G., ve Toroslu, I. H. (2012), "Sentiment Analysis of Turkish Political News". *In Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, Macau, **1**, 174-180.

Kaya, M., Fidan, G. ve Toroslu, I. H. (2013), "Transfer Learning Using Twitter Data for Improving Sentiment Classification of Turkish Political News", *In Information Sciences and Systems*, 139-148.

Lan, M., Tan, C. L., Su, J. ve Lu, Y. (2009), "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 721-735.

Li, S. ve Zong, C. (2008), "Multi-domain sentiment classification", *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, Columbus, Ohio, USA, 257-260.

Lin, C., He, Y., Everson, R., ve Ruger, S. (2012), "Weakly Supervised Joint Sentiment-Topic Detection From Text", *IEEE Transactions on Knowledge and Data Engineering*, , **24** (6), 1134-1145.

Liu, B. (2010), "Sentiment Analysis and Liu vity", *Handbook of Natural Language Processing*, **2**, 627-666.

Medhat, W., Hassan, A. ve Korashy, H. (2014), "Sentiment Analysis Algorithms and Applications: A Survey", *Ain Shams Engineering Journal*, **5** (4), 1093-1113.

Nizam, H. ve Akın, S.S. (2014), "Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması", *XIX. Türkiye'de İnternet Konferansı*, İzmir.

Özsert, C. M. ve Özgür, A. (2013), "Word Polarity Detection Using A Multilingual Approach", *In Computational Linguistics and Intelligent Text Processing*, 75-82.

Pang, B., Lee, L. ve Vaithyanathan, S. (2002), "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques", *In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, 79-86.

Pan, S. J., Ni, X., Sun, J. T., Yang, Q. ve Chen, Z. (2010), "Cross-Domain Sentiment Classification Via Spectral Feature Alignment", *In Proceedings of the 19th International Conference on World Wide Web*, New York, USA, 751-760.

Read, J. (2005), "Using Emoticons to Reduce Dependency in Machine Learning Techniques For Sentiment Classification", *In Proceedings of the ACL Student Research Workshop*, Ann Arbor, Michigan, 43-48.

Sevindi, B.İ. (2013), *Türkçe Metinlerde Denetimli Ve Sözlük Tabanlı Duygu Analizi Yaklaşımlarının Karşılaştırılması,* Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Sun, S., Shi, H. ve Wu, Y. (2015), "A Survey of Multi-source Domain Adaptation", *Information Fusion*, **24**, 84-92.

Taner, B. (2011), *Feature-Based Sentiment Analysis with Ontologies*, Yüksek Lisans Tezi, Sabancı Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Taşlıoğlu, H. (2014), *Irony Detection On Turkish Microblog Texts*, Yüksek Lisans Tezi, Orta Doğu Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Theodoridis, S. and Koutroumbas, K. (2008), "*Pattern Recognition*," Academic Press.

Tutar, K. (2013), *Sosyal Ağlar Üzerinde Ontoloji Tabanlı Sezgi Analizi İçin Bir Uygulama Çatısının Geliştirilmesi*, Yüksek Lisans Tezi, Ege Üniversitesi, Fen Bilimleri Enstitüsü, İzmir.

Uçan, A. (2014), *Otomatik Duygu Sözlüğü Çevirimi Ve Duygu Analizinde Kullanımı*, Yüksek Lisans Tezi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Vural, A.G., Cambazoglu, B.B., Senkul, P. ve Tokgoz, Z.O. (2013), "A Framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish", *Computer and Information Sciences III*, 437-445.

Vural, A.G. (2013), *Sentiment-Focused Web Crawling*, Doktora Tezi, Orta Doğu Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Whitehead, M. ve Yaeger, L. (2009), "Building A General Purpose Cross-Domain Sentiment Mining Model", *In Proceedings of the Computer Science and Information Engineering 2009 WRI World Congress on*, Los Angeles, **4**, 472-476.

Yang, Y. ve Pedersen, J. O. (1997), "A Comparative Study on Feature Selection in Text Categorization", *In Proceedings of the 14th International Conference on Machine Learning*, San Francisco, USA, 412-420.