T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KELİME YÖRÜNGELERİ İLE METİN SINIFLANDIRMA

İSLAM MAYDA

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
YRD. DOÇ. DR. M. FATİH AMASYALI

İSTANBUL, 2016

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

KELİME YÖRÜNGELERİ İLE METİN SINIFLANDIRMA

İslam MAYDA tarafından hazırlanan tez çalışması 15.04.2016 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Yrd. Doç. Dr. M. Fatih AMASYALI

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Yrd. Doç. Dr. M. Fatih AMASYALI

Yıldız Teknik Üniversitesi _____

Doç. Dr. Banu DİRİ

Yıldız Teknik Üniversitesi _____

Yrd. Doç. Dr. Murat ORHUN

İstanbul Bilgi Üniversitesi _____

ÖNSÖZ

Metin sınıflandırma, üzerinde en çok çalışma yapılan Doğal Dil İşleme konularından biridir. Metin sınıflandırma kısaca, yeni bir metnin çeşitli sınıflandırma algoritmaları kullanılarak daha önceden tanımlanmış sınıflardan birine atanması şeklinde tanımlanabilir. Günümüzde internetin de yaygınlaşmasıyla üretilen veri miktarının çok hızlı olarak artması, bu verinin sınıflandırmasını daha da önemli hale getirmiştir. "Spam" olarak adlandırılan istenmeyen e-postaların elenmesi, arama motorlarının daha önce girilen siteleri analiz ederek kullanıcının ilgili olduğu konuları arama sonuçlarında daha üst sıralarda göstermesi, bir haber sitesinin kullanıcının okuduğu haberlerin kategorilerini dikkate alarak benzer haberleri tavsiye etmesi, Başbakanlık İletişim Merkezi gibi ortak bir havuzda toplanan talep metinlerinin otomatik olarak ilgili birimlere yönlendirilmesi, metnin dilinin belirlenmesi, yazarının tanıması, yazarının yaşı, cinsiyeti veya başka bir özelliğinin tahmin edilmesi metin sınıflandırmanın uygulama alanlarından bazılarıdır. Metnin sınıflandırılabilmesi için öncelikle sayısal olarak ifade edilmesi, yani sayısallaştırılması gerekmektedir. Metin sayısallaştırmada da birçok farklı yöntem kullanılmaktadır. Bu tezde çeşitli metin temsil yöntemleri ele alınmış ve iki farklı veri kümesi üzerinde deneyler yapılmıştır. İlk veri kümesiyle yapılan çalışmada metnin yazarının psikolojik rahatsızlığı olup olmadığı tespit edilmeye çalışmış, ancak çok başarılı sonuçlar elde edilememiştir. İkinci veri kümesiyle yapılan çalışmada ise bir Twitter hesabının yazarı tespit edilmeye çalışmış ve yüksek başarı oranları yakalanmıştır.

Tüm tez çalışmalarım boyunca ilgisi ve desteğiyle yardımlarını hiç esirgemeyen danışmanım Sayın Yrd. Doç. Dr. M. Fatih AMASYALI Hocama en içten teşekkürlerimi sunarım.

Mart, 2016

İslam MAYDA

İÇİNDEKİLER

Sayfa

KISALTMA LİSTESİ vi

ŞEKİL LİSTESİ vii

ÇİZELGE LİSTESİ ix

ÖZET xi

ABSTRACT xii

BÖLÜM 1

GİRİŞ 1
- 1.1 Literatür Özeti 1
- 1.2 Tezin Amacı 4
- 1.3 Hipotez 4

BÖLÜM 2

METİN TEMSİL YÖNTEMLERİ 5
- 2.1 Kelimelerin Kendileri 5
- 2.2 Kelimelerin Kökleri 5
- 2.3 Karakter ve Kelime Ngramları 6
- 2.4 Frekans Hesaplamada Kullanılan Yöntemler 6
- 2.5 Anlamsal Uzay 7
- 2.6 Metin Yörüngeleri 7

BÖLÜM 3

VERİ SETLERİ 9
- 3.1 Psikolojik Hastaların Yazılarından Oluşan Veri Seti 9
- 3.2 Haber Metinleri 10
- 3.3 Köşe Yazarlarının Makaleleri ve Tweetlerinden Oluşan Veri Seti 10

BÖLÜM 4

DENEYSEL SONUÇLAR 13

4.1 Metinden Psikolojik Rahatsızlık Teşhisi 13
4.1.1 Kelimelerin Kökleri ile Üretilen Binary Matris 14
4.1.2 Kelimelerin Köklerinin Koordinatları ile Üretilen Matris 14
4.1.3 Metin Yörüngelerinin Görselleştirilmesi 15
4.1.4 Metin Yörüngelerinin Uzaklık ve Açı Değerlerinin Değişken Aralıklı Histogramları 17
4.1.5 Metin Yörüngelerinin Uzaklık ve Açı Değerlerinin Sabit Aralıklı Histogramları 20
4.1.6 Metin Yörüngelerinin Farklı Uzaklık ve Açı Değerlerinin Sabit Aralıklı Histogramları 22
4.1.7 Kelimelerin Kökleri ile Üretilen Binary Matrisin Zenginleştirilmesi 26
4.1.8 Sonuçların Yorumlanması 27
4.2 Yazar Tanımada Köşe Yazısı ve Tweet'lerin Çapraz Kullanımı 27
4.2.1 Her Tweet Tekil Değerlendirilerek Yapılan Yazar Tespiti 28
4.2.2 Birleştirilmiş Tweet Metinleri ile Yapılan Yazar Tespiti 30
4.2.3 Özellik Seçimi Yapılarak Testlerin Tekrarlanması 32
4.2.4 Sonuçların Yorumlanması 38

BÖLÜM 5

TARTIŞMA VE ÖNERİLER 43
5.1 Metinden Psikolojik Rahatsızlık Teşhisi 43
5.2 Yazar Tanımada Köşe Yazısı ve Tweet'lerin Çapraz Kullanımı 43

KAYNAKLAR 34

ÖZGEÇMİŞ 44

KISALTMA LİSTESİ

200D	İkiyüz boyutlu
2D	İki boyutlu
3D	Üç boyutlu
J48	Karar ağacı
NB	Naive Bayes
NBM	Multinomial Naive Bayes
NTF	Normalized Term Frequency
RF	Rastgele Orman
SMO	Karar Destek Makineleri
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency

ŞEKİL LİSTESİ

Sayfa

Şekil 2. 1 Örnek bir metin yörüngesi....8
Şekil 4.1 Haber verileriyle üretilen koordinatların sırasıyla 2 ve 3 boyutu kullanılarak çizilen birleştirilmiş metin yörüngeleri....15
Şekil 4.2 Hasta verileriyle üretilen koordinatların sırasıyla 2 ve 3 boyutu kullanılarak çizilen birleştirilmiş metin yörüngeleri....16
Şekil 4.3 Hasta verileriyle üretilen koordinatların sırasıyla 2 ve 3 boyutu kullanılarak çizilen ilk 3'er deneğin birleştirilmiş metin yörüngeleri....16
Şekil 4.4 Haber verileriyle üretilen koordinatların 2 ve 3 boyutu kullanılarak çizilen ilk 3'er deneğin birleştirilmiş metin yörüngeleri....17
Şekil 4.5 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 200 boyutta elde edilen histogramlar....21
Şekil 4.6 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 2 boyutta elde edilen histogramlar....21
Şekil 4.7 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 2 boyutta elde edilen histogramlar....21
Şekil 4.8 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 2 boyutta elde edilen histogramlar....22
Şekil 4.9 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 200 boyutta elde edilen histogramlar....23
Şekil 4.10 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 2 boyutta elde edilen histogramlar....24
Şekil 4.11 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 2 boyutta elde edilen histogramlar....25
Şekil 4.12 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 3 boyutta elde edilen histogramlar....25
Şekil 4.13 Makaleler eğitim, tekil tweetler test verisi olarak kullanıldığında elde edilen başarı oranları....39
Şekil 4.14 Tekil tweetler eğitim, Makaleler test verisi olarak kullanıldığında elde edilen başarı oranları....39
Şekil 4.15 Sadece makaleler ile 10 kez çapraz doğrulama yapıldığında elde edilen başarı oranları....39
Şekil 4.16 Sadece tekil tweetler ile 10 kez çapraz doğrulama yapıldığında elde edilen başarı oranları....40

Şekil 4.17 Makaleler eğitim, birleştirilmiş tweetler test verisi olarak kullanıldığında elde edilen başarı oranları.....41
Şekil 4.18 Birleştirilmiş tweetler eğitim, Makaleler test verisi olarak kullanıldığında elde edilen başarı oranları.....41
Şekil 4.19 Sadece birleştirilmiş tweetler ile 10 kez çapraz doğrulama yapıldığında elde edilen başarı oranları.....41

ÇİZELGE LİSTESİ

Sayfa

Çizelge 3. 1 Tweetleri ve makaleleri toplanan köşe yazarları.....10
Çizelge 3. 2 Tweetleri ve makaleleri toplanan köşe yazarları.....12
Çizelge 4. 1 Kelimelerin kökleri ile üretilen binary matris başarı oranları.....14
Çizelge 4. 2 Kelime köklerinin 200 boyutlu uzayda koordinatlarının başarı oranları.....15
Çizelge 4. 3 İki, üç ve ikiyüz boyutlu düzlemde uzaklıkların histogramlarıyla üretilen arff sonuçları.....18
Çizelge 4. 4 İki, üç ve ikiyüz boyutlu düzlemde açıların histogramlarıyla üretilen arff sonuçları.....18
Çizelge 4. 5 İki boyutlu düzlemde uzaklıkların histogramları ve açıların histogramları birlikte kullanılarak üretilen arff sonuçları.....18
Çizelge 4. 6 İki boyutlu düzlemde uzaklıkların histogramlarıyla (sadece sayılar ve sadece orta noktalar kullanılarak) üretilen arff sonuçları.....19
Çizelge 4. 7 İki boyutlu düzlemde uzaklıkların histogramları ve açıların histogramları birlikte ile (sadece sayılar ve sadece orta noktalar kullanılarak) üretilen arff sonuçları.....19
Çizelge 4. 8 Yörüngelerin uzaklık histogramlarıyla üretilen arff sonuçları 20
Çizelge 4. 9 Yörüngelerin açı histogramlarıyla üretilen arff sonuçları.....22
Çizelge 4. 10 Yörüngelerin uzaklık histogramlarıyla üretilen arff sonuçları 23
Çizelge 4. 11 Yörüngelerin açı histogramlarıyla üretilen arff sonuçları.....24
Çizelge 4. 12 Yörüngelerin açı histogramlarıyla üretilen arff sonuçları.....25
Çizelge 4. 13 Hasta verileri kullanılarak en yakın 3 ve 5 kelime ile zenginleştirilmiş binary matris sonuçları.....26
Çizelge 4. 14 Haber verileri kullanılarak yakınlık oranı 0.5 ve 0.7 üzeri olan kelimeler ile zenginleştirilmiş binary matris sonuçları.....26
Çizelge 4. 15 Makaleler eğitim, tekil tweetler test verisi olarak kullanıldığında sonuçlar.....28
Çizelge 4. 16 Tekil tweetler eğitim, makaleler test verisi olarak kullanıldığında sonuçlar.....29
Çizelge 4. 17 Sadece makaleler ile 10 kez çapraz doğrulama yapıldığında sonuçlar....29
Çizelge 4. 18 Sadece tekil tweetler ile 10 kez çapraz doğrulama yapıldığında sonuçlar.....30
Çizelge 4. 19 Makaleler eğitim, birleştirilmiş tweetler test verisi olarak kullanıldığında sonuçlar.....30

Çizelge 4. 20 Birleştirilmiş tweetler eğitim, makaleler test verisi olarak kullanıldığında sonuçlar.....31
Çizelge 4. 21 Sadece birleştirilmiş tweetler ile 10 kez çapraz doğrulama yapıldığında sonuçlar.....32
Çizelge 4. 22 Makaleler eğitim, tekil tweetler test verisi olarak kullanıldığında sonuçlar.....33
Çizelge 4. 23 Tekil tweetler eğitim, makaleler test verisi olarak kullanıldığında sonuçlar.....33
Çizelge 4. 24 Sadece makaleler ile 10 kez çapraz doğrulama yapıldığında sonuçlar....34
Çizelge 4. 25 Sadece tekil tweetler ile 10 kez çapraz doğrulama yapıldığında sonuçlar.....35
Çizelge 4. 26 Makaleler eğitim, birleştirilmiş tweetler test verisi olarak kullanıldığında sonuçlar.....35
Çizelge 4. 27 Birleştirilmiş tweetler eğitim, makaleler test verisi olarak kullanıldığında sonuçlar.....36
Çizelge 4. 28 Sadece makaleler ile 10 kez çapraz doğrulama yapıldığında sonuçlar....37
Çizelge 4. 29 Sadece birleştirilmiş tweetler ile 10 kez çapraz doğrulama sonuçları.....37

ÖZET

KELİME YÖRÜNGELERİ İLE METİN SINIFLANDIRMA

İslam MAYDA

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Yrd. Doç. Dr. M. Fatih AMASYALI

Bir kişi tarafından yazılan metinler analiz edilerek, metinlerin yazarına dair hangi çıkarımların yapılabileceği sorusunun popülerliği günümüzde artarak devam etmektedir. İnternetin yaygınlaşmasıyla, elektronik ortamda üretilen ve büyük bölümü metinlerden oluşan verilerin miktarı da çok hızlı bir şekilde artmaktadır. Şimdiye kadar metinden yola çıkarak, yazarın kim olduğunu, cinsiyetini, yaşını, karakterini, hatta politik eğilimini gibi birçok özelliğini tahmin etmeye yönelik çok sayıda çalışma yapılmış ve yapılmaya devam etmektedir. Bu tezde ise iki farklı veri kümesi kullanılarak iki farklı hipotez üzerinde çalışma yapılmıştır. Birinci çalışmadaki amaç metinden yola çıkarak yazarın anksiyete veya depresyon gibi bir psikolojik hastalığının olup olmadığının tahminin yapılmasıdır, ancak bu çalışmada istenen başarı oranları elde edilememiştir. İkinci çalışmadaki amaç ise, en popüler mikroblog sitesi olan Twitter'da kimliğini açıklamadan paylaşımlar yapan kullanıcıların kim olduğunun, aday yazarların makaleleri kullanılarak tahmin edilmesidir ve bu çalışmada yüksek başarı oranları elde edilmiştir. Yapılan çalışma sonucunda, bir Twitter hesabının kime ait olduğunu tespit etmeye çalışırken, bu hesaptan paylaşılan tweetlerin her birini birer metin olarak değerlendirmektense, belirli sayıda tweeti birleştirerek oluşturulan daha uzun tweet metinlerini kullanmanın çok daha başarılı sonuçlar verildiği görülmüştür. Metinleri sayısallaştırmada kelimelerin kendilerini, köklerini ve 3-gramları kullanılmıştır. Çeşitli sınıflandırıcılar arasından en başarılı sonuçları destek vektör makineleri ile elde edilmiştir.

Anahtar Kelimeler: Twitter, doğal dil işleme, yazar tanıma, metin sınıflandırma, metin yörüngeleri

ABSTRACT

TEXT CLASSIFICATION WITH WORD TRAJECTORIES

İslam MAYDA

Department of Computer Engineering

MSc. Thesis

Adviser: Assist. Prof. Dr. M. Fatih AMASYALI

The popularity of the question, "What can be deduced about the author by analyzing the text written by his/her?" is increasingly continuing in our day. With the proliferation of the Internet, the amount of the data which is generated online and mostly consisted of texts is rapidly increasing. Up to now, numerous studies have been done based on text such as author identification, guessing the gender, age, personality, and even political tendency of the author. In this thesis, two studies were done on two hypothesises by using two different data sets. The aim in the first study was guessing the psychological disorders of the author such as anxiety and depression based on his/her text, however the desired accuracy rates could not been obtained in this study. The aim in the second study was identification of the authors who makes sharing without revealing his/her identity on Twitter which is the most popular microblog web site, by using the articles of the candidate authors, and high accuracy rates were achieved in this study. As a result of the study conducted we perceived that, while trying to identify the author of a Twitter account, using the longer tweet texts which are obtained by concatenating the specific number of tweets gives more accurate results than using each tweet individually. The words themselves, their stems and 3-grams were used for digitizing of the texts. The most successful results were achieved with support vector machines from among several classifiers.

Keywords: Twitter, natural language processing, author identification, text classification, word trajectories

BÖLÜM 1

GİRİŞ

1.1 Literatür Özeti

Metin sınıflandırma adı altında metnin yazarına dair bazı çıkarımlar yapmayı amaçlayan çok sayıda çalışma yapılmıştır. Yazarın kim olduğu, yaşı, cinsiyeti ve diğer özelliklerini tespit etmeye yönelik araştırmalar hala popülerliğini sürdürmektedir. Web 2.0 teknolojisiyle geliştirilen kullanıcı merkezli siteler sayesinde kullanıcılar internette daha fazla paylaşım yaparak, zengin bilgi kaynakları oluşturmaya başlamışlardır [1]. Kullanıcılar tarafından üretilen bu veriler, metin sınıflandırma çalışmalarında da en çok başvurulan kaynaklardan biri haline gelmiştir.

Najib vd. [2] İngilizce, İspanyolca, İtalyanca ve Flemenkçe dillerinde yazılmış olan tweetlerden oluşan veri kümesi üzerinde yaptıkları çalışmada, yazarın yaşını, cinsiyetini ve kişilik özelliklerini tahmin etmeye çalışmışlardır. Yaş grupları olarak; 18-24, 25-34, 35-49 ve 50-XX şeklinde bir kategorizasyon yapılmıştır. Kişilik özellikleri bakımından ise psikolojide 5 büyük kişilik boyutu olarak tanımlanan açıklık, sorumluluk, dışadönüklük, uyumluluk ve duygusal denge özellikleri ele alınmıştır. Veri kümeleri üzerinde dört farklı sınıflandırıcı algoritması çalıştırılmıştır: Karar ağacı (J48), Rastgele Orman (RF), Karar Destek Makineleri (SMO) ve Naive Bayes (NB). Bu sınıflandırıcıların performansları değerlendirilerek sadece SMO sınıflandırıcının sonuçları sunulmuştur. Dil bazında yapılan testlerin sonuçlarına göre en yüksek başarı oranları; cinsiyet tahmininde %84 ile İspanyolca metinlerde, yaş tahmininde %66 ile İngilizce metinlerde elde edilmiştir.

Rao vd. [3] yaptıkları çalışmada Twitter kullanıcılarının verileri üzerinde SVM tabanlı sınıflandırma algoritmaları kullanarak, kullanıcıların cinsiyetini, yaşını, bölgesel kökenini ve siyasi eğilimini tespit etmeye çalışmışlardır. Yaş özelliği 30 yaş üstü ve 30 yaş altı olarak, bölgesel köken Kuzey Hindistan ve Güney Hindistan olarak, siyasi eğilim ise demokrat ve cumhuriyetçi olarak iki sınıflı olarak değerlendirilmişlerdir. Cinsiyet tahmininde %72.33, yaş tahmininde %74.11, bölgesel köken tahmininde %77.08 ve siyasi eğilim tahmininde %82.84 oranında başarı elde edilmiştir.

Román ve Cristóbal [4] İngilizce ve İspanyolca dillerinde yazılmış olan blog yazıları, tweetler, otel yorumları ve sosyal medya metinleri üzerinde cinsiyet ve yaş tahmini yapmaya çalışmışlardır. Yaş özelliğini 18-24, 25-34, 35-49 ve 50-64 olarak 5 sınıfta kategorize etmişlerdir. Multinomial Naive Bayes (NBM) sınıflandırıcıyı kullandıkları çalışmalarında ortalama olarak cinsiyet tahmininde İngilizce metinler üzerinde %59.52 oranında, yaş tahmininde ise İspanyolca metinler üzerinde %44.22 oranında daha başarılı sonuç almışlardır.

Yazarın yaşı ve cinsiyetini tespit etme üzerine çalışmaların yanı sıra yazarın kim olduğunu tahmin etmeye yönelik de çok sayıda çalışma yapılmıştır. Örneğin, Amasyalı vd. [5] yaptıkları araştırmada 18 farklı yazara ait 35'er gazete yazısı üzerinde SVM sınıflandırıcı ile %92.5 oranında yüksek bir başarı yakalamıştır. Üzerinde çalışılan metinlerin uzun olması başarı oranının yüksek çıkmasında önemli bir faktördür.

Twitter'da ise kullanıcılar tweet adı verilen en fazla 140 karakterden oluşan paylaşımlar yapabilmektedir. Paylaşılan metinlerin kısa olması bu metinlerin yazarlarını tanımayı zorlaştırmaktadır. Schwartz vd. [6] karakter n-gramları ve kelime n-gramları kullanarak SVM sınıflandırıcı ile 1000 yazarın 200'er tweeti üzerinde %30.3, 50 yazarın 50'şer tweeti üzerinde %50.7, 50 yazarın 1000'er tweeti üzerinde %71.2 başarı oranı elde etmişlerdir. Schwartz ve arkadaşları yaptıkları bu çalışmada her bir tweeti bir metin olarak ele alıp, 10 kat çapraz doğrulama ile test yapmışlardır.

Green ve Sheppard [7] finans sektöründeki 12 Twitter kullanıcısı üzerinde yaptıkları çalışmada, her yazar için 120'şer tweeti üzerinde SMO (Sequential Minimal Optimization) algoritmasını kullanarak 5x2 çapraz doğrulama ile test yapmış ve %40.5 doğruluk oranı yakalamışlardır. Green ve Sheppard bu araştırmalarında her bir tweetin

içerdiği karakterler, uzun kelimeler, boşluklar, noktalama işaretleri, köprüler, toplam karakter sayıları, toplam kelime sayıları ve bunların frekansları gibi toplam 86 özellikten oluşan özellik kümelerini çıkartmış ve sınıflandırmada bunları kullanılmışlardır.

Metinden yazarın kimliğini, yaşını, cinsiyetini tahmin edebildiğimiz gibi, herhangi bir psikolojik rahatsızlığı olup olmadığına dair de çıkarım yapabiliriz. Örneğin, Karmena vd. [8] çoğunlukla depresyon rahatsızlığı olan hastaların üye olduğu ve hastalıkları hakkında yazıştıkları bir forum sitesinden topladıkları metinler üzerinde çalışma yapmış ve üyelerin depresyon derecesini 4 farklı kategoride (çok depresif, kısmen depresif, az depresif ve depresif değil) tahmin etmeye çalışmışlardır. Veri alınan forum sitesinde 1988 yılından itibaren, 20 binin üzerinde kullanıcı tarafından 1 milyondan fazla mesaj paylaşılmıştır. Yapılan çalışmada öncelikle bu forumdan 2,110 gönderi toplanmış, bu gönderilerden 20 ila 200 kelime içeren 1,304 tanesiyle devam edilmiştir. Yazarların uyguladıkları algoritmaya göre bu mesajlar sıralanmış ve en sonunda içlerinden 122 tanesi seçilmiştir. Karmena vd. bu çalışmalarında %84 gibi bir başarı oranı elde etmişlerdir.

Orhan ve Mercan [9] ise anksiyete, depresyon hastaları ile sağlıklı kişilerden oluşan toplam 267 deneğin yazmış olduğu metinler üzerinde yaptıkları çalışmada yazarın anksiyete veya depresyon rahatsızlığı olup olmadığını tespit etmeye çalışmışlardır. Kelimelerin köklerinin metinde geçip geçmeme duruma göre ve Harvard Psikoloji Sözlüğünde 83 kategoriye ayrılan kelimelerden herhangi birinin metinde geçip geçmeme durumuna göre iki farklı binary matris üretilmiş ve NB, SMO ve J48 sınıflandırıcıları ile testler yapılmıştır. 49 depresyon hastasının ve 39 sağlıklı deneğin yazısını eğitim verisi olarak, 10'u depresyon hastası ve 10'u sağlıklı olmak üzere 20 deneğin yazısını test verisi olarak kullandıklarında en yüksek başarı kelime köklerinin özellik olarak kullanıldığı dosyanın NB sınıflandırıcıyla %80 olarak elde etmiştirler. Bu alt veri kümesinde yapılan rastgele bir seçimde despresyon hastasının yazısının seçilme ihtimali %55.6'dır. Anksiyete testinde ise, 54 anksiyete hastasının ve 39 sağlıklı deneğin yazısı eğitim verisi olarak, yine iki sınıftan da 10'ar olmak üzere 20 deneğin yazısını test verisi olarak kullandıklarında ise en yüksek başarı %70 ile kategorilerin özellik olarak kullanıldığı dosyanın J48 ile sınıflandırılmasında elde edilmiştir. Burada yapılacak olan rastgele bir seçimde anksiyete hastasının yazısının seçilme ihtimali ise

%58'dir. Sayılara bakıldığında, alt veri kümelerinin örnek sayılarının az olduğu görülmektedir. Ayrıca, testler seçilen alt veri kümeleri üzerinde sadece birer kez uygulanmış, herhangi bir çapraz doğrulama yapılmamıştır.

1.2 Tezin Amacı

Bu tezin amacı, bir metni çeşitli metin temsil yöntemleriyle sayısal olarak ifade ettikten sonra makine öğrenmesi algoritmalarını kullanarak metnin yazarına dair tespitlerde bulunmaktır. Bu genel amaç altında iki farklı çalışmaya özel amaçlar da mevcuttur.

İlk yapılan çalışmada, bir metni analiz ederek yazarın herhangi bir psikolojik rahatsızlığının olup olmadığının tespit edebilmektir. Farklı bir ifadeyle, bir psikolojik rahatsızlığın, özellikle anksiyete veya depresyonun, hastanın yazı yazma davranışına etki edip etmediğini görmek, eğer rahatsızlık yazıya etki ediyorsa bunun doğal dil işleme yöntemleri kullanılarak fark edilip edilemeyeceğini ölçmektir.

İkinci yapılan çalışmada ise, yazar tanımada bir kişinin tweetleri ile diğer yazdığı metinlerin çapraz olarak kullanılması amaçlanmaktadır. Bu şekilde, kim olduğu bilinmeyen bir Twitter hesabının kullanıcısının, aday yazarların diğer mecralarda yazdığı metinlerin kullanılarak tahmin edilebildiğini görmektir.

1.3 Hipotez

Bu tezde iki farklı hipotez öne sürülmüştür. Birinci hipotez şudur; bir metin analiz edilerek yazarının depresyon veya anksiyete rahatsızlığı olup olmadığı tespit edilebilir. İkinci hipotez ise; bir Twitter hesabının kullanıcısı, aday yazarların başka yazıları kullanılarak tespit edilebilir.

BÖLÜM 2

METİN TEMSİL YÖNTEMLERİ

Bu bölümde tez çalışmasında kullanılan çeşitli metin temsil yöntemleri alt başlıklar halinde anlatılmaktadır.

2.1 Kelimelerin Kendileri

Tüm veri kümesindeki farklı kelimeleri eklerine ve köküne ayırmadan tespit ettikten sonra, bu kelimeleri özellik seti olarak kullanıp, bir metinde kelimenin geçip geçmeme durumuna göre metni sayısal olarak ifade edebiliriz. Ancak, büyük veri kümelerinde bu yöntemi kullanmak hem başarı performansı bakımından hem de programın çalışma süresi bakımından çok fazla tercih edilmeyebilir. Büyük veri kümelerinde kelimelerin kendileri kullanıldığında özellikle Türkçe metinler için yüz binlerce kelimeden oluşan bir özellik seti ortaya çıkması muhtemeldir. Türkçe sondan eklemeli bir dil olduğu için bir isim veya fiil kelimesi durum ekleri, iyelik ekleri, çokluk eki, tamlama eki, kip ekleri gibi çekim ekleri alarak onlarca farklı şekilde yazılabilmektedir.

2.2 Kelimelerin Kökleri

Kelimelerin kökleri, metinlerin içerdikleri kelimelerle ifade edildikleri, literatürdeki en yaygın kullanılan özellik grubu olarak nitelendirilmektedir [10]. Kelimelerin kendisini kullanmak yerine kökleri kullanıldığında özellik setinin miktarı ciddi oranda azalmaktadır. Kelimelerin köklerini bulmada Zemberek [11] kütüphanesi kullanılmıştır. Zemberek, açık kaynak kodlu Türkçe doğal dil işleme kütüphanesidir. Tamamen Java ile geliştirilen kütüphane hecelemenin yanı sıra, yazım denetimi ve hatalı kelimeler için

öneri gibi farklı işlevlere de sahiptir [12]. Tez çalışmasında, metinde geçen kelimenin kökü Zemberek tarafından bulunamazsa, bu kelime de özellik grubuna dahil edilmiştir.

2.3 Karakter ve Kelime Ngramları

Karakter Ngramları, bir kelimenin yan yana N tane karakteriyle üretilen özellik kümesidir. Kelime Ngramları ise, bir metinde yan yana N tane kelime gruplarıyla üretilen özellik kümesidir. Özellikle yazarın yoğun olarak kullandığı N'li harf ve kelime öbeklerinin ortaya çıkarılmasını sağlar. Örneğin; "gelicem", "gidicem", "düşüncem" gibi gelecek zamanlı cümleleri sürekli bu şekilde bozarak yazan birini ayırt etmede "cem" 3gramı önemli bir özellik olabilir.

Tez çalışmasında da karakter 3gramları kullanılmıştır. Karakter 3gramları hesaplanırken, yazarı karakterize etmede fayda sağlayabilen noktalama işaretleri, boşluklar ve alfanümerik olmayan diğer karakterler de silinmemiştir. Örnek olarak "çok güzel!" cümlesinin tüm 3gramları aşağıdaki gibi çıkarılmıştır:

çok/ok /k g/ gü/güz/üze/zel/el!

2.4 Frekans Hesaplamada Kullanılan Yöntemler

Metinde geçen kelimelerin kendisi, kökü, türü, karakter Ngramları, kelime Ngramları gibi özellik setlerinin frekansları birkaç farklı yöntemde hesaplanabilir. Tez çalışmasında kullanılan terim frekansını hesaplama yöntemleri binary, TF, TF-IDF ve NTF'dir. Bu yöntemler aşağıdaki gibi tanımlanabilir:

- Binary : Terimin metinde geçip geçmeme durumuna göre 1 veya 0.
- TF (*Term Frequency*) : Terimin metinde toplam geçme sayısı.
- TF-IDF (*Term Frequency-Inverse Document Frequency*) : TF*IDF.

  IDF (*Inverse Document Frequency*) : $\log_e$(Toplam metin sayısı / Terimin geçtiği metin sayısı).
- NTF (*Normalized Term Frequency*) : (Terimin metinde toplam geçme sayısı / Metindeki toplam terimin sayısı).

2.5 Anlamsal Uzay

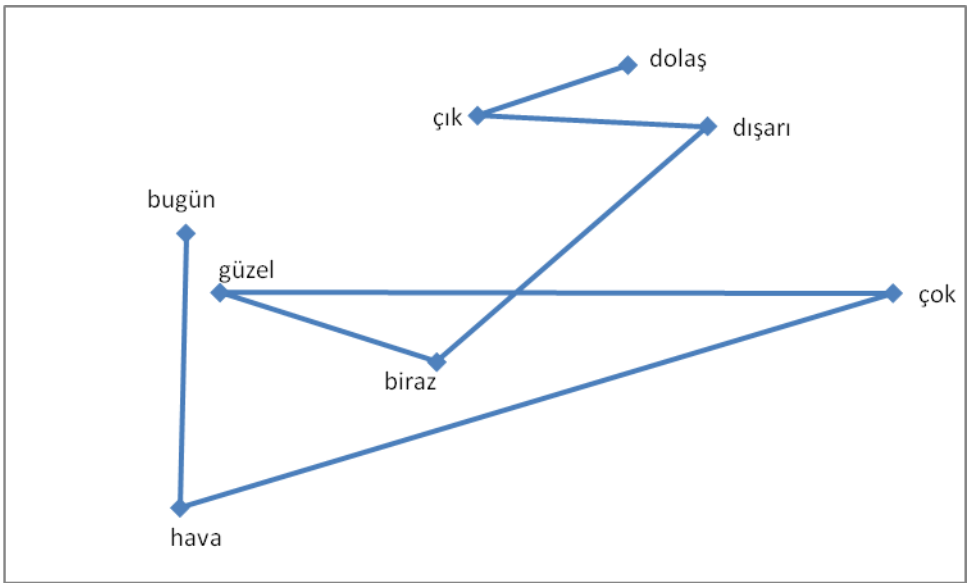
Bir metin, içinde geçen kelimelerin matematiksel uzaydaki koordinatları kullanılarak ifade edilebilir. Metnin içerdiği kelimelerin koordinatları çeşitli yöntemlerle hesaplanabilir. Tez çalışılmasında metinleri anlamsal uzayda ifade edebilmek için Word2Vec [13] aracı kullanılmıştır. Word2Vec C dilinde yazılmış, açık kaynak kodlu bir kütüphanedir. Girdi olarak aldığı metin içerisindeki kelimelerin belirli bir pencere boyutunda birlikte geçme frekanslarını kullanarak her kelimeyi anlamsal uzayda birer nokta olarak ifade eder ve koordinatlarını çıktı olarak verir. Ayrıca, ürettiği bu koordinatları kullanarak kelimelerin birbirleriyle arasındaki öklid uzaklığını hesaplar ve bir kelimeye en yakın olan belirli sayıdaki kelimeyi listeler. Kelimelerin birlikte geçme frekansları bir kelimenin anlamca hangi kelimelere daha yakın olduğunu gösterir. Yani, birlikte geçme frekansı yüksek olan iki kelime birbirine anlam bakımından yakındır ve koordinatları arasındaki öklid mesafesi de diğer kelimelerle olan uzaklığından küçüktür [10].

Word2Vec programı girdi ve çıktının dosyalarının yeri haricinde aşağıdaki parametreleri de almaktadır:

- Vektör boyutu
- Yanyana geçme frekansını bulurken kullanılan pencere boyutu
- Eğitim algoritması : hiyerarşik softmax veya negatif örnekleme
- Alt örnekleme (downsampling) için eşik frekans değeri
- Kullanılacak thread sayısı
- Çıktı dosyasının formatı : metin veya binary

2.6 Metin Yörüngeleri

Metin yörüngesi, metinde geçen kelimelerin anlamsal uzayda koordinatları hesaplandıktan sonra bu kelimeleri ifade eden noktaların kelimelerin metinde geçme sırasına göre birleştirilmesiyle çizilen yoldur. Örneğin; "Bugün hava çok güzel. Biraz dışarı çıkıp dolaşalım." yazısında geçen kelimelerin koordinatları bulunduktan sonra bu metnin yörüngesi Şekil 2.1'deki gibi çizilebilir.



Şekil 2. 1 Örnek bir metin yörüngesi

Veri kümesindeki metinlerin yörüngeleri elde edildikten sonra bu yörüngelerin çeşitli özellikleri kullanılabilir. Kelimeler arasındaki mesafeler ve açılar yörüngeleri ifade etmek için kullanılabilecek özelliklerden bazılarıdır. Yörünge üzerindeki ardışık her iki nokta arasındaki mesafeler kullanılabileceği gibi, her noktanın kendinden sonra gelen N'ninci nokta ile arasındaki mesafe de kullanılabilir. Açılar için de aynı şekilde, *(n, n+1, n+2)* gibi ardışık 3 noktanın oluşturduğu açının yanı sıra*, (n, n+1, n+3)* gibi üç noktanın oluşturduğu açı değeri yörüngeleri ifade etmek için kullanılabilir. İki farklı metnin yörüngeleri için hesaplanan uzaklık ve açı değerleri, bu iki metni birbirinden ayırt etmeyi sağlayabilir [14].

BÖLÜM 3

VERİ SETLERİ

Tez çalışmasında iki farklı hipotez için iki farklı veri seti grubu kullanılmıştır. Bunlardan birincisi psikolojik rahatsızlığı olan hastalar ile sağlıklı deneklerin yazdığı yazılardan oluşan veri kümesidir. Bu birinci çalışmada, ayrıca Word2Vec ile kelime koordinatlarını bulmak için haber metinleri de kullanılmıştır. İkinci veri kümesi ise, köşe yazarlarının makaleleri ve tweetlerinden oluşmaktadır. Bu bölümde bu iki veri seti hakkında detaylı bilgi verilmektedir.

3.1 Psikolojik Hastaların Yazılarından Oluşan Veri Seti

Metinden yazarın psikolojik rahatsızlığını tahmin etmeye yönelik olarak yapılan çalışmada kullanılan veri setidir. Bir nöropsikiyatri hastanesinden elde edilen veriler; bir kısmı depresyon ve/veya anksiyete rahatsızlığı olan hastalar ile diğer kısmı ise sağlıklı olan toplam 267 deneğin yazmış olduğu yazılardan ibarettir. Hasta ve sağlıklı deneklerden yazı toplayabilmek için, iki soru sorulmuş ve cevaplarını yazmaları istenmiştir. Bu sorulardan birincisi hayata ve geleceğe dair düşünceleri, ikincisi ise günlük aktiviteleridir. Cevaplarda verilen kişisel bilgiler gizlilik ilkesi gereği metinlerden çıkarılmıştır.

Veri kümesinde her yazarın yazdığı metin ve yazarın özellikleri birer satır olarak tutulmuştur. Yazarın özellikleri ile metin aralarına "/" karakteri konularak ayrılmıştır. Örnek bir satırın yapısı şu şekildedir:

Denek no / Cinsiyet / Yaş / Eğitim durumu / Teşhis1 / Teşhis2 / Teşhis3 / Yanıt1 / Yanıt2

3.2 Haber Metinleri

İnternetteki haber sitelerinden toplanan 47 bin haber metninden oluşan veri kümesidir. Metinlerde geçen kelimelerin köklerinin anlamsal uzayda koordinatlarını bulmak için kullandığımız Word2Vec programına girdi verisi olarak kullanılmıştır. Bu haber metinleri güncel, siyaset, ekonomi, spor, dünya, genel, magazin, kültür-sanat, planet, sağlık, Türkiye, sağlık, teknoloji ve yaşam olmak üzere 14 farklı kategoride yazılmış haberlerden oluşmaktadır.

3.3 Köşe Yazarlarının Makaleleri ve Tweetlerinden Oluşan Veri Seti

Twitter hesabının yazarını tanımaya yönelik olarak yapılan çalışmada kullanılan veri setidir. Öncelikle, sosyal medyayı aktif olarak kullanan, Türkiye'de en çok takip edilen Twitter kullanıcıları arasında yer alan, hesaplarında kişisel düşüncelerini sıkça paylaşan 10 köşe yazarı tespit edilmiştir. Bu köşe yazarlarının düşüncelerini yazarken genel olarak Türkçe dilbilgisi kurallarına uymaları, noktalama işaretlerine dikkat etmeleri tercih edilmelerinde önemli bir sebeptir. Belirlenen 10 yazarın 200'er tweeti ve 50'şer makalesi toplanmıştır. Çalışmada kullanılan yazarların ismi Çizelge 3.1'de verilmiştir.

Çizelge 3. 1 Tweetleri ve makaleleri toplanan köşe yazarları

Yazar Adı	Twitter Kullanıcı Adı
Cüneyt Özdemir	@cuneytozdemir
Can Dündar	@candundaradasi
Ece Temelkuran	@ETemelkuran
Uğur Dündar	@ugurdundarsozcu
İsmail Küçükkaya	@KucukkayaIsmail
Yekta Kopan	@yektakopan
İlber Ortaylı	@ILBERORTAYLIGSU
Mustafa Balbay	@mustafabalbay
Emre Uslu	@EmreUslu
Şamil Tayyar	@samiltayyar27

Çalışmada kullanılan makaleler Cumhuriyet (www.cumhuriyet.com.tr), Radikal (www.radikal.com.tr), Habertürk (www.haberturk.com), Taraf (www.taraf.com.tr), Milliyet (www.milliyet.com.tr), Sözcü (www.sozcu.com.tr) gibi yazarların çalıştıkları günlük gazetelerin web sitelerinin arşivinden toplanmıştır. Yazarların makaleleri toplanırken, en güncel yazılan makaleden geriye doğru ardışık tarihlerde yazdığı makalelerin herhangi biri elenmeden 50 tanesi alınmıştır.

Her bir yazarın 200'er tweeti toplanırken de yine en güncel paylaşılan tweetinden başlanarak geriye doğru gidilmiştir ve yazarın kendi görüşünü açık bir şekilde ifade etmeyen bazı "değersiz" tweetleri elenmiştir. Bu çalışma için "değersiz" olarak nitelendirilen ve elenen tweetlerin özellikleri şu şekildedir:

- Retweetler,
- Üç kelimeden az kelime sayısına sahip tweetler,
- Facebook, instagram gibi diğer sosyal medya hesaplarında yaptıkları paylaşımlarından sonra otomatik olarak atılan tweetler,
- Haber siteleri başta olmak üzere bir web sitesi sayfasında bulunan "Twitter'da paylaş" seçeneği ile paylaşılan tweetler,
- Yaptıkları programlar başta olmak üzere çeşitli programlar veya etkinlikler hakkındaki genel olarak yer, tarih ve saat bilgisi veren duyuruları içeren tweetler,
- Tek başına anlamlı olmayan, başka bir kullanıcıya yanıt olarak yazılan "mention" olarak adlandırılan sohbet tweetleri,
- Metin yerine imaj olarak paylaşılan tweetler,
- Bir özdeyişin paylaşıldığı, birebir alıntının yapıldığı tweetler.

Yapılan çalışmada tweetler ile makalelerin yazıldığı tarih aralığının sınıflandırmada bir etkisinin olup olmadığını da görebilmek için toplanan verilerin tarih aralığı da tutulmuştur. Çizelge 3.2'de görüldüğü gibi, Uğur Dündar ve Mustafa Balbay gibi bazı yazarların kullanılan tweetlerinin paylaşıldığı tarih aralığı ile kullanılan makalelerinin yazıldığı tarih aralığı büyük ölçüde örtüşürken, uzun süre önce günlük gazete yazarlığına ara veren Cüneyt Özdemir, Ece Temelkuran, İsmail Küçükkaya ve Şamil

Tayyar gibi yazarların kullanılan tweetlerinin paylaşıldığı tarih aralığı ile kullanılan makalelerinin yazıldığı tarih aralığı hiç örtüşmemektedir. Kimi yazarların tweetleri ile makaleleri yazdığı tarihler ise kısmen örtüşmektedir.

Çizelge 3. 2 Tweetleri ve makaleleri toplanan köşe yazarları

Köşe Yazarı	**Tweetlerin Paylaşıldığı Tarih Aralığı**	**Makalelerin Yazıldığı Tarih Aralığı**
Cüneyt Özdemir	11.01.2016 - 17.06.2015	15.06.2014 - 16.03.2014
Can Dündar	07.01.2016 - 03.04.2015	06.08.2015 - 29.04.2015
Ece Temelkuran	13.01.2016 - 03.08.2014	01.06.2014 - 12.08.2013
Uğur Dündar	18.01.2016 - 13.07.2015	20.01.2016 - 08.11.2015
İsmail Küçükkaya	21.01.2016 - 23.10.2015	26.06.2013 - 21.01.2013
Yekta Kopan	20.01.2016 - 13.11.2015	20.01.2016 - 31.12.2014
İlber Ortaylı	15.01.2016 - 29.09.2014	24.01.2016 - 08.03.2015
Mustafa Balbay	29.01.2016 - 29.10.2015	08.11.2015 - 31.08.2015
Emre Uslu	31.01.2016 - 12.01.2016	18.11.2015 - 28.08.2014
Şamil Tayyar	31.01.2016 - 05.09.2015	15.08.2012 - 03.11.2010

BÖLÜM 4

DENEYSEL SONUÇLAR

Üzerinde çalışılan iki farklı hipotez için yapılan deneylerden elde edilen sonuçlar ve yorumların yer aldığı bu bölümde, ilk olarak metinlerden psikolojik hastalık teşhisine yönelik yapılan testler ve yorumlar, daha sonra da bir Twitter hesabının kime ait olduğunun tespitine yönelik yapılan testler ve yorumlar sunulmuştur.

4.1 Metinden Psikolojik Rahatsızlık Teşhisi

Öncelikle psikolojik rahatsızlığı olan ve sağlıklı denekler tarafından yazılmış metinlerden oluşan veri setinde seçme işlemi yapılmıştır. Veri kümesindeki bir hastaya birden fazla tanı konulabilmektedir. Bu çalışmada amacımız, anksiyete hastası, depresyon hastası ve sağlıklı kişileri birbirinden ayırt edebilmek olduğu için, hem anksiyete hem de depresyon tanısı birlikte konmuş olan hastalar elenmiştir. 267 denekten oluşan veri setinde, seçme işlemi uygulandıktan sonra 225 denek kalmıştır. Bu deneklerin dağılımı ve toplam denek sayısına oranı şu şekildedir:

- Sağlıklı denem sayısı : 49, Oranı : 49 / 225 = %21,7
- Aksiyete olup da depresif olmayan denek sayısı : 40, Oranı 40 / 225 = %17,7
- Depresif olup da anksiyete olmayan denek sayısı : 136, Oranı 136 / 225 = %60,4

Yukarıdaki dağılıma bakıldığında, rastgele yapılacak olan bir seçimde depresif deneğin seçilme ihtimali %60,4'tür. Yani bizim için başarılı sonuç bu eşik değer üzerindeki başarı oranlarıdır.

Testlerin tamamında Word2Vec algoritması, yan yana geçme frekansını bulurken kullandığı pencere boyutu parametresi 5 değeri ile çalıştırılmıştır. Vektör boyutu olarak ise 200 değeri girilmiştir.

4.1.1 Kelimelerin Kökleri ile Üretilen Binary Matris

225 deneğin yazdığı metinler Zemberek yardımıyla köklerine ayrılmıştır. Kelimelerin köklerini kullanarak, kelimenin kökünün metinde geçip geçmeme duruma göre bir binary matris oluşturulmuştur. Bu binary matris kullanılarak üretilen arff dosyası WEKA aracındaki Bagging, NB, J48 ve SMO sınıflandırıcıları ile 10 kez çapraz doğrulama yapılarak test edilmiştir. Testin sonuçları Çizelge 4.1'de sunulmuştur.

Çizelge 4. 1 Kelimelerin kökleri ile üretilen binary matris başarı oranları

Sınıflandırıcı	**Başarı oranı (%)**
Bagging	57.77
Naive Bayes	55.55
J48	49.33
SMO	51.55

4.1.2 Kelimelerin Köklerinin Koordinatları ile Üretilen Matris

İlk olarak hem hasta verilerinin hem de haber metinlerinin köklerini bulmak için Zemberek aracı kullanılmış ve çalışmaya bu kelime kökleri ile devam edilmiştir. Toplam 12,376 kelimeden oluşan hasta metinlerinde 482 farklı kök geçtiği görülmüştür. Haber metinlerinde ise toplam 8,566,892 kelime mevcutken, farklı kelime sayısı 12,540'dır.

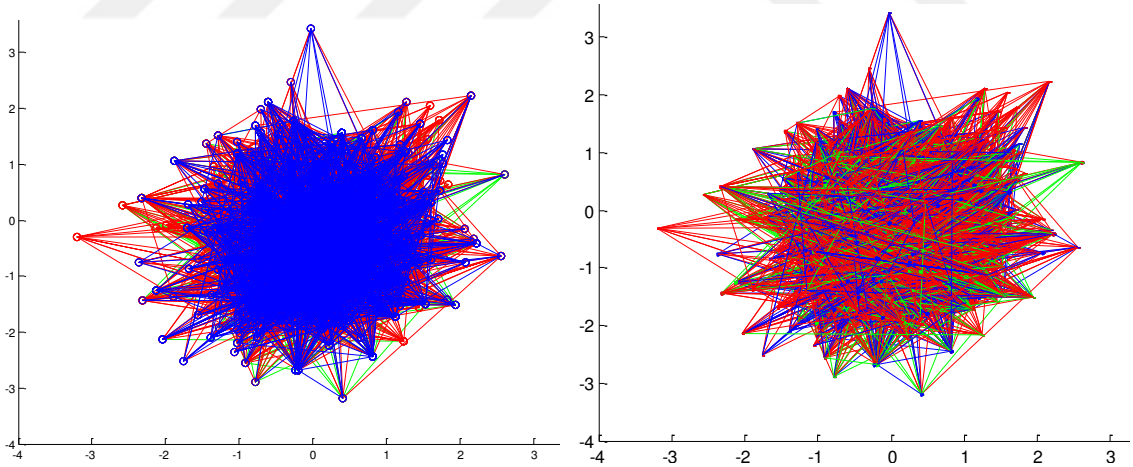
Kelimelerin köklerinin koordinatlarını bulabilmek için Word2Vec aracı kullanılmıştır. Word2Vec ile koordinatları hesaplarken öncelikle hastaların kendi yazdıkları metinler, daha sonra ise 47 bin haber metni kullanılmış ve bu koordinatlar ile üretilen arff dosyasının dört farklı sınıflandırıcıyla 10 kez çapraz doğrulama yapılarak elde edilen başarı değerleri Çizelge 4.2'de sunulmuştur.

Çizelge 4. 2 Kelime köklerinin 200 boyutlu uzayda koordinatlarının başarı oranları

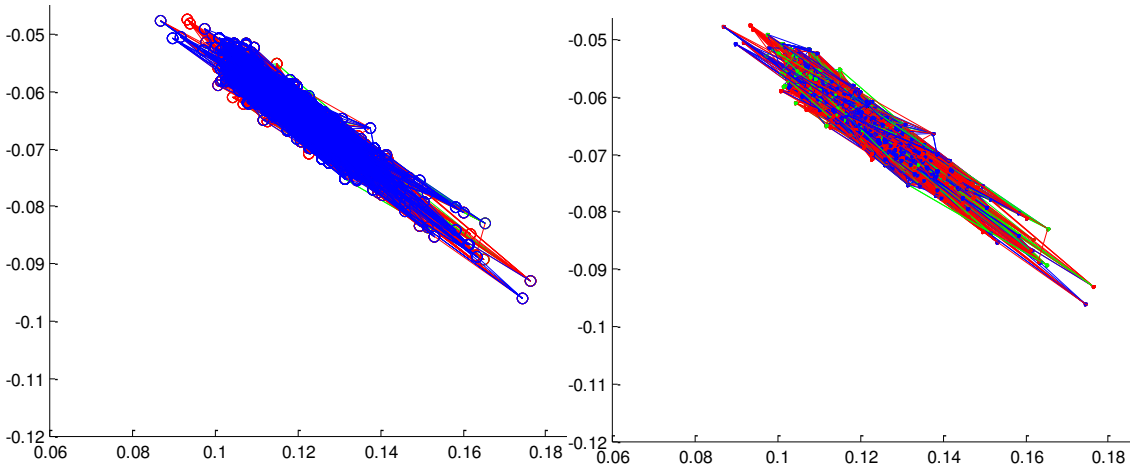
Sınıflandırıcı	Haber datalarıyla (%)	Hasta datalarıyla (%)
Bagging	59.11	59.55
Naive Bayes	34.66	28.00
J48	42.22	48.00
SMO	55.55	60.00

4.1.3 Metin Yörüngelerinin Görselleştirilmesi

Sınıfların metin yörüngelerinin birbirinden farklı olup olmadığını gözlemleyebilmek için, her sınıftaki deneklerin yazıları birleştirilerek elde edilen üç birleştirilmiş metin için üst üste üç farklı metin yörüngesi çizilmiştir. Metin yörüngelerini görselleştirmek için MATLAB aracı kullanılmıştır. Üretilen 200 boyuttaki koordinatların ilk 2 ve 3 boyutu kullanılarak çizilen yörüngeler Şekil 4.1, 4.2, 4.3 ve 4.4'te sunulmuştur. Anksiyete hastalarının metinleri yeşil, depresyon hastalarının metinleri kırmızı, sağlıklı deneklerin metinleri ise mavi renkte çizilmiştir.

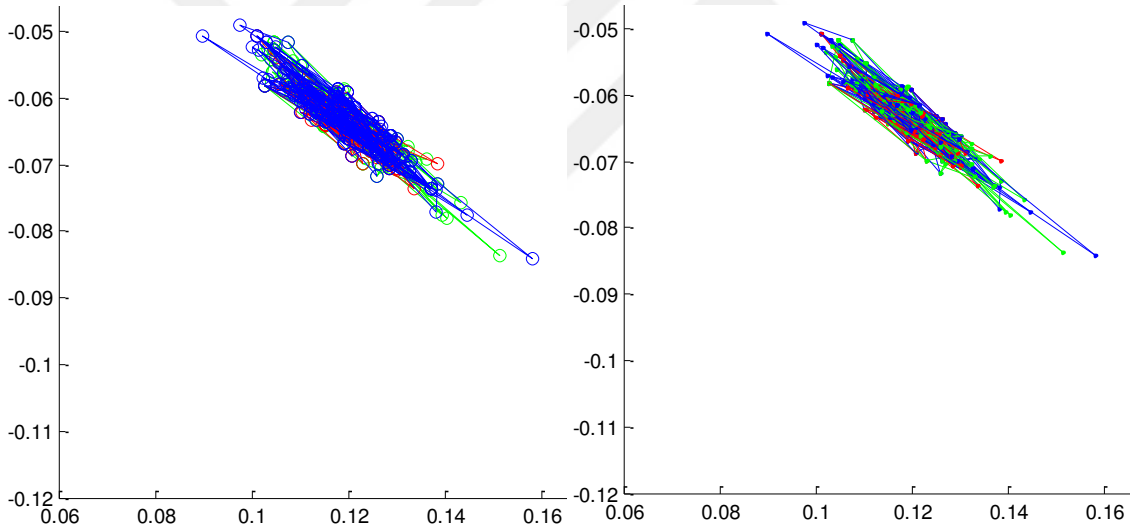


Şekil 4. 1 Haber verileriyle üretilen koordinatların sırasıyla 2 ve 3 boyutu kullanılarak çizilen birleştirilmiş metin yörüngeleri

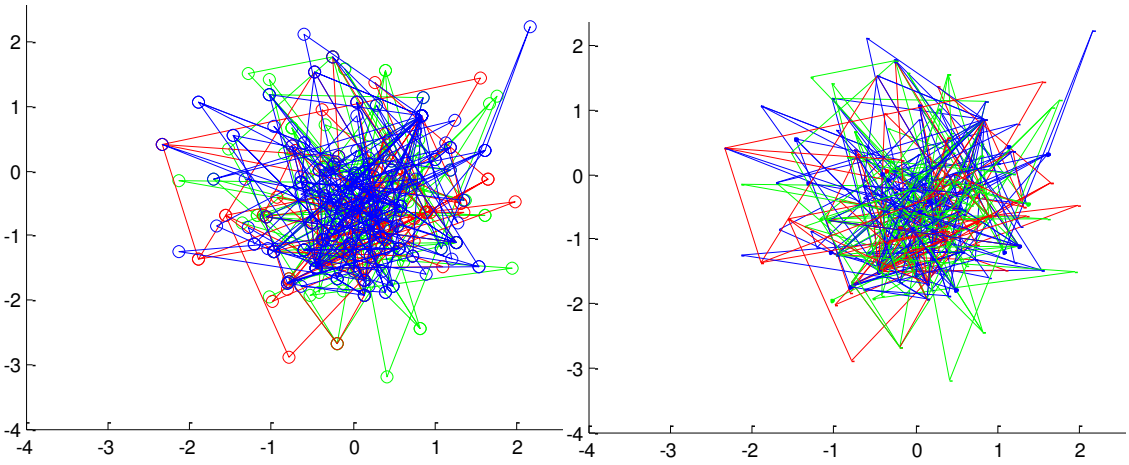


Şekil 4. 2 Hasta verileriyle üretilen koordinatların sırasıyla 2 ve 3 boyutu kullanılarak çizilen birleştirilmiş metin yörüngeleri

Tüm metinler kullanılarak üst üste çizdirilen görsellerde sınıflar arasında ayrım yapmanın zor olduğu görülmüş ve 3'er denek kullanılarak yörüngeler çizdirilmiştir.



Şekil 4. 3 Hasta verileriyle üretilen koordinatların sırasıyla 2 ve 3 boyutu kullanılarak çizilen ilk 3'er deneğin birleştirilmiş metin yörüngeleri



Şekil 4. 4 Haber verileriyle üretilen koordinatların 2 ve 3 boyutu kullanılarak çizilen ilk 3'er deneğin birleştirilmiş metin yörüngeleri

4.1.4 Metin Yörüngelerinin Uzaklık ve Açı Değerlerinin Değişken Aralıklı Histogramları

Word2Vec ile haber ve hasta metinleri ayrı ayrı kullanılarak elde edilen 200 boyuttaki koordinatların ilk 2 ve 3 ve 200 boyuttaki değerleri kullanılarak her deneğin yazdığı metin yörüngelerinin uzaklık ve açı histogramları MATLAB'da hesaplanmıştır. Histogram aralık sayısı 10 olarak verilmiş, aralık değerleri ise MATLAB'ın *hist* fonksiyonu tarafından eşit mesafeli olacak şekilde otomatik olarak oluşturulmuştur. Uzaklık değerleri ardışık iki noktanın birbirine olan öklid mesafesi hesaplanarak bulunmuştur. Ardışık olan üç noktanın oluşturduğu açı değeri hesaplanırken de birbirini kesen ve uzunluğu bilinen iki doğru parçasının arasındaki açıyı bulmaya yarayan kosinüs formülü kullanılmıştır. 2, 3 ve 200 boyutla elde edilen sonuçlar sırasıyla 2D, 3D ve 200D olarak çizelgede listelenmiştir.

4.1.4.1 Histogram Aralıklarına Düşen Değerlerin Hem Sayı Hem de Orta Noktalarının Kullanılması

Bir metinde geçen kelimelerin köklerinin koordinatlarıyla hesaplanan uzaklık ve açı değerlerinin histogram aralıklarına dağılım sayıları ve her aralıktaki değerlerin orta noktaları kullanılarak arff dosyaları oluşturulmuş ve dört sınıflandırıcı ile 10 kez çapraz doğrulama yapılarak çalıştırılmıştır. Uzunluk değerleri kullanılarak elde edilen sonuçlar Çizelge 4.3'te, açı değerleri kullanılarak elde edilen sonuçlar 4.4 ve ikisi bir arada kullanılarak elde edilen sonuçlar ise 4.5'te sunulmuştur.

Çizelge 4. 3 İki, üç ve ikiyüz boyutlu düzlemde uzaklıkların histogramlarıyla üretilen arff sonuçları

Sınıflandırıcı	**Haber datalarıyla (%)**			**Hasta datalarıyla (%)**		
	2D	**3D**	**200D**	**2D**	**3D**	**200D**
Bagging	57.33	58.22	58.66	54.66	56.00	55.55
Naive Bayes	27.55	31.11	31.11	21.33	22.66	24.44
J48	42.22	51.11	46.22	50.22	46.22	44.88
SMO	60.44	60.44	60.44	60.44	60.44	60.44

Çizelge 4. 4 İki, üç ve ikiyüz boyutlu düzlemde açıların histogramlarıyla üretilen arff sonuçları

Sınıflandırıcı	**Haber datalarıyla (%)**			**Hasta datalarıyla (%)**		
	2D	**3D**	**200D**	**2D**	**3D**	**200D**
Bagging	58.66	54.22	60.44	54.22	54.22	54.66
Naive Bayes	24.44	25.33	24.00	20.88	25.33	26.22
J48	43.55	47.55	60.44	44.00	47.55	43.11
SMO	60.44	60.44	60.44	60.44	60.44	60.44

Çizelge 4. 5 İki boyutlu düzlemde uzaklıkların histogramları ve açıların histogramları birlikte kullanılarak üretilen arff sonuçları

Sınıflandırıcı	**Haber datalarıyla (%)**
Bagging	54.66
Naive Bayes	25.77
J48	47.11
SMO	60.44

4.1.4.2 Histogram Aralıklarına Düşen Değerlerin Sadece Sayı veya Orta Noktalarının Kullanılması

Uzaklık ve açı değerlerinin histogram aralıklarına dağılım sayıları ve her aralıktaki değerlerin orta noktalarının tek başına daha başarılı olup olmadığını görmek için bu değerler ayrı ayrı kullanılarak arff dosyaları oluşturulmuş ve dört sınıflandırıcı ile 10 kez çapraz doğrulama yapılarak çalıştırılmıştır. Elde edilen sonuçlar Çizelge 4.6 ve 4.7'de sunulmuştur.

Çizelge 4. 6 İki boyutlu düzlemde uzaklıkların histogramlarıyla (sadece sayılar ve sadece orta noktalar kullanılarak) üretilen arff sonuçları

Sınıflandırıcı	**Haber datalarıyla (%)**		**Hasta datalarıyla (%)**	
	Sadece sayılarla	**Sadece orta noktalarla**	**Sadece sayılarla**	**Sadece orta noktalarla**
Bagging	56.44	55.55	53.33	56.88
Naive Bayes	36.44	27.11	47.11	20.00
J48	44.44	57.77	53.33	60.44
SMO	60.44	60.44	60.44	60.44

Çizelge 4. 7 İki boyutlu düzlemde uzaklıkların histogramları ve açıların histogramları birlikte ile (sadece sayılar ve sadece orta noktalar kullanılarak) üretilen arff sonuçları

Sınıflandırıcı	**Haber datalarıyla (%)**	
	Sadece sayılarla	**Sadece orta noktalarla**
Bagging	56.44	58.22
Naive Bayes	25.33	24.88
J48	47.11	54.22
SMO	60.44	60.44

4.1.5 Metin Yörüngelerinin Uzaklık ve Açı Değerlerinin Sabit Aralıklı Histogramları

Metin yörüngelerinden elde edilen uzaklık ve açı değerlerinin değişken aralıklarla üretilen histogramlarının farklı sınıflardan deneklerin metin yörüngeleri arasındaki farkı net olarak yansıtmayacağı düşünülerek, sabit aralıklarla histogramlar üretilmiştir. Yapılan testlerde MATLAB'ın *hist* fonksiyonuna istenen aralık değerleri doğrudan verilmiştir. Aralıkların orta noktaları tüm arff dosyalarında aynı olacağı için kullanılmamıştır. 2, 3 ve 200 boyutla elde edilen sonuçlar sırasıyla 2D, 3D ve 200D olarak çizelgede listelenmiştir.

4.1.5.1 Metin Yörüngelerinin Uzaklık Değerlerinin Sabit Aralıklı Histogramları

Word2Vec ile haber metinleriyle elde edilen 200 boyuttaki koordinatların, ilk 2, 3 boyutu ve tamamı kullanılarak elde edilen metin yörüngelerinin uzaklık değerlerinin sabit aralıklı histogramlarının sadece aralığa düşen uzaklık değerlerinin sayıları ile oluşturulan arff dosyalarının test sonuçları Çizelge 4.8'de sunulmuştur.

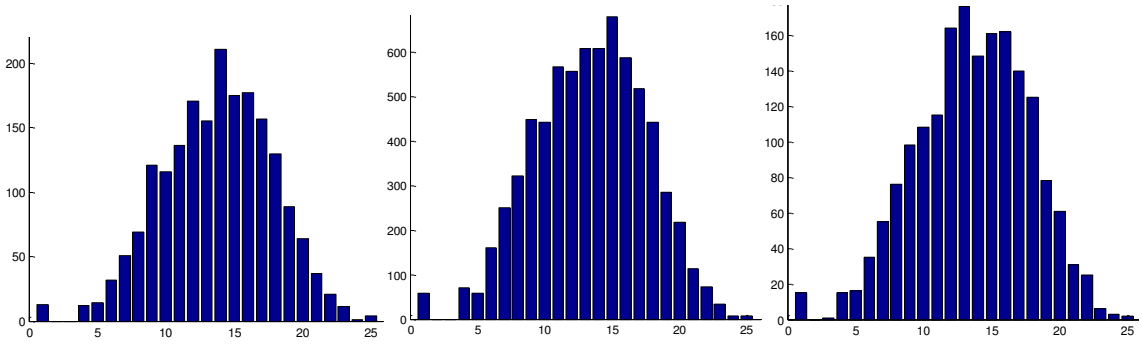
Çizelge 4. 8 Yörüngelerin uzaklık histogramlarıyla üretilen arff sonuçları

Sınıflandırıcı	**2D (%)**	**3D (%)**	**200D (%)**
Bagging	57.77	58.66	56.44
Naive Bayes	48.44	44.44	41.33
J48	44.88	45.33	50.66
SMO	60.44	60.44	41.33

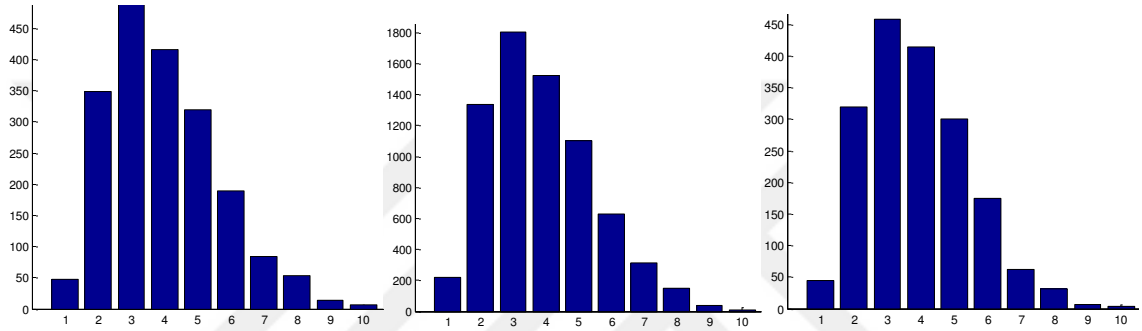
Çizelge 4.14'te kullanılan histogramların aralık değerleri şöyledir:

- 2D ve 3D için aralıklar = [0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5]
- 200D için aralıklar = [6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28]

Sınıfların histogramlarının görsel olarak birbirinden ayırt edilebilir olup olmadığını görmek için MATLAB'da sınıf bazlı histogramlar çizdirilmiştir. Şekil 4.5 ve 4.6'da sırasıyla 200 ve 2 boyuttaki yörüngelerin uzaklıklarının yukarıdaki aralık değerleri kullanılarak çizilen sınıf bazlı histogramlar sunulmuştur.

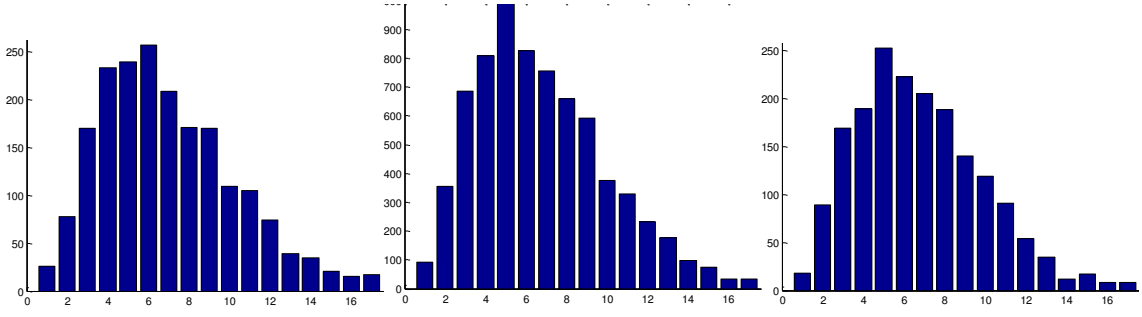


Şekil 4. 5 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 200 boyutta elde edilen histogramlar



Şekil 4. 6 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 2 boyutta elde edilen histogramlar

2 boyutta elde edilen histogramlarda aralık değerlerini daha da küçülterek sınıflar arasındaki fark görülmek istenmiştir ve Şekil 4.7'de şu aralık değerlerine göre histogramlar tekrar çizilmiştir: [0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 3.25, 3.5, 3.75, 4]



Şekil 4. 7 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 2 boyutta elde edilen histogramlar

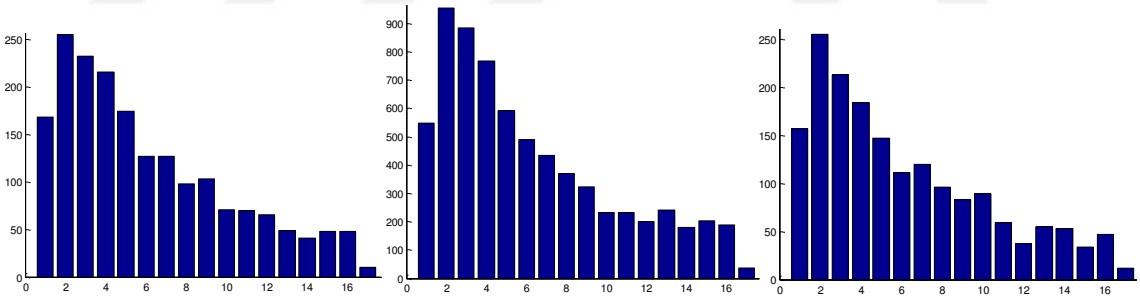
4.1.5.2 Metin Yörüngelerinin Açı Değerlerinin Sabit Aralıklı Histogramları

Haber metinleriyle 200 boyutta üretilen koordinatların ilk 2 ve 3 boyutu kullanılarak çizilen metin yörüngelerinin açı değerlerinin histogramları ile elde edilen arff dosyalarının dört farklı sınıflandırıcı ile yapılan test sonuçları Çizelge 4.9'da sunulmuştur.

Çizelge 4. 9 Yörüngelerin açı histogramlarıyla üretilen arff sonuçları

Sınıflandırıcı	**2D (%)**	**3D (%)**
Bagging	56.00	56.88
Naive Bayes	34.22	33.77
J48	46.66	39.55
SMO	60.44	60.44

İki boyutta çizilen yörüngelerin açı değerlerinin sınıf bazlı histogramları Şekil 4.8'deki gibi görselleştirilmiştir.



Şekil 4. 8 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 2 boyutta elde edilen histogramlar

4.1.6 Metin Yörüngelerinin Farklı Uzaklık ve Açı Değerlerinin Sabit Aralıklı Histogramları

Bir önceki bölümde *(n, n+1)* gibi ardışık iki noktanın arasındaki uzaklıklar ve *(n, n+1, n+2)* gibi ardışık üç noktanın oluşturduğu açılar hesaplanmıştı. Bu bölümde ise, farklı uzaklık ve açı değerleri hesaplanarak, testler tekrar yapılmıştır.

4.1.6.1 Metin Yörüngelerinin Farklı Uzaklık Değerlerinin Sabit Aralıklı Histogramları

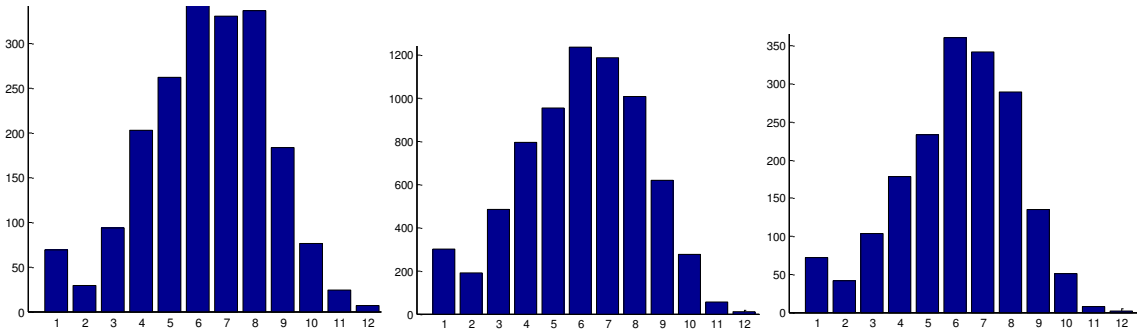
Bu bölümde, ardışık iki nokta arasındaki uzaklık yerine, bir noktanın kendisinden iki sonraki noktayla arasındaki mesafeye bakılmıştır. Örneğin, 1, 2, 3 4 ve 5 numaralı noktalar için 1-3, 2-4, 3-5 numaralı noktalar arasındaki mesafeler hesaplanmaktadır. Haber verileriyle elde edilen 200 boyuttaki koordinatların ilk 2, 3 ve 200 boyutu kullanılarak çizilen yörüngelerin anlatıldığı gibi hesaplanan uzaklıklarının histogramlarının dört sınıflandırıcı ile çalıştırılması ile alınan sonuçlar Çizelge 4.10'da sunulmuştur. Histogramlar için aşağıdaki aralık değerleri sabit olarak kullanılmıştır:

- 2D ve 3D için aralıklar = [0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5]
- 200D için aralıklar = [6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28]

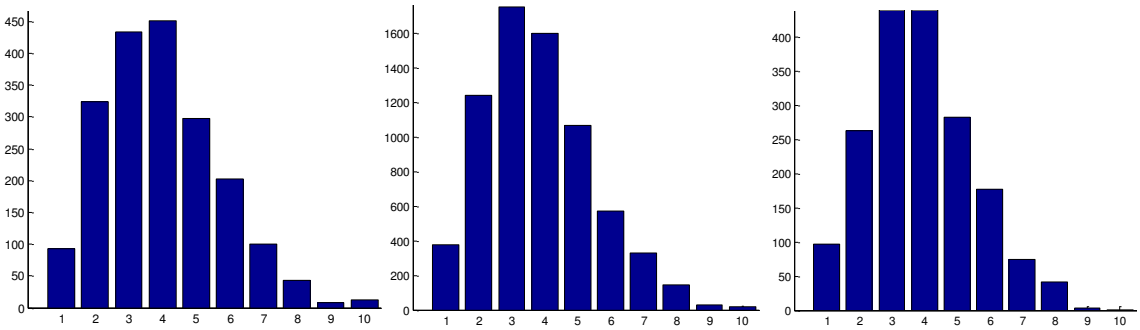
Çizelge 4. 10 Yörüngelerin uzaklık histogramlarıyla üretilen arff sonuçları

Sınıflandırıcı	**2D (%)**	**3D (%)**	**200D (%)**
Bagging	57.77	56.44	60.88
Naive Bayes	35.11	31.11	29.77
J48	56.00	51.55	50.22
SMO	60.00	60.44	60.44

200 ve 2 boyut kullanılarak elde edilen yörüngelerin uzaklıklarının histogramları sınıf bazlı olarak sırasıyla Şekil 4.9 ve 4.10'da sunulmuştur.



Şekil 4. 9 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 200 boyutta elde edilen histogramlar



Şekil 4. 10 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 2 boyutta elde edilen histogramlar

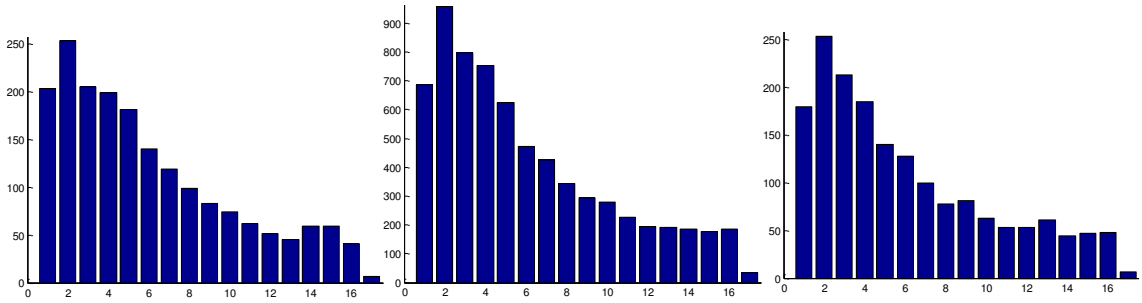
4.1.6.2 Metin Yörüngelerinin Farklı Açı Değerlerinin Sabit Aralıklı Histogramları

Bu bölümde, ardışık üç noktanın oluşturduğu açıyı kullanmak yerine, *(n, n+1, n+3)* gibi ardışık iki noktadan sonra bir nokta atlanarak dördüncü nokta arasındaki açı kullanılmıştır. Örneğin, 1, 2, 3 4 ve 5 numaralı noktalar için 1-2-4 ve 2-3-5 numaralı noktalar arasındaki açılar hesaplanmaktadır. Haber verileriyle elde edilen 200 boyuttaki koordinatların ilk 2, 3 boyutu kullanılarak çizilen yörüngelerin anlatıldığı şekilde hesaplanan açılarının histogramlarının dört sınıflandırıcı ile çalıştırılması ile alınan sonuçlar Çizelge 4.11'de sunulmuştur. Histogramlar bulunurken, hem 2D hem de 3D için aynı aralık değerleri kullanılmıştır: [0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2, 2.2, 2.4, 2.6, 2.8, 3, 3.2]

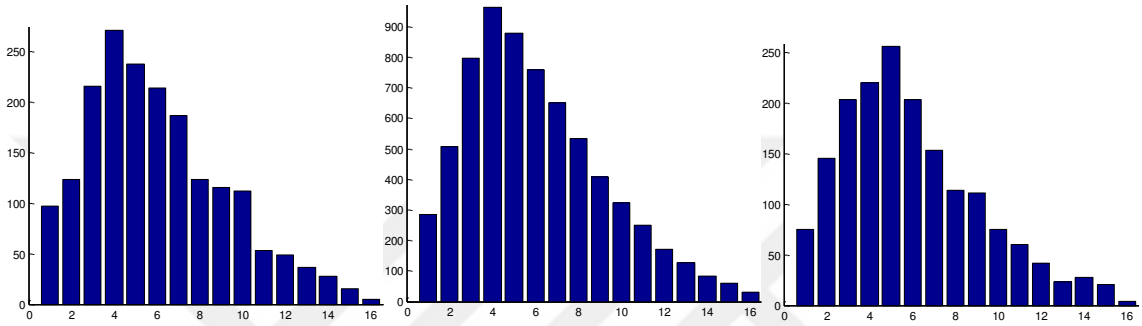
Çizelge 4. 11 Yörüngelerin açı histogramlarıyla üretilen arff sonuçları

Sınıflandırıcı	**2D (%)**	**3D (%)**
Bagging	56.44	57.33
Naive Bayes	30.66	32.44
J48	44.44	48.44
SMO	60.44	60.44

2 ve 3 boyut kullanılarak elde edilen yörüngelerin açılarının histogramları sınıf bazlı olarak sırasıyla Şekil 4.11 ve 4.12'de sunulmuştur.



Şekil 4. 11 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 2 boyutta elde edilen histogramlar



Şekil 4. 12 Sırasıyla anksiyete, depresyon hastalarının ve sağlıklı deneklerin metinlerinden 3 boyutta elde edilen histogramlar

Çizelge 4.17 haber metinleriyle üretilen 200 boyuttaki koordinatların ilk 2 ve 3 boyutu kullanılarak çizilen yörüngelerin testinin sonuçlarını içermekteydi. Bu aşamada ise, koordinatlar elde edilirken, Word2Vec'e girdi olarak haber metinleri yerine hastaların kendi yazıları verilmiş ve aynı testler tekrar yapılmıştır. Histogram aralık değerleri de aynı şekilde belirlenmiştir. Sonuçlar Çizelge 4.12'de sunulmuştur.

Çizelge 4. 12 Yörüngelerin açı histogramlarıyla üretilen arff sonuçları

Sınıflandırıcı	**2D (%)**	**3D (%)**
Bagging	56.44	57.33
Naive Bayes	30.66	32.44
J48	44.44	48.44
SMO	60.44	60.44

4.1.7 Kelimelerin Kökleri ile Üretilen Binary Matrisin Zenginleştirilmesi

Anksiyete, depresyon hastası ve sağlıklı denekler tarafından yazılan metin kısa olması nedeniyle kelimenin köklerinin metinde geçme durumuna göre üretilen binary matris çok büyük oranda sıfırlardan oluşmaktadır. Sıfırların oranını düşürmek amacıyla binary matris iki farklı yöntemde zenginleştirilmiştir. İki yöntemde de Word2Vec aracının bir kelimeye en yakın kelimeleri yakınlık oranıyla birlikte sıralama özelliği kullanılmıştır. İlk yöntemde binary matris oluşturulurken, metinde geçen bir kelimeyle birlikte bu kelimeye belirlenen sayıda en yakın olan kelimeler de metinde geçiyormuş gibi 1 olarak işaretlenmiştir. İkinci yöntemde ise, metinde geçen bir kelimeyle birlikte bu kelimeye belirlenen oranda yakın olan kelimeler de 1 olarak işaretlenmiştir. Dört farklı sınıflandırıcı ile testler tekrar yapılmış ve sonuçlar Çizelge 4.13 ve 4.14'de sunulmuştur.

Çizelge 4. 13 Hasta verileri kullanılarak en yakın 3 ve 5 kelime ile zenginleştirilmiş binary matris sonuçları

Sınıflandırıcı	**Binary (%)**	**Hasta datalarıyla (%)**		**Haber datalarıyla (%)**	
		En yakın 5	**En yakın 3**	**En yakın 5**	**En yakın 3**
Bagging	57.77	60.88	56.88	59.55	58.66
Naive Bayes	55.55	52.88	55.11	50.66	53.33
J48	49.33	42.22	51.55	52.44	48.00
SMO	51.55	50.22	49.33	58.66	52.44

Çizelge 4. 14 Haber verileri kullanılarak yakınlık oranı 0.5 ve 0.7 üzeri olan kelimeler ile zenginleştirilmiş binary matris sonuçları

Sınıflandırıcı	**Binary (%)**	**Yakınlık oranı 0.5 (%)**	**Yakınlık oranı 0.7 (%)**
Bagging	57.77	58.22	57.33
Naive Bayes	55.55	53.77	52.88
J48	49.33	51.55	51.11
SMO	51.55	52.88	50.22

4.1.8 Sonuçların Yorumlanması

225 deneğin yazısı üzerinde yapılan çalışmalarda gerek binary matris gerekse metin yörüngelerinin %60,4'ün üzerinde başarılı sonuçlar vermediği görülmüştür. Gerek binary matris ile yapılan gerekse, metin yörüngeleri ile yapılan testlerin sonuçları birbirine yakındır. Word2Vec ile kelimelerin yörüngeleri bulunurken, 225 deneğin yazılarının veya 47 bin haber metninin kullanılması arasında sonuçlarda ciddi bir farklılık olmasa da metin yörüngeleri çizdirildiğinde gözle görülür bir fark ortaya çıkmıştır. Hasta verileri, haber verilerine göre çok daha az olduğu için, kelimelerin koordinatlarını birbirine çok yakın olarak hesaplamaktadır. Metin yörüngelerindeki uzunluk ve açı değerleriyle çizdirilen histogramlara bakıldığında da sınıfların birbirinden ayırt edilemediği açıkca görülmektedir. Bu çalışmada, istenen oranda başarılı sonuçlar elde edilemediği için yapılan testlerin sonuçlarının detaylı incelemesi sunulmamaktadır.

4.2 Yazar Tanımada Köşe Yazısı ve Tweet'lerin Çapraz Kullanımı

Aday köşe yazarlarının makalelerini kullanarak bir Twitter hesabının yazarını tespit etmek üzere yapılan bu çalışmada iki farklı yöntem uygulanmıştır. Birinci yöntemde her bir tweet birer ayrı metin olarak değerlendirilmiş, ikinci yöntemde ise belirli sayıda tweet gruplandırılarak birleştirilmiş tweet metinleri oluşturulmuş ve bunlar sınıflandırılmada kullanılmıştır.

Bu bölümde yapılan çalışmaların tamamında, tweetlerde ve makalelerde geçen kelimelerin kendilerinin, Zemberek ile elde edilen köklerinin, 3-gramların metinde geçme duruma göre oluşturulan binary, TF, TF-IDF ve Normalized-TF (NTF) matrislerden elde edilen arff dosyaları dört farklı sınıflandırıcıda test edilmiştir. Kelimelerin kendilerinin kullanılarak üretilen arff 62742 özellik sayısına sahipken, bu özellik sayısı kelimelerin kökleri ile üretilen arff için 16255, 3-gramlarla üretilen arff için 18793'tür.

Üretilen arff dosyaları ile WEKA aracı kullanılarak dört farklı test uygulanmıştır:

1. Makaleler ile eğitim, tweetler ile test
2. Tweetler ile eğitim, makeleler ile test

3. Sadece makeleler ile 10 kez çapraz doğrulama

4. Sadece tweetler ile 10 kez çapraz doğrulama

4.2.1 Her Tweet Tekil Değerlendirilerek Yapılan Yazar Tespiti

Köşe yazarlarının Twitter hesaplarından toplanan 200'er tweeti ayrı birer metin olarak ele alınmıştır. Bu bölümde 10 yazar için 50'şer adet makale, 200'er adet tweet olmak üzere 2500 adetten oluşan bir veri kümesi kullanılmıştır.

Kelimelerin kendileri, kökleri ve 3-gramlar ile yapılan testlerin sonuçları Çizelge 4.15, 4.16, 4.17 ve 4.18'de sunulmuştur. Özellik ve/veya örnek sayısının çok fazla olması nedeniyle yeterli hafıza ayrılamadığı için, tekil tweetlerin eğitim olarak kullanıldığı Bagging ve J48 sınıflandırıcılarla yapılan testler sonuçlandırılamadığı için gösterilmemiştir.

Çizelge 4. 15 Makaleler eğitim, tekil tweetler test verisi olarak kullanıldığında sonuçlar

Terim	Sınıflandırıcı	Binary (%)	TF (%)	TF-IDF (%)	NTF (%)
Kelimenin kendisi	Bagging	11.05	11.30	11.00	12.80
	Naive Bayes	13.15	10.45	10.35	13.30
	J48	11.95	11.10	11.10	11.70
	SMO	12.00	10.05	10.05	31.15
Kök	Bagging	12.70	11.00	10.70	15.40
	Naive Bayes	13.65	10.25	10.25	14.30
	J48	12.45	12.60	12.80	12.55
	SMO	10.90	11.95	11.95	28.55
3-gram	Bagging	11.85	10.90	10.90	13.40
	Naive Bayes	10.20	10.25	10.25	11.20
	J48	9.85	11.15	11.15	14.00
	SMO	10.30	10.25	10.25	30.05

Çizelge 4. 16 Tekil tweetler eğitim, makaleler test verisi olarak kullanıldığında sonuçlar

Terim	Sınıflandırıcı	Binary (%)	TF (%)	TF-IDF (%)	NTF (%)
Kelimenin kendisi	Naive Bayes	25.6	17.0	17.0	15.0
	SMO	30.8	17.6	18.6	42.8
Kök	Naive Bayes	28.0	15.2	15.2	21.8
	SMO	32.2	12.0	13.4	42.8
3-gram	Naive Bayes	20.8	8.2	8.2	26.4
	SMO	22.6	26.6	26.6	51.6

Çizelge 4. 17 Sadece makaleler ile 10 kez çapraz doğrulama yapıldığında sonuçlar

Terim	Sınıflandırıcı	Binary (%)	TF (%)	TF-IDF (%)	NTF (%)
Kelimenin kendisi	Bagging	58.0	61.8	55.6	62.4
	Naive Bayes	91.8	75.6	76.0	55.2
	J48	50.2	54.6	52.8	50.8
	SMO	93.6	92.4	92.4	95.6
Kök	Bagging	51.8	61.2	55.4	62.6
	Naive Bayes	90.0	73.2	72.2	67.8
	J48	47.8	49.4	46.0	42.6
	SMO	91.8	92.4	93.0	93.6
3-gram	Bagging	67.6	72.8	72.0	71.6
	Naive Bayes	93.0	75.8	75.0	66.6
	J48	61.0	65.2	62.6	61.4
	SMO	96.2	95.0	95.0	97.8

Çizelge 4. 18 Sadece tekil tweetler ile 10 kez çapraz doğrulama yapıldığında sonuçlar

Terim	Sınıflandırıcı	Binary (%)	TF (%)	TF-IDF (%)	NTF (%)
Kelimenin kendisi	Naive Bayes	46.15	40.65	40.50	43.85
	SMO	45.05	51.70	52.00	53.10
Kök	Naive Bayes	42.90	40.25	40.15	44.15
	SMO	43.15	48.30	48.30	49.40
3-gram	Naive Bayes	49.60	41.50	41.45	44.65
	SMO	50.80	54.80	54.85	57.85

4.2.2 Birleştirilmiş Tweet Metinleri ile Yapılan Yazar Tespiti

Çalışmanın bu bölümünde, her tweet birer metin olarak değerlendirmek yerine, yazarların 200'er tweeti 40'ar adet olarak 5 eşit gruba bölünmüştür ve bu gruplardaki 40'ar tweet birleştirilerek bütün halinde tek bir metin olarak değerlendirilmiştir. Böylece her yazar için 5 adet birleştirilmiş tweet metni elde edilmiştir.

10 yazar için 50'şer makale ve 5'er birleştirilmiş tweet metni olmak üzere bu bölümde toplam 550 adetten oluşan bir veri kümesi üzerinde çalışılmıştır.

Kelimelerin kendileri, kökleri ve 3-gramlar ile yapılan testlerin sonuçları Çizelge 4.19, 4.20 ve 4.21'de sunulmuştur. Sadece makaleler ile 10 kez çapraz doğrulama sonucu bir önceki bölümdeki sonuçlarla aynı olduğu için bu bölümde tekrar gösterilmemiştir.

Çizelge 4. 19 Makaleler eğitim, birleştirilmiş tweetler test verisi olarak kullanıldığında sonuçlar

Terim	Sınıflandırıcı	Binary (%)	TF (%)	TF-IDF (%)	NTF (%)
Kelimenin kendisi	Bagging	46	42	48	42
	Naive Bayes	74	60	62	26
	J48	34	24	26	26
	SMO	88	88	86	84

Çizelge 4. 19 Makaleler eğitim, birleştirilmiş tweetler test verisi olarak kullanıldığında sonuçlar (devamı)

Kök	Bagging	48	34	42	48
	Naive Bayes	72	50	48	34
	J48	38	28	32	34
	SMO	80	82	82	82
3-gram	Bagging	18	20	22	26
	Naive Bayes	58	42	44	28
	J48	36	32	30	20
	SMO	74	62	62	76

Çizelge 4. 20 Birleştirilmiş tweetler eğitim, makaleler test verisi olarak kullanıldığında sonuçlar

Terim	**Sınıflandırıcı**	**Binary (%)**	**TF (%)**	**TF-IDF (%)**	**NTF (%)**
Kelimenin kendisi	Bagging	21.0	24.0	24.8	21.4
	Naive Bayes	52.8	32.6	30.4	19.4
	J48	18.2	18.0	18.0	18.6
	SMO	47.0	49.4	49.2	54.2
Kök	Bagging	18.4	20.0	18.4	21.2
	Naive Bayes	50.0	32.8	30.0	20.8
	J48	18.2	17.4	17.4	20.6
	SMO	46.0	46.8	44.8	53.0
3-gram	Bagging	17.8	24.8	20.8	24.0
	Naive Bayes	45.8	20.8	21.8	19.2
	J48	23.4	21.2	21.2	25.2
	SMO	39.0	40.6	40.4	46.6

Çizelge 4. 21 Sadece birleştirilmiş tweetler ile 10 kez çapraz doğrulama yapıldığında sonuçlar

Terim	Sınıflandırıcı	Binary (%)	TF (%)	TF-IDF (%)	NTF (%)
Kelimenin kendisi	Bagging	22	34	36	40
	Naive Bayes	86	54	52	16
	J48	44	30	30	42
	SMO	84	78	76	78
Kök	Bagging	24	32	24	46
	Naive Bayes	80	62	60	12
	J48	50	36	36	60
	SMO	84	88	86	86
3-gram	Bagging	26	28	24	20
	Naive Bayes	88	26	28	12
	J48	42	40	40	36
	SMO	94	92	92	96

4.2.3 Özellik Seçimi Yapılarak Testlerin Tekrarlanması

Özellik sayılarının çok fazla olması nedeniyle, özellik seçiminin testlerdeki başarı oranını yükseltip yükseltmeyeceğini görmek amacıyla WEKA aracı içerisinde yer alan ChiSquaredAttributeEval özellik seçme yöntemi kullanılarak tüm arff dosyaları için özellik sıralaması yapılmış ve ilk 1000 özellik seçilerek yeni arff dosyaları oluşturulmuş ve bu şekilde aynı testler tekrar yapılmıştır. Test sonuçları Çizelge 4.22'ten 4.29'a kadar sunulmuştur.

4.2.3.1 Tekil Tweetler Kullanılarak Yapılan Testlere Özellik Seçimi Uygulanması

Özellik seçimi yapıldıktan sonra tekil tweetler kullanılarak aynı testler tekrar edilmiş ve elde edilen sonuçlar Çizelge 4.22, 4.23, 4.24 ve 4.25'te sunulmuştur.

Çizelge 4. 22 Makaleler eğitim, tekil tweetler test verisi olarak kullanıldığında sonuçlar

Terim	Sınıflandırıcı	Binary (%)	TF (%)	TF-IDF (%)	NTF (%)
Kelimenin kendisi	Bagging	13.10	11.10	11.05	14.70
	Naive Bayes	14.55	10.10	10.10	14.20
	J48	12.20	11.95	11.95	12.10
	SMO	10.70	14.35	14.35	19.50
Kök	Bagging	13.30	11.00	11.00	14.70
	Naive Bayes	12.75	10.85	13.10	14.25
	J48	12.30	12.30	12.65	13.40
	SMO	10.80	10.35	11.40	18.15
3-gram	Bagging	10.70	8.95	8.95	14.85
	Naive Bayes	13.60	10.35	10.35	14.75
	J48	12.15	10.80	10.80	14.55
	SMO	12.15	10.25	10.25	24.65

Çizelge 4. 23 Tekil tweetler eğitim, makaleler test verisi olarak kullanıldığında sonuçlar

Terim	Sınıflandırıcı	Binary (%)	TF (%)	TF-IDF (%)	NTF (%)
Kelimenin kendisi	Bagging	17.8	16.8	18.8	11.4
	Naive Bayes	31.2	22.6	22.6	17.0
	J48	18.6	19.6	20.6	20.2
	SMO	39.8	10.8	11.2	29.4
Kök	Bagging	21.2	11.6	16.0	13.6
	Naive Bayes	35.4	19.8	20.2	18.4
	J48	19.2	15.6	15.6	17.6
	SMO	39.8	10.0	10.0	28.4

Çizelge 4. 23 Tekil tweetler eğitim, makaleler test verisi olarak kullanıldığında sonuçlar (devamı)

3-gram	Bagging	20.8	16.2	16.2	11.4
	Naive Bayes	39.0	19.6	19.6	22.4
	J48	25.2	18.2	18.2	18.8
	SMO	41.6	40.4	40.4	47.8

Çizelge 4. 24 Sadece makaleler ile 10 kez çapraz doğrulama yapıldığında sonuçlar

Terim	**Sınıflandırıcı**	**Binary (%)**	**TF (%)**	**TF-IDF (%)**	**NTF (%)**
Kelimenin kendisi	Bagging	61.6	63.4	64.6	66.8
	Naive Bayes	88.6	75.6	75.4	87.4
	J48	57.8	60.0	61.2	61.0
	SMO	86.8	85.6	85.2	89.4
Kök	Bagging	57.2	63.6	61.6	64.8
	Naive Bayes	84.4	69.8	69.8	82.0
	J48	54.4	53.6	54.6	53.0
	SMO	83.8	80.6	80.8	87.8
3-gram	Bagging	71.2	73.6	74.2	75.2
	Naive Bayes	93.0	85.2	85.0	93.6
	J48	70.0	70.8	70.8	67.0
	SMO	94.4	92.6	92.6	94.8

Çizelge 4. 25 Sadece tekil tweetler ile 10 kez çapraz doğrulama yapıldığında sonuçlar

Terim	Sınıflandırıcı	Binary (%)	TF (%)	TF-IDF (%)	NTF (%)
Kelimenin kendisi	Bagging	36.55	26.05	26.10	27.50
	Naive Bayes	36.35	37.85	37.65	42.05
	J48	32.30	29.20	28.70	33.65
	SMO	42.35	41.80	41.55	45.80
Kök	Bagging	32.40	25.95	27.05	26.70
	Naive Bayes	35.25	36.40	36.65	41.15
	J48	32.15	29.05	29.75	33.15
	SMO	40.25	43.00	42.05	45.40
3-gram	Bagging	34.35	30.20	30.15	32.40
	Naive Bayes	36.50	37.15	37.15	43.75
	J48	33.75	30.45	30.45	35.05
	SMO	40.55	40.85	40.80	48.30

4.2.3.2 Birleştirilmiş Tweetler Kullanılarak Yapılan Testlere Özellik Seçimi Uygulanması

Birleştirilmiş tweet metinleri kullanılarak üretilen arff dosyaları ile öncelikle özellik seçimi yapılmış ve testler tekrar edilmiştir.

Çizelge 4. 26 Makaleler eğitim, birleştirilmiş tweetler test verisi olarak kullanıldığında sonuçlar

Terim	Sınıflandırıcı	Binary (%)	TF (%)	TF-IDF (%)	NTF (%)
Kelimenin kendisi	Bagging	44	38	52	46
	Naive Bayes	84	60	54	70
	J48	42	36	46	34
	SMO	72	74	68	66

Çizelge 4. 26 Makaleler eğitim, birleştirilmiş tweetler test verisi olarak kullanıldığında sonuçlar (devamı)

Kök	Bagging	48	38	52	50
	Naive Bayes	64	64	58	70
	J48	38	30	34	50
	SMO	66	58	62	68
3-gram	Bagging	18	26	22	30
	Naive Bayes	38	30	34	34
	J48	28	36	20	14
	SMO	46	50	48	64

Çizelge 4. 27 Birleştirilmiş tweetler eğitim, makaleler test verisi olarak kullanıldığında sonuçlar

Terim	**Sınıflandırıcı**	**Binary (%)**	**TF (%)**	**TF-IDF (%)**	**NTF (%)**
Kelimenin kendisi	Bagging	27.6	25.8	24.2	26.2
	Naive Bayes	54.6	43.4	46.0	46.2
	J48	19.8	20.6	22.2	22.6
	SMO	50.8	46.6	48.4	52.8
Kök	Bagging	28.2	23.0	22.8	27.8
	Naive Bayes	47.4	39.2	41.4	46.8
	J48	21.6	16.2	23.6	26.0
	SMO	41.4	46.6	49.0	53.0
3-gram	Bagging	30.0	33.0	31.0	31.0
	Naive Bayes	36.8	32.8	30.4	36.8
	J48	27.2	27.6	27.6	27.6
	SMO	33.6	37.2	37.8	47.4

Çizelge 4. 28 Sadece makaleler ile 10 kez çapraz doğrulama yapıldığında sonuçlar

Terim	Sınıflandırıcı	Binary (%)	TF (%)	TF-IDF (%)	NTF (%)
Kelimenin kendisi	Bagging	61.6	63.2	59.6	67.8
	Naive Bayes	89.4	76.2	78.0	87.4
	J48	59.0	59.8	61.4	59.8
	SMO	88.2	87.6	88.0	90.6
Kök	Bagging	58.0	61.2	57.2	67.4
	Naive Bayes	87.2	69.8	68.2	82.6
	J48	53.8	57.2	55.6	54.2
	SMO	83.8	84.2	83.6	88.8
3-gram	Bagging	72.6	76.8	76.2	74.2
	Naive Bayes	93.4	85.6	85.0	94.0
	J48	66.8	72.2	67.6	67.2
	SMO	95.2	94.4	93.8	97.4

Çizelge 4. 29 Sadece birleştirilmiş tweetler ile 10 kez çapraz doğrulama sonuçları

Terim	Sınıflandırıcı	Binary (%)	TF (%)	TF-IDF (%)	NTF (%)
Kelimenin kendisi	Bagging	38	36	48	52
	Naive Bayes	84	68	74	68
	J48	52	56	58	44
	SMO	94	90	86	82
Kök	Bagging	38	46	28	40
	Naive Bayes	82	74	70	56
	J48	70	56	46	44
	SMO	84	84	78	88

Çizelge 4. 29 Sadece birleştirilmiş tweetler ile 10 kez çapraz doğrulama sonuçları (devamı)

3-gram	Bagging	48	48	50	52
	Naive Bayes	94	84	82	82
	J48	70	58	58	58
	SMO	94	100	98	96

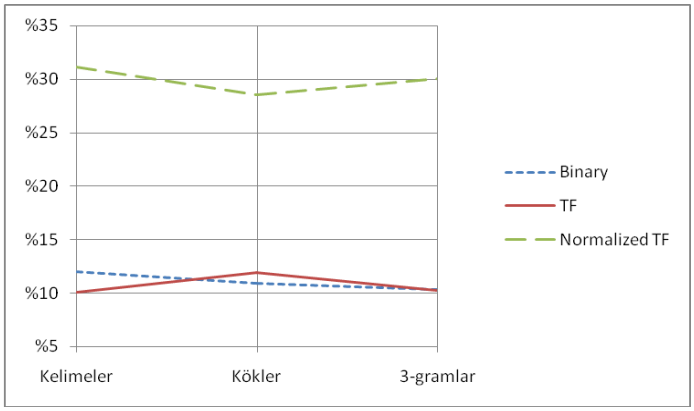
4.2.4 Sonuçların Yorumlanması

Yapılan tüm deneylerin hemen hemen hepsinde SMO sınıflandırıcı NB, J48 ve Bagging sınıflandırıcılarına göre, büyük çoğunluğunda ise NB sınıflandırıcı J48 ve Bagging sınıflandırıcılarına göre daha başarılı sonuçlar verirken J48 ile Bagging sınıflandırıcıları arasında net bir üstünlük görülmemektedir. En başarılı sonuçlar SMO sınıflandırıcı ile alındığı için sonuçların görselleştirilmesinde sadece bu sınıflandırıcıyla elde edilen değerler sunulmaktadır.

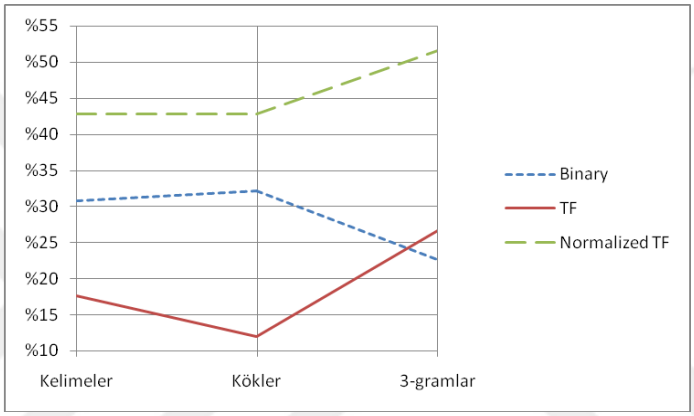
Yapılan tüm deneylerin genelinde TF ve TF-IDF ile yapılan deney sonuçlarının birbirine çok yakın olması nedeniyle de şekillerde sadece TF-IDF sonucu gösterilmemiştir.

Özellik seçimi ile yapılan testlerde başarı oranları hemen hemen hepsinde düşmüş, sadece birleştirilmiş tweetler ile 10 kez çapraz doğrulama yapıldığında daha yüksek başarı elde edilmiştir. Bu yüzden özellik seçimi ile yapılan testlerden sadece bu testin sonucu görsellerde karşılaştırılmaktadır. Şekillerde, tüm özellikler ile elde edilen sonuçlar Kelimeler, Kökler ve 3-gramlar şeklinde, özellik seçimi ile gerçekleştirilen testlerin sonuçları ise Kelimeler', Kökler' ve 3-gramlar' şeklinde ifade edilmiştir.

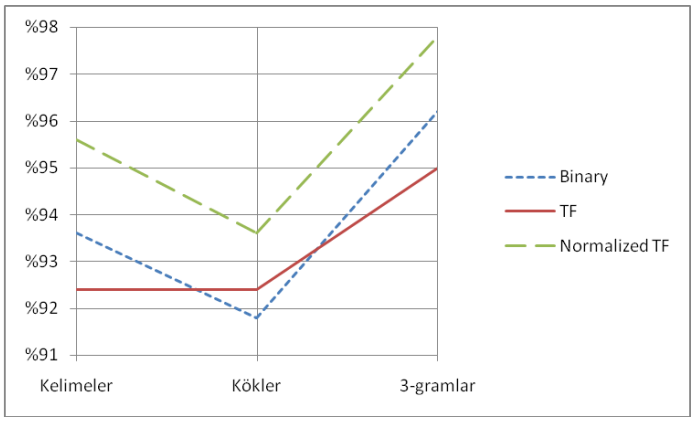
Her bir tweet birer metin olarak değerlendirildiğinde tüm özellikler kullanılarak ile elde edilen sonuçlar Şekil 4.13, 4.14, 4.15 ve 4.16'da görülmektedir.



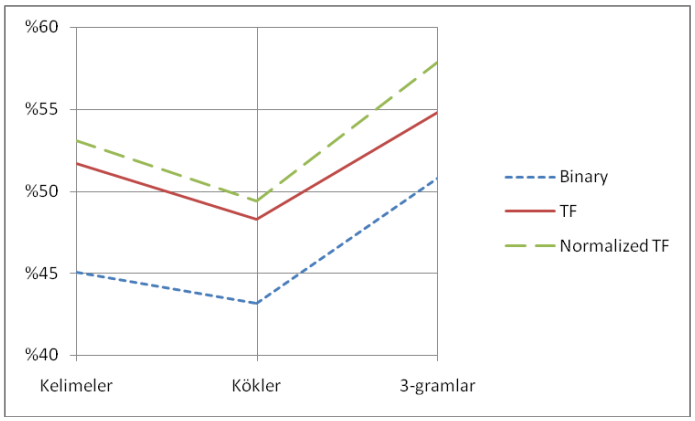
Şekil 4. 13 Makaleler eğitim, tekil tweetler test verisi olarak kullanıldığında elde edilen başarı oranları



Şekil 4. 14 Tekil tweetler eğitim, Makaleler test verisi olarak kullanıldığında elde edilen başarı oranları



Şekil 4. 15 Sadece makaleler ile 10 kez çapraz doğrulama yapıldığında elde edilen başarı oranları



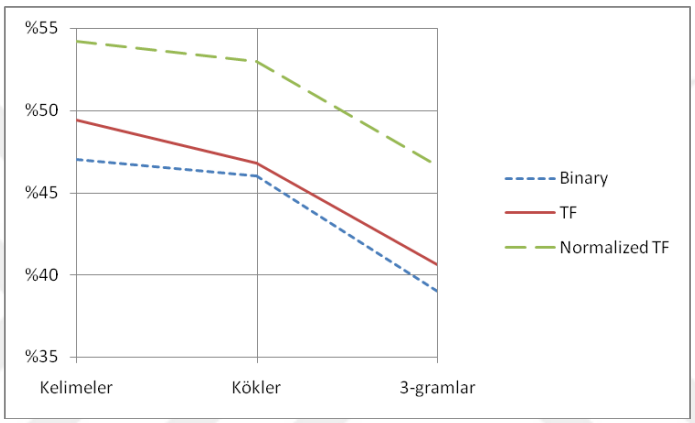
Şekil 4. 16 Sadece tekil tweetler ile 10 kez çapraz doğrulama yapıldığında elde edilen başarı oranları

Tekil tweetler ile yapılan bu bölümdeki testlerde en başarılı sonuçların Normalized-TF ile alındığı açıkça görülmektedir. Şekil 4.15'te görüldüğü üzere makaleler kendi arasında 10 kez çapraz doğrulama yapıldığında 3-gramlar ile %97.8 gibi çok yüksek bir başarı oranı elde edilmiştir. Bu sonuç uzun metinler kullanıldığında yazar tanıma testlerinde yüksek başarı elde edildiğini doğrulamaktadır. Şekil 4.16'da görüldüğü gibi, en fazla 160 karakterden oluşan tekil tweetler kendi arasında 10 kez çapraz doğrulama yapıldığında yine 3-gramlar ile %57.85'lik bir başarı oranı ortaya çıkmıştır. Şekil 4.14'de, tekil tweetler eğitim, makaleler test verisi olarak kullanıldığında 3-gramlar ile %51.6'lık bir başarı oranı yakalandığı görülürken, tam tersi şekilde makaleler eğitim, tekil tweetler test verisi olarak kullanıldığında ise en yüksek başarı oranının kelimelerin kendisi ile %31.15 olarak ortaya çıktığı Şekil 4.13'te görülmektedir. Bu sonuçlara göre, adayların makalelerinin tekil tweetlerin yazarını tahmin etmek için kullanılması çok başarılı netice vermemektedir.

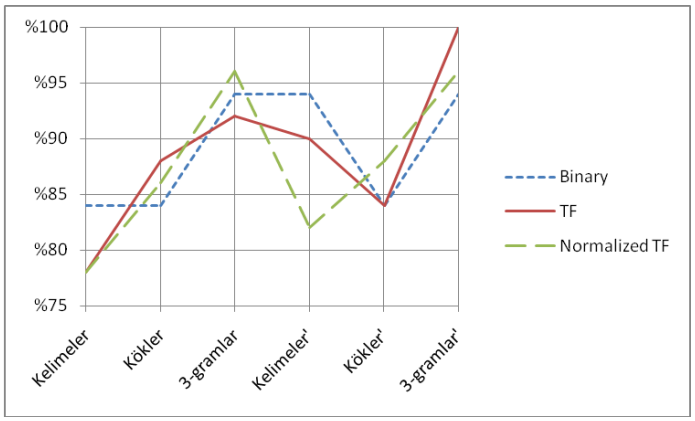
Birleştirilmiş tweetler kullanılarak yapılan testlerde elde edilen sonuçlar Şekil 4.17, 4.18 ve 4.19'da görülmektedir. Makaleler için 10 kez çapraz doğrulama Şekil 4.15'te verildiği için tekrar sunulmamıştır.



Şekil 4. 17 Makaleler eğitim, birleştirilmiş tweetler test verisi olarak kullanıldığında elde edilen başarı oranları



Şekil 4. 18 Birleştirilmiş tweetler eğitim, Makaleler test verisi olarak kullanıldığında elde edilen başarı oranları



Şekil 4. 19 Sadece birleştirilmiş tweetler ile 10 kez çapraz doğrulama yapıldığında elde edilen başarı oranları

Birleştirilmiş tweetlerle bu dölümdeki yapılan testlerde, aday yazarların makaleleri eğitim, birleştirilmiş tweetler test verisi olarak kullanıldığında bu tweetlerin paylaşıldığı Twitter hesabının yazarını %88 gibi yüksek bir oranda başarılı tahmin edebildiğimiz Şekil 4.17'de görülmektedir. Bu başarı oranı, makalelerde ve tweetlerde geçen

kelimelerin kendilerini kullanarak oluşturulan binary ve TF arff dosyalarının SMO sınıflandırıcı ile test edilmesiyle ortaya çıkmıştır.

Şekil 4.19'daki sonuçlardan anlaşıldığı gibi birleştirilmiş tweetler kendi arasında 10 kez çapraz doğrulamayla test edildiğinde gayet yüksek başarı oranları yakalanmış ve bu sonuç yine, uzun metinlerle yazar tanıma probleminin yüksek başarıyla çözüldüğünü bize kanıtlamıştır. Bu testte, özellik seçimi yapılmış 3-gramlar ile %100 gibi bir başarı oranı elde edilmesinin sebebi kullanılan verinin örnek sayısının düşük olmasıdır.

Şekil 4.18'de başarı oranlarının çok yüksek çıkmamasının sebebi de her yazar için 5'er birleştirmiş tweet metninin eğitim, 50'şer makalenin test olarak kullanılması dolayısıyla eğitim kümesinin çok küçük olmasıdır. Eğitim verisinin örnek sayısı artırıldığında bu testte de yüksek başarı elde edilmesi beklenebilir.

Ayrıca, sınıflandırma testleri sonrasında ortaya çıkan karışıklık matrisleri incelendiğinde, yazarların makaleleri yazdığı tarih aralığı ile Twitter hesaplarından yaptıkları paylaşımların tarih aralığının örtüşmesinin veya örtüşmemesinin sınıflandırma başarısında bir etkisi olmadığı görülmüştür.

Makaleler ile birleştirilmiş tweet metinleri eğitim ve test kümeleri olarak birbiriyle çaprazlama yapılan testlerde kelimelerin kendileri en yüksek sonuçları verirken, sadece birleştirilmiş tweetler ve sadece makaleler ile yapılan testlerde ise 3-gramlar daha yüksek sonuç vermiştir. Buradan, iki ayrı platformda kullanılan kendine has karakter öbekleri olduğu çıkarımı yapılabilir. Hem eğitim hem de test kümeleri aynı platformdan olduğu zaman bu karakter öbekleri yazar tanımada ayırt edici olurken, eğitim ve test kümeleri farklı platformlardan olduğu zaman ise kullanılan kelimelerin kendisi yazar tanımada daha ayırt edici nitelikte olmaktadır.

BÖLÜM 5

TARTIŞMA VE ÖNERİLER

Bu tezde yapılan iki farklı çalışmanın sonuçlarına dair tartışmalar ve öneriler alt başlıklarda sunulmaktadır.

5.1 Metinden Psikolojik Rahatsızlık Teşhisi

Çalışmada kullanılan metinlerin kısa olması ve sayılarının az olması önemli bir kısıttır. Daha uzun metinler üzerinde çalışılabilirse, daha başarılı sonuçlar alınması ihtimaldir. Bu çalışma anksiyete, depresyon hastaları ve sağlıklı denekleri birbirinden ayırt etmeye yönelik yapıldığı için, anksiyete ve depresyon rahatsızlığının hastaların yazma niteliklerine etki etmediği çıkarımı da yapılabilir. Zira, bu hastalıklar en yaygın psikolojik rahatsızlıklar arasında yer almaktadır [15]. Şizofreni gibi daha nadir görülen ağır hastalıkların kişinin yazı yazma yetisine etki etmesi daha muhtemeldir. Daha büyük bir veri kümesi üzerinde ve/veya farklı hastalık grubundan deneklerin yazılarıyla benzer bir çalışma yapılabilir.

5.2 Yazar Tanımada Köşe Yazısı ve Tweet'lerin Çapraz Kullanımı

Bu tez çalışmasında yapılan tüm testlere ve elde edilen sonuçlara göre birleştirilmiş bir tweet grubunun yazarını tespit etmenin tekil bir tweetin kim tarafından yazıldığını tespit etmekten çok daha kolay olduğu ortaya çıkarılmıştır. Kimliği belirli olmayan bir Twitter kullanıcısının paylaştığı tweetlerin her birinin birer metin olarak değerlendirmesinden ise belirli sayıda tweeti birleştirilerek bir bütün olarak ele alınmasıyla çok daha uzun tweet metinleri oluşturulabilir ve aday köşe yazarlarının

makaleleri ile karşılaştırılarak yüksek bir başarı oranıyla yazarın kimliğine dair tahmin yapılabilir. Ayrıca, bu şekilde yapılan testlerde birleştirilmiş tweetlerin kendi aralarında yapılan çapraz doğrulamalarda da çok başarılı sonuç vermesi, bir Twitter hesabının el değiştirip değiştirmediğinin anlaşılabileceğini de açıkça göstermektedir.

KAYNAKLAR

[1] Aslan, B., (2007). " Web 2.0, Teknikleri ve Uygulamaları", XII. "Türkiye'de İnternet" Konferansı, 8-10 Kasım 2007, Ankara.

[2] Najib, F., Cheema, W. A., and Nawab, R. M. A., (2015). "Author's Traits Prediction on Twitter Data using Content Based Approach", CLEF 2015, 8-11 September 2015, Toulouse.

[3] Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M., (2010). "Classifying Latent User Attributes in Twitter", 2nd International Workshop on Search and Mining User-generated Contents (SMUC), 26 - 30 October 2010, Toronto.

[4] Román, J. V., and Cristóbal, J. C. G., (2014). "DAEDALUS at PAN 2014: Guessing Tweet Author's Gender and Age", CLEF 2014, 15-18 September 2014, Sheffield.

[5] Amaysalı, M. F., Diri, B. ve Türkoğlu, F., (2006). "Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi", The Fifteenth Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN'2006), 21-24 June 2006, Muğla.

[6] Schwartz, R., Tsur, O., Rappoport, A., and Koppel, M., (2013). "Authorship Attribution of Micro-Messages", Conference on Empirical Methods in Natural Language Processing, 18-21 October 2013, Washington.

[7] Green, R. M., and Sheppard, J. W., (2013). "Comparing Frequency- and Style-Based Features for Twitter Author Identification", The Twenty-Sixth International Florida Artificial Intelligence Research Society Conference, 22–24 May 2013, Florida.

[8] Karmena, C., Hsiungb, R. C., and Wettera, T., (2015). "Screening Internet forum participants fordepression symptoms by assembling andenhancing multiple NLP methods", Computer Methods and Programs in Biomedicine, 120 (1): 27-36.

[9] Orhan, Z., and Mercan, M., (2014). "Turkish Text Analysis System for Automatic Detection of Psychiatric Disorders", The International Conference on Turkic Language Processing (TURKLANG), 6-7 November 2014, İstanbul.

[10] Amasyalı, M. F., Balcı, S., Varlı, E. N. ve Mete, E., (). "Türkçe Metinlerin Sınıflandırılmasında Metin Temsil Yöntemlerinin Performans Karşılaştırılması", EMO Bilimsel Dergi, 2 (4).

[11] Akın, M. D. ve Akın, A. A., "Türk Dilleri İçin Açık Kaynaklı Doğal Dil İşleme Kütüphanesi", Electricity Engineering – Elektrik Mühendisliği, 431: 38-44.

[12] Vikipedi, Zemberek, https://tr.wikipedia.org/wiki/Zemberek_(yazılım), 1 Mart 2016.

[13] Google Code Archive, word2vec, https://code.google.com/archive/p/word2vec/, 1 Mart 2016.

[14] Amasyalı, M. F., Yener, Y. ve Kaplan, S. S., (2012). "Kelime Yörüngeleri", 20. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU 2012), 18-20 Nisan 2012, Antalya.

[15] Psikoterapi, Psikolojik hastalıklar nelerdir? Psikolojik hastalıklar listesi, http://psikoterapi.info.tr/psikolojik-hastaliklar/, 1 Mart 2016.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : İslam MAYDA

Doğum Tarihi ve Yeri : 20.02.1990 Sivas

Yabancı Dili : İngilizce

E-posta : islam.mayda@stu.khas.edu.tr

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Y. Lisans	Bilgisayar Mühendisliği	Yıldız Teknik Üniversitesi	2016
Lisans	Bilgisayar Mühendisliği	Kadir Has Üniversitesi	2012
Lise	Fen Bilimleri	Kadriye Moroğlu (Y.D.A) Lisesi	2007

İŞ TECRÜBESİ

Yıl	Firma/Kurum	Görevi
2016	İstanbul Kemerburgaz Üniversitesi	Araştırma Görevlisi
2015	İstanbul Sabahattin Zaim Üniversitesi	Araştırma Görevlisi
2014	Fatih Sultan Mehmet Vakıf Üniversitesi	Araştırma Görevlisi

YAYINLARI

Makale

1. Aytekin, Ç. ve Mayda, İ., (2013). “Competitive Analysis on Social Media Comparison Task Oriented Opinion Mining Model”, Journal Academic Marketing Mysticism Online (JAMMO), 7(2): 414-425.

ÖDÜLLERİ

1. Bilgisayar Mühendisliği Bölümü Birinciliği (2012)
2. Mühendislik ve Doğa Bilimleri Fakültesi İkinciliği (2012)
3. Kadir Has Üniversitesi Üçüncülüğü (2012)
4. Onur Ödülü (2009, 2010, 2011, 2012)
5. %100 ÖSYM Bursu (2007)