BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MÜZİK ÜST-VERİ TAHMİNİ İÇİN TÜRKÇE ŞARKI SÖZÜ MADENCİLİĞİ

BAŞAR KIRMACI

YÜKSEK LİSANS TEZİ
2015

MÜZİK ÜST-VERİ TAHMİNİ İÇİN TÜRKÇE ŞARKI SÖZÜ MADENCİLİĞİ

TURKISH LYRICS MINING FOR MUSIC META-DATA ESTIMATION

BAŞAR KIRMACI

Başkent Üniversitesi
Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin
BİLGİSAYAR Mühendisliği Anabilim Dalı İçin Öngördüğü
YÜKSEK LİSANS TEZİ
olarak hazırlanmıştır.

2015

“Müzik Üst-Veri Tahmini İçin Türkçe Şarkı Sözü Madenciliği” başlıklı bu çalışma, jürimiz tarafından, 28 / 07 / 2015 tarihinde, **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI’nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan : Doç. Dr. İrem Soydal

Üye (Danışman) : Doç. Dr. Hasan Oğul

Üye : Yrd. Doç. Dr. Selda Güney

ONAY

..../..../.....

Prof. Dr. Emin AKATA

Fen Bilimleri Enstitüsü Müdürü

TEŞEKKÜR

Sayın Doç. Dr. Hasan Oğul'a bu tez çalışmasının planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda, engin bilgisi ve tecrübesi ile bana yardımcı olmasından dolayı teşekkür eder, saygılarımı sunarım.

Çalışmalarımda bana moral, destek ve anlayış gösteren aileme sonsuz teşekkür ederim.

ÖZ

MÜZİK ÜST-VERİ TAHMİNİ İÇİN TÜRKÇE ŞARKI SÖZÜ MADENCİLİĞİ

Başar KIRMACI

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Müzik geri getirimi, internet ve ilgili teknolojilerin eğlence amaçlı yaygın kullanımı ile birlikte önemli bir problem haline gelmiştir. Kullanıcının aradığı şarkıya daha kolay ulaşabilmesi, aradığı şarkıya benzer diğer şarkıları daha kolay bulabilmesi, dinlemek isteyebileceği şarkıları listeleyebilmesi için müzik geri getirim sistemleri geliştirilmiştir. Uygulanacak yöntem ne olursa olsun müzik nesnelerinin analiz edilmesi ve bu analizlere bağlı olarak müzik nesnelerinin anlamlandırılması gerekmektedir. Müzik analizi ile ilgili bu çalışmalar iki veri türü üzerine yoğunlaşmıştır. Bunlar; müzik geri getirim sistemleri için melodik ve aranjman özniteliklerin kullanıldığı içerik sinyali ve şarkının adı, türü, bestecisi gibi verilerin bulunduğu üst-veri bilgileridir. Şarkı sözü metninin kullanımı çok azdır. Bu çalışma müzik geri getirim uygulamalarında Türkçe şarkı sözü metninden müzik üst-verilerinin tahmin edilebilirliğine dayalı bir altyapı sağlamaktadır. Hazırlanan şarkı sözleri veri kümeleri üzerinden Türkçe metnine ve dilbilgisi yapısına göre öznitelikler seçilmiştir. Seçilen öznitelikler kullanılarak bir makine öğrenme algoritması ile şarkı sözü yazarını, türünü ve yayın tarihini tahmin edebilen bir sistem önerilmiş ve farklı tarzlardaki söz yazarlarından oluşturulan geniş bir şarkı veri kümesinde performansı değerlendirilmiştir. Elde edilen sonuçlar böyle bir yaklaşımın müzik veri madenciliği ve bilgi geri getirimi çalışmalarında faydalı olabileceğini göstermektedir.

ANAHTAR SÖZCÜKLER: Metin sınıflandırma, veri madenciliği, örüntü tanıma, müzik bilgisi geri getirimi, üst-veri analizi, şarkı sınıflandırma.

Danışman: Doç.Dr. Hasan OĞUL, Başkent Üniversitesi, Bilgisayar Mühendisliği Bölümü.

ABSTRACT

TURKISH LYRICS MINING FOR MUSIC META-DATA ESTIMATION

Başar KIRMACI

Başkent University Institue of Science and Engineering

Computer Engineering Department

Music retrieval has become an important problem with the widespread use of internet and related technologies for entertainment purposes. Music retrieval systems were developed for users to find songs they are looking for and similar ones in an easier manner, and list songs they might want to listen. Music objects should be analyzed and interpreted according to those analyses independent of the method that is going to be implemented. These studies on music analysis are mainly focused on two data types; content signal that is based on melodic and musical arrangement properties for music retrieval systems and meta-data information, such as name, genre, composer of the song. The use of lyrics text is very few. This study provides a basis for the prediction of meta-data of music from lyrics text in music retrieval applications. Features were chosen on the song lyrics data sets prepared according to the Turkish text and grammar structure. A system that can predict the writer, genre and relaese date of the song using the chosen features and a machine learning algorithm was presented and its performence on a large song data set generated from song writers with different styles was evaluated. Results show that this kind of an approach might be useful for music data mining and information retrieval studies.

KEYWORDS: Text classification, data mining, pattern recognition, music information retrieval, meta-data analysis, song classification.

Advisor: Assoc. Prof. Dr. Hasan OĞUL, Başkent University, Department of ComputerEngineering.

İÇİNDEKİLER LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

ÇİZELGELER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

SVM	Support Vector Machine
DVM	Destek Vektör Makineleri
DDVM	Doğrusal Destek Vektör Makineleri
NB	Naif Bayes
MNB	Multinom Naif Bayes
MYSQL	My Structured Query Language
WEKA	Waikato Environment for Knowledge Analysis
RTF	Radyal Tabanlı Fonksiyon
LIBSVM	A Library for Support Vector Machines
NLP	Natural Language Processing
FP	False Positive
FN	False Negative
TP	True Positive
TN	True Negative
ROC	Receiver Operating Characteristics
AUC	Area Under The Curve
ARFF	Attribute-Relation File Format
XML	Extensible Markup Language

1. GİRİŞ

İnternet ve ilgili teknolojilerin günlük hayatta kullanımının yaygınlaşması ile birlikte sunulan veri miktarının hızla artışı, var olan veriyi anlamlandırarak kullanıcıya sunabilen akıllı bilgi sistemlerine gereksinimleri artırmıştır. İnternetin en yaygın kullanıldığı alanlardan biri eğlence sektörüdür. Eğlence içeriği oyun, video, müzik gibi çoklu ortam verilerine erişim sağlar. Eğlence amaçlı içeriklere erişim ihtiyacına paralel olarak bilgisayar ve bilişim bilimlerinde bilgi geri-getirimi, veri madenciliği, tavsiye sistemleri gibi alanların yeni çözümler üretmelerine neden olmuştur.

Son zamanlarda insanların müzik dinleme alışkanlıklarındaki değişikliklere yanıt olarak müzik endüstrisindeki modada belirgin bir değişime rastlanmıştır. Bireysel albüm kayıtlarına kıyasla kolektif çevrimiçi mağazalar ve kütüphaneler artık daha popüler hale gelmiştir. Bunun sonucunda, internet üzerinden ulaşılabilir olan müzik veri miktarı son yıllarda açık bir şekilde artmıştır. Kullanıcıların ulaşabileceği, içerikten daha verimli ve rahat bir şekilde keyif alabileceği ve içerikle etkileşebileceği akıllı araçların geliştirilmesi oldukça gereklidir. Bu ihtiyaç mobil cihazların çevrimiçi müzik içeriğine ulaşmasıyla daha belirgin hale gelmiştir. Bu nedenle, son on yılda bu tarz zorlukları aşabilmek için müzik bilgi geri getirim ve öneri sistemleri üzerine yapılan araştırmalar önemli bir şekilde artmıştır [1 - 3].

Müzik çok içerikli bir yapıya sahiptir: ses sinyali, şarkı sözleri ve şarkıcı, besteci, yazar, tür, yayınlanma tarihi ve sosyal veri gibi girdiyle alakalı açıklayıcı bilgi sağlayan diğer metinsel dipnotları içerir. Bu metinsel veri genellikle üst-veri olarak adlandırılmaktadır ve çevrimiçi medyada içeriğe ulaşılması, içeriğin araştırılması ya da düzenlenmesi için müzik girdisinin öz ama faydalı bir şekilde temsil edilmesini sağlar. Müzik bilgisi geri getirimi dijital kütüphanelerde müzik içeriğine bir şekilde ulaşma üzerine yapılan bir çalışmadır. Bu çalışma genellikle müzik nesnelerini temsil etmek için bir özetleme tekniği ve ulaşılabilir depolardaki ilgili müzik girdilerini toplamak amacıyla bir kıyaslama modeli gerektirir. Müzik nesneleri için bu özetleme görevi üst-veri ya da ses içeriği yaklaşımı kullanılarak gerçekleştirilmektedir [4, 5]. Genellikle üst-veri yaklaşımını kullanmayı kısıtlayan pratikte iki temel neden bulunmaktadır. Birincisi, bazı özelliklerin veritabanı yöneticisi veya veriyi sunan kişi tarafından eksik ya da yanlış girilmiş olmasıdır. İkincisi, müzik girdisini düşünülen amaç doğrultusunda karakterize edebilmek için

var olan özniteliklerin yeterli olmamasıdır. Örnek olarak, eğer var olan üst-veri yapısı şarkının yayınlanma günü hakkında bir öznitelik sunmuyorsa kullanıcının belirli bir döneme ait benzer şarkılar edinmek istemesi durumunda bu üst-veri yapısı kullanışsız olmaktadır. Müzik geri getirim çalışmalarında kullanılan diğer bir yöntem; müzik içerisindeki ses sinyal içerikleri ile müziğin tanımlanmasıdır [6 - 9]. Ses, bir müzik girdisinin ana unsuru ve bir öznitelik üreticisi olmasına rağmen, müzik nesnesinin sınıflandırılması sırasında bazı kısıtları vardır. Örnek verilecek olursa, her bir enstrümantal ses, şarkıcının sesi ve arkaplan gürültüsü gibi çeşitli sinyalleri içermektedir. Bundan dolayı, ses içeriğini düzgün bir yapıda elde etmek zor bir görevdir [10]. Doğrusu bir şarkının ses içeriğinden, frekansa ait öznitelikleri hatasız ve kayıpsız bir şekilde elde edecek bir yöntem yoktur [3].

Genel olarak müzik algısı ses içeriğinden oluşan melodik ve akustik içeriklerle temsil edilse de, bir bütün olarak enstrümental olmayan şarkı algısı şarkı sözleri de dahil olmak üzere tüm yöntemleri göz önünde bulunduran bir yapı olarak açıklanabilir. Ses ve şarkı sözlerinin beyinde birbirinden bağımsız bir şekilde işlenerek algımızı tamamlandığı konusunda güçlü bir kanıt vardır [11]. Şarkı sözleri bazen “aşk şarkıları”, “protesto şarkısı” ve “okul şarkıları” gibi belirli türler için ses içeriğinden bağımsız bir içerik özgünlüğü sağlayabilir. Önceki bir çalışmada şarkı sözlerinin sese kıyasla sosyokültürel kavrayışı daha iyi yansıtabildiği de tartışılmıştır [12].

Birçok kavramı karakterize etme potansiyeline rağmen, şarkı sözü bazlı müzik bilgisi geri getirimi ve sınıflandırılması üzerine araştırma çabaları çok azdır. Şarkı sözlerinin belirli bir duygu durumunu vurgulayan sözcüksel öğeler içerebileceği ve aslında altında yatan duygusal durumunu teşhis edebilmek için kullanılabileceği varsayılmıştır [13]. Bu hipotez, duygu durumunun kelime tercihini etkilediğini ve sözcüksel öğelerin duygusal durumu ifade edebileceğini belirten daha eski bir çalışmayla kanıtlanmıştır [14]. Aslında, “mutlu”, “sinirli”, “gülümse” ve “ölü” gibi kelimelerin güçlü duygulu bir sesle hecelenmesine gerek yoktur. Bu bağlamda, şarkıları “mutlu”, “üzgün”, “depresif” ve “tutku” gibi birçok farklı duygu kategorilerine göre sınıflandırmakta şarkı sözleri kullanılmıştır [15, 16]. Bazı çalışmalarda benzer girişimler, şarkılara yumuşak kalpli ve sert kalpli gibi uygun his etiketleri atayan şarkı sözü bazlı şarkı his sınıflandırılması olarak adlandırılmıştır [17]. Şarkı sözünden şarkının türünün tahmin edilebilir olduğu

gösterilmiştir [18, 19]. Nordik dilinde yazılmış bir şarkı sözünden şarkının türünün anlaşılması için bir deneme çalışması sunulmuştur ve bu çalışma araştırmalarımız kapsamında literatürde İngilizce dışında başka bir dilde şarkı sözü bazlı şarkı sınıflandırmayı değerlendiren tek çalışmadır. Bazı çalışmalar şarkı sözü bilgisinin ses öznitelikleriyle birleştirildiğinde duygu durumuna ya da türe göre müzik sınıflandırılmasının netliğini geliştirebildiğini göstermiştir. Burada "Doğal Dil İşleme Kütüphaneleri" ve "Müzik Bilgi Geri Getirim" teknikleri birleştirilerek kullanılmaktadır. Sinirli ve rahat müzik tarzları için ses verisi tek başına belirleyici olabilmektedir; fakat mutlu ve üzgün müzik tarzlarında ses ve sözlerin beraber kullanılması performans üzerinde daha etkilidir [20]. Müziğe ait sembolik ve kültürel kaynaklarda, ses verisi ve şarkı sözleri ile müzik tarzının bulunmasına yönelik çalışmalar bulunmaktadır. Eskiden şarkı sözlerinin müzik tarzı sınıflandırmaya etkisi sembolik, kültürel ve ses verisine göre daha azdı. Ancak bazı yeni özniteliklerin bulunması ve bunların birleştirilmesiyle bu durum değişti. Ayrıca özniteliklerin belirlenmesi aşamasında internet üzerindeki kaynaklardan öznitelik çıkaran çeşitli araçlar bulunmaktadır. Örneğin; "LyricFetcher" ve "jLyrics" gibi uygulamalar internet üzerindeki şarkı sözlerinden öznitelik çıkarmaktadır. Diğer bir konu ise kaynaklarda bulunan şarkı sözlerinin XML ya da başka formlarda standart halde olmaması sebebi ile ortaya gürültü çıkmasıdır [21]. Ses verisi ile şarkı sözlerinin nasıl birleştirilebileceğine ait literatürdeki diğer bir yöntem de multi-modal sınıflandırmadır [22]. Yeni yöntemlerde şarkı tarzlarının sınıflandırılması için otomatik sistemlerde şarkı melodisinin tek başına tahminde yeterli olmayacağından ve genelde birçok farklı müzik bilgi geri getirim yöntemlerinin birleştirilerek kullanılmasının genel performansı artıracağından bahsediliyor. Ses verisinden çıkartılan özniteliklerin direk olarak tek başına kullanılması, şarkıların tarzlarının belirlenmesinde performans için zararlı bir durumdur. Bundan dolayı birden çok içerik tabanlı yöntem kullanılarak sınıflandırma performansı artırılmaktadır. Ayrıca birden çok öznitelik vektörünün beraber kullanılması, tek öznitelik vektörü kullanılan yöntemlere göre daha başarılıdır. Buradaki diğer bir konu ise kullanılacak olan öznitelik vektör gruplarının hepsinin kullanılmasının mı daha etkili olabileceği yoksa bu öznitelik vektör grupları içerisinden sınıflandırmayı etkileyebilecek olanlarının seçilmesi ve sadece o öznitelik gruplarının kullanılmasının mı daha etkili olabileceği sorusudur.

Özniteliklerin belirlenmesi işleminin gerçekleştirilmesi için uygulanan bir yöntem genetik tabanlı algoritmalardır. Genetik tabanlı algoritmalar, öznitelik vektörlerinin kısa ve etkili bir şekilde ifade edilmesi için de kullanılmaktadır. Bu şekilde hangi özniteliklerin sınıflandırma aşamasında daha önemli rol oynadığı belirlenecektir [23]. Metin verisi ile ses verisini birlikte kullanılması için gerçekleştirilen bir diğer yöntem ise, dilbilimsel yapı, yazım stili ve ses verisi üzerinden çıkartılan özniteliklerin bir arada kullanılmasıdır. Metin verisi ve ses verisinin bir arada kullanılması daha az örneklem ihtiyacı ve daha iyi performans sağlamaktadır. Bu yöntemler ile otomatik müzik tarzı sınıflandırma yöntemleri müzik kütüphanelerinde kullanılabilmektedir [24].

Şarkı sözlerinin sınıflandırılması bir metin dokümanının önceden belirlenen kategorilerden birine atanmasını gerektiren metin sınıflandırma probleminin özel bir durumu olarak tanımlanabilir. İnternet sayfası kategorizasyonu [25], spam tespiti [26] ve fikir madenciliği [27] gibi bu görevin farklı ortamlarda birçok örneğiyle karşı karşıya gelmekteyiz. Bizim çalışmamızla daha benzer bir amaç taşımakta olan yazar tanıma, yazılı bir metnin önceden bilinen yazarlardan hangisi olduğunu tahmin etmek için kullanılan diğer bir metin sınıflandırılması uygulamasıdır [28, 29]. Genel olarak, metin sınıflandırılması metin içeriğinin sabit sayıda sayısal özniteliklerle temsil edildiği ve veriyi önceden belirlenen sınıflardan birine atabilen bir makine öğrenme sınıflandırıcısının kurgulandığı bir altyapı üzerine kurulmuştur [30, 31].

Bu çalışmada, önceden belirtilen zorlukları ele almak için sadece şarkı sözlerinden yazar, tür ve yayınlanma tarihi gibi üst-veri niteliklerinin tahmini üzerine yoğunlaşmış bulunmaktayız. Şarkı sözlerinin sınıflandırılması için hazırlanan öznitelik kümeleri içerisinde bulunan bazı öznitelik grupları bu çalışmaya özgü hazırlanmış ve literatürde ilk defa kullanılmıştır. Literatürde şarkı sözlerinden tür ve duygu durumu sınıflandırılması üzerine birkaç girişim vardır. Ancak bu çalışma, araştırmalarımız kapsamında literatürde bulunan şarkı sözlerinden şarkının yazarı ve yayınlanma tarihinin tahmini için ilk girişimdir. Çalışmamızda ayrıca, şarkı sözlerini temsil ettiğine inandığımız çok sayıda yeni öznitelik önermekteyiz. Tez kapsamında aşağıdaki araştırma sorularına cevap aranmıştır:

- Şarkı sözleri müzik içeriğinin temsilinde ne kadar etkilidir?
- Metin içeriğinin temsilinde hangi öznitelikler faydalıdır?
- Kelime kökünün kullanılması temsili ne kadar güçlendirir?
- Şarkı sözlerinin sınıflandırılmasında öznitelik seçim algoritmaları sınıflandırmayı ne kadar güçlendirir?

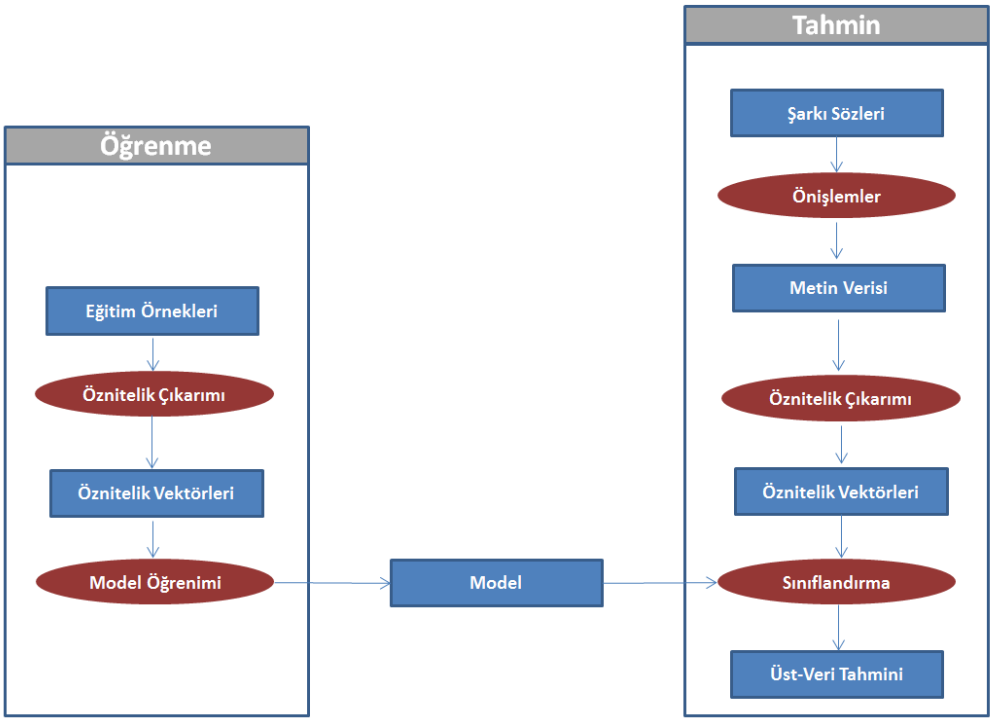
Yukarıda belirtilen bu sorular ile çalışmamızın diğer yöntemler ile olan farkına cevap bulmaya çalışmaktayız. Genellikle, daha önce gerçekleştirilen yöntemler müzik ses verisi ve metin verisini beraber kullanarak sadece müzik tarzının sınıflandırılması üzerine kurulmuştur. Biz ise bu çalışma kapsamında müzik ses verisi olmadan sadece şarkı sözü metin bilgisi ile şarkı sözleri yazarlarının, tarzının ve yıl aralıklarının tahmini üzerine bir çalışma sunmaktayız. Ayrıca yapılan araştırmalar doğrultusunda bu çalışma şarkı sözleri metni kullanarak Türkçe şarkı sözlerinin sınıflandırılması üzerine yapılan ilk çalışmadır. Yapılan araştırmalar sonucunda bu çalışma kapsamında bulunan ve kullanılan yeni öznitelik kümelerinin gelecek çalışmalarda da kullanılabilecek olması elde edinilen kazanımlardan bir tanesidir.

Tez kapsamında yapılan her bir deney adımı şarkı sözü metinlerinin sınıflandırılması için ayrı ayrı bilgi içermektedir. Yapılan çalışmalarda sınıflandırıcı algoritmalarının, öznitelik kümelerinin, kelime kök alma durumunun ve öznitelik seçim algoritmalarının sınıflandırma sonuçlarını ne yönde etkilediği tartışılmıştır. Deneyler sonucunda elde edilen veriler ile müzik bilgi geri getirim çalışmalarına yeni yöntemler ve bilgiler sunulmuştur. Bu çalışmada, müzik tarzının sınıflandırılmasında müzik ses ve metin verisinin bir arada kullanılması yerine sadece metin verisi ile nasıl çıktılar elde edilebileceği tartışılacak. Ayrıca yapılan araştırmalar sonucunda literatürde benzer bir çalışma bulunmayan ve bu çalışma kapsamında ilk defa gerçekleştirilen şarkı sözü metninden söz yazarı ve yıl aralığı sınıflandırmanın ve bu sınıflandırma deneyleri ile elde edilen sonuçların gözlemlenmesi gerçekleştirilecektir.

Bu çalışmada toplanan geniş bir Türkçe şarkı veri kümesi üzerinde titiz bir deneysel plan gerçekleştirip detaylı analizin sonuçlarını sunmaktayız. Deneysel sonuçlar önerilen tekniğin müzik bilgi geri getirimi uygulamalarında tamamlayıcı bir araç olarak kullanılabileceğini öne sürmektedir.

2. YÖNTEMLER

Tez kapsamında gerçekleştirilen ilk iş veri kümesinin hazırlanması olmuştur. İnternet üzerindeki çeşitli kaynaklardan, verilerin karşılaştırılıp doğrulanması ile 1048 adet şarkıdan oluşan veri kümesi hazırlanmıştır. Daha sonra hazırlanan veri kümesinin her bir elemanı için o şarkı öğesine ait bilgi dosyaları oluşturulmuştur. Şarkıya ait bilgi dosyaları içerisinde o şarkının çıktığı yıl, seslendiren, söz, müzik, vb. bilgileri bulunmaktadır. Şarkının bilgilerinin tutulduğu dosya her bir şarkı için hazırlandıktan sonra, şarkı sözü metinleri MySQL veritabanına aktarılmıştır. Şarkı sözlerinin veritabanına aktarılması sırasında Hibernate teknolojisi ile birlikte, tez kapsamında geliştirilen Java uygulamaları kullanılmıştır. Şarkı sözlerinin veritabanına aktarılmasından sonra, şarkı sözü metinleri ve bu şarkılara ait diğer bilgileri kullanarak sınıflandırma işlemi için diğer adımlara geçilmiştir. Şarkı sözlerinden üst-verinin tahmin edilmesi sırasında hazırlanan verileri ve gerçekleştirilen işlemleri gösteren ve sistemin genel yapısı hakkında bilgi veren şema Şekil 2.1'de gösterilmektedir.



Şekil 2.1 Şarkı sözünden söz yazarı tahmini genel görünümü

Hazırlanan şarkı sözleri ilk olarak önişlemlere alınmıştır. Her bir şarkı sözünün bilgileri (şarkı sözü yazarı, bestecisi, yılı, kategorisi, vb.) internet üzerindeki birçok kaynaktan kontrol edilerek, her bir şarkı için ayrı bir dosyada saklanmaktadır.

Daha sonra şarkıya ait bilgilerin bulunduğu dosyadaki veriler veritabanına aktarılmıştır. Şarkılara ait bilgiler veritabanına aktarıldıktan sonra, şarkı sözü metinleri üzerinden öznitelik çıkarım işlemine geçilmektedir. Burada Türkçe'nin yapısına uygun ve şarkı sözlerinin sınıflandırılmasında ayırt edici nitelikte olan öznitelik çıkarımına dikkat edilmiştir. Şarkı sözü metnine özgü olarak bu tez kapsamında bazı öznitelikler hazırlanmıştır. Bu öznitelikler şarkı sözlerinin karakteristik özellikleri hakkında bize bilgi verebilmektedir. Veri kümesi üzerinde incelemeler yapılmış ve bası söz yazarlarının kullandıkları metinsel yapının diğerlerinden farklı olduğu anlaşılmıştır. Örneğin veri seti içerisinde bulunan "Pop", "Rock" ve "Arabesk-Fantezi" müzik türlerini birbirleri arasında satır uzunlukları açısından farklı davranışlar sergilemektedir. "Rock" kategorisindeki şarkı sözleri birkaç kelimeden oluşan satırlardan oluşabilirken, "Arabesk-Fantezi" kategorisindeki şarkılar genelde uzun satır uzunlukları içermektedir. Bu gibi "Şarkı sözü yazarı", "Yıl aralığı" ve "Kategori" gibi sınıflar için belirleyici olabilecek öznitelik gruplar tez kapsamında düşünülmüş ve deneylerde uygulanmıştır.

Her bir şarkı sözü için özniteliklere karar verildikten sonra bu öznitelikler veritabanına aktarılmıştır. Örneğin; öznitelik vektöründe kullanılacak olan n-gram öznitelikleri bu aşamada şarkı sözü metinlerinden tek tek çıkartılıp, veritabanına aktarılmıştır. Böylece her bir şarkı için veritabanında kendisine ait öznitelikler hali hazırda bulunmaktadır. Bu sayede gerçekleştirilecek her olası sınıflandırma işlemi için performans kazancı sağlanmıştır. Veritabanı üzerinden her bir şarkıya ait öznitelikler ilişkili tablolar arasında tutulmakta ve gerektiği durumda hızlı bir şekilde bu öznitelikler işleme alınmaktadır.

Öğrenme aşamasında, Weka üzerinde o deney için gerçekleştirilecek adımlar sırasıyla hazırlanan öznitelik vektörüne uygulanarak bir model ortaya çıkarılır. Buradaki önişlemler; öznitelik seçim yöntemleri, sınıflandırıcı algoritması gibi adımlardır. Uygulanacak adımlar seçildikten sonra, ilgili öznitelik vektörleri için bu deneyler sınıflar üzerinde uygulanır. Burada bahsedilen sınıflar bu çalışma için; söz yazarı, kategori ve yıl aralığıdır. Sınıflandırıcılar ilgili sınıfa (söz yazarı, kategori, tarih) uygulanır ve ortaya çıkan sonuç değerlendirilir. Bu işlem birden çok öznitelik kümesi için gerçekleştirilir ve hangi öznitelik vektörlerinin hangi sınıflar için belirleyici olduğu değerlendirilir.

Bu çalışma kapsamında, sınıflandırma işlemine alınan şarkı sayısının çok olması ile doğru orantılı olarak, her bir şarkı için çıkartılacak olan öznitelik sayısı da fazladır. Bu sebeple deneye alınacak öznitelik veri kümesi boyutu arttıkça sınıflandırma işleminin de süresi artmaktadır. Dolayısıyla eklenen her bir öznitelik için deney seti çok daha fazla olasılık içermektedir. Ayrıca sınıflandırıcıların kendi aralarında da çalışma süresi farkı bulunmaktadır.

Tez çalışması kapsamında, hazırlanan veri kümesi üzerinden öznitelik vektörleri hazırlanmıştır. Öznitelik vektörleri kümeleri içerisinde kelimenin kökü, karakter n-gramlar, sonek n-gramlar, global istatistikler ve satır uzunluğu istatistikleri gibi farklı öznitelik kümeleri bulunmaktadır. Öznitelik vektörleri oluşturulduktan sonra, bu vektörler üzerinde öznitelik seçimi yöntemleri uygulanmıştır. Bu yöntemler Ki-Kare ve ReliefF algoritmalarıdır. Öznitelik seçimi aşamasından sonra ise sınıflandırma aşamasına geçilmiştir. Bu çalışmada Naif Bayes ve Destek Vektör Makineleri sınıflandırma yöntemleri olarak seçilmiştir. Naif Bayes algoritmasının temel hali ve Multinom versiyonu; Destek Vektör Makinelerinin ise Lineer Fonksiyon ve Radyal Tabanlı Fonksiyonları bu çalışmada sınıflandırma aşamasında kullanılmıştır. Sınıflandırma algoritmaları uygulandıktan sonra elde edilen sonuçlar doğruluk-hata oranı, anma, duyarlılık, özgüllük, f-ölçütü ve roc eğrisi gibi model başarım ölçütleri ile değerlendirilmiştir ve elde edilen veriler bu kriterler üzerinden değerlendirilmiştir.

2.1 Sınıflandırma

2.1.1 Multinom Naif Bayes

Adını İngiliz matematikçi Thomas Bayes'ten alan ve Bayes istatistiğine dayanan olasılıkçı bir sınıflandırıcıdır. Naif Bayes sınıflandırıcısı olasılık ilkelerine göre tanımlanmış bir dizi hesaplama ile, problemdeki verilerin sınıflarını tespit etmeyi amaçlar.

Verilen öznitelik vektörlerindeki elemanlarının gerçekte birbiri ile ilişkisi olmasına rağmen, çözüm sırasında bu elemanlar her biri bağımsız şekilde işleme alınır. Bu şekilde her elemanın problemin çözümüne geri kalan diğer elemanlardan bağımsız olarak katkı sağladığı farz eder. Naif (Naive) Bayes ismindeki "naive" kelimesi bu kabulden dolayı sıfat olarak eklenmiştir. Yöntemdeki "naive" varsayımına rağmen,

büyük veriler üzerinde gerçekleştirilen problemlerde sınıflandırma performansı yüksektir.

Bayes kuralı şu şekilde formülüze edilebilir;

$$P(C_j|x) = \frac{p(x|C_j)P(C_j)}{p(x)} \quad (2.1)$$

- $p(x|C_j)$: Sınıf j'den bir örneğin x olma olasılığı.
- $P(C_j)$: Sınıf j'nin ilk olasılığı
- $p(x)$: Herhangi bir örneğin x olma olasılığı
- $P(C_j|x)$: x olan bir örneğin sınıf j'den olma olasılığı

Bir örnek üzerinden Naif Bayes sınıflandırıcısının nasıl çalıştığı aşağıda açıklanmıştır:

Çizelge 2.1 Naif Bayes için örnek veri kümesi

Yaş	Gelir	Öğrenci Mi?	Kredi Durumu	Bilgisayar Alabilir Mi?
<= 30	Yüksek	Hayır	Makul	Hayır
< = 30	Yüksek	Hayır	Mükemmel	Hayır
31-40	Yüksek	Hayır	Makul	Evet
>40	Orta	Hayır	Makul	Evet
>40	Az	Evet	Makul	Evet
>40	Az	Evet	Mükemmel	Hayır
31-40	Az	Evet	Mükemmel	Evet
<=30	Orta	Hayır	Makul	Hayır
<=30	Az	Evet	Makul	Evet
>40	Orta	Evet	Makul	Evet
<=30	Orta	Evet	Mükemmel	Evet
31-40	Orta	Hayır	Mükemmel	Evet
31-40	Yüksek	Evet	Makul	Evet
>40	Orta	Hayır	Mükemmel	Hayır

Çizelge 2.1'de verilen veri kümesinde yaş, gelir, öğrencilik durumu ve kredi durumu bilgilerine bağlı olarak bireylerin bilgisayar alıp alamayacağını gösteren tablo verilmiştir. Bu tabloya göre Naif Bayes yöntemiyle aşağıdaki örneğin bilgisayar alabileceğini ya da alamayacağını hesaplayacak olursak;

- Örnek veri: X= (yaş = genç, gelir=Az, öğrenci mi? = Hayır, kredi durumu=Mükemmel) Bilgisayar alabilir mi?

Çizelge 2.2 Naif Bayes örneği için özniteliklerin sınıflara göre dağılımı

Yaş	**PC Alır**	**PC Alamaz**	**Gelir**	**PC Alır**	**PC Alamaz**
<=30	2/9	3/5	Yüksek	2/9	2/5
31-40	4/9	0/5	Orta	4/9	2/5
>40	3/9	2/5	Az	3/9	1/5
Öğrenci Mi?	**PC Alır**	**PC Alamaz**	**Kredi Durumu**	**PC Alır**	**PC Alamaz**
Evet	6/9	1/5	Makul	6/9	2/5
Hayır	3/9	4/5	Mükemmel	3/9	3/5

PC	
Alır	**Alamaz**
9/14	5/14

$$P(\text{Evet}) = = \text{P}(\text{C}_1) = \frac{2}{9} * \frac{3}{9} * \frac{3}{9} * \frac{3}{9} = 0.0082$$

$$\text{P}(\text{Hayır}) = = \text{P}(\text{C}_2) = \frac{3}{5} * \frac{1}{5} * \frac{4}{5} * \frac{3}{5} = 0.0577$$

Çizelge 2.2'de bütün öznitelikler iki sınıfta hesaplanmıştır. Daha sonrasında verilen örnekteki öznitelikler Çizelge 2.2'deki değerlerine göre "PC Alabilir" ve "PC Alamaz" sınıfları için hesaplanmıştır. Sıradaki işlem ise her sınıfın toplam olasılığı hesaba katılır ve özniteliklerin olasılıkları ile çarpılır.

$$P(Evet) = 0.0082 * \frac{9}{14} = 0.0053$$

$$P(Hayır) = 00577 * \frac{5}{14} = 0.0206$$

Burada $\text{P}(\text{Hayır})$ değeri daha çok çıktığı için, verilen örnekteki bilgisayar alabilir mi sorusunun cevabı "Hayır" olarak etiketlenir.

Naif Bayes modellerde parametre tahminini en yüksek olasılık (maximumlikehood) kullanılarak yapılır. Multinom Naif Bayes (MNB) yönteminde, multinom olasılık dağılımı olduğu kabul edilmiştir.

Bayes Teoremi diğer bir şekilde ifade edilirse;

$\text{X} = \{\text{x}_1, \text{x}_2, \dots\dots, \text{x}_\text{d}\}$ veri kümesi üzerinden, olası $\text{C} = \{\text{c}_1, \text{c}_2, \dots, \text{c}_\text{d}\}$ sınıfları arasından C_i sınıfının sonsal olasılığı (posterior probability) oluşturulmak istenmektedir. Daha iyi bir şekilde ifade edilirse, X öznitelik veri kümesini, C ise sınıfların kümesini temsil etmektedir. Bayes kuralı ile yazılan aşağıdaki ifade:

$$\text{p}(\text{C}_\text{i}|\text{x}_1, \text{x}_2, \dots., \text{x}_\text{d}) = \text{p}(\text{x}_1, \text{x}_2, \dots., \text{x}_\text{d}|\text{C}_\text{i})\ \text{p}(\text{C}_\text{i}) \qquad (2.2)$$

$p(C_i|x_1, x_2, \ldots, x_d)$ sınıfına ait olma ile ilgili sonsal olasılıktır, bu da X'in C_i sınıfına ait olma olasılığı demektir. Naif Bayes her bir bağımsız değişkenin koşullu olasılıklarını istatistiksel olarak bağımsız kabul ettiğinden, olasılık terimler çarpımına çevrilebilir:

$$p(X|C_i) = \prod_{k=1}^{d} p(x_k|C_i) \quad (2.3)$$

Bu durumda sonsal olasılık aşağıdaki şekilde yeniden yazılabilir:

$$p(C_i|X) = p(C_i) \prod_{k=1}^{d} p(x_k|C_i) \quad (2.4)$$

Yukarıdaki Bayes kuralını kullanarak, bir X örneğini en yüksek sonsal olasılığa ulaşan C_i sınıfı ile atayabiliriz. Değişkenlerin her birinin bağımsız olması varsayımı her zaman doğru olmamakla birlikte, sınıflandırma işlemini büyük oranda kolaylaştırmaktadır. Bunun sebebi $p(x_k|C_i)$ ifadesinin her değişken için tekrar hesaplanmasına izin vermesidir. Bu sayede çok boyutlu iş tek boyutlu bir işe çevrilir. Bunun ötesinde, varsayım sonsal olasılıkları büyük ölçüde değiştirmediğinden, sınıflandırma işini etkilemez.

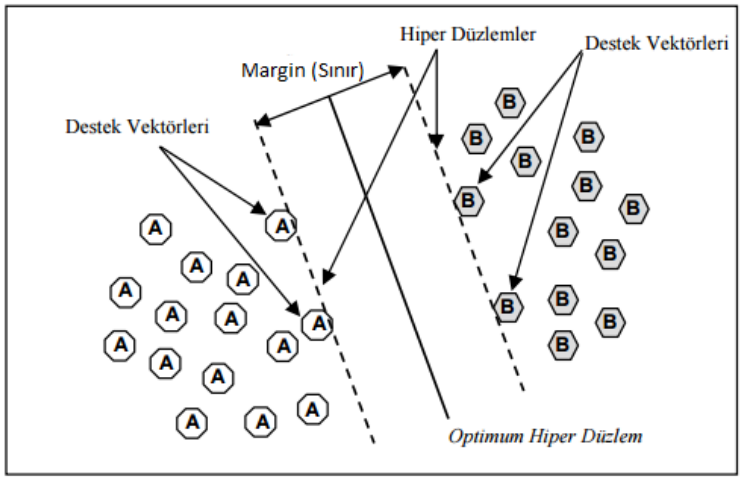
2.1.2 Destek Vektör Makinesi

Destek Vektör Makineleri (DVM), veri madenciliğinin kapsadığı alanlardan birisi olan sınıflandırma probleminin çözümü için geliştirilmiş bir makine öğrenme algoritmasıdır. Alexey Chervonenkis ve Vladimir Vapnik tarafından 1960'lı yıllarda başlatıp 1970'li yıllarda geliştirilen bir yöntem olan Destek Vektör Makineleri, başlangıçta iki sınıflı doğrusal veriler üzerinde sınıflandırma işlemleri için tasarlanmışken, daha sonrasında çok sınıflı ve doğrusal olmayan problemler için genişletilmiştir.

DVM'leri sınıflandırma işlemlerinde yüksek performans göstermesi bakımından oldukça kullanışlıdır. DVM'lerinde işleme alınacak örnek sayısının bir önemi yoktur ve DVM'leri bu açıdan genelleştirebilme özelliğine sahiptir. Diğer tekniklere göre DVM'lerinin bu genelleştirebilme özelliği, DVM'lerini iyi bir alternatif yöntem yapmaktadır. DVM'lerinin iyi bir alternatif olmasından dolayı, örüntü tanıma, görüntü işleme, arttırılmış gerçeklik, biyoloji, tıp, gen analizleri, veri madenciliği gibi birçok alanda verilerin sınıflandırılmasında DVM yoğun olarak kullanılmaktadır.

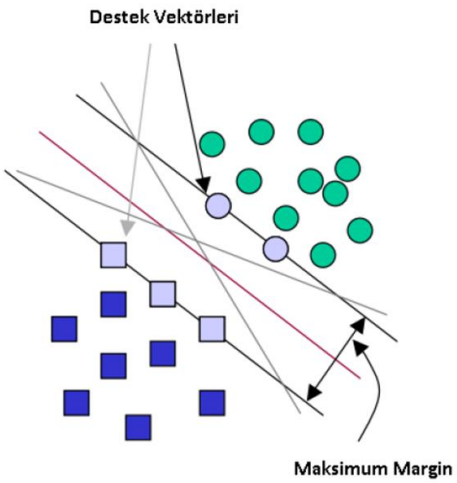
DVM için en temel sınıflandırma problemi, doğrusal olarak ayrılabilen iki sınıflı bir verinin sınıflandırılmasıdır. Destek Vektör Makineleri, bu problemin çözümü için

verilen iki sınıf arasındaki ayrımı en optimize şekilde yapan ve sınıfları birbirinden ayıran sınırın maksimum olduğu bir hiper-düzlemi belirlemeye çalışır. Verinin iki boyutlu olmasından dolayı hiper-düzlem bir çizgidir. Şekil 2.2'de iki sınıfı birbirinden ayıran optimal hiper-düzlem görülmektedir.



Şekil 2.2 İki sınıfı birbirinden ayıran optimum hiper-düzlem ve Destek Vektörleri

Şekil 2.2'deki gibi iki sınıfı birbirinden ayıran tek bir hiper düzlem yerine yine bu iki sınıfı birbirinden ayıran başka hiper düzlemler de çizilebilir. Hangi düzlemin daha iyi olduğunu ve optimal düzlemin nasıl bulunacağı önemli bir problemdir. Destek Vektör Makinelerindeki amaç Şekil 2.3'de görüldüğü gibi optimal ayrımı yapan hiper-düzlemi bulabilmektir.



Şekil 2.3 Maksimum margininin hesaplandığı Destek Vektör Makinesi

Verileri sınıflandırma sırasında en iyi hiper-düzlemi bulabilmek için, her iki sınıfın verilerine en yakın şekilde geçecek olan Şekil2.2'de görüldüğü gibi hiper-düzlemler çizilir. Bu iki hiper-düzlem birbirlerine paraleldir ve bu hiper-düzlemler arasındaki

mesafe optimum hiper-düzlemin başarısını belirlemektedir. DVM'leri bu aşamada iki sınıf arasındaki sınırı belirlemede ve optimum hiper-düzlemin tanımlanmasında kullanılır. Destek vektörler, hiper düzlemler, optimum hiper düzlem ve margin Şekil 2.2'de gösterilmiştir.

Sınıfa ait veriler deney için DVM'ları tarafından işleme alınır. Bu işlem sonucunda elde edilen çıktı test edilen verinin ayırt edici skorudur. Elde edilen sonuç pozitif bir değer ise verinin o sınıfa ait olduğuna işaret eder. Ortaya çıkan değer sıfırdan büyük ise bu, sistem için iyi bir skor olarak kabul edilir.

Destek Vektör Makinelerinde iki durum ile karşılaşılabilir, bunlardan birincisi sınıflandırma işlemi gerçekleştirilirken verilerin lineer olarak ayrılabilmesi durumu, diğeri ise verilerin lineer bir şekilde ayrılamayacak durumda olması sonucunda ortaya çıkan durumdur. Lineer olarak ayrılmış verilerin bulunduğu durumda Şekil2.3'de de görülebileceği gibi maksimum marginin hesaplanması kolaydır; fakat lineer olarak ayrılamayan veriler lineer olarak sınıflandırılabilecekleri başka bir uzaya aktarılmalıdırlar.

DVM'leri matematiksel olarak aşağıdaki gibi tanımlanır:

DVM yöntemi ile sınıflandırma işlemleri 2. dereceden bir denklemin çözümü ile gerçekleştirilir. Sırasıyla X_i destek vektörleri $y_i$ait oldukları sınıf etiketleri olmak üzere, $X_i \in R^T$ ve $y_i = \{+1, -1\}$ N uzunluğundaki (x_i, y_i) çiftine bağlı optimum hiper düzlem denklemi şu şekilde tanımlanır:

$$w * x + b = 0 \tag{2.5}$$

Formülde belirtilen (w, b) en iyi ayırıcı düzlem parametreleridir.

w, ağırlık vektörüdür ve hiper düzleme dik bir vektör tanımıdır.

B, eğilim değerlerini ifade etmektedir.

DVM'lerinin sınıflandırma fonksiyonu olarak bu formülü kullanırsak,

$$f(x) = \text{sign}(w * x + b) \tag{2.6}$$

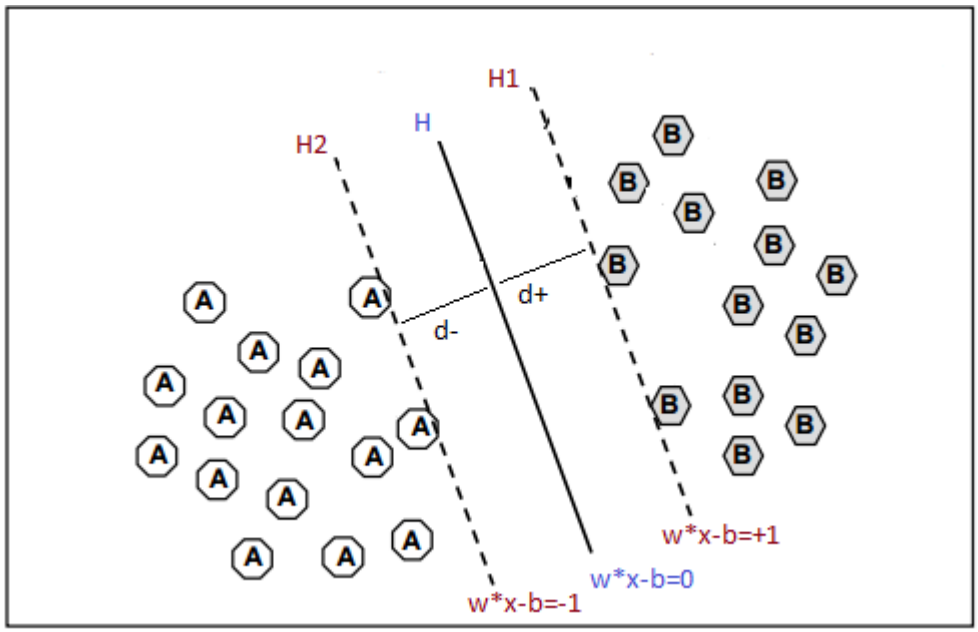
formülüne ulaşırız.

Veri kümesinde bulunan bir örnek olan x_i formülde yerine koyulursa Şekil 2.4'te de görüldüğü optimum hiper-düzlemin belirlenmesi için bu düzleme paralel olan ve

düzlemin sınırlarını belirleyen iki hiper-düzlem belirlenir. Bu iki düzlem için aşağıdaki gibi bir sonuç ortaya çıkar:

$$x_i * w + b \geq +1 \; ise \; y_i = +1 \tag{2.7}$$

$$x_i * w + b \leq -1 \; ise \; y_i = -1 \tag{2.8}$$



Şekil 2.4 Formüller üzerinden hiper düzlemler

Düzlem formüllerini daha basit şekilde ifade edecek olursak;

$$y_i(x_i * w + b) \geq 1 \tag{2.9}$$

denklemi sınıflandırma için kullanılacak olan veri kümesi içerisindeki her örnek için doğru olur.

x_i noktasının geometrik olarak hiper düzleme olan uzaklığını hesaplarken w'nin değeri normalize edilir. Böylece x_i noktasının hiper düzleme olan uzaklığı şu şekilde ifade edilebilir:

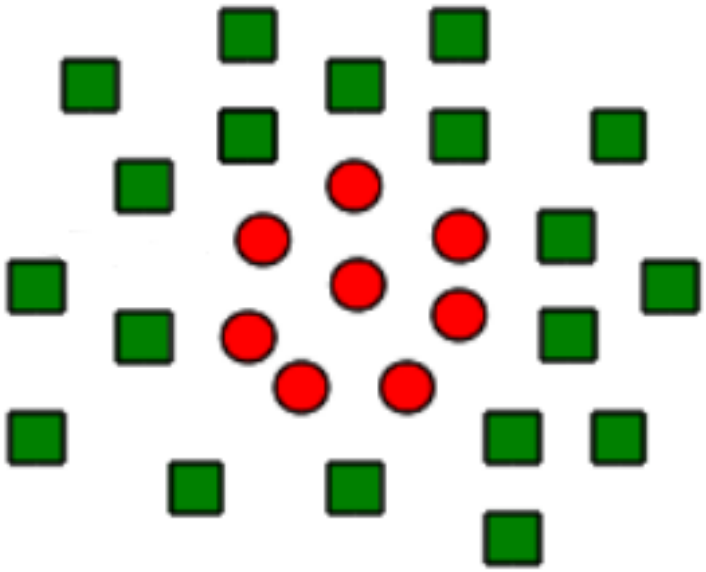
$$d\big((w,b),x_i\big) = \frac{y_i(x_i * w + b)}{\|w\|} \geq \frac{1}{\|w\|} \tag{2.10}$$

x_i noktasının hiper düzleme uzaklığı maksimize edilmek istenildiği için yukarıdaki formüldeki $\|w\|$ ifadesinin minimize edilmesi gerekir. Bunun için kullanılan başlıca yöntem Vapnik'te de belirtildiği gibi Lagrange çarpanlarıdır [32]. Bu yöntem kullanılarak ifade aşağıdaki ifadenin minimize edilmesine dönüştürülür.

$$W(\alpha) = -\sum_{i=1}^{N} \alpha_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j (x_i^T * x_j) \tag{2.11}$$

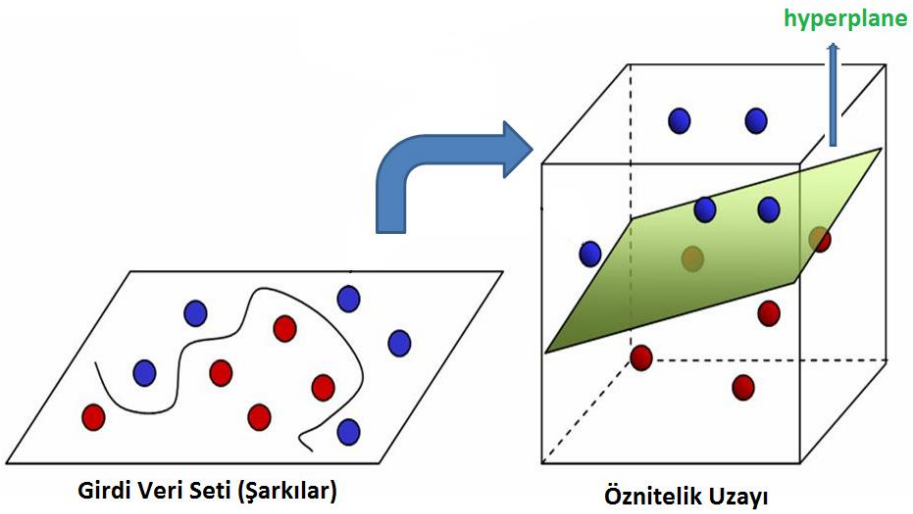
Yukardaki ifade ile her bir veri için bir tane olmak üzere toplam L tane α değeri bulunur [33]. Bulunan alfa değerlerinden sıfırdan büyük olanlar destek vektörleri olarak tanımlanmıştır. Örnek olarak 1000 verilik bir eğitim setinde çıkan α değerlerinin birçoğu sıfır olacaktır [33]. Bu noktalar veriyi ayıran maksimum margin ile tanımlamış hiper düzlemin dışında kalan noktalardır. Fakat α_i değeri sıfırdan büyük ise bu değerin ait olduğu xi vektörü destek vektörü olarak tanımlanır. Destek vektörlerinin bulunması ile doğrusal olarak ayrılan veriler için maksimum margine sahip hiper düzlem bulunmuş olur.

Yukarıda bahsedilen veri kümesinin DVM ile doğrusal olarak ayrılabildiği varsayılmıştır; fakat sınıflandırma problemlerinde veri kümesi genel olarak doğrusal ayrılamaz. Şekil 2.5'te doğrusal (lineer) olarak ayrılamayan bir veri kümesi gösterilmektedir.



Şekil 2.5 Destek Vektör Makineleri için doğrusal ayrılamayan veri kümesi

Uygun bir Φ fonksiyonu ile veri kümesinin doğrusal olarak ayrılabileceği yüksek boyutlu bir sisteme taşındığı farz edilirse, yeni oluşan çok boyutlu uzay öznitelik uzayı H olarak adlandırabilir. Bu uzayda bulunan bir hiper düzlem ile mevcut veriler doğrusal olarak ayrılacaktır [34] (Şekil 2.6).



Şekil 2.6 Veri kümesinin hiper düzlemde doğrusal olarak ayrılması

Doğrusal (Lineer) olarak ayrılamayan veriler için elde edilen optimum hiper düzlemin formülü, doğrusal olarak ayrılabilen veri kümesi için olan formül ile birebir aynıdır. Tek fark formüldeki x_i vektörlerinin d boyut olması yerine, Φ(xi) vektörünün sonsuz boyut gibi daha yüksek boyutta olmasıdır.

$$W(\alpha) = -\sum_{i=1}^{N} \alpha_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j (\phi(x_i^T) * \phi(x_j)) \quad (2.12)$$

Formül incelendiği zaman fark edilen en önemli nokta çok boyutlu uzaydaki vektörlerin nokta çarpımı ile ilgilidir. Vektörlerin yüksek boyutlu uzaya taşınmış halindeki $\phi(x_i^T) * \phi(x_j)$ nokta çarpımını yüksek boyutlu uzayda yapılması çok maliyetli bir işlemdir [33]. Verinin sonsuz boyutlu uzaya taşınması durumda bu formülü gerçekleştirmek imkansız duruma gelmektedir. Böyle bir durum gerçekleştiği zaman çekirdek fonksiyonları veri kümesinin aktarılmış uzaydaki nokta çarpımlarını verirler. Çekirdek fonksiyonlar sayesinde verinin aktarıldığı uzay hakkında bilgi olmamasına rağmen bu uzaylar kullanılabilmektedirler. Bu durum;

K çekirdek fonksiyonu ve Φ vektörleri yüksek boyuta taşıma fonksiyonu olmak üzere

$$K(x, y) = \phi(x) * \phi(y) \quad (2.13)$$

formülü ile ifade edilebilir.

Büyük boyuttaki vektörleri çok sayılı boyuta taşıyan fonksiyon hakkında hiçbir bilgi bilinmemesine rağmen destek vektör makineleri bu fonksiyonları verimli bir şekilde

kullanabilirler [35]. DVM yönteminde kullanılan çekirdek fonksiyonları aşağıda verilmiştir:

- Doğrusal Fonksiyon
- Radyal Tabanlı Fonksiyon
- Polinomiyal Fonksiyon
- Sigmoid Fonksiyon

Doğrusal Fonksiyon: Doğrusal çekirdek, sınıflandırma işlemini doğrular çizerek tanımlar. Vektörlerin iç çarpımlarına sabit bir değer ekleyerek bulunması sonucunda bu fonksiyon ortaya çıkar. Veri kümesinin doğrusal olarak düzgün bir şekilde ayrılamayacağı durumlar için doğru bir seçim değildir. Aşağıdaki gibi formülüze edilir [36];

$$\mathrm{K}(\mathrm{x},\mathrm{y}) = \mathrm{x}^{\mathrm{T}} * \mathrm{y} + \mathrm{c} \tag{2.14}$$

Radyal Tabanlı Fonksiyon: Doğrusal olmayan veriyi daha yüksek boyutlu bir uzaya taşıyarak sınıflandırma işlemini gerçekleştirir. Doğrusal fonksiyonun aksine verilerin doğrusal şekilde sınıflandırılamayacağı durumlarda verimli bir şekilde çalışabilir. Öznitelik vektörünün sayısının çok fazla olması durumlarında kullanılması tavsiye edilmez. Radyal Tabanlı Fonksiyon, doğrusal çekirdek ile ceza parametresinin birleşmiş halidir. Aşağıdaki gibi formülüze edilir [36] :

$$\mathrm{K}(\mathrm{x},\mathrm{y}) = \exp\left(-\alpha\left\|\mathrm{x_i} - \mathrm{x_j}\right\|^2\right), \alpha > 0 \tag{2.15}$$

Polinomiyal Fonksiyon: Radyal Tabanlı Fonksiyona göre daha fazla parametre içerir. Bu sebepten dolayı, RTF çekirdeğinin daha az sayısal zorlukları bulunmaktadır. Eğitim veri kümesindeki tüm değerlerin normalize edildiği problemlerin kullanımında tercih edilebilir. Aşağıdaki gibi formülüze edilir [36] :

$$\mathrm{K}(\mathrm{x},\mathrm{y}) = (\alpha \mathrm{x}^{\mathrm{T}}\mathrm{y} + \mathrm{c})^{\mathrm{d}}, \alpha > 0 \tag{2.16}$$

Sigmoid Fonksiyon:

Aşağıdaki gibi formülüze edilir [36]:

$$\mathrm{K}(\mathrm{x},\mathrm{y}) = \tanh(\alpha \mathrm{x}^{\mathrm{T}}\mathrm{y} + \mathrm{c}) \tag{2.17}$$

Tez çalışmasındaki şarkı sözlerinin sınıflandırılması işlemi sırasında, Doğrusal (Linear) fonksiyon ve Radyal Tabanlı Fonksiyon yöntemleri uygulanmıştır. Ayrıca tez kapsamında, LIBSVM [37] uygulamasındaki varsayılan parametreler

kullanılmıştır. Bu parametreler “-s” için 0 değeri; “-t” için ise Doğrusal yöntem için 0, Radyal Tabanlı yöntem için ise 2 olarak kullanılmıştır.

2.2 Öznitelikler

Metin verisinden yazar tanıma çalışmalarında yazarların, kategorilerin ve benzeri sınıfların ayırt edilebilmesi için öznitelik seçimi önemli bir aşamadır. Özniteliklerin doğru bir şekilde seçilmesinden sonra bu özniteliklerin metin verisinden çıkartılması ve işlenmesi aşaması gelmektedir. Yazar tanıma, kategori tanıma gibi çalışmalarda çok çeşitli öznitelik kümeleri kullanılmaktadır. Daha önceki çalışmalarda [38] belirtildiği gibi bine yakın özniteliğin kullanılmasına rağmen henüz metin üzerinden yazar tanıma, kategori tanıma gibi sınıf tanıma çalışmalarında kullanılan uzlaşılmış ve net kabul görmüş bir öznitelik kümesi bulunmamaktadır. Bunun sebeplerinden birisi de her dilin kendine özgü bir dil bilgisinin bulunmasıdır. Örneğin; Türkçe eklemeli bir dil iken, İngilizce çekimli bir dildir. Bundan dolayı dilin kökenine bağlı olarak çıkartılacak öznitelikler başarı ölçütünü olumlu etkileyebildiği gibi olumsuz da etkileyebilecektir.

Metin üzerinden sınıf tanıma yöntemlerinde ilk yapılan çalışmalarda genellikle tek bir öznitelik kümesi üzerinde durulmuştur. Sözcük uzunluklarının [39] ve cümle uzunluklarının [40] öznitelik olarak kullanılması buna örnek verilebilir. Daha sonraki çalışmalarda birden çok öznitelik kümesinin birbirleriyle kombinasyonu kullanılarak çok çeşitli bir öznitelik kümesi kullanımı yaygınlaşmıştır. Kullanılan çeşitli öznitelik kümelerinin yanı sıra istatistiksel çözümler de öznitelik kümelerine eklenerek geniş bir öznitelik kombinasyonu elde edilmiştir. Öznitelik kümelerinin genişlemesi ile birlikte, kullanılan kümelerin daha anlaşılır bir şekilde aktarılabilmesi için öznitelikler belirli kurallara göre gruplara ayrıştırılmıştır. Yaygın olarak kullanılan beş tür öznitelik grubu bulunmaktadır.

- Sözcüksel Öznitelikler (Lexical Features): Metin içerisinde bulunan kelime ve harf verilerine dayalı istatistiksel öznitelik kümesidir. Örneğin; sözcük sayısı, farklı sözcük sayısı, harf sayısı, toplam harf sayısı, ortalama kelime uzunluğu, v.b.
- Sözdizimsel Öznitelikler (Syntactic Features): Tür tabanlı istatistiksel öznitelik kümesidir. Örneğin; Sözcük dizileri (N-Gram), noktalama işaretleri, sözcük türleri, v.b.

- Yapısal Öznitelikler (Structural Features): Metin verisinin genel yapısına ilişkin öznitelik kümeleridir. Metindeki başlık kullanımı, yazı tipi özellikleri, metin içerisindeki resim ya da bağlantılar, v.b. bu öznitelik kümesine örnek olarak gösterilebilir.
- İçeriğe Özgü Öznitelikler (Content-Specific Features): Metin üzerinden sınıf tanıma çalışmalarında sınıflandırmaya bağlı olarak metin içerisindeki bazı kelime ya da cümleler kullanılma nedeni, metin içerisinde geçme sıklığı gibi sebeplerden dolayı diğer kelime ya da cümlelere göre daha önem taşıyabilmektedir. Bu kelime ya da cümlelerin sayıları gibi istatistiksel veriler bu öznitelik kümesinde kullanılabilir.
- Kişiye Özgü Öznitelikler (Idiosyncratic Features): Metinin sahibi olan yazarın kullandığı yanlış sözcük kullanımları ya da gramer hataları gibi veriler bu öznitelik kümesine aittir.

Türkçe şarkı sözü madenciliği çalışmasında, kullanılan şarkı sözleri üzerinde yukarıda belirtilen öznitelik kümelerinin tümü ya da öznitelik kümeleri içerisindeki özniteliklerin bazıları aşağıda belirtilen sebeplerden dolayı kullanılmamıştır;

- Sözdizimsel öznitelik kümesi içerisindeki noktalama işaretleri özniteliği şarkı sözlerinin bulunduğu kaynağa aktarılması sırasında aktaran kişiye bağlı olarak kullanılan noktalama işaretlerinin değişkenlik göstermesinden dolayı kullanılmamıştır.
- Yapısal öznitelik kümesi şarkı sözü madenciliğinde kullanılamaz. Bunun sebebi; şarkı sözlerinin içerisinde başlık bilgisi, resim ya da bağlantı bilgilerinin bulunmaması ve ayrıca şarkı sözü için yazı tipinin bir anlam ifade etmemesidir.
- Kişiye özgü öznitelik kümesi, şarkıya ait metnin, bulunduğu kaynağa aktarılması sırasında değişkenlik gösterebileceği için bu çalışmada kullanılmamıştır.

Tez çalışması sırasında yukarıda bahsedilen durumlar dikkate alınarak hızlı, doğru ve etkili şekilde uygulanabilecek, sonuçların elde edilmesi sırasında en doğru ve anlaşılır çıktıları verecek, verilerin hazırlanması sırasında kullanılan kaynaklara göre (internet gibi) değişiklik göstermeyecek ve Türkçe'ye ait dil kullanım şekillerini kapsayacak öznitelik kümeleri seçilmiştir. Özniteliklerin seçilmesi ve çıkartılması

sırasında bazı öznitelikler için Zemberek [41] adlı Doğal Dil İşleme (Natural Language Processing – NLP) kütüphanesinden faydalanan bir uygulama geliştirilmiştir. Bu uygulama sayesinde metin içerisindeki kelimelerin kökleri elde edilmiş ve tez çalışması kapsamında kullanılmıştır.

Tez kapsamında kullanılan öznitelik kümeleri Çizelge 2.3'te görülmektedir.

Çizelge 2.3 Öznitelik tanımları ve kısaltmaları

Öznitelik Adı Kısaltması	**Öznitelik**
KK34	Kelime Kökü + 3Gram + 4Gram
HS	Harf Sayısı
KS	Kelime Sayısı
FKS	Farklı Kelime Sayısı
OKU	Ortalama Kelime Uzunluğu
MMF	Satır İçin Maksimum Minimum Farkı
MED	Medyan
OSU	Ortalama Satır Uzunluğu
SS	Standart Sapma
S23	Kelime Sonu 2 Gram ve 3 Gram
E23	Satır Sonu 2 Gram ve 3 Gram

Tez çalışmasında sonuçların daha iyi analiz edilebilmesi ve karmaşıklığın önlenmesi için Çizelge 2.3'de belirtilen özniteliklerden oluşan öznitelik kümelerine kısaltmalar verilmiştir. Tez çalışmasının bütününde de kullanılacak olan bu kısaltmalar Çizelge 2.4'te verilmiştir.

Çizelge 2.4 Öznitelik kümeleri ve kısaltmaları

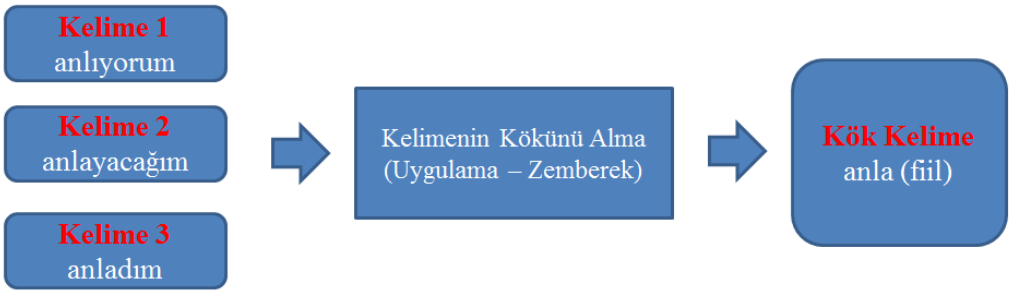
A	KK34
B	KK34 + HS
C	KK34 + KS
D	KK34 + FKS
E	KK34 + HS + KS
F	KK34 + HS + FKS
G	KK34 + KS + FKS
H	KK34 + HS + KS + FKS
I	KK34 + HS + KS + FKS + OKU
J	KK34 + HS + KS + FKS + MMF
K	KK34 + HS + KS + FKS + MED
L	KK34 + HS + KS + FKS + OSU
M	KK34 + HS + KS + FKS + SS
N	KK34 + HS + KS + FKS + VAR
O	KK34 + HS + KS + FKS + OKU + MMF + MED + OSU + SS + VAR
P	Kelime Kökü
R	Kelime Kökü + S23 + E23
S	KK34 + S23 + E23 + HS + KS + FKS + OKU + MMF + MED + OSU + SS +
T	Kelime Kökü + S23 + E23
U	KK34 + S23 + E23
V	KK34 + S23 + E23 + HS + KS + FKS + OKU + MMF + MED + OSU + SS +

2.2.1 Öznitelik grupları

2.2.1.1 Kelimenin kökü

Şarkı metninden alınan kelimeler, Zemberek Doğal Dil İşleme Kütüphanesi'nden de yararlanılarak tez çalışması kapsamında geliştirilen bir uygulama sayesinde köklerine ayrıştırılmıştır. Kelimenin kendisi yerine o kelimenin köklerini kullanmanın daha verimli olduğu görülmüştür. Bunun sebebi kullanılan bir

kelimenin, Türkçe'nin yapısından dolayı farklı çekimler, ekler gibi yapısal değişiklikler sonucunda farklı bir kelime gibi davranmasının önüne geçilmesidir. Örneğin; şarkıcı belirli bir kelimeyi çok sık kullanmaktadır; ama kelimenin üzerine belirli ekler geldiği zaman sınıflandırma yöntemlerinde kullanılacak öznitelik kümesinde bu kullanılan kelime farklı kelime gibi davranacaktır. Böylece şarkıcı tarafından kullanılan aynı kelime olmasına rağmen bu kelime öznitelik kümesinde farklı kelimeler gibi algılanacaktır. Bunun önüne geçmek için metin içerisindeki kelimelerin kökleri alınmıştır. Şekil 2.7'de tek bir kök kelimenin birden farklı şekilde kullanabileceği ve bu kelimelerin kökleri alındıktan sonra aynı kök kelimeyi gösterdiği gösterilmiştir.



Şekil 2.7 Kelimelerin köklerinin alınması

2.2.1.2 Karakter N-Gramlar

N-gram, bir karakter katarının n adet karakter dilimidir. N-gram tabanlı sınıflandırma yöntemi, şarkı metni içerisindeki karakter tabanlı n-gram'ların kullanım sıklığına dayalı bir işlemdir [42]. Bu çalışmada, n-gram'ın farklı birkaç uzunluğu alınarak 2-, 3- ve 4-gram'lar kullanılmıştır. N-gram'ların elde edilmesinde izlenen yolu bir örnek ile açıklayacak olursak: Örnekte boşluk karakterini göstermek için "_" altçizgi karakteri kullanılmıştır.

Cümlemiz "Şarkı Tanıma" ise, bu cümlenin ngram'ları;

2-gram'lar: "Şa", "ar", "rk", "kı", "ı_", "_T", "Ta", "an", "nı", "ım", "ma"

3-gram'lar: "Şar", "ark", "rkı", "kı_", "ı_T", "_Ta", "Tan", "anı", "nım", "ıma"

4-gram'lar: "Şark", "arkı", "rkı_", "kı_T", "ı_Ta_", "_Tan", "Tanı", "anım", "nıma" şeklinde çıkarılır.

N-gram yöntemi, metinlerin benzerliklerinin incelenmesinde ve kümeleme çalışmalarında kullanıldığı gibi genelde büyük boyutlu metinlere uygulanır ve metin

içinde kullanılan her kelimenin olasılıkları hesaplanarak elde edilen sonuçlar, takip eden kelimelerin görülme olasılıklarına yansıtılır.

Şarkı sözü madenciliği ve genel olarak metin sınıflandırmada N-gram yöntemi basit ve güvenilir bir yöntem olarak kullanılmaktadır. N-gram yöntemi ile elde edilen özniteliklerin kullanılmasındaki bir diğer neden ise N-gram özniteliklerinin dilden bağımsız bir şekilde çalışmasıdır. Ayrıca sınıflandırma işlemi metin içerisindeki karakterlerin kullanım sıklığından yararlanılarak yapıldığı için, örneğin içerik bir aşk şarkısı ise kelimenin ilgili formaları için ("sev", "sevmek", "seviyorum", "seveceksin", "sevsen") elde ettiğimiz n-gram'ların sıklığı ile sınıflandırma işlemini kolayca yapabiliriz. Özet olarak N-gram özniteliğinin en büyük avantajları dilden bağımsız olması ve metin içerisinde kullanılan, ekler ve çekimler yüzünden farklı formlarda ifade edilmesine rağmen aynı kök kelimenin sınıflandırılmasında kolaylık sağlamasıdır.

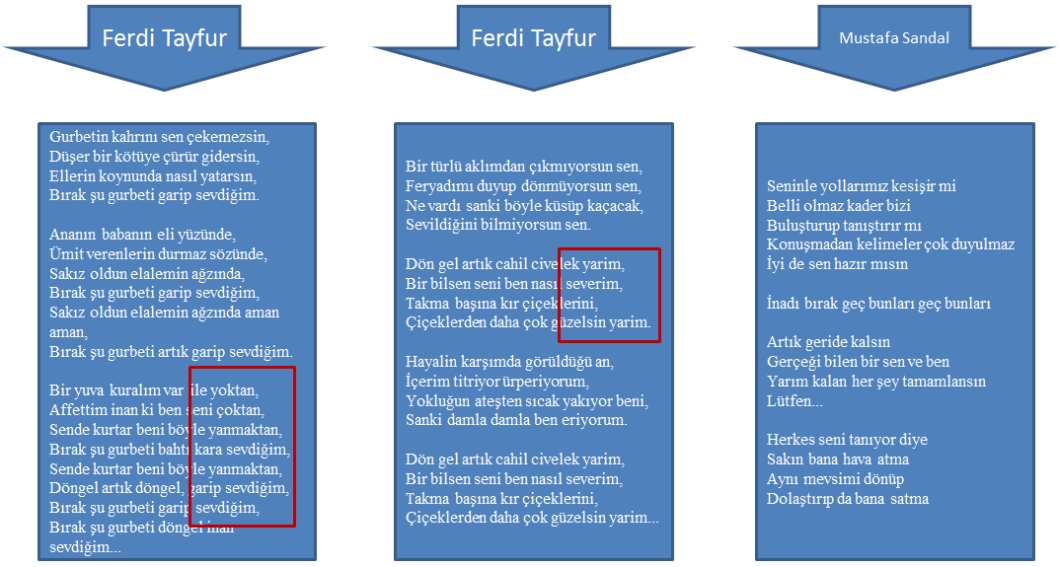
Bu çalışmada, şarkı sözü metninden elde edilen 2-gram, 3-gram ve 4-gram'lar öznitelik vektörüne eklenmiştir ve kullanılmıştır.

2.2.1.3 Sonek N-Gramlar

Sonek N-gram'lar, bu çalışma kapsamında karakter N-gram'ların özelleştirilmesiyle elde edilmiştir. Sonek N-gram'ları iki kategoride kullanılmıştır; bunlardan ilki kelime sonu sonek N-gram'ları, diğeri ise satır sonu sonek N-gram'larıdır.

Türkçe eklemeli bir dil olduğu için kelime sonundaki ekler, şarkı sözü sınıflandırmada belirleyici bir rol oynamaktadır. Örneğin; bir şarkı sözü yazarı geçmiş zamanı çok kullanırken, diğer bir şarkı sözü yazarı gelecek zamanı daha sık kullanabilmektedir veya bir şarkı sözü yazarı ilgi eklerini sıklıkla kullanırken, diğer bir şarkı sözü yazarı iyelik eklerini daha sık kullanabilmektedir. Bu sebepten dolayı, kelime sonu sonek N-gram'ları şarkı sözlerini sınıflandırmada etkili olabilecek özniteliklerdir.

Türkçe şarkı sözlerinde kullanılabilecek en önemli özniteliklerden bir tanesi satır sonundaki kafiyelerdir. Kullanılan kafiyeler şarkıcı ve kategori için çok belirleyici bir öznitelik olabileceği için, satır sonu N-gram'lar öznitelik vektör kümesine dahil edilmiştir. Şekil 2-8'de satır sonu sonek N-gram'lara örnek verilmiştir.



Şekil 2.8 Satır sonu sonek N-Gram

Şekil 2.8'de de görüldüğü gibi Ferdi Tayfur adlı şarkı sözü yazarı iki şarkısında da kafiyeler kullanmıştır. Birinci şarkısında "yoktan", "çoktan", "yanmaktan" gibi kelimeler kullanırken, ikinci şarkısında ise "yârim", "severim" gibi kafiye oluşturacak kelimeleri seçmiştir. Mustafa Sandal adlı şarkı sözü yazarının şarkısında ise herhangi bir kafiyeye rastlanmamıştır. Şekil 2.8'deki örnekte görüldüğü gibi, satır sonundaki ekler ve kafiyeler sınıflandırma işlemi gerçekleştirilirken kullanılacak olan öznitelik veri kümesine eklenmesi, şarkı sözü yazarlarının yazım tarzları hakkında bilgi verebileceği için sınıflandırma sonuçlarını olumlu yönde etkileyecektir.

Tez kapsamında, "Karakter N-Gramlar" belirlenirken ve kullanılırken 3 Gramdan 4 Grama geçildiği zaman öznitelik sayısında bariz bir artış gözlenmiştir. Bu sayı 4 Gramdan 5 Grama geçiş yapılırken daha da bariz artacaktır. Her bir "Karakter N-Gram" öznitelikleri sayısındaki bu artış doğru orantılı olarak veritabanı boyutunu ve sınıflandırma performansını etkilemektedir. Yapılan deneyler sonucunda, "Karakter N-Gram" öznitelikleri için 2, 3 ve 4 Gramların kullanılmasının etkili ve yeterli olduğu gözlemlenmiştir. Bu nedenlerden dolayı bu çalışma kapsamında "Karakter N-Gramlar" için 2, 3 ve 4 Gramlar tercih edilmiştir. "Sonek N-Gram" öznitelikleri için ise metin içerisindeki ekleri ve kafiyeleri belirlemede 2 Gram ve 3 Gramlar yeterli olmaktadır. Bu sebepten dolayı tez kapsamında "Sonek N-Gram" öznitelikleri seçiminde 2 ve 3 Gramlar tercih edilmiştir.

2.2.1.4 Global istatistikler

Global istatistikler, daha önce bahsedilen sözcüksel öznitelikler grubuna girmektedir. Şarkı sözü metnindeki harf sayısı, kelime sayısı, farklı kelime sayısı ve ortalama kelime uzunluğu bu çalışmadaki global istatistiksel özniteliklerdir.

Harf Sayısı: Şarkı sözü metnindeki tüm harflerin toplam sayısıdır.

Kelime Sayısı: Şarkı sözü metnindeki toplam kelime sayısıdır.

Farklı Kelime Sayısı: Şarkı sözü yazarının şarkı sözü metninde kullandığı farklı kelimelerin sayısıdır.

Ortalama Kelime Uzunluğu: Şarkı sözü yazarının şarkı sözü metninde kullandığı kelimelerin ortalama uzunluğudur.

Global istatistiksel öznitelikler şarkı sözü yazarlarının sözcükleri kullanım özellikleri hakkında bilgi vermektedir. Örneğin; bir şarkı sözü yazarı söz yazarken uzun kelimeleri ve/veya uzun şarkı sözlerini tercih ederken, diğer bir şarkı sözü yazarı daha kısa kelimeler kullanarak daha kısa şarkılar yazabilmektedir. Bu sebeplerden dolayı şarkı sözü sınıflandırmada bahsedilen global istatistiksel öznitelikler önem kazanmaktadır.

2.2.1.5 Satır uzunluğu istatistikleri

Satır uzunluğu istatistikleri, global istatistiklere benzer bir şekilde bu proje kapsamında geliştirilmiş özniteliklerdir. Şarkı sözü sınıflandırmada, şarkı sözü yazarlarının sözleri yazarken satırları farklı şekilde kullandıkları gözlemlenmiştir. Şarkı sözü metinlerinde kısa cümlelerden oluşan satırlar bulunduğu gibi uzun cümlelerden oluşan satırlar da bulunmaktadır. Örneğin; Barış Manço'nun yazdığı şarkılarda satır uzunlukları fazlayken Teoman'ın yazdığı şarkıların satır uzunlukları Barış Manço şarkılarına göre nispeten daha kısadır. Bu çalışma kapsamında satır uzunluğu istatistik öznitelikleri için düşünülen değerler şunlardır;

Satır İçin Maksimum Minimum Farkı: Şarkı metni içerisindeki en uzun satır ile en kısa satırın farkıdır.

Medyan: Medyan, bir sayısal veri serisi sıralandığında ortada kalan sayıdır. Bu çalışmada şarkı sözü metnindeki tüm satır uzunlukları bir diziye atılmaktadır, daha sonra bu dizi üzerindeki sayısal değerler sıralanıp ortada olan değer medyan özniteliği olarak kullanılmaktadır.

Ortalama Satır Uzunluğu: Ortalama, bir sayı serisindeki sayıların toplamının serinin eleman sayısına bölünmesi sonucu elde edilen değerdir. Şarkı sözü içerisindeki satır uzunlukları bir diziye atıldıktan sonra bu dizi üzerindeki sayısal değerin ortalama satır uzunluğu (mean) değeri hesaplanmıştır ve bir öznitelik kullanılmıştır.

Standart Sapma: Standart sapma, bir sayı serisindeki sayıların, serinin aritmetik ortalamasından farklarının karelerinin toplamının dizinin eleman sayısının bir eksiğine bölümünün karekköüdür. Bu tez kapsamında standart sapma hesaplamak için;

- Ortalama satır uzunlukları hesaplanır.
- Her bir satır uzunluğunun ortalama satır uzunluğundan farkı bulunur.
- Bulunan farkların her birinin karesi hesaplanır.
- Farkların kareleri toplanır.
- Elde edilen toplam, satır uzunluklarının atıldığı serinin eleman sayısının bir eksiğine bölünür.
- Bulunan sayının karekökü alınır.

Standart sapma ile satır uzunluklarının ne kadarının ortalamaya yakın olduğunu buluruz. Eğer standart sapma küçükse satır uzunlukları ortalamaya yakın yerlerde dağılmışlardır. Bunun tersi olarak standart sapma büyükse satır uzunlukları ortalamadan uzak yerlerde dağılmışlardır. Bütün veri değerleri aynı olursa standart sapma sıfır olur. Standart sapma şarkı sözü içerisindeki satır sayısı arttıkça ve daha büyük diziler elde edildikçe daha anlamlı veriler kullanılacaktır.

Bu çalışma kapsamında kullanılan satır uzunluğu istatistiksel öznitelikler sınıflandırma yöntemleri kullanırken, şarkı sözü yazarları hakkında bilgi vermektedir. Bazı yazarlar şarkı içerisinde çok uzun ve çok kısa satırlar halinde söz yazabilmektedirler. Aynı şekilde bazı şarkı sözü yazarlarının yazdıkları şarkılarda satır uzunlukları yaklaşık olarak benzerdir.

Satır uzunluğu ile ilgili istatistiksel öznitelikler, bu çalışma kapsamında şarkı sözlerini kategorize etmede de önem teşkil etmektedir. “Rock” türündeki şarkılar uzun cümleler içeren satırlardan oluşurken, “Pop” şarkıları nispeten daha kısa satır uzunluklarına sahiptir. Bu yüzden “Rock”, “Arabesk-Fantezi” ve “Pop” şarkılarını sınıflandırırken bu tür istatistiksel öznitelikler önem kazanmaktadır.

2.2.2 Öznitelik vektörü

Öğrenme ve sınıflandırma aşamalarında kullanılacak olan öznitelik vektörleri daha önce bahsedilen öznitelik gruplarının hepsi ya da bazılarının bir araya gelmesinden oluşmaktadır.

Tez kapsamında gerçekleştirilen deneylerin farklılık göstermesinden dolayı kullanılan öznitelikler de farklılık göstermektedir. Bu sebepten dolayı öznitelik vektörleri içerisinde kullanılan öznitelik kümesi de deneye bağlı olarak değişebilmektedir. Tez çalışmasının ilerlemesiyle birlikte öznitelik vektörü de güncellenerek, büyümüştür. Tezin ilk deneylerinde kullanılan toplam öznitelik çeşitliliği ve doğru orantılı olarak öznitelik sayısı, tez çalışması kapsamı genişledikçe ve ilerledikçe artmıştır. Çizelge 2.5'te örnek bir deneyde kullanılan öznitelik kümesi ve bu özniteliklerin her bir ayrı şarkı metni içerisinde kullanılma sayıları gösterilmektedir.

Çizelge 2.5 Şarkı sözü sınıflandırılması sırasında kullanılan metin tabanlı örnek öznitelik kümesi

Öznitelik Kümesi	**Öznitelik Sayısı**
Kök kelimenin	4084
Karakter 3-Gram	5713
Karakter 4-Gram	23651
Kelime sonu 2-Gram	321
Kelime sonu 3-Gram	1733
Satır sonu 2-Gram	234
Satır sonu 3-Gram	1042
Toplam harf sayısı	1
Toplam kelime sayısı	1
Farklı kelime sayısı	1
Ortalama kelime uzunluğu	1
Satır İçin Maksimum Minimum Farkı	1
Medyan	1
Ortalama Satır Uzunluğu	1
Standart Sapma	1
Toplam	36786

Tez kapsamında yapılan deneylerin en son versiyonlarında, her bir şarkı metni öğesi için 36786 adet öznitelik kullanılmıştır. Çalışmalarda kullanılan ve daha önceden hazırlanan veri kümesinde toplam 1048 şarkı bulunmaktadır ve deneylerin büyük bir kısmında her bir şarkı için 36786 öznitelik kullanılmaktadır; bu da hazırlanan öznitelik vektör setinde toplam olarak 38551728 adet özniteliğe denk gelmektedir. Yaklaşık olarak 40 milyon adet öznitelik WekaTool aracılığı ile öğrenme ve sınıflandırma işlemine alınmaktadır. Her bir deney için yapılan bu çalışma, sonuç üretilmesi ve sonuçların yorumlanması sırasında zaman almaktadır. Özniteliklerin sayısının çok olmasının sebebi; 1048 adet şarkıda geçen toplam kelime kökleri ve bütün kelimeler için hesaplanan karakter 3-gram, karakter 4-gram, kelime sonu 2-gram, kelime sonu 3-gram, satır sonu 2-gram ve satır sonu 3-gram öznitelik sayılarının çok olmasıdır. 2-gram öznitelikleri ile 3-gram öznitelikleri arasında fark ve ya 3-gram öznitelikleri ve 4-gram öznitelikleri arsındaki farkın bu kadar fazla olmasının sebebi n-gram hesaplamaları yapılırken *n* değerinin artması ile üretilen sonuçların kombinasyonun artmasıdır. Örneğin; “seviyorsun” kelimesinin kelime sonu 2-gram’ı “un” kısmıdır. Kelimenin “un” kısmı bir başka şarkıda geçen “coştun” veya “sordun” kelimelerinin “un” kısmı ile aynı olduğu için öznitelik vektöründe bir adet öznitelik olarak temsil edilmektedir; fakat “seviyorsun”, “coştun” ve “sordun” kelimelerinin kelime sonu 3-gramları sırası ile “sun”, “tun” ve “dun” kısımlarıdır. Farklı kelimelerin, kelime sonu 2-gram’ları “un” ile bitebilirken, kelime sonu 3-gram’larıı çok farklı şekilde davranabilir. Bu sebepten dolayı, bu çalışma kapsamında kullanılan n-gram’ların *n* değeri arttıkça kullanılan öznitelik sayısı da artmaktadır.

Çizelge 2.6’da öznitelik vektörlerinin daha net anlaşılması için bir örnek verilmiştir. Öznitelik vektörünün büyüklüğünden dolayı her öznitelik Çizelge 2.6’da verilmemiştir, bunun yerine belirli öznitelik gruplarından bazıları seçilmiştir.

Çizelge 2.6 Örnek öznitelik vektörü

Şarkı	Ö1	...	Ö2	Ö3	...	Ö4	Ö5	...	Ö6	Ö7	...	Ö8	Ö9	Ö10	Ö11	Ö12	Yazar
1	2	-	0	1	-	2	0	-	1	0	-	0	496	78	59	6.36	**Sezen Aksu**
2	0	-	0	0	-	4	0	-	0	0	-	0	677	127	106	5,33	**Mustafa Sandal**
3	0	-	2	3	-	0	3	-	1	0	-	1	527	100	55	5,27	**Şebnem Ferah**
4	0	-	0	6	-	0	0	-	2	1	-	2	697	147	79	4,74	**Hakan Altun**
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**-**
1048	3	-	4	0	-	0	6	-	1	0	-	2	364	83	47	7,74	**Ferdi Tayfur**

Kısaltmalar
Ö1: Kelime kökü 1. Öznitelik
Ö2: Kelime kökü 4084. Öznitelik
Ö3: Karakter 3-Gram 1. Öznitelik
Ö4: Karakter 3-Gram 5713. Öznitelik
Ö5: Kelime sonu 2-Gram 1. Öznitelik
Ö6: Kelime sonu 2-Gram 321. Öznitelik
Ö7: Satır sonu 2-Gram 1. Öznitelik
Ö8: Satır sonu 2-Gram 234. Öznitelik
Ö9: Toplam Harf Sayısı Özniteliği
Ö10: Toplam Kelime Sayısı Özniteliği
Ö11: Farklı Kelime Sayısı Özniteliği
Ö12: Ortalama Kelime Uzunluğu Özniteliği

Çizelge 2.6'da Ö1 ve Ö2 olarak verilen kelime kökü öznitelikleri, şarkı metinlerindeki kelimelerin köklerinden alınmış iki kök kelimenin şarkılarda geçme frekansıdır. Örnek üzerinden açıklayacak olursak, Ö1 özniteliğinde frekans değeri hesaplanan kelime kökü, "anlıyorum", "anlamak", vb. kelimelerin kökü olan "anla" kelimesi olsun. Çizelge 2.6'da bakacak olursak, "anla" kelime kökü 1.şarkıda 2 kere 1048. şarkı da ise 3 kere geçmiştir. Burada ayrıca Ö1ile Ö2 öznitelikleri arasında toplamda 4082 adet kelime kökü frekansı bulunmaktadır. Ö1 ile Ö2 öznitelikleri arasında her biri birbirinden farklı kelime kökü için; bu kelime köklerinin şarkılarda geçen frekans değerlerini içermektedir. Ö3 ile Ö4 dahil olmak üzere toplam 5713 adet karakter 3-gram öznitelikleri bulunmaktadır. Örnek verecek olursak, Ö3 ve Ö4 dahil olmak üzere 5713 adet öznitelik, "çuy", "raz", "ücu" gibi her

biri birbirinden farklı 3-gram'ların şarkılarda geçme frekansını içermektedir. Ö3 için, "çuy" 3-gram'ı 3. şarkıda 3 kere geçerken, 4. şarkıda 6 kere geçmektedir. Aynı şekilde, Ö5-Ö6 arası kelime sonu 2-gram ve Ö7-Ö8 arası satır sonu 2-gram öznitelikleri de şarkı metni içerisindeki kendi frekans değerlerini temsil etmektedir. Öznitelik vektörünün çok büyük olmasından dolayı karakter 4-gram, kelime sonu 3-gram, satır sonu 3-gram ve satır uzunluğu istatistiksel öznitelikleri Çizelge 2.6'da belirtilmemiştir. Ö9, Ö10, Ö11 ve Ö12'de belirtilen öznitelikler global istatistiksel özniteliklerdir. Ö9 özniteliği, bir şarkı metnindeki toplam harf sayısı değeridir. Örneğin; 2. şarkıdaki metin içeriği toplam 677 adet harften oluşurken, 3. şarkıdaki metin içeriğinde toplamda 527 harf bulunmaktadır. Ö10, her bir şarkı metni için toplam kelime sayısını; Ö11 her bir şarkı metni içerisinde kullanılan farklı kelime sayısı değerlerini belirtmektedir. Ö12 ise bir şarkı metni içerisinde kullanılan kelimelerin ortalama uzunluğunu temsil eder. Ö12 ile belirtilen "Ortalama Kelime Uzunluğu" özniteliği, bir söz yazarının şarkılarında uzun ya da kısa kelimeler kullanıp kullanmadığını anlamada yardımcı olur. WekaTool'a girdi olarak verilen bu öznitelik vektöründe en son değer ize sınıfı temsil etmektedir. Çizelge 2.6'da sınıf, yazar olarak tanımlanmıştır. Tez kapsamında yapılan deneylerde belirtilen sınıf değer, "yazar", "kategori" ve "yıl-aralığı" olarak değişmektedir.

Şarkı sözü sınıflandırma çalışmasında, her bir öznitelik vektörü girdi olarak WekaTool'a verilmiştir. Her bir deney için, farklı sınıflandırma algoritmaları, farklı öznitelik seçim algoritmaları denenmiş ve gerçekleştirilen sınıflandırma işlemlerinden sonra çıktı olarak üretilen sonuçlar gözlemlenmiştir. Böylece hangi öznitelik vektörünün şarkı sözü sınıflandırmada daha etkili ve verimli olduğu deneyler sonucunda anlaşılmıştır.

2.3 Öznitelik Seçimi

Öznitelik seçim yöntemleri, belgeleri terimlerin bileşkesini alarak daha düşük boyutlu yeni bir uzayda kaynaştırılmış yeni özniteliklerle ifade eder. Bu sayede veri, sayıca daha az ve orijinallerinden bağımsız özniteliklerle ifade edilmiş olur. Öznitelik seçimi doğruluk ve ölçeklenebilirlik gibi sebeplerden dolayı kullanılır. Veri kümesi içerisindeki bazı öznitelikler sonucu direk etkileyebilecek gürültüye sahip olabilirler ve bu nedenle bu özniteliklerin veri kümesi içerisinden çıkartılması sonucun doğruluğunun artmasında etkili olacaktır. Ayrıca öznitelik seçimi sonrasında algoritmalarda kullanılacak olan verinin boyutu da azalacağı için işlem

gücü, hafıza ihtiyacı ve depolama gibi işlem süreci üzerinde de performans kazancı sağlayacaktır.

Öznitelik seçimi yöntemleri filtreleme ve sarmalama olarak iki gruba ayrılır. Sarmalama yöntemleri en iyi öznitelik seçiminde sınıflandırma yöntemleri kullanarak en iyi öznitelik kümesini bulmaya çalışır. Filtreleme yöntemleri ise, öznitelik ilişkilerinin belirlenmesinde istatistiksel yöntemlere başvurur ve belirlenen bir eşik değerinin üzerinde kalan öznitelik kümesini seçer. Özniteliklerin değerlendirilmesi ve gereksiz özniteliklerin elenmesi esnasında sarmalama yöntemi tüm veri kümesi üzerinde uygulanırken, filtreleme yöntemleri ise her özniteliği bağımsız olarak değerlendirir [43].

Bu çalışmada, her bir özniteliğin diğer özniteliklere göre değerlendirilmesi ve hangi özniteliklerin sınıflandırma aşamasında daha etkili olduğunu belirleyebilmek için Ki-kare ve ReliefF öznitelik seçim yöntemleri uygulanmıştır.

2.3.1 Ki-Kare (Chi-square)

Ki-kare testi (x^2) iki değişken arasındaki ilişkinin bağımlı veya bağımsız olduğunu belirlemeye yarayan ayrık veriler için kullanılan bir hipotez test yöntemidir [43]. Ki-kare istatistiğine dayalı öznitelik seçimi metodu iki adımı içermektedir. Yöntemin ilk kısmında özniteliklerin sınıflara göre ki-kare istatistikleri hesaplanır. İkinci kısımda serbestlik derecesi ve belirlenen önemlilik seviyesine göre ki-kaynaşımı (chi-merge) prensibi ile ki-kare değerlerine bakılarak veri kümesi içerisindeki tutarsız özniteliklerin bulunana kadar art arda özniteliklerin ayrıştırılmasıdır [44]. Ki-kaynaşımı algoritması 1992 yılında RandyKerber tarafından yazılmış ve 1995 yılında HuanLiu ve RudySetiono tarafından ise yeniden düzenlemiştir. Veri kümesi içinde yer alan bir faktör için hesaplanan ki-kare değeri, o faktörün sınıf içerisindeki bağımlılığını ölçmektedir. Sıfır değerine sahip bir faktör o küme içinde bağımsız olduğunu gösterir. Yüksek bir ki-kare değerine sahip olan faktör, veri kümesi için daha tanımlayıcıdır. Ki-kare değerinin hesaplanmasında kullanılan genel eşitlik aşağıda verilmiştir.

$$x^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{\left(A_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{2.18}$$

Bu eşiklikte k sınıf sayısı, A_{ij}gözlenen frekans değeri (i satır, j sütün)ve E_{ij}ise beklenen (teorik) frekans değeridir.

2.3.2 ReliefF

Relief yöntemi, özniteliklerin değerini aralarında bulunan ya da bulunmayan bağımlılıkları ortaya çıkarmaya çalışarak bulmayı hedefler [45]. Relief algoritmasının ana fikri, k en yakın komşuluk algoritmasının temel kuralına benzemektedir. Yani verilen bir mesafeye daha yakın uzaklıkların aynı sınıfa ait olmaları çok daha olasıdır. Eğer bir öznitelik yararlı ise, aynı sınıfın en yakın mesafeleri o öznitelik boyunca verilen aralığa, diğer tüm sınıfların en yakın mesafelerinden daha yakın olması beklenecektir. Dolayısıyla, verilen *i* özniteliğinin önem derecesi [46] de hesapladığı gibi;

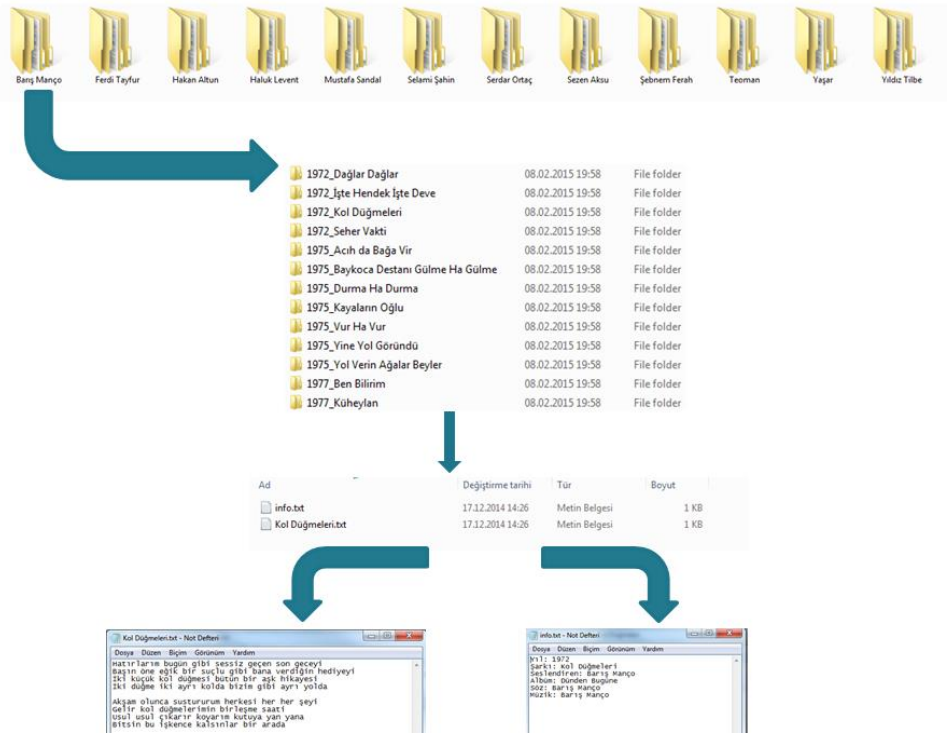
$$\left[\sum_{j=1}^{m} -diff\left(x_{ij}, nearhit_{ij}\right) + diff\left(x_{ij}, nearmiss_{ij}\right)\right] / m \qquad (2.19)$$

biçiminde verilir. Burada m örnek boyutu (eğitim setinden rasgele seçilmiş bir alt küme) ve $diff\left(x_{ij}, nearhit_{ij}\right)$, rasgele seçilmiş *j* mesafesi içindeki *i* özniteliğinin değeri ile aynı sınıfa sahip en yakın eğitim örneğindeki *i* özniteliğinin $(nearhit_{ij})$ değeri arasındaki farktır. Paralel olarak, $nearmiss_{ij}$ farklı bir sınıfa sahip en yakın eğitim örneğindeki *i*'nin değeri olarak tanımlanır. Yararlı öznitelikler için x_{ij} ve $nearhit_{ij}$ değerlerinin çok yakın olması beklenir. Eğer bir öznitelik yararlı değil ise, her iki farklılığın da hemen hemen aynı dağılımı almaları beklenir [47].

3. SONUÇLAR

3.1 Veri Kümesi

Türkçe şarkı sözü madenciliği çalışmasında farklı cinsiyetlerden ve farklı müzik tarzlarından 12 şarkı sözü yazarına ait toplam 1048 şarkıdan oluşan bir veri kümesi oluşturulmuştur. Şarkı sözleri verileri hazırlanırken, verilerin doğruluğunu kontrol edebilmek için şarkı sözü yazarlarının kendi siteleri öncelikli olarak kullanılmıştır. Eğer şarkı sözü yazarının resmi internet sitesinde, kendisine ait şarkı sözleri bulunmuyor ise birçok farklı kaynaktan kontrol edilerek şarkı sözleri hazırlanmıştır. Hazırlanan şarkı sözleri daha sonra lokal dosya veritabanı olarak hazırlanmış ve daha sonrasında bu veriler MySQL veritabanına aktarılmıştır. Hazırlanan veri kümesinin lokal dosya tabanlı saklandığı yapı Şekil 3.1'de gözükmektedir.



Şekil 3.1 Veri kümesinin dosya tabanlı tutulduğu yapı

Şekil 3.1'de belirtilen yapıda tutulan şarkılar bazı önişlemlerden geçirilerek ve öznitelikleri çıkartılarak daha hızlı bir şekilde işlem yapabilmek amacı ile MySQL veritabanına aktarılmıştır. Çalışmaya yeni bir öznitelik eklenmesi durumunda veritabanı üzerinde şarkıların bulunduğu tabloya, bir ilişki tablosu üzerinden eklenen yeni özniteliklerin tablosu bağlanarak işlemler gerçekleştirilebilmektedir. Bu sayede deneylerde kullanılacak olan her bir şarkının kendisine ait bilgileri ve

öznitelikleri daha kolay bir şekilde elde edilmekte ve performans açısından kazanç sağlanmaktadır.

Veri kümesi hazırlanırken Türkiye'de popüler olan müzik türleri dikkate alınmıştır. Şarkı sözü yazarları "Pop", "Rock" ve "Arabesk-Fantezi" olmak üzere üç kategoriye ayrılmıştır. Veri kümesinde bulunan söz yazarları ve bu söz yazarlarının bulunduğu kategoriler eşit sayıda olacak şekilde hazırlanmıştır. Her kategori için dört adet şarkı sözü yazarı bulunmaktadır. Veri kümesine dahil olan söz yazarlarının ve söz yazarlarının ait olduğu müzik tarzları Çizelge 3.1'de verilmiştir.

Çizelge 3.1 Veri kümesindeki şarkı sözü yazarlarının ait olduğu kategoriler

Söz Yazarı	**Tür**
Sezen Aksu	Pop
Serdar Ortaç	Pop
Yaşar	Pop
Mustafa Sandal	Pop
Teoman	Rock
Haluk Levent	Rock
Barış Manço	Rock
Şebnem Ferah	Rock
Selami Şahin	Arabesk-Fantezi
Yıldız Tilbe	Arabesk-Fantezi
Ferdi Tayfur	Arabesk-Fantezi
Hakan Altun	Arabesk-Fantezi

Veri kümesinde bulunan 1048 adet şarkı sözü dört adet kategori ve üç adet yıl aralığına ayrılırken, her gruba yakın sayıda şarkı atanmasına dikkat edilmiştir. Çizelge 3.2'de tüm veri kümesinin yıl aralıkları, kategoriler ve şarkı sözü yazarları üzerindeki dağılımları gösterilmiştir. Ayrıca Çizelge 3.2'de her bir alan için kaç adet şarkı sözü bulunduğu ve bu alanların oluşturduğu grupların toplamda kaç adet şarkı sözü içerdiği belirtilmiştir.

Çizelge 3.2 Veri kümesinin gruplara ve şarkı sözü yazarlarına göre dağılımı

Şarkıcılar	1972-1993	1994-2006	2007-2014	TOPLAM	
Sezen Aksu	53	55	16	124	**POP 365**
Serdar Ortaç	0	75	59	134	
Yaşar	0	38	11	49	
Mustafa	0	42	16	58	
Teoman	0	55	13	68	**ROCK 266**
Haluk	9	45	3	57	
Barış Manço	67	9	0	76	
Şebnem	0	44	21	65	
Selami	52	37	3	92	**FANTEZİ 417**
Yıldız Tilbe	0	39	18	57	
Ferdi Tayfur	112	65	0	177	
Hakan Altun	0	59	32	91	
	293	563	192	**1048**	

Şarkı sözleri yıl aralıklarına ayrılırken Türkiye'deki sosyokültürel değişim göz önüne alınmıştır. Bu yöntem ile şarkı sözlerindeki yapısal değişiklikler 1993 öncesi, 1994 yılı ile 2006 yılı arası ve 2007 yılı sonrası olarak ön görülmüştür. 1993 yılı Türkçe pop müziğin yükselmeye başladığı yıldır. Şehirlerde dinlenen müzik türlerinde, özellikle arabesk müzik ile karşılaştırılırsa 1993 yılından sonra Türkçe pop müzik giderek popülerliğini arttırmıştır. Türk Pop Müziği, 1995'lerden sonra, alaturka müzikle beslenmeye başlamıştır. Türk Pop Müziği, dünya pop çizgisinden kaymış ancak kendi çizgisini nispeten de olsa oluşturabilmiştir. Bu sonuçlar doğrultusunda daha eski eserlerdeki muhafazakar ve artistik modalara kıyasla, hem ezgi hem de sözlerde daha serbest tarzda bir ortaya çıkmasını sağlamıştır. Türkiye'de 2000'li yılların ortalarında, yeni nesli önemli ölçüde etkileyecek birçok rock müzik grubu ortaya çıkmaya başlamıştır. Rock müziğin popülerliği genel olarak şarkı sözlerindeki yapıyı da etkilemiştir. Günlük

konuşmalardaki kelime ve ifadeler ile yazılan şarkı sözleri yerine daha yaratıcı şarkı sözleri ortaya çıkmıştır. Yukarıdaki gibi belirtilen kriterlerden dolayı şarkı sözlerinde ortaya çıkan yapısal değişikliklerden dolayı şarkı sözleri üç adet gruba ayrılmıştır.

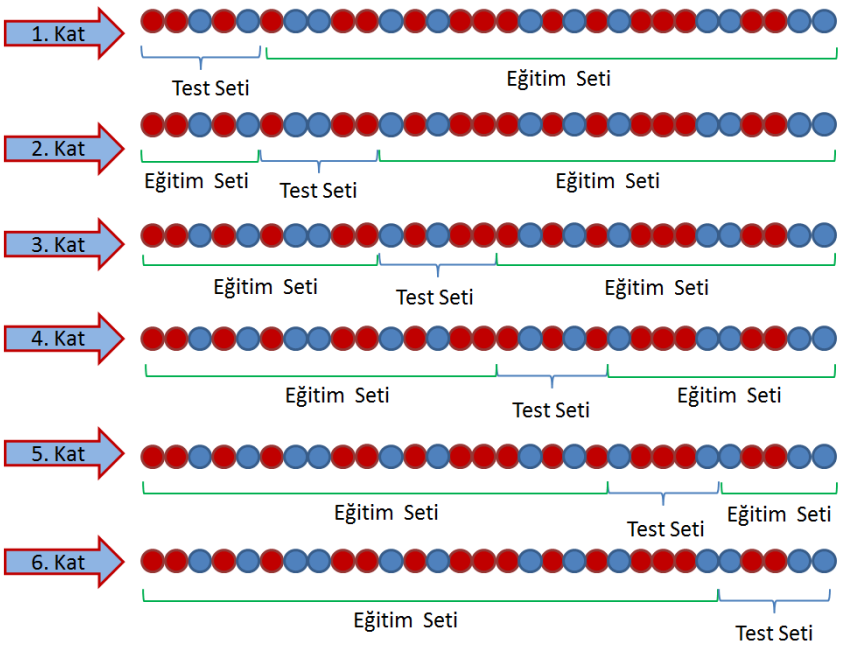
Tez çalışması kapsamında veri kümesi hazırlanırken yıl aralıklarına ve kategorilere göre şarkı sözlerinin seçilmesine ve dağılımına dikkat edilmiştir. Veri kümesinde, kategorilere göre bakacak olursak, "Pop" türü için 365; "Rock" türü için 266 ve "Arabesk-Fantezi" türü için ise 417 adet şarkı bulunmaktadır. Aynı şekilde yıl aralıklarına göre, 1972 ile 1993 yılları arasında 293; 1994 ile 2006 yılları arasında 563; 2007 ile 2014 yılları arasında ise 192 şarkı bulunmaktadır. Çizelge 3.2'de şarkı sözlerinin kategorilerinin yıllara göre gösterdiği değişimde net bir şekilde gözükmektedir. Elde edilen veri kümesindeki her türe ait yazar sayısı eşittir ve her kategorideki örnek sayısı birbirine çok yakındır. Yayınlanma tarihlerini baz alarak şarkıları sınıflara atarken, o tarihlerde ülkede meydana gelen sosyo-kültürel değişiklikler konusuyla ilgilendik. 1993, Türkçe pop müziğin özellikle kentsel topluluklarda büyük bir patlama yaşadığı ve arabeske kıyasla çok popüler hale geldiği bir yıl olmuştur. Bu durum eski eserlerdeki daha tutucu ve artistik modaya kıyaslandığında hem melodi de hem de şarkı sözünde özgür tarzda bir içerikle sonuçlanmıştır. 2000'li yılların ortasında ülkede yeni neslin dinleme tarzını önemli derecede etkileyen birçok rock grubu ortaya çıktı. Rock müziğin popülerliği şarkı sözü tarzını genel anlamda etkilemiştir. Günlük diyaloglardan sözcük ve tabirler kullanılması yerine, şarkı sözü yazarları şarkı sözlerinde yaratıcı dilsel tasarılar kullanmaya zorlanmaktaydı. "Arabesk-Fantezi" türündeki şarkı sözleri 1972 ile 1993 yılları arasında diğer müzik türlerine göre daha fazlayken, "Pop" müzik türü ise 2007 ile 2014 yılları arasında diğer kategorilere göre fazla sayıda veri içermektedir. 1994 ile 2006 yılları arasında ise her müzik türüne ait şarkı sözü birbirlerine yakındır. Şarkı sözü madenciliğinde, sınıflandırma çalışmalarına başlanmadan önce veri kümesinin hazırlanması aşamasında yukarıda bahsedilen verilerin dağılımına, deneylerin doğru ve etkili bir şekilde gerçekleştirilmesi için özellikle dikkat edilmiştir.

Türkçe şarkı sözü madenciliği çalışmasında kullanılan tüm veri kümesi "www.baskent.edu.tr/~hogul/lyrics" adresinde bulunmaktadır.

3.2 Deney Düzeneği

3.2.1 N-Kat çapraz doğrulama yöntemi (N-Fold cross validation)

Tez kapsamında kullanılan şarkı sözü sınıflandırma algoritmalarının performansı genellikle tahmin hatası ile ölçülmektedir. Sınıflandırma işlemlerinin çoğunda hata tam olarak hesaplanamamaktadır ve bu hata tahmin edilmelidir. Dolayısıyla bu aşamada uygun bir tahmin edici yönteminin seçilmesi önemlidir. N-kat çapraz doğrulamada veri kümesi n adet eşit parçaya bölünmektedir ve bir sınıflandırıcı n-1 parça ile eğitilmektedir ve geri kalan parçada sınıflandırıcı ile veri kümesi üzerinde test edilerek bir hata değeri hesaplanmaktadır. Şekil 3.2'de örnek olarak 6 kat çapraz doğrulama modeli gösterilmiştir. N-kat çapraz doğrulama işlemi gerçekleştirilirken hata tahmini, her parçadaki elde edilen hatanın ortalama değeri olarak elde edilmektedir. N-kat çapraz doğrulama işlemi gerçekleştirilirken hata tahmin edici yöntem iki etmene bağlıdır; birisi eğitim setiyken, diğer etmen ise bölünen parçalardır.



Şekil 3.2 6 Kat çapraz doğrulama modeli

Şarkı sözlerini modellerken performansı hesaplamak için kestirim hatasının tahmini gerekmektedir. N-kat çapraz doğruluma, kestirim hatalarının tahmin edilmesi aşamasında çoğunlukla kullanılmaktadır. N-kat çapraz doğrulamadaki n değeri küçük olduğunda gerçek veri analizinde bir problem galine gelebilmektedir. N-kat çapraz doğrulama, veri kümesi üzerinde model seçimi için kullanılmaktadır

fakat daha iyi bir kestirim hatası tahmini yöntemi, daha iyi bir veri kümesi üzerinde model seçim kriterine götürmek zorunda değildir.

Bu çalışmada, sınıflandırma performansı değerlendirmeleri 10-kat çaprazlama deneyleri ile yapılmıştır. Bunun için, her defasında veri kümesinin farklı bir onda dokuzu eğitim kümesi kalan onda biri de test kümesi seçilerek, toplam on kez sınıflandırma algoritması uygulanmıştır. Böylece her bir şarkı nesnesi bir kez sınıflandırmaya tabi tutulmuştur. Bu işlem seçilen her öznitelik kombinasyonu için tekrarlanmıştır.

3.2.2 Model başarım ölçütleri

Model başarısı değerlendirilirken kullanılan ölçütler arasında doğruluk-hata oranı (accuracy-error rate), anma (recall), duyarlılık (precision) ve F-ölçütü (F-measure) yer almaktadır. Şarkı sözü sınıflandırmada modelin başarısı, doğru sınıfa atanan örnek sayısı ve yanlış sınıfa atanan örnek sayısı nicelikleriyle ilgilidir.

Model üzerindeki başarı testinin yapılması sırasında, başarı oranını doğru ölçebilmek için veri kümesindeki öğrenme veri kümesini işleme almamak gerekir. Veri kümesindeki öğrenme veri kümesi dışında kalan verilere test veri kümesi denilmektedir. Öğrenme veri kümesi üzerinde model oluşturulduktan sonra oluşan model test veri kümesinde sınanır.

Test sonucunda ulaşılan sonuçların başarılılık bilgileri hata matrisi ile ifade edilebilir. Hata matrisinde satırlar test kümesindeki örneklere ait gerçek sayıları, kolonlar ise modelin tahminlemesini ifade eder. Çizelge 3.3'de iki sınıflı bir veri kümesinde oluşturulmuş bir modelin hata matrisi verilmiştir. Sınıf sayısı ikiden fazla olduğunda 2x2 boyutundaki bu matris, n sınıf sayısı olmak üzere, n x n boyutlarında genişletilmiş bir matris şeklini alacaktır. TP ve TN değerleri doğru sınıflandırılmış örnek sayısıdır. False Pozitif (FP), aslında 0 (negatif) sınıfındayken 1 (pozitif) olarak tahminlenmiş örneklerin sayısıdır. FalseNegative (FN) ise 1 (pozitif) sınıfındayken 0 (negatif) olarak tahminlenmiş örneklerin sayısını ifade eder. Genel olarak n x n boyutlarındaki bir hata matrisinde ana köşegen doğru tahminlenmiş örnek sayılarını; ana köşegen dışında kalan matris elemanları ise hatalı sonuçları ifade etmektedir [48]. Hata matrisi Çizelge 3.3'de gösterilmektedir.

Çizelge 3.3 Hata matrisi

		Öngörülen Sınıf	
		Sınıf = 1	Sınıf = 0
Doğru Sınıf	Sınıf = 1	a	b
	Sınıf = 0	c	d
a: TP (Doğru Kabul - True Pozitif) b: FN (Yanlış Red -False Negatif) c: FP (Yanlış Kabul - False Pozitif) d: TN (Doğru Red - True Negatif)			

Belirtilen bu tanımlar için başarım ölçütleri hesaplanır. Buna göre;

3.2.2.1 Doğruluk-hata oranı (Accuracy-error rate)

Model başarımının ölçülmesinde kullanılan en popüler ve basit yöntem, modele ait doğruluk oranıdır.

Doğru sınıflandırılmış örnek sayısının $(\mathrm{TP}+\mathrm{TN})$, toplam örnek sayısına $(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})$ oranıdır. Doğruluk oranı denklem şeklinde ifade edilirse:

$$\text{Doğruluk} = \frac{\mathrm{a+d}}{\mathrm{a+b+c+d}} = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+FN+TN}} \tag{3.1}$$

Hata oranı ise bu değerin 1'e tamlayanıdır. Diğer bir ifadeyle yanlış sınıflandırılmış örnek sayısının $(\mathrm{FP}+\mathrm{FN})$, toplam örnek sayısına $(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})$ oranıdır. Hata oranı denklem şeklinde ifade edilirse:

$$\text{Hata Oranı} = \frac{\mathrm{b+c}}{\mathrm{a+b+c+d}} = \frac{\mathrm{FP+FN}}{\mathrm{TP+FP+FN+TN}} \tag{3.2}$$

3.2.2.2 Anma (Recall)

Doğru sınıflandırılmış pozitif örnek (TP) sayısının, toplam pozitif örnek sayısına $(TP+FN)$ oranıdır. Duyarlılık denklem şeklinde ifade edilirse:

$$\text{Duyarlılık} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{3.3}$$

3.2.2.3 Duyarlılık (Precision)

Kesinlik, sınıfı 1 olarak tahmin edilmiş True Pozitif (TP) örnek sayısının, sınıfı 1 olarak tahmin edilmiş tüm örnek sayısına ($TP + FP$)oranıdır. Kesinlik denklem şeklinde ifade edilirse:

$$\text{Kesinlik} = \frac{\text{TP}}{\text{TP+FP}} \tag{3.4}$$

3.2.2.4 Özgüllük

Özgüllük, Doğru Red (TN) örnek sayısının, sınıfı negatif tahmin edilmiş tüm örnek sayısına $(\text{TN} + \text{FN})$ oranıdır. Özgüllük denklem şeklinde ifade edilirse:

$$\text{Özgüllük} = \frac{\text{TN}}{\text{TN+FN}} \tag{3.5}$$

3.2.2.5 F-ölçütü (F-measure)

Kesinlik ve duyarlılık ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli değildir. Her iki ölçütü beraber değerlendirmek daha doğru sonuçlar verir. Bunun için f-ölçütü tanımlanmıştır. F-ölçütü, kesinlik ve duyarlılığın harmonik ortalamasıdır [49]. Bu değer maksimum 1 olabilir. İlgili sınıfın, sınıflandırıcı algoritma tarafından ne derece doğru sınıflandırma tahmini yapabileceğini gösterir. Değer 1'e yaklaştıkça ilgili sınıfın algoritma tarafından öğrenilmesi o derece başarılı anlamına gelir.

$$\text{F} - \text{ölçütü} = \frac{2\times\text{Duyarlılık}\times\text{Kesinlik}}{\text{Duyarlılık+Kesinlik}} \tag{3.6}$$

F-ölçütü özellikle veri kümesi üzerindeki eğitim kümesinin hazırlanışında, uygulanacak sınıflandırıcının performansını artırmak için önemli bir ölçüt olmaktadır. F-ölçütü, hem yanlış pozitif hem de yanlış negatif değerlerini bünyesinde barındırdığından, başarım ölçütü olarak daha başarılıdır ve tez kapsamında kullanılan sınıflandırma algoritmalarının birbirleri ile karşılaştırılmasında daha verimlidir.

3.2.2.6 ROC (receiver operating characteristics) eğrisi

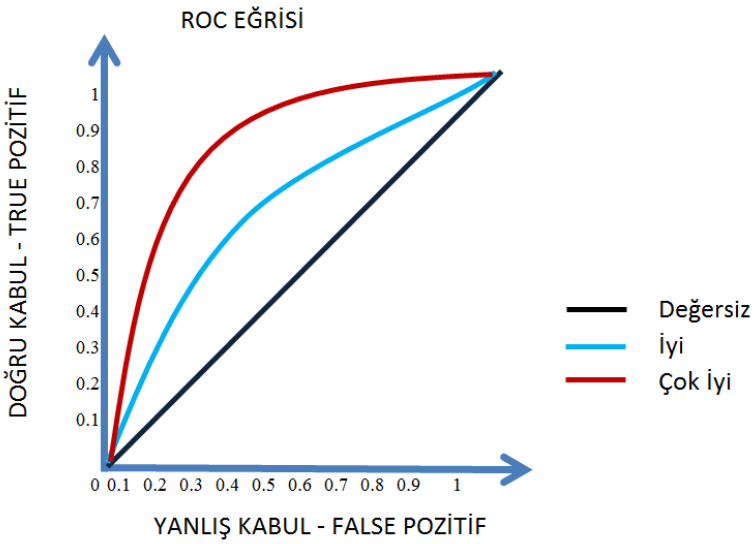
İlk ROC eğrisi 2. Dünya Savaşı'nda radar sinyallerinin analizi için kullanılmıştır. Düşman uçaklarını, radar sinyallerini kullanarak daha doğru bir şekilde saptamak amacıyla araştırmalara başlanmıştır [50, 51]. Sinyal algılama çalışmalarında kullanılan ROC eğrileri daha sonra tanı testi performanslarının belirlenmesinde,

finansal çalışmalarda kredi skorlamada ve veri madenciliğinde yaygın olarak kullanılmaktadır. Bu çalışmada, şarkı sözlerinin yazar, kategori ve yıl aralığı için sınıflandırılması aşamasında ROC analizi kullanılmıştır.

ROC eğrisi, ikili sınıflandırma sistemlerinde ayrım eşik değerinin farklılık gösterdiği durumlarda, doğru pozitiflerin yanlış pozitiflere olan kesri olarak da ifade edilebilir. ROC eğrisini grafiksel olarak açıklayacak olursak, farklı değerler için dikey eksen üzerinde doğru kabul (true pozitif) değerlerinin, yatay eksen üzerinde ise yanlış kabul (false pozitif) değerlerinin oranlarının yer aldığı bir eğridir. ROC eğrisi üzerindeki her nokta, farklı eşik değerlerine karşılık gelen duyarlılık ve özgüllük değerlerini ortaya koyar. Genelde düşük yanlış pozitiflik oranlarını veren eşik değerleri, düşük doğru pozitiflik oranına da sahiptir. Gerçekleştirilen deneyin başarılı olarak adlandırılabilmesi için; doğru pozitif oranın yüksek, yanlış pozitif oranın ise düşük olması gerekmektedir. ROC eğrisi y eksenine (yanlış pozitif) yaklaştıkça başarının seviyesi düşer. Sistemin başarısının tek bir değer ile ifade edilmesi gerekirse bu da ROC eğrisinin altında kalan alan ile ifade edilmelidir. ROC eğrisinin altındaki bu alan ne kadar büyük ise sistemin güvenilirliği ve sonuçların başarısı da o kadar yüksek olur. Sonuçlar üzerinde karar verebilmek için en doğru durum, doğru pozitiflik oranı yüksek ve yanlış pozitiflik oranı düşük olma durumudur.

Area Under The Curve (AUC): ROC eğrisinin altında kalan alana verilen isimdir ve sınıflandırma üstünlüğü için bir karşılaştırma ölçütü olarak kullanılır. ROC eğrisi altında kalan alan büyük ise, eğri x (doğru pozitif) eksenine daha yakın demektir, bu da sınıflandırma durumunun tahmin edilmesinde söz konusu deney, iyi sonuç alınmış bir deneydir. Bu çalışmada ROC eğrisi ile AUC (Area Under The Curve) değerleri analiz edilmiş, yorumlanmış ve bu sonuçlara göre sınıflandırma işlemleri gerçekleştirilirken izlenen adımları üzerinde değerlendirmeler yapılmıştır.

İdeal ve kötü performans gösteren testlere ilişkin ROC eğrileri Şekil 3.3'te verilmiştir.



Şekil 3.3 Eğrisi performans değerlendirmesi

Eğri altındaki alanın değerinin yorumlanmasında derecelendirmeler kullanılabilir:

- 0.80 – 0.90 = iyi
- 0.70 – 0.80 = orta
- 0.60 – 0.70 = zayıf
- 0.50 – 0.60 = başarısız

Rasgele durumda, ROC eğrisinin altında kalan alan 0.50 değerine eşittir. Sınıflandırma sırasında doğruluk oranı arttıkça, sıfır yanlış pozitif ve sıfır yanlış negatif değerleri elde edildikçe alanın değeri 1.00 olacaktır. Bu da mükemmel sonuç anlamına gelmektedir.

Sınıflandırma işlemleri Weka (Waikato Environment for Knowledge Analysis) aracı kullanılarak test edilmiştir.

Weka, Yeni Zelanda Waikato Üniversitesi'nde geliştirilen bir veri madenciliği ve makine öğrenmesi yazılımıdır. Weka yazılımı nesneye yönelik programlama dillerinden olan Java ile geliştirilmiştir; çünkü Java kullanımı birçok değişik öğrenme algoritmalarının düzenli ve etkili bir biçimde kullanılabildiği bir platformdur. Weka birçok sınıflandırma tekniğini bünyesinde bulundurduğu için sınıflandırma çalışmalarında güçlü bir platformdur. Ayrıca, Weka çalışmalara komut girilerek de çalışmasına olanak sağladığı için tercih edilmektedir. Weka'nın içerisinde Veri İşleme, Veri Sınıflandırma, Veri Kümeleme, Veri İlişkilendirme özellikleri mevcuttur.

Çalışma kapsamında gerçekleştirilebilecek bazı Weka yetenekleri şunlardır:

- Kullanılacak olan veri kümesi üzerinde öğrenme metotları uygulayabilir
- Veri kümesi için istatistiksel ve sayısal bilgiler edinebilir
- Veri kümesi üzerinde kullanılan özniteliklerin seçimi için çeşitli öznitelik seçim algoritmaları uygulanabilir
- Veri kümesi deneyler ve çalışma doğrultusunda tekrar yapılandırılabilir
- Veriler görselleştirilebilir.
- Sınıflandırma işleminde kullanılacak eğitim ve test veri kümeleri için ilgili parametreler ayarlanabilir.

Projenin amacına göre açılan sayfadaki uygun tabdaki (Sınıflandırma, Kümeleme, İlişkilendirme) uygun algoritma veya algoritmalar seçilerek veriler üzerine uygulanmakta ve en doğru sonucu veren algoritma seçilebilmektedir.

Bu çalışma kapsamında, veri kümesi üzerinden ilgili sınıflar (yazar, kategori ve yıl aralığı) için Weka programında çalıştırabilmek için ARFF (Attribute-Relation File Format) dosyaları üretilmiştir. Weka'ya girdi olarak verilen bu ARFF dosyaları her deney için beri kümesi ayrı bir şekilde düzenlenmiştir. Weka aracı ile ARFF dosyası içinde bulunan veri kümesine tez kapsamında kullanılan öznitelik seçme algoritmaları, sınıflandırma algoritmaları gibi çeşitli yöntemler uygulanmış ve sonuçlar detaylı bir şekilde incelenmiştir.

3.3 Deneysel Sonuçlar

Deneyler üzerinde gerçekleştirilen sınıflandırma performans değerlendirmeleri 10-kat çaprazlama deneyleri ile yapılmıştır. Bunun için, her defasında veri kümesinin farklı bir onda dokuzu eğitim kümesi kalan onda biri de test kümesi seçilerek, toplam on kez Naif Bayes algoritması uygulanmıştır. Böylece her bir şarkı nesnesi bir kez sınıflandırmaya tabi tutulmuştur. Bu işlem seçilen her öznitelik kombinasyonu için tekrarlanmıştır.

3.3.1 Sınıflandırma algoritmalarına göre sonuçlar

Bu çalışma kapsamında gerçekleştirilen deneyler, öncelikle hangi sınıflandırma algoritmasının daha başarılı olduğunu bulmak için yapılmıştır. Veri kümesi üzerinde öncelikle "O" öznitelik veri kümesi için dört adet sınıflandırıcı denenmiştir. "O" öznitelik kümesi Çizelge 2.3'te Öznitelik Kümeleri ve Kısaltmaları tablosunda

hangi özniteliklerden oluştuğu belirtilmiştir. Seçilen “O” öznitelik kümesi üzerinde gerçekleştirilen deneyler ROC eğrisinin altında kalan alan (AUC) değerlerine göre değerlendirilmiştir. Çizelge 3.4’te “O” öznitelik kümesine uygulanan sınıflandırıcıların performansları gözükmektedir.

Çizelge 3.4 Öznitelik kümesi üzerinde sınıflandırıcı performansları

Sınıflandırma Algoritması	Öznitelik Kümesi	Yazar Sınıfı için AUC Değerleri	Kategori Sınıfı için AUC Değerleri	Yıl Aralığı Sınıfı için AUC Değerleri
Radyal Tabanlı Destek Vektör Makinesi (LibSVM RBF)	O	0.579	0.678	0.514
Doğrusal Destek Vektör Makinesi (LibSVM Linear)	O	0.739	0.770	0.610
Naif Bayes	O	0.679	0.687	0,578
Multinom Naif Bayes	O	0.860	0.873	0.674

Çizelge 3.4’te görüldüğü gibi dört adet sınıflandırma yöntemi bulunmaktadır. Bu dört adet sınıflandırma yöntemi temel iki algoritmaya aittir. Bunlar; Destek Vektör Makinesi ve Naif Bayes algoritmalarıdır. Çizelge 3.4’te Doğrusal DVM’sı yönteminin Radyal Tabanlı DVM’sine göre daha başarılı olduğu; Multinom Naif Bayes yönteminin ise Gaussian Naif Bayes yöntemine göre daha başarılı olduğu görülmektedir. “O” öznitelik kümesi üzerinde gerçekleştirilen sınıflandırıcı performansları üç sınıf için gerçekleştirilmiştir. Bunlar; söz yazarları, kategoriler ve yıl aralığı sınıflarıdır. Her üç sınıf için de “Multinom Naif Bayes” algoritması uygulandığı zaman, diğer üç algoritmaya göre daha başarılı sonuçlar elde edilmiştir. Uygulanan sınıflandırma algoritmaları sonucunda kategorilerin diğer sınıflara göre daha başarılı sınıflandırıldığı görülmektedir. Çizelge 3.4’te belirtilen deneyler sonucunda söz yazarları ise yıl aralıklarına göre daha başarılı şekilde sınıflandırılmıştır. Bu deney ile elde edilen sonuçlar ile giriş kısmında bahsedilen

“Şarkı sözleri müzik içeriğinin temsilinde ne kadar etkilidir?” sorusuna ilk cevap olarak düşünülebilir. İlk adımı gerçekleştirilen deneyler sonucunda dört yöntemin kullanılması performans kaybettireceği için; geri kalan çalışmada “Doğrusal DVM” ve “Multinom Naif Bayes” olmak üzere en iyi iki sınıflandırma algoritması kullanılmıştır.

3.3.2 Öznitelik kümelerine göre sonuçlar

En iyi iki sınıflandırma algoritması seçildikten sonra, öznitelik kümelerinin “Doğrusal Destek Vektör Makineleri” ile gerçekleştirilen sınıflandırma üzerindeki etkisi gözlemlenmiştir. Çizelge 3.5’te özniteliklerin “Yazar”, “Kategori” ve “Yıl Aralığı” sınıfları üzerindeki performansları belirtilmiştir.

Çizelge 3.5 Öznitelik kümelerinin Doğrusal Destek Vektör Makinleri sınıflandırıcısı ile sınıflar üzerindeki etkileri

Sınıflandırma Algoritması	Öznitelik Kümesi	Yazar Sınıfı için AUC Değerleri	Kategori Sınıfı için AUC Değerleri	Yıl Aralığı Sınıfı için AUC Değerleri
Doğrusal Destek Vektör Makineleri (LibSVM Linear)	P	0.689	0.704	0.590
	A	0.713	0.754	0.581
	B	0.721	0.749	0.583
	C	0.719	0.750	0.594
	D	0.717	0.756	0.593
	E	0.725	0.750	0.595
	F	0.724	0.755	0.600
	G	0.724	0.756	0.599
	H	0.722	0.754	0.599
	I	0.722	0.753	0.599
	J	0.732	0.771	0.604
	K	0.728	0.760	0.607
	L	0.739	0.760	0.602
	M	0.697	0.769	0.600
	N	0.728	0.758	0.601
	O	0.739	0.770	0.610
	R	0.697	0.722	0.600
	S	0.733	0.770	0.604
	T	0.697	0.723	0.599
	U	0.717	0.745	0.593
	V	0.733	0.770	0.604

Tez kapsamında yapılan ilk deneylerde her dil için uygun olacak temel öznitelikler kullanılmıştır. Bunlar A ile I dahil olmak üzere bu iki öznitelik kümesi arasındaki öznitelik kümeleridir. Bu öznitelik kümeleri yapılan araştırmalar ve çalışmalar

sonucunda ortaya çıkan kelimenin kökü, kelime 3 gram, kelime 4 gram ve temel istatistiksel özniteliklerdir. Şarkı metninde geçen “Harf Sayısı”, “Kelime Sayısı”, “Farklı Kelime Sayısı” ve “Ortalama Kelime Uzunluğu” gibi istatistiksel verilerin kombinasyonu ile A-I öznitelik vektör kümeleri oluşturulmuş ve sınıflandırma üzerindeki performansları çalışma kapsamında gerçekleştirilen ilk deneyler üzerinde gözlemlenmiştir. Daha sonraki çalışmalarda, yukarıda belirtilen sözcüksel özniteliklere şarkı sözü metnindeki satır bilgilerini içeren istatistiksel öznitelikler eklenmiştir. Satır uzunluğuna bağlı özniteliklerin bu çalışma kapsamında kullanılmasının nedeni; bu öznitelik kümelerinin şarkı sözü yazarlarının sınıflandırılmasında önemli bir etmen olabileceğinin düşünülmesidir. Satır uzunluğu ile ilgili kullanılan bu öznitelikler; satır için maksimum minimum farkı, medyan, ortalama satır uzunluğu ve standart sapmadır. Bu öznitelikler Çizelge 3.5’te J ile O dahil olmak üzere bu iki öznitelik kümeleri arasındaki öznitelik kümeleridir. Gerçekleştirilen bu aşamaya kadar olan deneyler arasında her bir sınıf için en iyi seçimin O öznitelik kümesi olduğu görülmektedir. Kelime 3 gram ve kelime 4 gram’ın sınıflandırma üzerindeki etkisinin görülmesi için P öznitelik kümesi (Sadece Kelimenin Kökü) ile A öznitelik kümesinin (Kelimenin Kökü + Kelime 3 Gram + Kelime 4 Gram) karşılaştırılması gerçekleştirilmiştir ve söz yazarlarının sınıflandırılmasında A öznitelik kümesinin P öznitelik kümesine göre daha başarılı bir küme olduğu gözlemlenmiştir. Satır istatistiksel özniteliklerinin sınıflandırma üzerindeki başarılı etkisi gözlemlendikten sonra kelime ve satır sonundaki N-Gram’ların sınıflandırma üzerindeki etkisi gözlemlenmiştir. Kelime sonu 2 gram ve kelime sonu 3 gramların söz yazarlarının kelimeler üzerinde kullandıkları ekler hakkında; satır sonu 2 gram ve satır sonu 3 gramların ise şarkı sözlerinde kullanılan kafiyelerin belirlemesinde yol göstereceği düşünülmüştür. R ile V kümesi dahil olmak üzere bu iki küme arasındaki bütün öznitelik kümeleri ile gerçekleştirilen deneyler O öznitelik kümesi ile gerçekleştirilen sınıflandırma deneylerini geçememiştir. Sonuç olarak, öznitelik kümelerinin sınıflandırma üzerindeki etkilerinin gözlemlendiği Doğrusal Destek Vektörleri sınıflandırıcı ile gerçekleştirilen bu deneyler sonucunda O öznitelik kümesinin diğer öznitelik kümelerinden daha başarılı olduğu gözlemlenmiştir.

Çizelge 3.6 Öznitelik kümelerinin Multinom Naif Bayes Sınıflandırıcısı ile sınıflar üzerindeki etkileri

Sınıflandırma Algoritması	Öznitelik Kümesi	Yazar Sınıfı için AUC Değerleri	Kategori Sınıfı için AUC Değerleri	Yıl Aralığı Sınıfı için AUC Değerleri
Multinom Naif Bayes	P	0.839	0.822	0.650
	A	0.862	0.865	0.665
	B	0.854	0.865	0.663
	C	0.861	0.865	0.664
	D	0.863	0.868	0.666
	E	0.853	0.865	0.663
	F	0.856	0.867	0.665
	G	0.862	0.868	0.667
	H	0.855	0.867	0.664
	I	0.856	0.868	0.664
	J	0.856	0.869	0.668
	K	0.856	0.869	0.664
	L	0.856	0.869	0.664
	M	0.855	0.868	0.664
	N	0.856	0.869	0.670
	O	0.860	0.873	0.674
	R	0.849	0.839	0.638
	S	0.862	0.872	0.670

Öznitelik vektörlerinin “Doğrusal Destek Vektör Makineleri” sınıflandırıcı üzerinde etkisi gözlemlendikten sonra; öznitelik vektörlerinin diğer iyi sınıflandırıcı olan “Multinom Naif Bayes” sınıflandırıcı üzerindeki etkileri için deneyler gerçekleştirilmiştir. Bu deneylerin sonuçları Çizelge 3.6’da görülmektedir. A, G ve S öznitelik kümeleri “Söz Yazarı” sınıf için en iyi seçim olduğu görülmektedir. “Multinom Naif Bayes” sınıflandırıcısı için G ve S öznitelik kümeleri yerine A öznitelik kümesinin disk alanı ve hız açısından kazanç sağlaması açısından sonraki deneylerde kullanılmasına karar verilmiştir. “Kategori” ve “Yıl Aralık” sınıfları için O öznitelik kümesinin, “Multinom Naif Bayes” sınıflandırıcısı için de aynı “Doğrusal Destek Vektör Makineleri”nde olduğu gibi diğer öznitelik kümelerinden daha başarılı olduğu görülmektedir.

Çizelge 3.5 ve Çizelge 3.6’da elde edilen sonuçlar neticesinde hangi sınıf ve sınıflandırıcı için hangi öznitelik kümelerinin daha başarılı olduğu belirtilmiştir. Bu adımda gerçekleştirilen deneyler ile giriş kısmında belirtilen “Metin içeriğinin temsilinde hangi öznitelikler faydalıdır?” sorusuna cevap verilmiştir.

3.3.3 Kelime kökü alınma durumuna göre sonuçlar

Öznitelik kümelerinin seçiminden sonra sınıflandırma algoritmaları üzerinde kelimelerin orijinal hali ile kelimelerin köklerinin alınmış halinin etkisi araştırılmıştır. Kelimelerin köklerinin alınması işleminin "Yazar", "Kategori" ve "Yıl Aralığı" sınıflandırılması üzerindeki etkisi Çizelge 3.7'de verilmiştir.

Çizelge 3.7 Kelime kökü alınma durumunun sınıflandırma üzerindeki etkisi

AUC	**Doğrusal Destek Vektör Makinesi**		**Multinom Naif Bayes**	
Öznitelik Kümesi: O	Kelimenin Kökü Alınmış Mı?		Kelimenin Kökü Alınmış Mı?	
	Evet	Hayır	Evet	Hayır
Yazar	0.739	0.718	0.860	0.841
Kategori	0.770	0.751	0.873	0.875
Yıl Aralığı	0.610	0.595	0.674	0.654

Şarkı sözü metnin de geçen kelimelerin başka çekimler ve ekler ile kullanılması, aslında aynı kelimenin birden fazla kelime gibi davranması hem öznitelik vektörünü büyütmekte hem de söz yazarlarının sık kullandıkları kelimeler hakkında yanlış bilgiler verebilmektedir. Bu sebeple kelimelerin kökleri Türkçe Doğal Dil İşleme Kütüphanesi olan Zemberek Kütüphanesi ile alınmış ve sınıflandırma üzerindeki etkisi incelenmiştir. Çizelge 3.7'de görüldüğü üzere iki sadece MNB sınıflandırıcı ile "Kategori" sınıfı üzerinde gerçekleştirilen deney haricinde, kelimenin köklerinin alınması sınıflandırma performansını artırmaktadır. MNB sınıflandırıcı ile "Kategori" sınıfı üzerinde gerçekleştirilen çalışma sonucunda elde edilen iki sonuç arasındaki fark göz ardı edilebileceği için ve genel olarak diğer sınıflar ve sınıflandırma algoritmaları üzerinde kelime kökünün alınması halinde performans açısından bariz bir artış olması sebebi ile geri kalan çalışmalarda öznitelik kümelerinde kelimelerin köklerinin kullanılmasına karar verilmiştir. Bu adımda gerçekleştirilen deneyler ile giriş kısmında belirtilen "Kelime kökünün kullanılması temsili ne kadar güçlendirir?" sorusuna cevap verilmiştir.

3.3.4 Öznitelik seçim yöntemlerine göre sonuçlar

Gerçekleştirilen deneylerde kullanılan öznitelik vektörleri içerisinde çok fazla öznitelik bulunabilmektedir; fakat bu özniteliklerin bir kısmı sınıflandırma için etkili bir rol oynamamakta ve hem disk alanı açısından hem de işlem hızı açısından negatif etki yaratmaktadırlar. Öznitelik vektörleri içerisinde kullanılan öznitelikler arasında kaliteli ve sınıflandırma yöntemlerini gerçekten olumlu yönde etkileyecek

olanlarını seçmek amacıyla bazı öznitelik seçme algoritmaları kullanılmıştır. Bu çalışma kapsamında kullanılan öznitelik seçim algoritmaları “Ki-kare” ve “ReliefF” algoritmalarıdır. Ayrıca bu öznitelik seçim algoritmalarının kullanılması eğer performansa olumlu yönde etki ediyorsa; deneylerde kullanılan Arff dosyalarının da boyutunda önemli bir azalmaya sebep olmaktadır ve böylece deneylerin işlem hızlı artmaktadır. Çizelge 3.8’de öznitelik seçim algoritmalarının sınıflandırmaya olan etkisi görülmektedir.

Çizelge 3.8 Öznitelik seçim algoritmalarının sınıflandırma üzerindeki etkisi

Sınıflandırma Algoritması	Sınıf	Yok (AUC Değerleri)	Ki-kare (AUC Değerleri)	ReliefF (AUC Değerleri)
Doğrusal Destek Vektör Makinesi	Yazar	0.739	0.679	0.718
	Kategori	0.770	0.789	0.720
	Yıl Aralığı	0.610	0.712	0.574
Multinom Naif Bayes	Yazar	0.862	0.784	0.854
	Kategori	0.873	0.925	0.796
	Yıl Aralığı	0.674	0.818	0.675

Çizelge 3.8’de görüldüğü gibi “Yazar” sınıfı için öznitelik seçim algoritmalarının bir etkisi olmamıştır. Aynı şekilde ReliefF algoritmasının “MNB” sınıflandırıcı üzerinde gerçekleştirilen “Yıl Aralığı” sınıflandırma deneyi dışında hiçbir sınıflandırma deneyi üzerinde olumlu etkisi yoktur. Ancak Ki-kare algoritması “Kategori” ve “Yıl Aralığı” sınıfları üzerinde bariz bir şekilde performansı arttırmıştır. Ki-kare algoritması ile öznitelik vektörleri üzerinde kaliteli özniteliklerin seçilmesi, “Kategori” ve “Yıl Aralık”larının sınıflandırılmasını olumlu etkilemiştir.

Ki-kare öznitelik seçim algoritması sonucunda; “Kategori” sınıfı için kullanılan 33458 adet öznitelik sayısı 2142’te düşmüş, “Yıl Aralığı” sınıfı için ise bu sayı 33458’ten 650 özniteliğe düşmüştür. En iyi deney setlerini hazırlanması aşamasında bundan sonra “Yazar” sınıfı için öznitelik seçim algoritmaları uygulanmazken, diğeri iki sınıf için Ki-kare algoritması kullanılmıştır. Bu deney ile elde edilen sonuçlar ile giriş kısmında belirtilen “Şarkı sözlerinin sınıflandırılmasında öznitelik seçim algoritmaları sınıflandırmayı ne kadar güçlendirir?” sorusuna da cevap verilmiştir.

Öznitelik seçim algoritmaları öznitelikleri sınıflandırma üzerindeki etiklerine göre sıralamaktadır. Ki-kare algoritması öznitelik vektörü üzerindeki her bir özniteliğe puan vererek özniteliklerin kalitelerini belirtmektedir. Bu çalışmada Ki-kare algoritması ile gerçekleştirilen öznitelik seçim işlemi sonucunda her bir sınıf için en kaliteli 12 adet öznitelik Çizelge 3.9'da verilmiştir

Çizelge 3.9 Her bir sınıf için en açıklayıcı öznitelikler

Yazar	Kategori	Yıl Aralığı
Maksimum Satır Uzunluğu ile Minimum Satır Uzunluğu Farkı	Harf Sayısı	Maksimum Satır Uzunluğu ile Minimum Satır Uzunluğu Farkı
Kelime Sayısı	Maksimum Satır Uzunluğu ile Minimum Satır Uzunluğu Farkı	Ortalama Satır Uzunluğu
Harf Sayısı	Farklı Kelime Sayısı	"Aşk" Kelimesinin Frekansı
Farklı Kelime Sayısı	"Bir" Kelimesinin Frekansı	"Sevgi" Kelimesinin Frekansı
"Barış" Kelimesinin Frekansı	"he" N-gramının Frekansı	"urb" N-gramının Frekansı
"Bir" Kelimesinin Frekansı	"da" N-gramının Frekansı	"Kalp" Kelimesinin Frekansı
Ortalama Satır Uzunluğu	Ortalama Satır Uzunluğu	"Ümit" Kelimesinin Frekansı
"Vücut" Kelimesinin Frekansı	"nda" N-gramının Frekansı	"ik" N-gramının Frekansı
"lar" N-gramının Frekansı	"Gönül" Kelimesinin Frekansı	"dim" N-gramının Frekansı
"Aşk" Kelimesinin Frekansı	"lar" N-gramının Frekansı	"Barış" Kelimesinin Frekansı
"her" N-gramının Frekansı	"dü" N-gramının Frekansı	"Yıl" Kelimesinin Frekansı
"da" N-gramının Frekansı	"miş" N-gramının Frekansı	"yad" N-gramının Frekansı

Çizelge 3.9'a göre en bilgi verici özniteliklerin satır ile ilgili istatistiksel öznitelikler olduğu görülmektedir. Bu tez kapsamına özgü olarak düşünülen en uzun satır uzunluğu ile en kısa satır uzunluğu farkının hesaplandığı "Maksimum Satır Uzunluğu ile Minimum Satır Uzunluğu Farkı" isimli öznitelik, genel olarak sınıflandırma işlemleri sırasında en belirleyici ve en önemli öznitelik olarak dikkat çekmektedir. Öznitelik seçim yöntemleri sonucunda elde edilen en etkili ve

sınıflandırmada belirleyici rol oynayan özniteliklerin içerisinde, bu çalışma kapsamında düşünülen satır uzunlukları ile ilgili özniteliklerin bulunması dikkat çekmektedir. “Harf Sayısı”, “Kelime Sayısı” ve “Farklı Kelime Sayısı” gibi global istatistiksel öznitelikler ise ikinci en çok katkıda bulunan öznitelik kümeleridir. Kelime köklerine bakacak olursak, “Aşk”, “Sevgi”, “Vücut”, “Kalp”, “Barış”, “Ümit”, “Bir” ve “Yıl” kelimeleri diğer kelime köklerine göre sınıflandırma üzerinde daha belirleyicidir. Bu seçilen kelime kökleri, şarkı sözü yazarlarının yazım tercihleri ve hangi ortak kelimeleri daha sık kullandıkları hakkında bilgi vermektedir. En çok bilgi verici N-Gramlar ise; “lar”, “ik”, “dim”, “dü”, “miş” ve “da” N-Gramlarıdır. Bu N-Gramlar, çoğul eki (“ler”), birinci çoğul şahıs için geçmiş zaman eki (“ik”), birinci tekil şahıs için geçmiş zaman eki (“dim”), üçüncü tekil şahıs için geçmiş zaman eki (“dü”), üçüncü tekil şahıs için öğrenilen geçmiş zaman eki (“miş”) ve bulunma hali (“da”) ekleridir. Seçilen bu karakter N-gramlar, bazı kelime eklerinin dağılımını ve böylece söz yazarlarının şarkı sözleri yazarken kullandıkları gramer tercihlerini belirlerler.

3.3.5 En başarılı deney setinin seçilmesi

Öznitelik seçimi yöntemlerinin uygulanması veya uygulanmaması aşamalarına karar verildikten sonra, en başarılı sonucu veren deney setine karar verilmiştir. Şarkı sözlerinin sınıflandırılması çalışmasında, en başarılı sonucu veren deneyler kümesi bahsedilen adımlar ile oluşturulmuş ve bu nihai deney seti üzerinden elde edilen sonuçlar bu tez kapsamında elde edilen en başarılı sonuçlar olarak kaydedilmiştir. İki sınıflandırma yöntemi (DDVM, MNB) ve üç sınıf (“Yazar”, “Kategori”, “Yıl Aralığı”) için en başarılı deney setleri Şekil 3.4’te verilmiştir.

Şekil 3.4 Her bir sınıf için en başarılı deney setleri

Gerçekleştirilen bütün deney kümeleri için kelimelerin köklerinin alınması işlemi gerçekleştirilmiştir. DDVM'leri için en iyi öznitelik kümesi "O" öznitelik kümesidir. NMB sınıflandırıcı için "Kategori" ve "Yıl Aralık" sınıfları için en iyi öznitelik kümesi aynı şekilde "O" öznitelik kümesiyken, "Yazar" sınıfı için ise bu "A" öznitelik kümesi olarak değişmektedir. "Yazar" sınıfına her iki sınıflandırma algoritması da uygulanırken, öznitelik seçimi algoritmalarının uygulanmaması daha doğru bir seçimken, "Kategori" ve "Yıl Aralığı" sınıfları için Ki-kare öznitelik seçim yöntemlerinin uygulanması daha doğrudur.

Şekil 3.4'da belirtilen en başarılı deney adımları gözlemlenip, hazırlandıktan sonra ilgili deneylerin sonuçları elde edilmiş ve yorumlanmıştır. Çizelge 3.10'da DDVM ve NMB iki sınıflandırma algoritması için elde edilen en iyi sonuçlar verilmiştir.

Çizelge 3.10 DDVM ve MNB sınıflandırıcıları için elde edilen en iyi sonuçlar

	Sınıf	Doğru Sınıflandırma	Yanlış Sınıflandırma	Duyarlılık	Kesinlik	F-Ölçütü	AUC
Doğrusal Destek Vektör Makineleri	Yazar	53,435	46,564	0,534	0,523	0,521	0,739
	Kategori	72,805	27,194	0,728	0,730	0,728	0,789
	Yıl Aralığı	67,652	32,347	0,677	0,670	0,670	0,712
Multinom Naif Bayes	Yazar	49,141	50,858	0,491	0,593	0,434	0,862
	Kategori	81,488	18,511	0,815	0,817	0,815	0,925
	Yıl Aralığı	71,374	28,626	0,697	0,698	0,695	0,818

Çizelge 3.10'da görüldüğü gibi, DVM veri kümesi üzerinde en iyi deney seti ile birlikte uygulandığı zaman verilerin ~%53'ü doğru sınıflandırılmaktadır. 12 adet söz yazarının bulunduğu bir kümede, söz yazarlarının ~%53 gibi bir oranla doğru sınıflandırılması, neredeyse iki söz yazarından bir tanesinin doğru sınıflandırıldığı anlamına gelmektedir. Veri kümesinde toplam üç adet kategori bulunmaktadır ve bu sınıflar için "Doğru Sınıflandırma" değeri ~%73'lerde görülmektedir. Üç adet yıl aralık sınıfının bulunduğu veri kümesinde ise bu değer ~%68 oranlarındadır. "Duyarlılık", "Kesinlik" ve "F-Ölçütü" değerleri için en başarılı sonuçlar

“Kategori”lerin sınıflandırılmasında elde edilmiştir. Çizelge 3.10’da tez kapsamında temel model başarım ölçütü olarak kullanılan “AUC” değeri; “Yazar” sınıfı için 0.739, “Kategori” sınıfı için 0.789 ve “Yıl Aralığı” sınıfı için 0.712 değerleri elde edilmiştir.

MNB sınıflandırıcı ile elde edilen sınıflandırma değerleri DDVM sınıflandırıcısı ile elde edilen değerlere göre daha başarılıdır. Veriler “Yazar” sınıfına göre sınıflandırılırken “Doğru Sınıflandırma” değeri ~%49’ken, bu değer “Kategori” sınıfı için ~%81, “Yıl Aralığı” sınıfı için ise ~%71 olarak hesaplanmıştır. “Duyarlık”, “Kesinlik” ve “F-Ölçütü” model başarım ölçütleri için hesaplanan değerler genel olarak DDVM sınıflandırıcısına göre daha başarılı çıkmıştır. Temel sınıflandırma değerlendirme kriteri olan “AUC” değerlerinde ise her üç sınıf için de MNB sınıflandırıcısı DDVM sınıflandırıcısına göre daha başarılıdır. “AUC” değeri “Yazar”lar sınıflandırılırken 0.862, “Kategoriler” sınıflandırılırken 0.925 ve “Yıl Aralık”ları sınıflandırılırken ise bu değer 0.818’e çıkmaktadır.

3.3.6 Sınıflara ait model başarım ölçütü ve karışıklık matrisi sonuçları

Her bir şarkı sözü yazarının MNB sınıflandırıcı algoritması ile sınıflandırılması sonucunda elde edilen “Model Başarım Ölçütleri” Çizelge 3.11’de verilmiştir.

Çizelge 3.11 Her bir söz yazarı için elde edilen model başarım ölçütü değerleri

Sınıf = Yazar	Duyarlılık	Kesinlik	F-Ölçütü	AUC
Sezen Aksu	0,629	0,426	0,508	0,851
Serdar Ortaç	0,821	0,415	0,551	0,908
Yaşar	0,020	0,333	0,038	0,722
Mustafa Sandal	0,017	1,000	0,034	0,892
Teoman	0,691	0,758	0,723	0,980
Haluk Levent	0,140	0,800	0,239	0,827
Barış Manço	0,487	0,725	0,583	0,797
Şebnem Ferah	0,462	0,652	0,541	0,909
Selami Şahin	0,217	0,690	0,331	0,830
Yıldız Tilbe	0,053	0,750	0,098	0,717
Ferdi Tayfur	0,893	0,438	0,587	0,891
Hakan Altun	0,242	0,667	0,355	0,893
Ortalama Ağırlık	**0,491**	**0,593**	**0,434**	**0,862**

Çizelge 3.11‘de, “Serdar Ortaç” ve “Şebnem Ferah” isimli şarkı sözü yazarlarının değerlerinin diğer şarkı sözü yazarlarına göre daha iyi sınıflandırıldığı görülmektedir. Bu sonuçlar “Serdar Ortaç”ın şarkılarının kendi içerisinde benzer sözcüksel ve metinsel yapıda olduğu; fakat diğer şarkılar ile farklılık gösterdiği şeklinde yorumlanabilir. Aynı şekilde, “Yıldız Tilbe”nin sınıflandırılma sonuçları düşük çıktığı için; “Yıldız Tilbe”nin yazdığı şarkı sözlerinin diğer şarkı sözleri ile benzerlik gösterdiği olarak yorumlanabilir.

Çizelge 3.12’de her bir şarkı sözü yazarının sınıflandırılması sonucu elde edilen “Karışıklık Matrisi” (Confusion Matrix) verilmiştir.

Çizelge 3.12 Şarkı sözü yazarlarıının sınıflandırılması sonucu elde edilen karışıklık matrisi

Tahmin / Gerçek Değer	SA	SO	Y	MS	T	HL	BM	ŞF	SŞ	YT	FT	HA
Sezen Aksu	78	20	0	0	2	0	5	4	2	0	12	1
Serdar Ortaç	1	110	0	0	1	0	0	1	0	0	19	2
Yaşar	17	10	1	0	1	0	1	0	0	0	18	1
Mustafa Sandal	10	25	0	1	1	0	1	3	0	0	15	2
Teoman	9	7	0	0	47	0	0	3	0	0	2	0
Haluk Levent	17	6	0	0	2	8	4	2	2	0	15	1
Barış Manço	13	9	1	0	1	1	37	2	1	0	11	0
Şebnem Ferah	16	10	0	0	5	0	0	30	1	0	3	0
Selami Şahin	6	10	0	0	0	0	1	1	20	0	50	4
Yıldız Tilbe	9	28	0	0	0	1	0	0	1	3	15	0
Ferdi Tayfur	4	9	1	0	0	0	2	0	2	1	158	0
Hakan Altun	3	21	0	0	2	0	0	0	0	0	43	22

Karışıklık matrisi ile belirtilen tabloda, sınıflandırma işlemi sonucunda her bir şarkı sözü yazarına ait şarkıların, kendisi ve diğer şarkı sözü yazarlarına ait olan dağılımı verilmektedir. Örneğin; “Serdar Ortaç” a ait 134 adet şarkının, 1 tanesi “Sezen Aksun”nun yazdığı, 1 tanesi “Teoman”ın yazdığı, 1 tanesi “Şebnem Ferah”ın yazdığı, 19 tanesi “Ferdi Tayfur”un yazdığı ve 2 tanesi “Hakan Altun” un

yazdığı şarkı olarak yanlış sınıflandırılmıştır. Fakat 134 adet “Serdar Ortaç” şarkısının 110 tanesi doğru olarak sınıflandırılmış ve “Serdar Ortaç” şarkısı olarak Çizelge 3.12’de belirtilmiştir.

Çizelge 3.11’e benzer olarak Çizelge 3.13’de de “Kategori” sınıfının MNB sınıflandırıcı algoritması ile sınıflandırılması sonucunda elde edilen “Model Başarım Ölçütleri” verilmiştir.

Çizelge 3.13 Her bir müzik kategorisi için elde edilen model başarım ölçütü değerleri

Sınıf = Kategori	Duyarlılık	Kesinlik	F- Ölçütü	AUC
Pop	0,767	0,782	0,775	0,901
Rock	0,786	0,882	0,831	0,950
Arabesk-Fantezi	0,875	0,806	0,839	0,929
Ortalama Ağırlık	**0,815**	**0,817**	**0,815**	**0,925**

Çizelge 3.13’de AUC değerlerine bakacak olursak, en iyi sınıflandırma değerleri “Rock” müzik kategorisine aittir. En başarısız sonuçlar ise “Pop” müzik şarkı sözleri kategorisinde elde edilmiştir. Bu sonuçları yorumlayacak olursak; “Rock” türüne ait şarkı sözleri birbiri içerisinde benzerlik gösterirken, diğer türlere ait şarkı sözlerinden farklılık göstermektedir.

Çizelge 3.14’de müzik kategorilerine ait şarkı sözlerini karışıklık matris sonuçları verilmiştir.

Çizelge 3.14 Müzik kategorilerinin sınıflandırılması sonucu elde edilen karışıklık matrisi

Tahmin / Gerçek Değer	Pop	Rock	Arabesk-Fantezi
Pop	280	23	62
Rock	31	209	26
Arabesk-Fantezi	47	5	365

Çizelge 3.14’de görüldüğü gibi, “Rock” müzik türüne ait 266 adet şarkı sözünün, 31 tanesi “Pop” müzik türüne, 26 tanesi de “Arabesk-Fantezi” müzik türüne ait şarkı sözlerine aitmiş gibi yanlış sınıflandırılmıştır. Toplam 1048 adet şarkı sözünden, 280 tanesi “Pop”, “209” tanesi “Rock” ve 365 tanesi “Arabesk-Fantezi” olmak üzere toplamda 854 adet şarkı sözü türü doğru şekilde sınıflandırılmıştır.

“Yıl Aralığı” sınıflarının MNB sınıflandırıcı algoritması ile “Model Başarım Ölçütleri”ne göre elde edilen sonuçları Çizelge 3.15’te verilmiştir.

Çizelge 3.15 Her bir yıl aralığı için elde edilen model başarım ölçütü değerleri

Sınıf = Yıl Aralığı	Duyarlılık	Kesinlik	F-Ölçütü	AUC
1972-1993	0,747	0,638	0,689	0,849
1994-2006	0,730	0,751	0,740	0,793
2007-2014	0,523	0,635	0,574	0,845
Ortalama Ağırlık	**0,697**	**0,698**	**0,695**	**0,818**

“Yıl Aralığı” sınıfının sonuçları Çizelge 3.15’te görüldüğü gibi “Kategori” sınıfının sınıflandırılma sonuçlarına göre daha başarısızdır. “Kategori” sınıfına ait “AUC Ortalama Ağırlık” değeri 0,925 iken bu değer “Yıl Aralığı” sınıfı için 0,818’dir. Çizelge 3.15’te görüldüğü gibi şarkı sözleri yıl aralıklarına göre sınıflandırılırken en kötü değerler “1994-2006” yılları arasında elde edilmiştir. Bu da “1994-2006” yılları arasında yazılan şarkıların kendi içlerinde belirleyici özellikler içermediği ve diğer yıl aralıklarında yazılan şarkı sözleri ile benzerlik gösterdiği şeklinde yorumlanabilir. Örneğin; “1972-1993” yılları arasında yazılan şarkılar kendi içlerinde benzerlik gösterirken, diğer yıl aralıklarında yazılan şarkılara göre farklılık göstermektedir.

NMB sınıflandırma algoritması ile “Yıl Aralığı” sınıfı üzerinde gerçekleştirilen deney sonucunda elde edilen karışıklık matrisi Çizelge 3.16’da verilmiştir.

Çizelge 3.16 Yıl aralıklarının sınıflandırılması ile elde edilen karışıklık matrisi

Tahmin / Gerçek Değer	1972-1993	1994-2006	2007-2014
1972-1993	219	65	9
1994-2006	103	410	49
2007-2014	21	71	101

Çizelge 3.16’da toplam 1048 adet şarkı sözünün yıl aralıklarına göre kendi içlerinde nasıl sınıflandırıldığının dağılımı verilmiştir. Örneğin; 1972 ile 1993 yılları arasında yazılan 293 adet şarkının 65 tanesi 1994-2006 arasındaki yıllara ait, 9 tanesi ise 2007-2014 yılları arasındaki şarkı sözlerine ait olarak yanlış sınıflandırılmıştır. Fakat 293 adet “1972-1993” sınıfına ait şarkı sözünden 219 tanesi doğru şekilde sınıflandırılmıştır. Toplamda ise 1048 adet şarkı sözünden 730 tane şarkı sözü doğru yıl aralıklarına ait olarak sınıflandırılmıştır.

4. TARTIŞMA VE GELECEK ÇALIŞMALAR

Türkçe Şarkı sözleri kullanılarak söz yazarının, kategorisinin ve yıl aralığının sınıflandırılmasına yönelik bir çalışma yapılmış ve bu amaçla geliştirilen sınıflandırma modeli, bilgi geri-geri getirimi bakış açısıyla değerlendirilmiştir.

Tez sonuçları göstermektedir ki; şarkı sözü metin verisinin otomatik analizi ve sınıflandırılması, dijital ortamlarda bulunan müzik içeriklerinin erişilebilirliği, organize edilebilirliği ve yönetilebilirliği açısından umut verici bir aşamadır.

Türkçe şarkı sözlerinin ve bu şarkı sözlerine ait üst bilgilerin bulunduğu geniş bir veri kümesi hazırlanmıştır. Hazırlanan bu veri kümesi üzerinde, Türkçe dil bilgisi yapısına uygun olarak yapısal ve istatistiksel öznitelik vektörleri hazırlanmıştır. Hazırlanan bu öznitelik vektörlerine eklenen her bir yeni öznitelik ve öznitelik grubu için şarkı sözlerinin sınıflandırılması gerçekleştirilmiştir. Bu sayede öznitelik kümesine eklenen her bir yeni özniteliğin, şarkı sözlerinin sınıflandırılmasını olumlu ya da olumsuz yönde etkileyebileceği gözlemlenmiştir. Bu çalışma kapsamında kullanılan öznitelikler arasından en iyi kombinasyonun seçilmesi ile ideal öznitelik vektörü oluşturulmuş ve oluşturulan bu öznitelik vektörünün performansı nasıl doğrudan olumlu etkilediği gözlemlenmiştir.

Şarkı sözleri metinlerinin sınıflandırılması sırasında geliştirilen öznitelik grupları içerisindeki “Satır uzunluğu istatistikleri” öznitelik grubu, gelecek çalışmalarda yol gösterici bir öznitelik grubu olarak ön plana çıkmaktadır. Her bir kategoriye ait şarkı sözü metinlerinde kullanılan satır uzunluğu yapısı belirleyici rol oynamaktadır. Tez kapsamında araştırılan ve gözlemlenen bu durum deneylerde kullanılmıştır ve deneyler sonucunda elde edilen verilere göre, bu teze özgü geliştirilen bu tip öznitelik grupları sınıflandırma sırasında belirleyici ve etkilidir. Özellikle öznitelik seçim aşamasında, en başarılı öznitelik olarak göze çarpan “Maksimum Satır Uzunluğu ile Minimum Satır Uzunluğu Farkı” özniteliği bu tez kapsamında bulunmuş ve uygulanmış “Satır uzunluğu istatistikleri” öznitelik grubuna ait bir özniteliktir. Bu kapsamda geliştirilen bazı öznitelik grupları gelecek çalışmalarda yol gösterici rol oynayabilirler. Bu durum, bu tezin literatüre kazandırabileceği en önemli özelliklerden bir tanesi olarak ön plana çıkmaktadır.

Türkçe şarkı sözü madenciliği ile ilgili gelecek çalışmalar içerisinde, bu çalışma kapsamında kullanılan veri kümesinin ve şarkı sözlerinin sınıflandırılması

aşamasında kullanılan öznitelik vektörleri genişletilebilir. Bu çalışmada kullanılan 12 adet şarkı sözü yazarının yazdığı toplam 1048 adet şarkının genişletilmesi ile gerçekleştirilen deneyler sonucunda elde edilen çıktılar yeni yöntemlerin ve çalışmaların ortaya çıkmasına neden olabilir. Ayrıca veri kümesinin genişletilmesi ile paralel olarak, bu çalışma kapsamında Türkçe şarkı sözü metinlerinden çıkartılan öznitelik kümeleri genişletilmesi planlanmaktadır. Özellikle Türkçe'ye ve şarkı sözü metinlerine uygun yeni özniteliklerin çalışmaya dahil edilmesiyle sınıflandırma başarısı arttırılabilecektir. Her bir müzik türü için kullanılan jargon birbirinden farklı olduğu için, müzik türleri arasındaki jargonu temel alarak çıkartılan yeni öznitelikler performansı doğrudan etkileyecektir. Ayrıca bu çalışma kapsamında şarkı sözü yazarlarının yaklaşımlarının yaklaşık olarak her 10 yıllık dönemlerde değişebildiği tespit edilmiştir. Dönemler içerisindeki bu değişiklik her bir şarkı sözü yazarına ve müzik türlerine ait tarzların değişmesine neden olmaktadır. Daha sonra yapılacak çalışmalarda, yıllara göre şarkı sözlerindeki değişen bu üslubun dikkate alınması ile hazırlanacak yeni öznitelik kümeleri performansı doğrudan olumlu yönde etkileyecektir.

Şarkı sözlerinin sınıflandırılması sırasında, sınıflandırma ve öznitelik seçim algoritmalarının performans üzerindeki etkisini inceleyen bir alt yapı sunulmuştur. Bu alt yapı ile sistemin en iyi şekilde çalışmasını sağlayan algoritmaların seçilmesini sağlamaktadır. Yapılan çalışmalarda kullanılan DVM, DDVM, NB ve MNB algoritmaları içerisinden bu sistem için en iyisi olan MNB sınıflandırma algoritması seçilmiştir. Aynı şekilde çok büyük boyutlu olan öznitelik vektörü içerisinden, en kaliteli ve sınıflandırmayı en çok etkileyecek özniteliklerin seçilmesini sağlayan öznitelik seçim algoritmalarının sistem performansını nasıl etkilediği gözlemlenmiş ve en başarılı öznitelik seçim algoritmasının en iyi deney setinde kullanılmasına karar verilmiştir. Şarkı sözü yazarlarının sınıflandırılmasında herhangi bir öznitelik seçim algoritmasının sistemi olumlu yönde etkilememesinden dolayı söz yazarlarını sınıflandırırken en iyi deney seti içerisinde öznitelik seçim algoritması kullanılmamıştır. Müzik türlerini ve yıl aralıklarını sınıflandırırken Ki-kare ve ReliefF algoritmaları kullanılmış ve bu iki sınıf için Ki-kare öznitelik seçim algoritması sistemin performansını artırmıştır. Bu adımlar gerçekleştirilmesi ve en iyi yöntemlerin sırası ile seçilerek sisteme dahil edilmesinin şarkı sözlerini sınıflandırılmasında doğrudan önemli bir rol oynadığı

gözlemlenmiştir. Elde edinilen sonuçlara göre başlangıçta belirlediğimiz araştırma soruları da kanıta dayalı olarak cevap bulmuştur. Buna göre; şarkı sözlerinin müzik içeriğinin temsilinde etkili bir rol oynayabileceği gözlemlenmiştir. Yapılan deneyler ile elde edilen sonuçlara göre; "Şarkı sözü yazarı", "kategori" ve "yıl aralığı" sınıflarının sınıflandırılmasında şarkı sözlerinin etkili olduğu gözlemlenmiştir. Sınıflandırma aşamasında kullanılacak hangi özniteliklerin faydalı olabileceği sorusu da bu çalışma kapsamında cevap bulmuştur. Şarkı metinlerinde kullanılan kelimelerin kökleri öznitelikleri, şarkı sözlerinin metin yapısı hakkında bilgi veren, kelime ek bilgisini içeren ve şarkı metni içerisindeki kafiye kullanımı hakkında bilgi veren n-gram öznitelikleri, şarkı metinleri hakkında genel bilgi veren global ve satır uzunluğu istatistiksel öznitelikleri şarkı sözlerinin sınıflandırmasında önemli bir oynamıştır. Özellikle bu çalışma kapsamında geliştirilen "Satır uzunluğu istatistikleri" öznitelik grubu en dikkat çekici öznitelik gruplarından bir tanesidir. Öznitelik seçim yöntemleri ile elde edilen sonuçlara göre, sınıflandırma performansını etkileyecek en önemli öznitelikler "Satır uzunluğu istatistikler" gurubuna aittir. Bu çalışma kapsamında kelimelerin köklerinin alınmasının sınıflandırma performansını nasıl etkilediği gözlemlenmiştir. Öznitelik olarak kelimeler yerine kelime köklerinin kullanılmasının sınıflandırma performansını arttırdığı deneyler ile ispatlanmıştır. Ayrıca kullanılan özniteliklerin hepsinin kullanılmasının performansı olumlu ya da olumsuz etkilediği de öznitelik seçim yöntemleri üzerinde gerçekleştirilen deneyler ile cevap bulmuştur. "Yazar" sınıfı için öznitelik seçim yöntemi uygulanması performansı olumsuz yönde etkilerken, "kategori" ve "yıl aralığı" sınıfları için ise Ki-kare öznitelik seçim yönteminin kullanılması sınıflandırma performansını arttırmıştır. Böylece bu çalışma kapsamında gerçekleştirilen deneyler ve elde edilen sonuçlar ile giriş kısmında belirtilen sorulara cevap verilmiştir.

Gelecek çalışmalarda, yeni sınıflandırma ve öznitelik seçim algoritmalarının denenmesi ve bu algoritmaların Türkçe şarkı sözü sınıflandırma üzerindeki etkilerinin incelenmesi planlanmaktadır. Yeni denenecek sınıflandırma algoritmalarının sistemin performansını artırması durumunda yanlış sınıflandırılan şarkı sözlerinin de doğru şekilde sınıflandırılması hedeflenmektedir. Büyük boyutlu öznitelik vektörleri üzerinden kalitesiz özniteliklerin elenmesini sağlayan yeni öznitelik seçim yöntemlerinin sisteme entegrasyonu, hem işlem hızının hem de

sınıflandırma performansının artmasını sağlayacaktır. Türkçe dil yapısına uygun ve istatistiksel öznitelikleri sınıflandırma da ve seçiminde kullanılacak yeni yöntemlerin entegrasyonu ile sistemin genel performansının artırılması planlanmaktadır.

Araştırmalarımız kapsamında gerçekleştirilen bu çalışma şarkı sözü metninden şarkı sözü yazarının tahmin edilebilirliği için yapılan ilk çalışmadır. Birçok ilgili ve farklı kategorideki öznitelik kümeleri şarkı sözü metinlerini ifade edebilmek için değerlendirilerek ve birleştirilerek tek bir öznitelik vektörü haline getirilmiştir. Ayrıca, şarkı sözü sınıflandırma işlemi ilk kez Türkçe bağlamında ele alınmaktadır. Sınıflandırma performansının doğru bir şekilde değerlendirilmesi için geniş bir veri kümesi hazırlanmış ve herkese açık hale getirilmiştir. Bilgi geri getirim araştırmacılarının, doğal dil işleme kütüphaneleri geliştiricilerinin ve müzik analiz topluluklarının hazırladığımız veri kümesi ve gerçekleştirdiğimiz çalışmadan yararlanacaklarına inanıyoruz. Önerilen modelin, hali hazırda geliştirilmiş müzik tavsiye sistemlerinde yardımcı araç olarak kullanılabileceği düşünülmektedir. Bu çalışma ile elde edilen kazanımlar, müzik bilgi geri getirim alanı ve tavsiye sistemleri uygulamalarında tamamlayıcı bir araç olarak kullanabilir.

KAYNAKLAR LİSTESİ

[1] Orio N., “Music retrieval: a tutorial and review,” Foundations and Trends in Information Retrieval, vol.1, no.1, s.1-90, 2006.

[2] Song Y., Dixon S. and Pearce M., “A survey of music recommendation systems and future perspectives,” 9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012), 2012.

[3] Schedl M., Gómez E. and Urbano J., "Music information retrieval: recent developments and applications," Foundations and Trends in Information Retrieval, vol.8, no.2-3, s.127-261, 2014.

[4] Moutselakis E.V. and Karakos A.S., “Semantic web multimedia metadata retrieval: a music approach,” 13th Panhellenic Conference on Informatics, PCI’09, s.43-47, 2009.

[5] Debaecker J. and Mustafa el Hadi W., “Music indexing and retrieval: evaluating the social production of music metadata and its use,” ISKO UK Conference 2011: Facets of Knowledge Organization, s.353-363, 2011.

[6] Costa C. H. L., Valle Jr. J. D. and Koerich, A.L., "Automatic classification of audio data," 2004 IEEE International Conference on Systems, Man and Cybernetics, vol.1, s.562-567, 2004.

[7] Qian Y. Z. Dou H. and Feng Y., “A novel algorithm for audio information retrieval based on audio fingerprint,” 2010 International Conference on Information Networking and Automation (ICINA), vol.1, s.266-270, 2010.

[8] Su L., Yeh C., Liu J. Y., Wang J. C. and Yang Y. H. "A systematic evaluation of the bag-of-frames representation for music information retrieval," IEEE Transactions on Multimedia, vol.16, no.5, s.1188-1200, 2014.

[9] Sarkar R. and Saha S. K., “Music genre classification using EMD and pitch based feature,” 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR), s.1-6, 2015.

[10] Byrd D. and Crawford T., “Problems of music information retrieval in the real world,” Information Processing and Management, vol. 38, no.2, s. 249-272, 2002.

[11] Besson M., Fata F., Peretz I., Bonnel A. M. and Requin J., “Singing in the brain: Independence of lyrics and tunes,” Psychological Science, vol.9, no., s.6494–6498, 1998.

[12] Frith S., “Music for pleasure: essays in the sociology of pop,” Routledge, New York, 1988.

[13] van Zaannen M. and Kanters P., “Automatic mood classification using TF*IDF based on lyrics,” Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010), s.75-80, 2010.

[14] Beukeboom C. J. and Semin G. R., “How mood turns on language,” Journal of Experimental Social Psychology, vol.42, no.5, s.553–566, 2005.

[15] Xiao H., Downie J., and Ehman A. F., “Lyric text mining in music mood classification,” 10th International Society for Music Information Retrieval Conference (ISMIR 2009), s.411–416, 2009.

[16] He H., Jin J., Xiong H., Chen B., Sun W. and Zhao L., “Language feature mining for music emotion classification via supervised learning from lyrics,” Advances in Computation and Intelligence, Lecture Notes in Computer Science, vol.5370, s.426–435, 2008.

[17] Xia Y., Wang L., Wong K.F. and Xu M., “Sentiment vector space model for lyric-based song sentiment classification,” Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, s.133-136, 2008.

[18] Mayer R., Neumayer R. and Rauber A., “Rhyme and style features for musical genre classification by song lyrics,” Proceedings of the 9th International Conference on Music Information Retrieval, s.337–342, 2008.

[19] Fell M., and Sporleder C., “Lyrics-based analysis and classification of music,” COLING, s.620-631, 2014.

[20] Laurier C., Grivolla J. and Herrera, P., “Multimodal music mood classification using audio and lyrics,” Proceedings of the 7th International Conference on Machine Learning and Applications, s.688–693, 2008.

[21] Mckay J., Burgoyne J. A., Hockman J., Smith J. B. L., Vigliensoni G. and Fujinaga I., “Evaluating the genre classification performance of lyrical features relative to audio, symbolic and cultural features,” International Society for Music Information Retrieval Conference, 2010.

[22] Mayer R. and Rauber A., “Musical genre classification by ensembles of audio and lyrics features,” Proceedings of International Conference on Music Information Retrieval, s.675–680, 2011.

[23] Silla Jr. C. N., Koerich A. L. and Kaestner C. A. A., “Improving automatic music genre classification with hybrid content-based feature vectors,” Proceedings of the 2010 ACM Symposium on Applied Computing, s.1702–1707, 2010.

[24] Hu X. and Downie J. S., “Improving mood classification in music digital libraries by combining lyrics and audio,” Proceedings of the 10th Annual Joint Conference on Digital Libraries, s.159–168, 2010.

[25] Qi X. and Davison B. D., “Web page classification: features and algorithms,” ACM Computing Surveys, vol.41, no.2, article 12, 2009.

[26] Idrisa I., Selamata A., Nguyenc N. T., Omatud S., Krejcare O., Kucae K. And Penhakerf M., “A combined negative selection algorithm–particle swarm optimization for an email spam detection system,” Engineering Applications of Artificial Intelligence, vol.39, s.33–44, 2015.

[27] Petza G., Karpowicza M., Fürschußa H., Auingera A., Stříteskýb A. And Holzingerc A., “Computational approaches for mining user’s opinions on the Web 2.0,” Information Processing and Management, vol.50, no.6, s.899–908, 2014.

[28] Stamatatos E., “A survey of modern authorship attribution methods,” Journal of the American Society for Information Science and Technology, vol.60, no.3, s.538-556, 2009.

[29] Koppel M., Schler J. and Argamon S., “Computational methods in authorship attribution,” Journal of the American Society for Information Science and Technology, vol.60, no.1 s.9-26, 2009.

[30] Sebastiani F., “Machine learning in automated text categorization,” ACM Computing Surveys, vol.34, no.1, s.1–47, 2002.

[31] Badawi D. and Altınçay H., “A novel framework for termset selection and weighting in binary text classification,” Engineering Applications of Artificial Intelligence, vol.35, s.38–53, 2014.

[32] Vapnik V. and Cortes C., “Support-vector networks,” Machine Learning, vol.20, no.3, s.273-297, 1995.

[33] Abu-Mostafa Y., “Lecture 14 – Support Vector Machines,” http://www.youtube.com/watch?v=eHsErlPJWUU, 2012.

[34] Abu-Mostafa Y., “Lecture 15 - KernelMethods,” http://www.youtube.com/watch?v=XUj5JbQihlU&feature=relmfu%20kernel%20methods, 2012.

[35] Huson D., “SVMs and kernel functions,” Algorithms in Bioinformatics II, SoSe107, ZBIT, s.265, 2007.

[36] Hsu C., Chang C. and Lin C., “A practical guide to support vector classification,” National Taiwan University, Department of Computer Science, 2003.

[37] Chang C. and Lin C., “LIBSVM: a library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, vol.2, no.27, s.1-27, 2011.

[38] Rudman J., “The state of authorship attribution studies: some problems and solutions,” Computers and the Humanities, vol.31, no.4, s.351-365, 1998.

[39] Mendenhall T. C., “The characteristic curves of composition,” Science, vol.9, s.237-246, 1887.

[40] Yule G. U., “On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship,” Biometrika, vol.30, no.3/4, s.363-390, 1938.

[41] Zemberek, “Open Source NLP Library for Turkic Languages,” https://code.google.com/p/zemberek/, 2014.

[42] Doğan S. ve Diri B., “Türkçe Dokümanlar İçin N-gram Tabanlı Yeni Bir Sınıflandırma (Ng-ind): Yazar, Tür ve Cinsiyet,”, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği, 2006.

[43] Kavzoğlu T., Şahin E. K. ve Çölkesen İ., “Heyelan duyarlılık analizinde ki-kare testine dayalı faktör seçimi,” V. Uzaktan Algılama ve Coğrafi Bilgi Sistemleri Sempozyumu (UZALCBS 2014), 2014.

[44] Liu H. And Setiono R., “Chi2: Feature selection and discretization of numeric attributes,” Proceedings of the IEEE 7th International Conference on Tools with Artificial Intelligence, s.388-391, 1995.

[45] Biricik G., “Sınıf bilgisini kullanan boyut indirgeme yöntemlerinin metin sınıflandırmadaki etkilerinin karşılaştırılması, IEEE, 2012.

[46] K. and Rendell L. A., “The feature selection problem: traditional methods and a new algorithm,” AAAI’92 Proceedings of the tenth national conference on Artificial intelligence, s.129-134, 1992.

[47] Altun H., Polat E., Polat G. ve Güneş T., “İnsan-bilgisayar etkileşimini geliştirmek için ses ve yüz görüntü işaretlerinden çok kipli biometrik özniteliklerin belirlenmesi ve etkin birleştirilmesi,” TÜBİTAK, Elektrik, Elektronik ve Enformatik Araştırma Grubu, 2007.

[48] Coşkun C., “Veri madenciliği algoritmaları karşılaştırılması,” Dicle Üniversitesi Fen Bilimleri Enstitüsü, Matematik Anabilim Dalı, 2010.

[49] Aydın F., “Kalp ritim bozukluğu olan hastaların tedavi süreçlerini desteklemek amaçlı makine öğrenmesine dayalı bir sistemin geliştirilmesi,” Trakya Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, 2011.

[50] Mason S. J. and Graham N. E., “Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation,” Quarterly Journal of the Royal Meteorological Society, vol.128, s.2145–66, 2002.

[51] Hayran M., “ROC analizi, sağlık araştırmaları için temel istatistikler,” Omega Araştırma Yayınları, s.403-416, 2011.