KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

FACEBOOK'DA YORUM MADENCİLİĞİ KULLANARAK KİŞİLERİN CİNSİYET, YAŞ VE EĞİTİM DÜZEYLERİNİN TANIMLANMASI

YÜKSEK LİSANS TEZİ

Bilgisayar Müh. Masoud TALEBİ

HAZİRAN 2013
TRABZON

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

FACEBOOK'DA YORUM MADENCİLİĞİ KULLANARAK KİŞİLERİN CİNSİYET, YAŞ VE EĞİTİM DÜZEYLERİNİN TANIMLANMASI

Bilgisayar Müh. Masoud TALEBİ

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsünce
"BİLGİSAYAR YÜKSEK MÜHENDİSİ"
Unvanı Verilmesi İçin Kabul Edilen Tezdir.

Tezin Enstitüye Verildiği Tarih : 21/05/2013
Tezin Savunma Tarihi : 10/06/2013

Tez Danışmanı : Doç. Dr. Cemal KÖSE

Trabzon 2013

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalında
Masoud TALEBİ tarafından hazırlanan

FACEBOOK'DA YORUM MADENCİLİĞİ KULLANARAK KİŞİLERİN CİNSİYET, YAŞ VE EĞİTİM DÜZEYLERİNİN TANIMLANMASI

başlıklı bu çalışma, Enstitü Yönetim Kurulunun 21/05/2013 gün ve 1506 sayılı kararıyla oluşturulan jüri tarafından yapılan sınavda

YÜKSEK LİSANS TEZİ

olarak kabul edilmiştir.

Jüri Üyeleri

Başkan : Prof. Dr. İsmail HAKKI ÇAVDAR ----------

Üye : Doç. Dr. Cemal KÖSE ----------

Üye : Yrd. Doç. Dr. Bekir DİZDAROĞLU ----------

Prof. Dr. Sadettin KORKMAZ

Enstitü Müdürü

ÖNSÖZ

Sosyal medyanın, hayatımızın vazgeçilmez bir parçası olduğu gerçeği herkesçe bilinmektedir. Gerçek dünyada olduğu gibi bu sanal dünyada da sahtekarlar ve dolandırıcılar bulunmaktadır. Bunların birçoğu gerçek olmayan kimliklerle internet ortamında ve sosyal ağlarda dolaşmaktadır. Sahtekarların internet ortamındaki gerçek kimliklerinin tespit edilmesiyle, e-suç işlenmesi büyük ölçüde önlenebilir. Bunun için en önemli adımlardan biri, bu kişilerin gerçek cinsiyet, yaş ve eğitim düzeyini belirleyebilmektir.

Böylesine güncel ve önemli bir konuyu seçmemde bana destek olan ve tüm çalışma boyunca desteklerini esirgemeyen saygıdeğer hocam Doç.Dr.Cemal Köse'ye çok teşekkür eder, şükranlarımı sunarım. Bu çalışmada yardımı geçen herkese özellikle Elif Üçüncü, Nimet Özdemir, Canan kepenek, Zeynep Köroğlu, Semra Karakullukçu ve Filiz Demir'e çok teşekkür ederim.

Ayrıca hayatım boyunca her an yanımda olan sevgili anneme ve saygıdeğer babama ve son olarak, her zaman pozitif enerjisiyle beni destekleyip arkamda duran ağabeyim Saeed Talebi'ye minnetlerimi sunar teşekkürü bir borç bilirim.

Masoud TALEBİ
Trabzon 2013

TEZ BEYANNAMESİ

Yüksek Lisans Tezi olarak sunduğum "Facebook'da Yorum Madenciliği Kullanarak Kişilerin Cinsiyet, Yaş ve Eğitim Düzeylerinin Tanımlanması" başlıklı bu çalışmayı baştan sona kadar danışmanım Doç. Dr. Cemal KÖSE'nin sorumluluğunda tamamladığımı, verileri/örnekleri kendim topladığımı, deneyleri/analizleri ilgili laboratuvarlarda yaptığımı/yaptırdığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim. 10/06/2013

Masoud TALEBİ

İÇİNDEKİLER

SayfaNo

Yüksek Lisans

ÖZET

FACEBOOK'DA YORUM MADENCİLİĞİ KULLANARAK KİŞİLERİN CİNSİYET, YAŞ VE EĞİTİM DÜZEYLERİNİN TANIMLANMASI

Masoud TALEBİ

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Doç. Dr. Cemal Köse
2013, 79 Sayfa, 4 Ek Sayfa

Sosyal medyanın günden güne toplumun farklı tabakalarına yayılmasıyla beraber, yararlı işlevlerinin yanında, bireysel hak ihlallerine yol açan kötüye kullanımları da artmaktadır. Bazı kişiler gerçek olmayan bilgilerle sahte hesaplar açarak, insanları hem maddi hem de manevi yönde istismar etmektedir. Bazen de sahtekarlar, daha küçük yaşta olduklarını belirterek çocuk istismarını hedeflemektedirler ya da kendilerini karşı cins veya olduğundan daha yüksek eğitim seviyesinde göstererek insanları aldatmaktadırlar. Sosyal medyadaki insanların gerçek kimliklerini tespit etmek, e-suç oluşumunun engellenmesinde büyük bir etki yaratabilir. Bu çalışmada Facebook kullanıcılarının yorumlarını analiz ederek, Cinsiyet, Yaş ve Eğitim düzeyini belirlemek için bir sistem geliştirilmiştir. Bu çalışmada kullanılan veri kümesi, Türkçe Facebook sayfalarından toplanan yorumlardan oluşturulmuştur. Genel metin madenciliği işlemleri uygulandıktan sonra "Significance Analysis of Microarrays" (SAM) yöntemi, nitelik ağırlıklandırmak ve boyut azaltma işlemi için kullanılmıştır. Kişileri sınıflandırmak için Naive Bayes, "Support Vector Machine" (SVM) ve K-en yakın komşuluk (KNN) sınıflandırma yöntemleri kullanılmış ve elde edilen sonuçlar karşılaştırılmıştır. Ayrıca terim vektörü için farklı ağırlıklandırma yöntemleri karşılaştırılmıştır.

Anahtar Kelimeler: Sosyal Medya, Facebook, Cinsiyet, Yaş, Eğitim Düzeyi, Metin madenciliği, SAM.

Master Thesis

SUMMERY

COMMENT MINING FOR IDENTIFYING GENDER, AGE AND EDUCATION LEVEL ON FACEBOOK

Masoud TALEBİ

Karadeniz Technical University
The Graduate School of Science
Computer Engineering Graduate Program
Supervisor: Assoc. Prof. Dr. Cemal KÖSE
2013, 79 Pages, 4 Pages Appendix

As the use of social media rises in different layers of the society, alongside of useful applications, some abuses of these media are also rising. The people who create fake accounts with a profile details differ from what they really are and pretending opposite gender, lower age, or higher education level for cheating people or hiding real identity are some of the abuses of the social media. Identifying real identity of a person in social media can be vital for preventing crimes.

In this study, a system implemented to identify Gender, Age and Education level of a Facebook user by analyzing the comments he/she wrote on different shares of the Facebook pages in Turkish. The data used in this study is collected from Facebook social media. After applying common text mining operations on data, Significant Analysis of Microarrays (SAM) method is employed for weighting and dimension reduction of the data. The Naïve Bayesian, Support vector machine (SVM) and K-Nearest Neighbors (KNN) classification methods used for classifying. In addition, the effect of using different term weights in vector space model investigated.

Key Words: Social Media, Facebook, Gender, Age, Education level, Text mining, SAM.

ŞEKİLLER DİZİNİ

Sayfa No

TABLOLAR DİZİNİ

Sayfa No

SEMBOLLER DİZİNİ

BE : Bilgi Erişim

BKK : Büyük Karakter Kullanımı

CSV : Comma Seperated Vector

DDS : Devrik Döküman Sıklığı

DN : Doğru Negatif

DP : Doğru Pozitif

DVM : Destek Vektör Makineleri

IDF : Inverse Document Frequency

KEYK : K-En Yakın Komşuluk

KNN : K-Nearest Neighbor

OYU : Ortalama Terim Uzunluğu

SAM : Significance Analysis of Microarrays

SAS : Sosyal Ağ Siteleri

SVM : Support Vector Machine

TF : Term Frequency

TS : Terim Sıklığı

YN : Yanlış Negatif

YP : Yanlış Pozitif

1. GİRİŞ

Günümüzde sosyal medya hayatımızın ayrılmaz bir parçası olmuştur, herkes en önemli anılarını, duygularını, resimlerini ve videolarını bu medyalarda paylaşarak dünyaya yaymaktadır. Bunun dışında sosyal medya, pazarlama ve ürün tanıtımı için de yeni bir ortam oluşturmuştur ve artık büyük üreticiler ve firmalar, müşteri ilişkileri yönetimlerini ağırlıklı olarak sosyal medyalardan yönetmektedirler. Birçok kurum ve kuruluş duyurularını, reklamlarını ve satışlarını sosyal medyadan hizmet vererek gerçekleştirmektedirler. Alexa'ya göre dünyanın en çok kullanılan ikinci sitesi Facebook sosyal medyasıdır [1] ve bu, sosyal medyanın ne kadar çok kullanıldığının bir kanıtıdır. Sosyal medya kullanımının artışıyla birlikte büyük bir veri yığınağı oluşmaktadır. Bu da araştırmacıların birçok yönden ilgisini çekmektedir, bu yüzden geçmiş yıllarda sosyal medya üzerine birçok araştırma yapılmıştır. Tüm bunlara rağmen sosyal medya aynı zamanda sahtekarlık ve e-suç için de yeni bir ortam yaratmıştır. Sahtekarlıkların bir kısmını şöyle sıralayabiliriz: ünlülerin resmini ve bilgilerini kullanarak sahte hesaplar oluşturmak, çocuk istismarı amaçlayarak kendilerini küçük yaşta göstermek, maddi ve manevi karlar hedefleyerek kendilerini gerçekte olmadıkları cinsiyette göstermek vb. Öyle ki İnternet Suçları Şikayet Merkezi, 2011 yılı raporunda gelen şikayet sayısını 314000 civarında bildirmiştir [2].

Sahtekarlar yukarıda belirtilen suçlara karışırken genelde sahte hesaplar ve bazen daha da ileri giderek, internetteki adreslerini (IP) gizlemek için tanımsız sunucular kullanmaktadırlar [3]. Bilindiği üzere genel olarak sosyal medyalara kayıt yapılırken, kişiler gerçek kimliklerini belirtmek zorunda değiller ve herhangi bir ad, yaş ve resmi kendi profil bilgileri olarak seçebilirler [4].

Kullanıcıların verdikleri bilgilere güvenerek bir profilin sahte olup olmadığını belirleyemeyeceğimize göre, onların yazdıklarını çözümlememiz gerekmektedir. Ancak Facebook gibi sosyal medyalarda kullanıcıların aralarındaki özel yazışmalara erişim imkanımız olmadığından, kişilerin sadece farklı sayfalardaki paylaşımlara yaptıkları yorumları incelemeye alabiliriz. Zaten sahtekarlar, kurbanlarını genel olarak umumi olarak kullanılan sayfalardan seçerler. Bunun nedenini şu şekilde açıklayabiliriz; Facebook'ta

kullanıcılar, sayfalardaki paylaşımların yorumlar kısmını sohbet amaçlı kullanmakta ve orada biraz yazıştıktan sonra özel olarak görüşüp, arkadaşlıklarını ilerletmektedirler.

Bu çalışmada amaç; kişinin yorumlarının analiziyle cinsiyeti, yaşı ve eğitim seviyesi hakkında tahminde bulunmaktır. Bu amaca ulaşmak için ise aşağıdaki sorulara cevap aranmaktadır:

1- Bir kişinin sadece yorumlarının incelenmesi, o kişinin cinsiyeti, yaşı ve eğitim düzeyi hakkında ne derece doğru bilgi sunabilir?
2- Hangi sınıflandırma yöntemleri büyük verilerde seri ve doğru sonuçlar üretebilir?
3- Kelime dizilerinin hangi yöntemle ağırlıklandırılması en iyi sonuçları üretmektedir?

2. GENEL BİLGİLER

2.1. Sosyal Medya

SAS (Sosyal Ağ Siteleri) hayatımıza girdiği günden beri milyonlarca insanın ilgisini çekmiştir. Bazı insanlar için bu siteler hayatlarının bir parçası haline gelmiştir. Öyle ki bu siteleri her gün en az bir kere ziyaret etmektedirler. SAS'ların kuruluş sebepleri ve insanları birbiriyle tanıştırma politikaları açısından farklılıklar göstermektedir. Bazı SAS'lar gerçek hayatta zaten mevcut olan ağları canlı tutmayı, bazılarıysa ilgi alanları veya politik görüşleri aynı olan ya da aynı aktiviteleri seven yabancı bireyleri tanıştırmayı hedeflemektedirler. Bazı siteler değişik kitlelerden insanları hedeflerken, diğerleriyse aynı dili kullanan, aynı ırk, cinsiyet, dine sahip veya aynı milliyetten bireyleri hedeflemektedir [5].

Ellison [5] SAS'ları web tabanlı servisler olarak tanımlamıştır. Bu servisler farklı kullanıcılara aşağıdaki imkanları ve daha fazlasını sunmaktadır:

1) Umumi veya yari-umumi bir profil oluşturmak.
2) Farklı kullanıcılarla bağlanmak.
3) Bağlandığı kişilerin bağlantılarını görmek veya o kişilerle bir şeyler paylaşmak.

Kullanıcılar sosyal medyaya katıldıktan sonra önlerine çıkan formları doldurmaktadırlar. Bu formlarda, kişinin bazı sorulara cevap vermesi istenmekte ve bu sorulara verilen cevaplar kullanılarak kişinin profili oluşturulmaktadır. Kişilerin profil bilgilerinin gizlilik politikası, SAS'lara ve kullanıcıların kendi tercihlerine bağlıdır. SAS'ları birbirinden ayrı kılan en önemli etkenler, gizlilik ve erişim politikalarıdır.

Kullanıcılar SAS'larda kayıtlarının tamamlanmasıyla birlikte çevrelerini oluşturmaya ve ilişkilerini kurmaya davet edilmektedirler, bu çevre veya bağlantılar farklı SAS'larda farklı isimlerle adlandırılır. Bu isimlerden en ünlüleri "Arkadaşlar", "Bağlantılar", "Takipçiler" ve "Taraftarlar"dır. SAS'ların bir çoğunda bağlantı kurmak için iki tarafın da onayı gerekmektedir.

İnsanların SAS'lara kayıt olma nedenleri farklıdır; bazıları eski arkadaşlarıyla tekrar bir araya gelmeyi hedeflerken bazıları da farklı kültürden insanlarla tanışarak çevrelerini

genişletmeyi hedeflemektedir. Bazı insanlar kendisiyle aynı düşünce tarzına ya da aynı diplomatik fikirlere sahip kişilerle tanışmayı düşünerek SAS'lara üye olmaktadırlar.

SAS'ların kullanıcılarına sunduğu başka imkanlar da vardır, bunlar resim ve video paylaşımı, özel ve grup olarak mesajlaşma ve sohbete katılma, paylaşılanları beğenme veya paylaşım hakkında yorum yazma gibi imkanlardır.

2.1.1. Facebook

Facebook orjinal olarak birkaç kolej öğrencisinin birbirine bağlanabilmesi için tasarlanmış, ilk olarak 2004 yılında ve sadece Harvard'da kullanıma açılmıştır. Bu SAS'a üye olabilmek için kullanıcının "*harvard.edu*" uzantılı e-posta adresine sahip olması gerekmekteydi. Facebook daha sonra diğer okulları da destekleyerek, kullanıcıların bir üniversite uzantılı e-posta ile üye olabilmesine olanak sunmuştur. Böylece Facebook, üniversiteliler arasında kalmaya devam etmiştir [5].

Diğer SAS'ların aksine, Facebook kullanıcılarının profil bilgileri, başlangıçta umuma açık değildi ve bu sitenin farklı kılınmasını sağlayan bir başka özellik ise kullanıcılara uygulama geliştirme yetkisi vermesidir.

Yukarıda bahsedilen tüm özelliklerin yanısıra, araştırmacılar açısından SAS'lar, aynı zamanda çok zengin bir veri kaynağıdır. Birçok araştırmada kullanılan bu veriler, otomatik veri toplama teknikleri veya sitelerin kendilerinin sunduğu veri kümeleri olarak elde edilebilir.

Sosyal medya veya SAS'lar hakkında 2007 yılında, Lenhart vd. [6] tarafından yapılan anketlere göre internet kullanan gençlerin %55'inin en az bir SAS hesabı bulunmaktadır, bu kişilerin %66'sının profil bigileri erişime kapalı olarak ayarlanmıştır. Erişime açık kullanıcılarının %46'sının profilinde ise en az bir adet yanlış bilgi bulunmaktadır.

Wolak vd.'ye [7] göre Her 7 gençten biri internette istenmeyen cinsel içerikli mesaj veya davetler almaktadır ve bu mesaj veya davetlerin sadece %9'u 25 yaş ve üstü kişilerden gönderilmektedir.

2.2. Veri Madenciliği

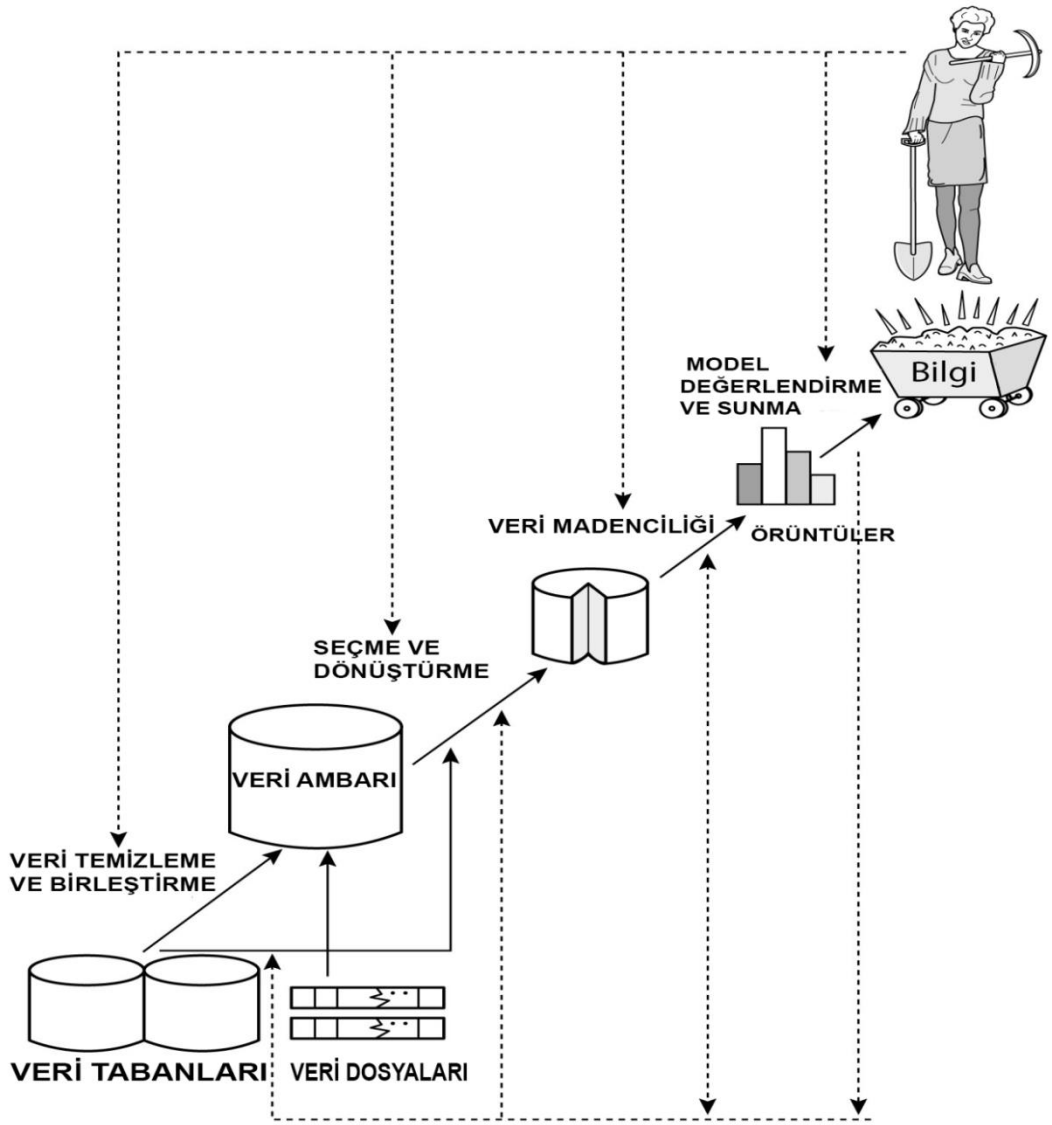
Bilgisayar ve internet teknolojisinin gelişimiyle veri tabanlarında tutulan bilgiler çok büyük bir seviyeye gelmiştir. Örneğin sadece Facebook sitesi, kullanıcılar tarafından üretilen verileri tutmak için "*Hadoop Distributed File System*" (HDFS) teknolojili 100 petabyte'ın üzerinde sabit diskler kullandığını iddia etmiştir [8].

Bu kadar büyük bir miktarda veriden kullanışlı bir bilgiye ulaşma işlemine Veri Madenciliği veya Bilgi Keşfi denilmektedir. Bu işlemler 7 adımdan oluşmaktadır: Şekil 1'de bilgi keşfinin adımları görsel olarak gösterilmiştir [9].

1- Veri temizleme: Gürültülü ve tutarsız verilerin arındırılması.
2- Veri birleştirme: Farklı kaynaklardan gelen veriyi birleştirmek.
3- Veri seçme: Veri tabanından analizle bağlantılı olan verilerin seçilmesi.
4- Veri dönüştürme: Farklı kaynaklardan toplanan verilerin belli bir formata dönüştürülüp veri madenciliğine uygun hale getirilmesi.
5- Veri madenciliği: Bir veri madenciliği tekniğinin seçilmesi ve örüntülerin bulunması
6- Örüntülerin değerlendirilmesi: Bulunan örüntü veya modelin bilgi keşfindeki etkisinin bazı ölçme yöntemleriyle değerlendirilmesi.
7- Bilginin yorumlanması: Bulunan bilginin görsel veya başka bir teknikle kullanıcılara yorumlanması.

Yukarıdaki aşamaların ilk dört tanesi Veri Önişleme aşamalarıdır. Bu 4 aşamada veriler, veri madenciliği için hazırlanır ve bilgi keşfinin en büyük bölümünü bu dördü oluşturmaktadır.

Şekil 1'de de görüldüğü üzere veri madenciliği, bilgi keşfinin sadece bir aşaması olmasına rağmen, bilgi keşfi yerine veri madenciliği kullanmak çok daha popülerdir ve bu tezde de veri madenciliği kullanılmıştır.



Şekil 1. Bilgi keşfinin aşamaları

2.2.1. Veri Madenciliği Algoritmaları

Veri madenciliği algoritmalarının amacı, veriyi belli bir modele uydurmaktır. Böylece model oluşturulurken, yeni veya kullanılmayan verileri bu modelde sınayarak veriden bilgi keşfine erişim sağlanır. Veri madenciliği algoritmaları Gözetimli (*Supervised*) ve Gözetimsiz (*Unsupervised*) olarak ikiye ayrılır. Gözetimli algoritmalarda grup isimleri ve grup sayısı bilinmekte olup, veri madenciliği algoritması uygulandıktan sonra verinin

sınıf etiketi belirli etiketlerden biri olarak tanımlanacaktır. Bu çalışmada Gözetimli veri madenciliği yöntemlerinden biri olan Sınıflandırma (*Classification*) yöntemi kullanılmıştır. Gözetimsiz algoritmalarda ise grup isimleri belirli olmadığı gibi bazen grup sayısı da bilinmemektedir. Demetleme (*Clustering*) yöntemi, Gözetimsiz veri madenciliği algoritmalarına örnektir. Sonraki bölümlerde bu çalışmada kullanılan sınıflandırma yöntemleri hakkında kısa bilgi sunulmuştur.

2.2.1.1. K-En Yakın Komşuluk Sınıflandırıcı

KEYK (K en Yakın Komşuluk) sınıflandırıcısı, İngilizce ismiyle *K-Nearest Neighbor*, veri madenciliğinde KNN olarak İngilizce kısaltmasıyla tanınmaktadır. Bu yöntem Tembel algoritmalar (*Lazy Learners*) arasında yer almaktadır. Tembel algoritmalar, sınıf etiketini komşularından öğrendiklerinden bu ismi almışlardır ve adlarından da anlaşıldığı üzre sınıflandırılacak olan örneği, ona en yakın olan komşuyla aynı sınıfa atmaktadırlar.

Bu yöntemin isminin başında gelen K harfi ise, bir örneğin A sınıfına atanması için en az "K" adet en yakın komşusunun aynı sınıfa ait olması gerektiğini göstermektedir. K sayısı genellikle random olarak atanır ve algoritma birkaç kez çalıştırılarak her defasında çıkan sonuçlar ve K değeri karşılaştırılır. En iyi sonuç hangi K değerine göre elde edilmişse o sayı seçilir.

KEYK yönteminde, her örnek n-boyutlu uzayda bir nokta gibi düşünülür ve böylece sınıf etiketi olmayan her nokta için K en yakın komşusunun sınıf etiketlerine bakılır. Bu yöntemdeki bahsi geçen yakınlık kavramı metrik sistemde bir mesafedir. Bu mesafe çeşitli uzay noktalarını ölçme yöntemleriyle belirlenebilir. Örneğin Oklid uzaklığı veya bu çalışmada kullanılan Kosinus Benzerliği. Bu iki ölçütün hesaplanma şekli aşağıda verilmiştir:

- Oklid Uzaklığı: n boyutlu bir Oklid uzayında P(p_1, p_2, p_3, ... , p_n) ve Q(q_1, q_2, q_3, ... , q_n) noktaları arasındaki mesafeye denilir ve Formül 1'le hesaplanır.

$$d = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (1)$$

- Kosinüs Benzerliği: Bu ölçüt iki metin dosyası arasındaki farkı ölçmek için kullanılır. Vektör uzayı modelinde, iki döküman arasındaki benzerliği hesaplamak için kullanılır. İki döküman birbirine ne kadar benzerse bu dökümanların vektörleri de birbirine o kadar yakındır. Bu ölçüt Formül 2'yle hesaplanır.

$$Cos(d_1, d_2) = \frac{d_1 \cdot d_2}{|d_1||d_2|} \quad (2)$$

2.2.1.2. Destek Vektör Makineleri

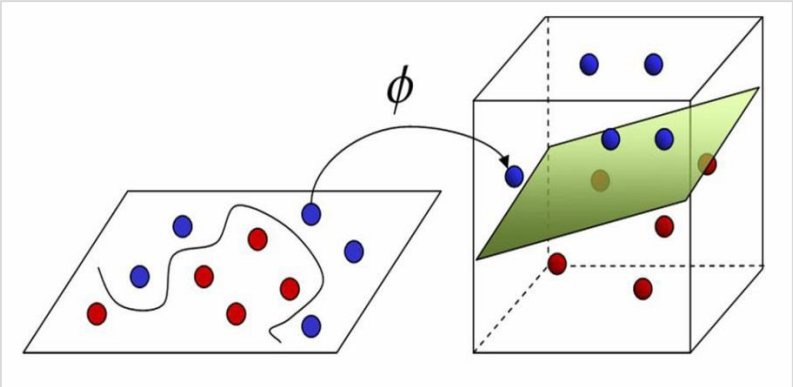
DVM (Destek Vektör Makineleri) veya ingilizce ismiyle SVM (*Support Vector Machine*); sınıflandırma yöntemlerinde en çok kullanılan yöntemlerden biridir, doğrusal ve doğrusal olmayan veri üzerinde sınıflandırma yapabilmektedir. Bu yöntem doğrusal olmayan bir eşleme kullanarak, orjinal eğitim verilerini daha yüksek bir boyuta taşır, daha sonra bu yeni yüksek boyut içinde doğrusal bir ayrıştırıcı hiper düzlem bulmaya çalışır. Uygun bir doğrusal olmayan eşleştirme ile, iki sınıftan oluşan bir veriyi, her zaman yeterince yüksek bir boyutta hiper düzlemle ayırmak mümkündür. DVM, hiper düzlem, destek vektörler ve onların kenar uzaklıklarıyla veriyi sınıflandırır.

Şekil 2 DVM'nin veriyi yüksek bir boyuta taşıyarak hiper düzlemle nasıl ayırdığını göstermektedir.

2.2.1.3.Naive Bayes Sınıflandırıcı

Bayes sınıflandırıcıları istatistiksel sınıflandırıcılardır ve bir sınıfın üyelik derecesini tahmin edebilirler. Örneğin bir Facebook kullanıcısının ne derece erkek sınıfına ait olması gibi. Naive Bayes sınıflandırıcısı, Bayes sınıflandırıcılarının en basit olanıdır.

Bayes sınıflandırıcılar büyük veri kümelerinde dikkat ve hız açısından iyi derecededirler.



Şekil 2. DVM ile veri sınıflandırma

Naive Bayes sınıflandırıcısı, bir nitelik değerinin sınıf etiketindeki etkisinin, diğer niteliklerin değerinden bağımsız olduğunu varsayar. Bu bağımsızlığa sınıf koşullu bağımsızlık denilir.

Bayes Teoremi: X'in, bir Facebook kullanıcısının veri vektörü olduğunu varsayalım, H bir hipotez ve X'in, C sınıfına ait olduğunu kabul edersek, sınıflandırma problemleri için P(H | X)'yi (X'in H hipotezini doğrulama olasılığını) hesaplamamız gerekmektedir. Bu olasılık, Formül 3 ile hesaplanmakta ve Bayes Teoremi olarak adlandırılmaktadır.

$$P(H\,|X) = \frac{P(X\,|H)P(H)}{P(X)} \qquad (3)$$

Bu çalışmayla ilgili bir örnek verelim, terim vektörünün iki nitelikten oluştuğunu varsayalım. Bu nitelikler sırasıyla yorum içeriğinde gülücük kullanmak (':)') ve yorum içeriğinde "çoook tatlııı" ibaresini kullanmak olsun. X bir Facebook kullanıcısıdır ve yorum içerisinde her iki niteliği de kullanmıştır. H hipotezi bu kişinin KADIN sınıfına ait olduğunu söylüyor. Böylece P(H | X), X kullanıcısının KADIN sınıfına ait olma olasılığını göstermektedir, eğer bu değer 0,5'in üzerindeyse hipotez doğrudur ve kullanıcı kadındır.

P(H), nitelik değerlerine bakmaksızın her kullanıcının KADIN olma olasılığıdır.

P(X | H), X kullanıcısının KADIN olduğunu varsayarak bu kullanıcının yorumda gülücük ve "çoook tatlııı" ibaresini kullanma olasılığıdır.

P(X), X kullanıcısının yorumda gülücük ve "çoook tatlııı" ibaresini kullanma olasılığıdır.

2.2.2. Sınıflandırmada Doğruluk Ölçütleri

Herhangi bir sınıflandırma algoritmasının oluşturduğu modelin doğruluğunu değerlendirmek için veri kümesi, öğrenme (*train set*) ve sınama (*test set*) kümeleri olarak iki ayrı alt kümeye ayrılır. Model, öğrenme kümesiyle oluşturulduktan sonra sınama kümesine uygulanarak doğruluğu ölçülür. Öğrenme kümesi, en basit şekilde verinin %70'lik bölümünden oluşmakta, geri kalan %30'luk bölüm ise sınama (*test set*) kümesi olarak adlandırılmaktadır. Sınama kümesi, oluşturulan modele uygulanarak karmaşıklık matrisi elde edilir. İki sınıflı bir veri için karmaşıklık matrisi Tablo 1'de gösterilmiştir. Bu tablo'da 1. Sınıf "C1" ile gösterilmiş ve asıl sınıf etiketi olduğunu belirmek için pozitif sınıf olarak işaretlenmiştir. 2. Sınıf ise "C2" ile gösterilmiştir. Sınıflandırma işlemlerinde genelde örnek sayısı daha az olan sınıf pozitif sınıf olarak işaretlenir. Örneğin, sınama kümesinde toplamda 100 örnek olduğunu varsayarsak ve C1'in sadece 40 adet olduğunu düşünürsek, bu sınıfta olan örnek sayısı daha az olduğundan C1 pozitif olarak işaretlenir.

Tablo 1. Karmaşıklık matrisi

		Modelin tahmin ettiği sınıf etiketleri	
		C1 (pozitif)	C2 (negatif)
Gerçek sınıf etiketleri	C1	DP	YN
	C2	YP	DN

Tablo 1'de görüldüğü üzere gerçek sınıf etiketlerini model ile tahmin edilen sınıf etiketleri karşılaştırılmıştır. bu karşılaştırmadan çıkan sonuç 4 şekilde olabilir:

- DP (Doğru Pozitif): Doğru olarak Pozitif etiketlenen örneklerin sayısı
- YN (Yanlış Negatif): Yanlış olarak Negatif etiketlenen örneklerin sayısı (örnekler gerçekte pozitiftir)
- YP (Yanlış Pozitif): Yanlış olarak Pozitif etiketlenen örneklerin sayısı (örnekler gerçekte negatiftir)
- DN (Doğru Negatif): Doğru olarak Negatif etiketlenen örneklerin sayısı

Karmaşıklık matrisinden elde edilen sonuçlara göre sınıflandırma algoritmasının başarısı üç parametreyle ölçülebilir.

- Doğruluk (*Accuracy*): Doğru etikelenen sınıfların yüzdesini gösterir ve Formül (4) ile ölçülür.

$$doğruluk = \frac{DP+DN}{DP+DN+YP+YN} \quad (4)$$

- Kesinlik (*Precision*): Bu ölçüt, pozitif olarak etiketlenen örneklerin gerçekte yüzde kaçının pozitif olduğunu gösterir ve Formül (5) ile hesaplanır.

$$kesinlik = \frac{DP}{DP+YP} \quad (5)$$

- Anma (Recall): Pozitif olarak etiketlenen örnekler, gerçekte pozitif olan örneklerin yüzde kaçı olduğunu gösterir ve Formül (6) ile hesaplanır.

$$anma = \frac{DP}{DP+YN} \quad (6)$$

2.2.3. Sınıflandırma Yöntemlerinde Doğrulama

2.2.3.1.K-Kat Çapraz Doğrulama

Sınıflandırma algoritmalarının performansını ölçmek için, veri kümesi farklı şekillerde öğrenme ve sınama kümelerine bölünür. Bunlardan en önemlisi ve bu çalışmada da kullanılan K-kat çapraz doğrulama (K-fold Cross Validation) yöntemidir.

Bu yöntemde başlangıçta "D" veri kümesi, random olarak "D_1, D_2, ... , D_k" şeklinde k adet eşit bölüme ayrılır. i'inci iterasyonda "D_i" bölümü sınama kümesi olarak ayrılır ve i'inci modeli oluşturmak için kalan diğer bölümler öğrenme kümesi olarak kullanılır. Ardından i+1'inci iterasyonda "D_{i+1}" sınama kümesi olarak ayrılır, kalan bölümler kullanılarak i+1 modeli oluşturulur ve i=k iterasyona kadar bu işlem devam eder. İterasyonların sonunda tüm örnekler hem öğrenme hem de sınama için kullanılmış olur,

bununla beraber sınıflandırma doğruluğu ise tüm iterasyonlardaki doğru sınıflandırılan örneklerin, toplam örnek sayısına oranıyla elde edilir.

Çapraz Doğrulama esnasında katlar, tabakalı örneklemeye (*stratified sampling*) göre seçilirse bu uygulamaya Tabakalı Çapraz Doğrulama (*Stratified Cross Validation*) denir. Tabakalı örneklemede, sınıf etiketinin dağılımı korunacak şekilde örnekler seçilir. Tabakalı örnekleme için bir örnek Şekil 3'te verilmiştir. Bu yöntemde şekilde de görüldüğü üzere seçilen örneklerin sınıf etiketleri başlangıçtaki veriyle aynı oranda seçilmiştir [9].

T38	ÇOCUK
T256	ÇOCUK
T307	ÇOCUK
T369	ÇOCUK
T12	GENÇ
T69	GENÇ
T156	GENÇ
T190	GENÇ
T237	GENÇ
T260	GENÇ
T301	GENÇ
T337	GENÇ
T18	ORTA YAŞ
T193	ORTA YAŞ

T38	ÇOCUK
T307	ÇOCUK
T69	GENÇ
T190	GENÇ
T237	GENÇ
T337	GENÇ
T18	ORTA YAŞ

Şekil 3. Yaş sınıf etiketine göre tabakalı örnekleme

2.3. Metin Madenciliği

Veri madenciliği işlemi için verinin, yapısal ve matris formunda olması gerekmektedir; ancak gerçek hayatta veri her zaman veritabanı sistemlerinde olduğu gibi yapısal değildir. Günümüzde verilerin büyük bir bölümü metin olarak mevcuttur, özellikle internetin gelişimi ve kullanımının artışıyla metin şeklindeki veri kaynakları günden güne artmaktadır. Metinsel verilere örnek olarak haber metinleri, web siteleri içeriği, akademik yazılar, e-posta mesajları ve bu çalışmada yer alan sosyal medyada yazılan yorumlar gösterilebilir.

Metinsel veriler genelde yarı-yapısal şekilde mevcuttur, bu veriler veritabanı sistemlerinde olduğu gibi tam-yapısal değildir ancak az da olsa yapısal bilgi içermektedirler. Örneğin bir akademik makalede, makalenin başlığı, yazarları, yayın yeri,

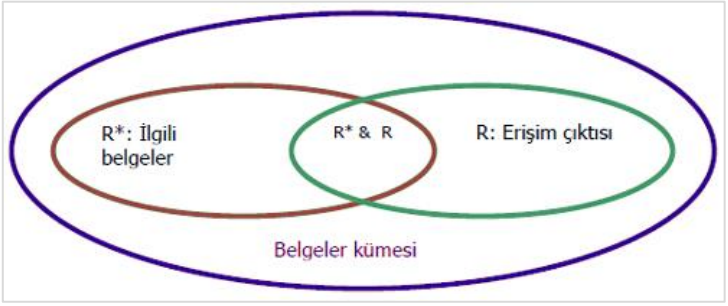
yayın tarihi vb. gibi bilgiler yapısal bir şekilde mevcuttur fakat makalenin içeriği yapısal değildir.

Metin madenciliği işlemlerini kolaylaştırmak amacıyla bir çok BE (Bilgi Erişim) (*Information Retrieval*) teknikleri geliştirilmiştir.

Veritabanı sistemlerinin yapısal veriler üzerindeki hareket (*Transaction*) ve sorgu işlemlerinin aksine, BE sistemleri, metin dökümanlarının yapısı ve onlardan bilgi elde etmek üzerine yoğunlaşmıştır.

2.3.1. BE Sistemlerinde Performans Ölçüm Teknikleri

BE sistemlerinde elde edilen verinin ne kadar doğru ve dikkatli olduğunu ölçmek için Kesinlik (Precision) ve Anma (Recall) hesaplanır. Sorguyla gerçekten ilişkili olan dökümanlara {ilgili} (Relevant) ve sorguyla elde edilen dökümanlara ise {Erişim çıktısı} denir, böylece bir sorgu için elde edilen ilişkili Erişim çıktısı {ilgili} ∩ {Erişim çıktısı} formülüyle hesaplanır. Bu iki ölçütün ilişkisini göstermek için Venn diyagramı kullanılır. (Şekil 4)



Şekil 4. Venn diyagramı

Kesinlik: Elde edilen dökümanların yüzde kaçının gerçekten sorguyla ilişkili olduğunu gösterir ve Formül 7 ile hesaplanır.

$$Kesinlik = \frac{|\{ilgili\} \cap \{Erişim\ çıktısı\}|}{|\{Erişim\ çıktısı\}|} \quad (7)$$

Anma: Gerçekte sorguyla ilişkisi olan dökümanların yüzde kaçının elde edildiği Formül 8'le hesaplanır.

$$Anma = \frac{|\{ilgili\} \cap \{Erişim\ çıktısı\}|}{|\{ilgili\}|} \quad (8)$$

BE sistemlerinde bazen Kesinlik için Anma'yı bazen de Anma için Kesinliği göz ardı etme gereği duyulur. Bu iki ölçütün ortalamasını hesaplamak için F-Olçütü (*F-Score or F-measure*) kullanılır. Bu ölçüt Kesinlik ve Anma'nın Harmony Ortalamalarını hesaplamaktadır. (Formül 9)

$$F\_ölçütü = \frac{2 \times Anma \times Kesinlik}{(Anma + Kesinlik)} \quad (9)$$

Bu hesaplamayla Kesinlik ölçütünün Anma ölçütüne veya Anma ölçütünün Kesinlik ölçütüne galip olması büyük bir ölçüde engellenmektedir. Böylelikle F-Ölçütünün büyük olması Kesinlik ve Anma ölçütlerinin birbirlerine yakın olmalarının garantisidir.

2.3.2. Metinsel Verilerde BE Teknikleri

BE teknikleri Döküman seçme yöntemi ve Döküman puanlama yöntemi olarak ikiye ayrılmaktadır.

Döküman seçme yönteminde, ilgili dökümanları seçmek için, sorguda geçen terimler, belirtici arama anahtarı olarak kullanılmaktadır. Bu kategoride genel yöntem olarak mantıksal erişim modeli (Boolean Retrieval Model) örnek verilebilir. Bu modelde doküman, bir anahtar kelime kümesi olarak gösterilir ve ilgili dökümana erişmek için kullanıcı mantıksal ibarelerden oluşan anahtar kelimeleri sağlar, örneğin, "araba ve yedek parça" veya "Çay veya Kahve" gibi. BE sistemi bu anahtar kelimelerden yola çıkarak ilgili dökümanları bulmaya çalışır.

Döküman puanlama yöntemindeyse, sorguda geçen terimler kullanılarak tüm dökümanlar, ilgililik açısından puanlanır. Bu yöntem döküman seçme yöntemine göre daha sık kullanılan bir yöntemdir.

Tüm bu yöntemlerin arkasındaki ortak sezgi, bir sorguda geçen anahtar kelimeleri dökümanlardaki kelimelere eşleştirerek, bu dökümanları gerçekleşen eşleşmenin doğruluğuna göre puanlamaktır.

2.3.3. Vektör Uzay Modeli

Metinsel bir veriyi veri madenciliği işlemine uygun hale getirmek için ilk adım, metinsel veriyi matris haline donüştürmektir. Bu işlem için dökümanlar, vektör uzayına taşınmalıdır. Bunun için metindeki kelimeler veya terimler ayrıştırılmalıdır. Dökümandaki her terim vektör uzayında bir boyutu temsil etmektedir ve bu yüzden metinsel veriler genelde çok yüksek boyutlu uzay modellerinde gösterilmektedir.

Dökümanları vektör uzayına taşırken, birinci adımda, döküman değersiz kelimelerden arındırılır. Değersiz kelimeler çok sık rastlanan ve veri hakkında bilgi vermeyen kelimelerdir: "ve", "veya", "yada" vs. İkinci adımdaysa ortak köklü kelimler eklerinden arındırılarak, kelimenin sadece kökü terimler vektöründe yer alır, örneğin "gidiyor", "gidecekler" vs. gibi kelimeler yerine sadece "gitmek" kelimesi vektörde kullanılır.

Vektör uzayındaki terimleri ağırlıklandırmak için bir çok yöntem vardır. Bunlara TS (Terim Sıklığı) (*Term Frequency*) yöntemleri denir ve "TF" ile gösterilir. Bu yöntemlerden biri mantıksal ağırlıklandırmadır. Eğer bir terim, dökümanda mevcut ise 1, mevcut değilse 0 değerini alır [9]. Mantıksal vektör uzayı için aşağıda örnek verilmiştir.

Dökümanlar kümesi d={d1, d2, d3} olsun, dökümanlar ve dökümanlardaki terimler Tablo 2'de gösterilmiştir:

Bu dökümanların mantıksal vektörünü oluştumak için v tüm terimler vektörü olarak şöyle tanımlanır: v= {internet, ağ, net, sayfa, site}.

Tablo 2. Dökümanlar ve dökümanlar kümesinde geçen terimler

Döküman	Metin	Terimler
d1	internet ağ	internet, ağ
d2	ağ internet net ağ net	internet, ağ, net
d3	sayfa internet net site	internet, net, sayfa, site

daha sonra her döküman için mantıksal vektör Formul 10'a göre bulunur.

$$\begin{cases} 1 & eğer\ terim\ \in\ v \\ 0 & eğer\ terim\ \notin\ v \end{cases} \quad (10)$$

Yani:

d1={1,1,0,0,0}

d2={1,1,1,0,0}

d3={1,0,1,1,1}

Bir başka ağırlıklandırma yöntemiyse Terim Sayma yöntemidir. t teriminin d dökümanındaki geçme sayısı, terim ağırlığı olarak hesaplanır ve *freq(d,t)* olarak gösterilir. Ancak bu yöntemde, uzun metinlerde daha fazla terim olacağından her biri için daha fazla ağırlık verilecektir. Bu da uzun metinler için bir ayrıcalık olacaktır. Bunu önlemek için Terim Sayma yöntemini normalleştirmek gerekmektedir.

Terim Sayma yöntemini normalleştirmek için kullanılan bir yöntem İlgili Terim Sıklığıdır (*Relative Term Frequency*). Terim sıklığı, dökümandaki terim sayısının toplam dökümanlardaki terim sayısına oranı olarak hesaplanır. Bir başka normalleştirme yöntemiyse Normalleşmiş Sıklık yöntemidir. Bu yöntemde ise terimin geçme sayısı, dökümandaki en çok tekrarlanan herhangi bir terimin sayısına bölünür.

Cornell SMART sistemi ise Terim Saymayı normalleştirmek için Formül 11'i kulanmaktadır.

$$\begin{cases} 0 & eğer\ terim\ \notin\ v \\ 1 + \log\left(1 + log\left(freq(d,t)\right)\right) & eğer\ terim\ \in v \end{cases} \quad (11)$$

Terim ağırlıklandırma ölçütlerinin yanı sıra DDS (Devrik Döküman Sıklığı) (*IDF* (*Inverse document Frequency*)) bir başka önemli ölçüttür ve bir terimin ölçekleme faktörü veya terimin önemini göstermektedir. Böylece bir terimin dokümanlar arasında nadir olup olmadığını tespit etmektedir. Eğer t terimi, bir çok dökümanda geçmişse; o terimin önemi azalır. Çünkü bir çok dökümanda geçtiği için çok fazla ayrıştırıcı gücü yoktur. Aynı *Cornell SMART* sistemine göre bir terimin DDS'si (IDF(t)) Formül 12'ye göre hesaplanır.

$$IDF(t) = \log \frac{1+|d|}{|d_t|} \quad (12)$$

Bu Formülde ‘d’ dökümanlar kümesidir ve ‘d_t’, ‘t’ teriminin mevcut olduğu dökümanlar kümesidir.

Eğer $|d_t| << |d|$ ise, t teriminin DDS’si çok büyük olacaktır ve bu ‘t’ teriminin ayrıştırıcı gücünün yüksek olduğunu gösterecektir; ancak eğer t teriminin ayrıştırıcı gücü çok düşük olursa, o zaman da t teriminin DDS’si sıfıra yakınlaşacaktır.

TS ve DDS’yi birleştirerek bir tam vektör uzay modeli oluşturabilir, bu yöntem TF-IDF adıyla tanımlanır ve Formül 13 gibi hesaplanır [9].

$$TF - IDF(d,t) = TF(d,t) \times IDF(t) \quad (13)$$

2.3.4. Nitelik Seçme ve Boyut Azaltma

Metin madenciliği işlemlerinde nitelik sayısı genellikle fazla olduğundan, elde edilen terim vektörleri çok sayıda niteliğe sahiptirler. Böylesine büyük bir verinin işlenmesinin hesaplama ve bellek maliyeti çok fazladır. Ayrıca çoğu zaman bazı niteliklerin sınıf etiketiyle ilişkisi yoktur veya hesaplama ve bellek kullanımı maliyetine karşın önemsenmeyecek kadar azdır. Bu yüzden büyük terim vektörlerinde sadece daha önemli nitelikleri model oluşturmaya dahil ederek boyut azaltmak için birçok algoritma vardır.

Boyut azaltma işlemleri genel olarak niteliklerin ağırlıklarını ölçerek, ağırlıkları belirli bir sınırın altında olan nitelikleri elemeyi amaçlamaktadır.

Büyük veriler için nitelik ağırlıklandırma işlemi aynı zamanda çok hızlı ve etkin olmalıdır. Bu çalışmada nitelik ağırlıklandırmak için SAM (*Significance Analysis of Microarrays*) [10] yöntemi kullanılmıştır.

2.3.4.1. SAM nedir?

SAM bir istatistiksel yöntemdir ve başlangıçta gen ekspresyonundaki değişikliğin istatistiksel önemini göstermek için oluşturulmuştur. Bu yöntem gen ekspresyonu ve cevap değişkeninin ilişkisini ölçmek için kullanılmıştır.

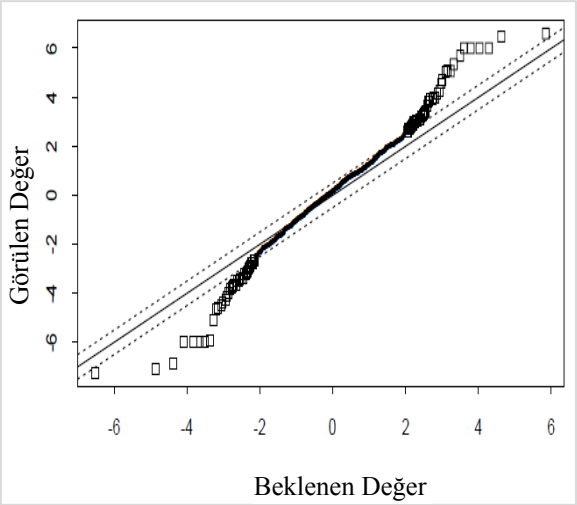
Bu çalışmada, gen ekspresyonu, terim vektörü ve cevap değişkeni, sınıf etiketi olarak kullanılmıştır. Böylece terim vektörünün terimlerindeki değişikliğin önemi, sınıf etiketine karşı ölçülmüştür.

Bu yöntemde, her Facebook kullanıcısının terim vektörü ile sınıf etiketi arasındaki bağlantının anlamlı olup olmadığını ölçmek için, verilerin tekrarlanan permütasyonları kullanılmıştır.

SAM, terim vektörlerindeki bağlantılı değişikliklerin ve vektör verilerinin permütasyon analizine dayanarak bir test istatistiği hesaplar. Ayrıca Yanlış Keşif Oranını da (*False Discovery Rate*) hesaplar. SAM için basit bir algoritma şu aşamalara sahiptir:

1- Hesaplanan test istatistiklerini sırala
2- Her permütasyon için sıralı etkisiz(null) puanları hesapla
3- Sıralı test istatistiğinin, beklenen etkisiz puanlara karşı grafiğini çiz (Şekil 5)
4- Eğer bir niteliğin test istatistiğinin mutlak değeri ile o niteliğin ortalama test istatistiği arasındaki fark, belirlenen sınırdan büyük veya eşit ise o nitelik önemlidir.
5- Beklenen değerlere karşın görülen değerlere dayalı olarak yanlış keşif oranı tahmin edilir.

Şekil 5'te görünen küçük kareler hesaplanan test istatistikleridir, kesik çizgili hat ise algoritmanın dördüncü aşamasında bahsedilen sınırı göstermektedir. Böylece her nitelik için test istatistiğin mutlak değeri belirli bir sınırın altında ise, o nitelik önemsiz olarak işaretlenir veya puanlanır. Daha sonra belirli bir puanın altında olan tüm nitelikler elenir ve kalan nitelikler model oluşturmak için veri madenciliği algoritmasına yönlendirilir.



Şekil 5. Beklenen değere karşı görünen değer grafiği

2.4. RapidMiner: Veri Madenciliği Aracı

Rapidminer [11] makine öğrenme, veri madenciliği, metin madenciliği, öngörü analizi ve iş zekası projeleri için kullanışlı bir geliştirme alanı sunmaktadır.

Bu program, KDNuggets tarafından düzenlenen en iyi veri madenciliği araçları oylamasında 2009 yılında ikinci en iyi yazılım [12] ve 2010, 2011 yılında ise 1. en iyi veri madenciliği aracı ünvanını kazanmıştır [13, 14]. Bu program 2004 yılından beri açık kaynak kodlu olarak sunulmaktadır.

Rapidminer projesine 2001 yılında Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer tarafından Dortmund Teknik Universitesi, yapay zeka ünitesinde başlanılmıştır [11].

Bu program veri madenciliği ve makine öğrenme işlemlerini bir araya getirmek için interaktif bir arayüz sunmaktadır. Bu program veri toplama, veri dönüştürme, veri önişleme işlemleri (temizleme, boyut azaltma, nitelik seçme), sınıflandırma, eğri uydurma, demetleme, öngörü analizi, metin madenciliği, multimedya madenciliği vs. için birçok algoritma sunmaktadır ve tüm bu algoritmalara rağmen WEKA [15] veri madenciliği aracının da tüm algoritmalarını içermektedir.

Bu program WEKA'yla [15] karışılaştırıldığında bellek kullanımını büyük bir ölçüde düşürmüştür ve interaktif arayüzüyle de çok kolay bir kullanım sağlamaktadır. Bu program

sayesinde işlem yapılacak veri, farklı algoritmalarla çalıştırılarak sonuçlar karşılaştırılabilinir. Ayrıca veri analizi de çeşitli grafiklerle kolay bir hale getirilmektedir.

2.5. Türkçenin Sohbet (*Chat*) Dili

Bu çalışma Türk dilinde yapıldığından öncelikle Türkçe morfoloji özellikleri ve chat ortamlarındaki kullanım özelliklerinin ele alınması gerekmektedir.

Facebook'ta paylaşılan yorumların analiziyle, aşırı derecede chat dili[1] kullanıldığını görüyoruz. Yanlış yazılan kelimeler, kasıtlı olarak hatalı yazılan kelimeler, sesli harflerin tamamını veya bazısını silerek kısaltmaya çalışılan kelimeler, alfasayısal olmayan karakter kullanımları ve akronimler gibi değişken yazı tarzları görünen problemlerdir. Bahsedilen problemler daha detaylı açıklamalarla aşağıda verilmiştir.

2.5.1. Kasıtlı Yazım Hataları

Her kullanıcı, yorumlarında istemeyerek de olsa hatalı harf yazma ihtimaline sahiptir. Örneğin "selam" yazmaya çalışırken yanlışlıkla "n" tuşuna dokunup "selan" yazılabilir ve bir imla kontrolü uygulamasıyla bu çözülebilir. Ancak bazen kullanıcılar kasıtlı olarak yanlış yazmış olabilirler. Aşağıda kasıtlı olarak yapılan yanlışlıklar açıklanmıştır.

1 Sohbet ortamlarında kullanılan yazı stili

2.5.1.1. Bir Kelime için Farklı Formlar

Türkçe chat dilinde, bazı kelimelerin farklı yazım biçimleri vardır. Yazar, ya alışkanlık yüzünden ya da kendi tercihiyle bir yazım biçimini kullanmaktadır. Tablo 3'te "gidiyorum" ve "geleceğim" kelimeleri için veri kümesinde rastlanan farklı formlar gösterilmiştir.

Tablo 3. Gidiyorum ve Geleceğim kelimesi için farklı formlar

Gidiyorum	Geleceğim
Gidiom	Gelcem
Gidiorm	Gelecem
Gidiyorm	Gelcm
Gidyom	Glcm
Gidym	
Gdiom	

2.5.1.2. Türkçeye Özel Harfler Yerine İngilizce Harfler Kullanmak

Türkçede, İngilizcede olmayan bazı karakterler vardır. Bu karakterler "Ç","Ğ","Ş","ı","İ","Ö" ve "Ü"dür. Bazı yorumlarda yazar, (alışkanlıktan ya da cep telefonu kullandığı ve cep telefonunun Türkçe tuş takımının olmamasından dolayı) bu harflerin İngilizce eşdeğerini, yani sırasıyla "C","G","S","i","I","O" ve "U" harflerini kullanmaktadır.

2.5.1.3.Alternatif Karakterler

Kullanıcılar, bazen alışkanlıklarından, bazen cep telefonu kullandıklarından, bazen de daha çok ilgi çekebilmek adına belli karakterlere özgü alternatif karakterler kullanmaktadırlar. Bu konuda en çok kullanılan alternatiflerden "w" karakterinin "v" yerine ve "q" karakterinin "g" ve "k" yerine kullanılmasıdır. Örneğin "Evet" kelimesinin

yerine “Ewet” yazmak gibi. Tablo 4’te veri kümesinde rastlanan birbirinin yerine kullanılan tüm alternatif karakterler listelenmiştir.

Tablo 4. Alternatif karakterler listesi

Karakter	Alternatifleri
ö	o, ò, ô,ó, õ
ü	u, û, ù,ú
ı	i, Ì, ¡, Ï,Í,Î, í, İ
ç	C
ş	s, $
ğ	g, q
k	g,q
b	ß
a	â
e	£, è,é
v	w
?	¿
“	“
‘	`, ´, ’

2.5.1.4.Bazı Harflerin Tekrarı

Bazı kullanıcılar, yazdıklarını vurgulamak için kelimelerin son karakterini veya sesli harflerini tekrarlamaktadırlar. Bunun herhangi bir kuralı yoktur, kullanıcı hangi harfi, kaç kez tekrarlayacağına kendisi karar vermektedir. Örneğin “merhabalaaarrrrrrr” kelimesinde “a” sesli harfi ve “r” sessiz harfi tekrarlanmıştır.

2.5.1.5. Sesli Harflerin Tamamını veya Bazısını Silerek Kısaltmalar

Kullanıcıların büyük bir bölümü, alışkanlık yüzünden veya daha hızlı yazmak uğruna kelimedeki sesli harflerin bazılarını silerek kısaltmaya çalışmaktadırlar. Kullanıcılar hangi sesli harfleri sileceklerine kendileri karar vermektedirler, herhangi bir kural yoktur.

Örneğin “gelecektim” kelimesi “gelcektm” veya “glcktm” olarak yazılabilir. Bu tipteki kısaltma yönteminin, 2.5.1.1 bölümündeki anlatılanlardan farkı şöyle açıklanabilir: 2.5.1.1 bölümünde sesli veya sessiz karakterler ikiside yazılmayabilir, örneğin “Geleceğim” kelimesi “gelcem” olarak yazıldığında “ğ” karakteri bir sessiz harftir ve yazılmamıştır. Veya “gidiyorum” kelimesinde “y” ve “r” harfi yazılmayarak kelime “gidiom” şeklinde yazılabilir. Bu bölümde anlatılmak istenen, kelimenin başında veya sonunda sesli harf mevcut ise, bu sesli harfler her zaman yazılmaktadır. Örneğin “ekonomi” kelimesi kısaltılırken “eknmi” ya da “ekonmi” şeklinde kısaltılır, ilk ve son sesli harf her zaman yazılmaktadır.

2.5.1.6. Tanınmış ve Alışılmış Kısaltmalar

Bazı kısaltmalar 2.5.1.5 ve 2.5.1.1 kısımlarında bahsedilen tarzdan farklıdır. Bu tarz kısaltmalar alışılmış ve çok kullanılan kısaltmalardır. Bu kısaltmalardan bazıları Tablo 5’te gösterilmiştir.

Tablo 5.Alışılmışkısaltmalardan bazıları

Kelime	Kısaltması
Merhaba	Mrb
İnşallah	İnş
Arkadaş	Arki
Ağabey	Abi

2.5.2. Akronimler (*Acronyms*)

Akronimler, kısaltmalardan farklı olarak kelimelerin sadece baş harfleri yazılarak oluşturulur. Örneğin “Allaha Emanet Ol” ibaresinin akronimi “AEO” olarak yazılır. Bu akronimlerden bazısı Tablo 6’da gösterilmiştir. Akronimler, açılımları da dahil olmak

üzere chat dilinde farklı şekillerde kullanılmaktadır. Örneğin “AEO” akronimi “A.E.O”, “aeo”, “a.e.o”, “A.E olun”, A e ol” vb. şekillerde kullanılmaktadır.

Tablo 6. Akronimler ve açılımı

Akronim	Açılımı
AEO	Allaha Emanet Ol
KİB	Kendine İyi Bak
HG	Hoş Geldin
HB	Hoş Buldum
SA	Selamün Aleyküm
AS	Aleyküm Selam
BB	Bye Bye

2.5.3. Yüz İfadeleri

Kullanıcılar, daha önce kullandıkları chat programlarından alışkın oldukları yüz ifadelerini kullanmaktadırlar. Ancak her kullanıcı farklı chat programına alışkın olduğu için aynı anlamı taşıyan, her bir yüz ifadesi için çeşitli formlar yorumlarda mevcuttur. Yüz ifadeleri, anlamları ve alternatif kullanımları Tablo 7’de gösterilmiştir. Alternatiflerin dışında kullanıcılar, yüz ifadelerini kullanırken o andaki duygularını yansıtmak için son karakterini tekrarlamaktadırlar. Örneğin gülmek yüz ifadesi “:))” şeklinde yazılır, ancak kullanıcı çok fazla güldüğünü belirtmek “:)))))” şeklinde kullanmaktadır. Burada tekrarlanma sayısı hakkında herhangi bir kural yoktur ve tamamen kullanıcıya bağlıdır.

2.6. N-Gram

Bilişimsel dilbilim ve olasılıkta, metinde veya konuşmada ard arda geçen N adet terim veya karakterlere n-gram denir [16]. N-gramlar örtüşen veya örtüşmeyen şeklinde kullanılabilir, Şekil 6’da örtüşen ve örtüşmeyen n-gramlar için bir örnek gösterilmiştir. N-gram’lar ard arda gelen ses birimi (*phoneme*), harf, karakter veya kelime olabilirler.

Tablo 7. Yüz ifadelerinin listesi ve alternatifleri

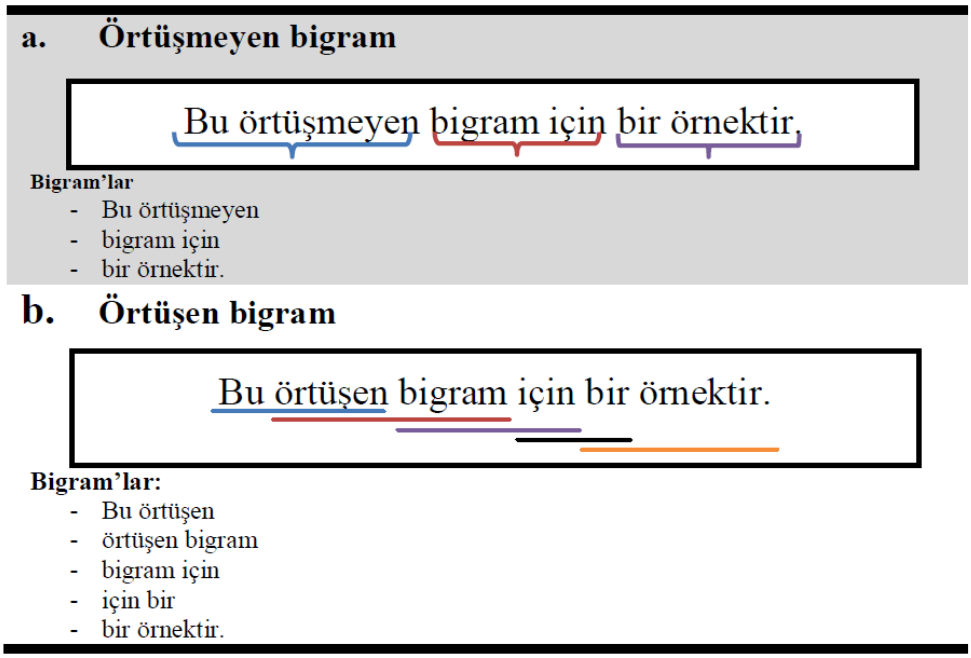
Yüz ifadesi	Alternatifleri	Anlamı
:)	:-) :=)	Gülümsemek
:))	:-)) :=))	Gülmek
;)	;-) ;=)	Göz kırpmak
:(	:-(:=(	Üzgün
:((	:-((:=((	Çok Üzgün veya Ağlamak
:D	:-D :D 8-D 8D x-D xD X-D XD =D .d .D	Sesli gülmek veya Sırıtmak
:O	:-O o_O	Surpriz olmak veya şok olmak
:*	:-*	Öpmek
:P	:-P	Dil çıkarmak
<3	♥	Sevmek veya kalp
</3		Kırık kalp
:s	.s	Utanmak veya stres olmak

N-gramlar genel olarak İstatistiksel Doğal Dil İşleme'de kullanılır. Ayrıca ses birimleri ve ses birim n-gramları, ses tanıma işlemlerinde kullanılır.

Her bir terimin teker teker geçmesine unigram denilir, eğer n=2 ise o zaman n-gram'a bigram denir ve arda arda geçen ikişer terimler bigram olarak adlandırılır. Böylece N=3'e trigram olarak belirtilir ve N sayısı arttıkça "*sayı*-gram" şeklinde gösterilir, örneğin N=4 için 4-gram veya ingilizcesi olarak "four-gram" ve N=5 için 5-gram şeklinde gösterilmektedir.

2.7. Metin Madenciliği ile Cinsiyet ve Yaş Belirleme

Bir yazıyı analiz ederek yazarın, cinsiyetini ve yaşını belirleme işlemi geçmiş yıllarda birçok araştırmacının ilgisini çekmiştir. Metin Madenciliği yöntemlerini kullanarak, farklı medyalarda yazarın cinsiyetini ve/veya yaşını belirlemek için birçok araştırma yapılmıştır. Bu çalışmalarda genel olarak sohbet medyaları, bloglar ve sosyal medya veri kaynağı olarak kullanılmıştır.



Şekil 6. Örtüşen ve Örtüşmeyen n-gramlar için birer örnek

2.7.1. İlgili Çalışmalar

Sohbet ortamlarında yapılan çalışmalarda, Köse vd., [17, 18] özel bir sohbet odası uygulaması geliştirerek, 140 kişi arasında geçen 300 adet diyaloğu analiz ederek kullanıcıların cinsiyetini belirlemeyi amaçlamıştır. Bu araştırmada bir ayrıştırıcı fonksiyon kullanılarak, konuşmalar semantik olarak analiz edilip, kullanıcıların cinsiyeti tahmin edilmeye çalışılmıştır. DVM ve Naive Bayes sınıflandırma algoritmaları kullanılarak kişiler cinsiyetlerine göre sınıflandırılmıştır. Bu iki sınıflandırıcıdan ve yazarlar tarafından geliştirilen ayrıştırıcı yöntemden çıkan sonuçlar karşılaştırılmıştır. Yazarların geliştirdiği ayrıştırıcı fonksiyon %92,9 doğruluk oranıyla en iyi sonuçları üretmiştir.

Hariharan [19] sohbet ortamlarındaki diyalogları analiz ederek, kullanıcının cinsiyetini belirlemek için otomatik bir sistem geliştirmiştir. Bu çalışmada yazar, kullanıcıların yazı tarzlarını göz önünde bulundurmuştur ve herhangi bir semantik analizi yapılmamıştır. Çalışma İngilizce dilinde yapılmış olup, iki kişinin karşılıklı konuşmasını analiz ederek kişilerin cinsiyetini tahmin etmeye çalışmıştır. Bu çalışmada elde edilen sonuçlar şöyledir:

- erkek – erkek diyaloğu: %82,5,
- kadın – kadın diyaloğu: %87,5
- kadın-erkek diyaloğu: %72,5

Blog yazarının cinsiyetini belirlemek için Kobayashi vd. [20] her cinsiyete özgü olan kelimeleri ayırıp, DVM'yi kullanarak blog yazarlarını sınıflandırmıştır. Bu çalışmada veri kümesi olarak "Doblog" blog sunucusunda yayınlanan bloglar kullanılmıştır ve çalışma Japonca veriler üzerine yapılmıştır. Veri filtreleme aşamasında, kısa içerikli bloglar ve aynı seviyede erkeğe veya kadına özgü kelimeler kullanılan bloglar kümeden çıkarılmıştır. Blogların sınıf etiketleri için, blog sunucusuna kaydolurken, blog yazarlarının kendileri tarafından doldurdukları formlar kullanılmıştır. Bu çalışmada en iyi sonuç %90 doğrulukla veri kümesinin %84'ü kullanılarak elde edilmiştir.

Mukherjee ve Bing [21], blog yazarının cinsiyetini belirlemek için kelime kategorileri, n-gramlar ve konuşmanın bir parçasının (*part of speech*) öz niteliklerini kullanmanın yanında, sıralı örüntü madenciliği (*sequence pattern mining*) yaparak elde edilen konuşmanın değişken uzunluklu parçalarını da niteliklere eklemiştir. Ayrıca nitelik seçme yöntemlerinin bir karışımını da kullanmıştır. Bu çalışma ingilizce bloglar üzerinde yapılmıştır ve en iyi sonuç %88,56 elde edilmiştir.

Ranjendra Prasath [22], chat yazı stilindeki kelimelerin beraber kullanımını analiz ederek blog yazarının cinsiyetini ve yaşını aynı zamanda belirlemeye çalışmıştır.

Blog yazarının cinsiyeti ve/veya yaşını tahmin etmek için birçok çalışma daha yapılmıştır. Genel olarak bu çalışmalarda chat yazı stilindeki kelimlerin beraber kullanımı ve kullanıcıların alışkın oldukları yazı stili analiz edilmiştir [23, 24, 25, 26].

Sosyal medyada cinsiyet ve/veya yaş belirlemek için yapılan çalışmalarda ise Peersman vd. [27] Netlog adlı bir SAS'da yapılan sohbeti kullanarak, kullanıcının cinsiyetini ve yaşını tahmin etmeye çalışmıştır.

Burger vd. [28] ise Twitter sosyal medyasında sadece cinsiyet belirlemeye çalışmıştır, ancak bunun için kullanıcının bazı profil bilgilerini de kullanmıştır. Bu çalışmada kullanılan veri kümesi yaklaşık 147,000 farklı dili konuşan kişinin attığı 3,280,532 twit'ten oluşmaktadır. Öz nitelikler kümesini, karakter ve kelime n-gramlarının yanı sıra bazı profil bilgileri oluşturmaktadır. Elde edilen en iyi sonuç ise %92 doğru tahmindir.

Fink vd. [29], Twitter sosyal medyasında, cinsiyet tahmini için atılan twitleri incelemeye almıştır ve Burger vd.'nin [28] aksine herhangi bir profil bilgisini kullanmamıştır. Bu çalışmada f-ölçütü %80 ile en iyi sonuç elde edilmiştir. Doğruluk değeri hakkında herhangi bir bilgi sunulmamıştır.

Bamman vd. [30], 14000 Twitter kullanıcısının twitlerinden oluşan bir veri kümesiyle kişilerin cinsiyetlerini belirlemeye çalışmıştır. Kullanıcının cinsiyeti ve dilbilimsel stili arasındaki ilişki analiz edilmiştir. En iyi sonuç olarak %88 elde edilmiştir.

Deitrick vd. [31], akış tabanlı basit bir sinir ağı algoritması kullanarak Twitter'da cinsiyet belirlemeye çalışmıştır. Bu çalışmada en iyi sonuç olarak %98,51 oranında doğruluk elde edilmiştir.

Cheng vd. [4] Cinsiyet belirlemek için veri kümesi olarak "Reuters" haber grubu ve "Enron" e-posta kümesini kullanmıştır. Bu çalışmada, Erkek ve Kadın kullanıcıların yazı stilleri arasındaki farklılıkları belirleyip, cinsiyeti tahmin etmek için 545 adet psikolojik ve dilbilimsel öz nitelik hesaplanmıştır ve en iyi sonuç %85,1 oranında elde edilmiştir.

3. YAPILAN ÇALIŞMALAR, BULGULAR VE İRDELEME

Bu çalışmada Facebook kullanıcılarının Facebook sayfalarında yazdıkları yorumların analiziyle kişinin Cinsiyeti, Yaşı ve Eğitim düzeyini belirlemek için bir sistem geliştirilmiştir. Her veri madenciliği işleminde olduğu gibi ilk olarak veri toplama ve dönüştürme işlemi yapılmıştır. Bir sonraki bölümde verinin Facebook sayfalarından toplanma yöntemi bahsedilmiştir.

3.1. Veri Toplama

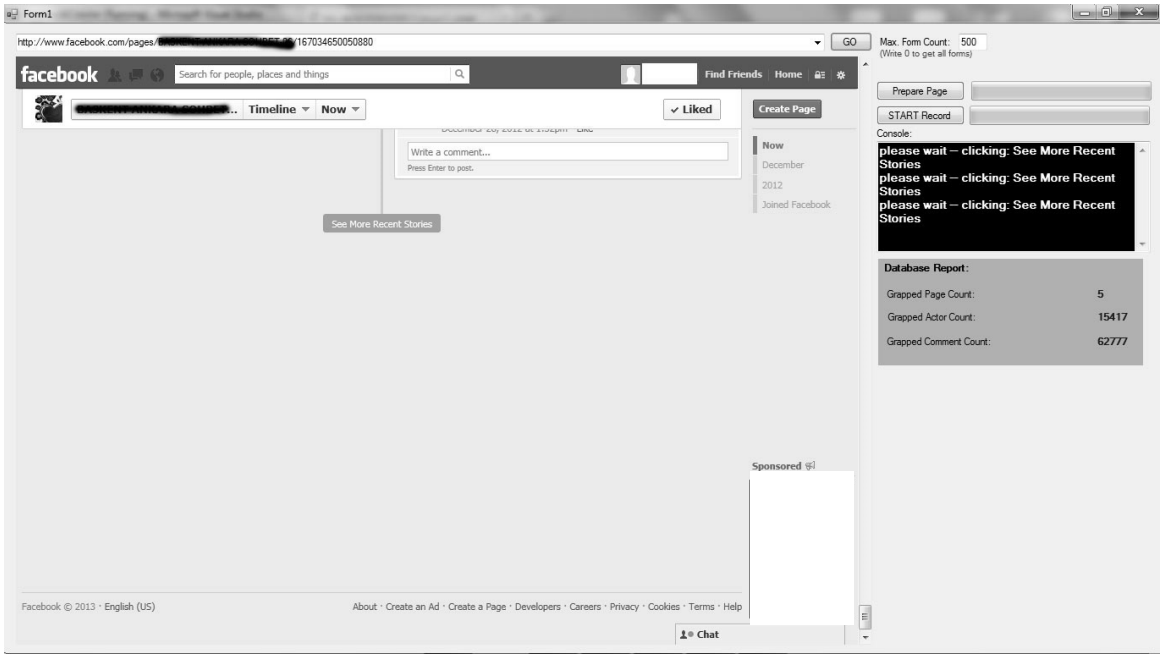
Türkçe Facebook sayfalarındaki yorumları toplayıp bir veritabanına eklemek için otomatik bir veri toplama böceği geliştirilmiştir. Bunun için öncelikle Facebook'ta rastgele önümüze çıkan Türkçe sayfalardan, paylaşımlarının altında daha çok yorum olduğu düşünülen 5 tanesi veri kaynağı olarak seçilmiştir.

Şekil 7'de geliştirilen böceğin çalışırken arayüzü görüntülenmektedir. Bu böcek 2 geçişte veriyi toplamaya çalışır; Sayfa hazırlama geçişi ve Veri toplama geçişi.

Facebook sayfaları iki şekilde mevcuttur: birincisi Zaman Tüneli şekli, ikincisi ise Facebook'un eskiden kullandığı klasik sayfa şeklidir. Aşağıda bahsedilen sayfa hazırlama ve veri toplama geçişleri her iki sayfa şeklinde de birçok benzerliğe sahip ve ayrıca Zaman Tüneli, klasik sayfa şeklinin güncellenmiş hali olduğundan, aynı böcek her iki türde de problemsiz olarak çalışabilmektedir. Geçişlerin detaylı açıklaması aşağıdadır:

3.1.1. Birinci Geçiş – Sayfa Hazırlama Geçişi

Facebook sitesinin yapısı nedeniyle sayfalardaki paylaşımlar tarihe göre yukardan aşağı doğru sıralanır. Facebook, sayfaların yüklenme süresini daha kısa tutmak için paylaşımların görünümünde bazı sınırlamalar koymuştur. Bu sınırlamalar şöyledir:



Şekil 7. Veri toplama böceği Facebook sayfasını indiriyor.

- En güncel paylaşımlar yüklenir: ilk kez sayfaya istek gönderildiğinde, Facebook sadece en güncel paylaşımları kullanıcının bilgisayarına yükler ve böylece büyük bir verinin kullanıcının bilgisayarına aktarılırken kaybolacak zamanı önler. Kullanıcı daha sonra, eğer daha fazla paylaşım görüntülemek isterse tarayıcının dikey çubuğunu aşağı doğru kaydırır, Facebook otomatik olarak daha fazla paylaşımı bilgisayarına indirmiş olur. Bu işlem otomatik veri toplama böceğinde aynen uygulanmıştır, böylece Facebook sayfasına ilk istek gönderilir, sayfanın tarayıcıda yükleme işleminin bitmesi beklenir, yükleme biter bitmez, yüklemenin bittiğine dair bir sinyal üretilir, böcek sinyali alarak kaydırma çubuğunu son aşamaya kaydırır ve bu işlem sayfanın tamamı indirilene kadar devam ettirilir.
- Daha az ilgi gören paylaşımlar kapalı tutuluyor: Eğer paylaşım miktarı çok fazlaysa, kaydırma çubuğu aşağı doğru çekilse bile daha az ilgi gören paylaşımlar (Beğenme sayısı, yorum sayısı ve paylaşım sayısı az olan paylaşımlar daha az ilgi gören paylaşım olarak kabul edilir.) indirilmiyor ve bu paylaşımları da indirmek için bir Şekil 8'de gösterilen "Yakınlardaki Diğer Haberleri Gör" butonunun tıklanması gerekmektedir. Bu tıklamanın gerçekleşmesi için otomatik böcek kaydırma çubuğunu aşağı doğru indirirken, bu butona rastladığı anda butona

tıklar ve böylece bir sonraki kaydırma aşamasına gelmeden önce tüm paylaşımları indirmiş olmaktadır.

Şekil 8. Kapatılan paylaşımları açmak için kullanılan buton

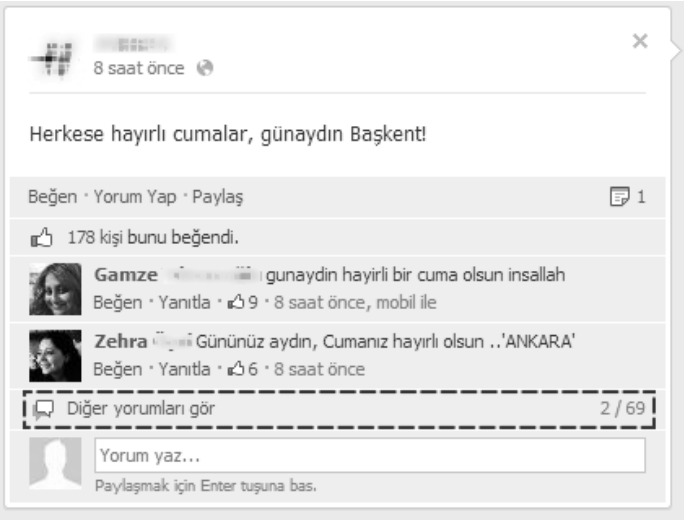
- Belirli bir sayının üzerindeki paylaşımlar kapalı tutuluyor: Paylaşım sayısı belirli bir sayıyı geçtiği zaman (bu sayı bilinmiyor) geriye kalan paylaşımlar iki şekilde kapalı tutuluyor; birincisi Şekil 9'da da görüldüğü üzere bir yıl için geçmiş paylaşımların kapatılması, ikincisi bir ay için geçmiş paylaşımların kapatılması. Her iki durum için Şekil 7'de gösterildiği gibi bir butonun tıklanması gerekmektedir. Ayrıca ikinci durum için buton, Şekil 7'deki buton tıklandıktan sonra görünebilir. Önceki bölümde olduğu gibi otomatik böcek kaydırma çubuğunu kaydırırken, karşılaştığında bu butonları tıklar. Sonuç olarak toplam iki geçiş aşamasında veriler elde edilebildiğinden, daha fazla sayıda geçiş basamağına gereksinimi önlemektedir.

Şekil 9. Yıl içerisinde ki tüm paylaşımları görme butonu

3.1.2. İkinci Geçiş – Veri Toplama Geçişi

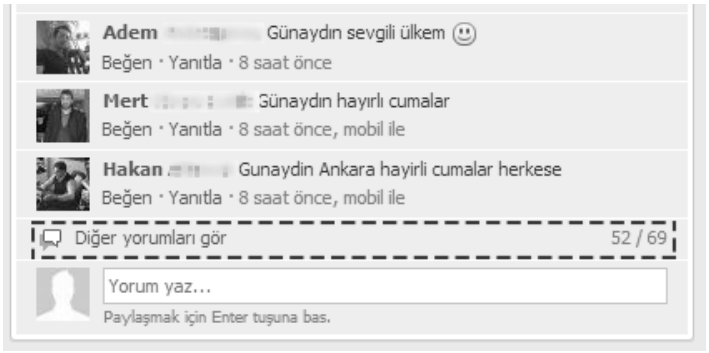
Facebook sayfalarındaki paylaşımların altında yapılan yorumlar, başlangıçta kapalıdır veya birkaç tanesi açıktır. Yorumları toplamadan önce tüm paylaşımlardaki

yorumların görünmesi için Şekil 10'da görünen "Diğer yorumları gör" butonuna tıklamak gerekmektedir. Bu butona her tıklandığında 50 adet daha fazla yorumu gösterir. Şekil 11'de, Şekil 10'da gösterilen aynı paylaşımın "Diğer yorumları gör" butonuna bir kere tıklandıktan sonraki durumu gösterilmiştir. Tüm yorumların gösterilmesi için aynı işlem, bahsedilen buton göründüğü sürece yapılması gerekmektedir.



Şekil 10. Yorumların hepsi görünmüyor.

Eğer bir paylaşımın yorumlarının sayısı 50'den az ise; bahsi geçen butonda, görünen yorumların dışında görünmeyen kaç yorum kalmış ise o sayı, butonun yazısının başına eklenerek gösterilir. (Şekil 12)

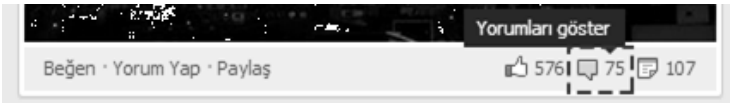


Şekil 11. "Diğer yorumlar" butonu bir kere tıklanmıştır.

Buradaki istisnai durum ise, eski paylaşımlarda başlangıçta hiçbir yorum görünmemektedir ve bahsedilen buton farklı isimde ve farklı şekilde mevcuttur. Bu durum Şekil 13'te gösterilmiştir. Otomatik böceğin her iki şekildeki paylaşımlardan yorumları toplaması gerekmektir. Bunun için otomatik böcek her paylaşımda ilk olarak Şekil 10'da ve Şekil 11'de gösterilen butonları bulmaya çalışır ve bulduğu anda tıklayarak yorumların açılmasını sağlar, eğer bulamazsa eski paylaşım olduğunu düşünerek Şekil 13'deki butonu bulmaya çalışır ve başarılı olursa yorumların açılması için tıklar. Eğer yorumları gösteren butonların hiçbirini bulamazsa paylaşımı yorumsuz sayıp bir sonraki paylaşıma geçer.

Otomatik böcek tüm yorumları toplayabilmek için yukarıda bahsedilen işlemleri, sayfadaki tüm yorumlar gösterilene kadar devam eder.

Şekil 12. Yorum sayısı 50'den azdır.



Şekil 13. Eski paylaşımlarda yorumların gösterilmesini sağlayan buton

3.1.2.1.Verilerin Veritabanına Kaydolması

Facebook sayfalarındaki her paylaşım, yazılımsal açıdan bir Form öğesidir. Buna göre otomatik böcek, her form için tüm yorumların önceki bölümde bahsedilen yöntemle indirilmesini sağlar. Daha sonra o formun yorumlarını kaydetmek üzre, her Facebook kullanıcısı farklı birer ID ile kayıt edilip, sayfalardaki yorumları bir veritabanına eklenir. Kullanıcıların kişisel haklarının korunması nedeniyle kişiler sadece ID ile veritabanına eklenmiştir. İsim ya da herhangi başka kişisel bilgiler kaydolmamıştır. Öte yandan eğer bir veri, sahibi tarafından umuma açık bir şekilde ve arama motorlarıyla aranabilir halde

internette yayınlanmışsa onu kullanmanın herhangi bir kişisel hak ihlali oluşturması söz konusu değildir.

Uzun metin içeren yorumların sadece bir kısmı gözükmektedir ve sonuna "Daha fazlasını gör" link butonu eklenmiştir. Otomatik böcek verileri veritabanına eklemeden önce tüm yorumlarda eğer basedilen link buton mevcutsa onu tıklayarak yazının tamamının gözükmesini sağlayıp, yorumları veritabanına eklemektedir.

Beş adet Türkçe Facebook sayfasından toplanan yorum sayısı ve kaydedilen kullanıcı sayısı Tablo 8'de gösterilmiştir.

Tablo 8. Veri kümesindeki kullanıcı ve onlara ait yorum sayısı

Toplam Facebook Kullanıcısı	15417
Toplam Yorum Sayısı	62777

3.2. Veri Filtreleme

Verileri etiketlemeye sunmadan önce daha doğru ve etkili uygulama için gereksiz ve değersiz yorumların çıkarılması gerekmektedir. Bunun için birinci aşamada verideki tüm spam ve link içeren yorumlar filtrelenmiştir. Facebook yorumları iki şekilde link içerebilir: 1- Bir başka facebook sayfası veya profiline, 2- dış bağlantı linklerine. Her iki link şekli de basitçe veri kümesinden kaldırılmıştır.

Veri filtrelemenin ikinci aşamasında düzenli ifadeler (*Regular Expression*) kullanılarak belirli bazı reklam yorumları filtrelenmiştir. Örneğin; zaman tünelinden kurtulmak için yapılan reklam yorumlarını veri kümesinden elemek için "Zaman Tüneli" sözcükleri geçen tüm yorumlar silinmiştir.

Veri filtrelemenin üçüncü aşamasında aynı yorumu defalarca tekrarlayan kullanıcı için bunlardan sadece biri veri kümesinde kalacak şekilde diğerleri elenmiştir.

Son olarak en az iki yorumu bulunmayan veya yorumlarında toplamda en az 50 karakter olmayan tüm Facebook kullanıcıları çıkarılmıştır. Böylece filtreleme bittiğinde veri kümesindeki kalan kullanıcı ve yorum sayısı Tablo 9'da gösterilmiştir.

Tablo 9. Filtreleme işleminden sonra veri kümesinde kalan Kullanıcı ve onlara ait yorum Sayısı

Toplam Facebook Kullanıcısı	4151
Toplam Yorum Sayısı	50096

3.3. Veri Etiketleme

Facebook kullanıcılarını Cinsiyet, Eğitim düzeyi ve Yaş kategorilerinde etiketlemek için özel bir web sitesi geliştirilerek, veri kümesi online bir veritabanına eklenmiştir. Online Veritabanında üç tablo oluşturulmuştur, bu üç tablo Şekil 14'te gösterilmiştir.

Şekil 14'te göründüğü üzere, kullanıcıları etiketleyecek jurinin kullanıcı adı ve şifresi "tblJuri" tablosunda tutulmaktadır. Juri bu bilgileri kullanarak web sitesine giriş yaparak kullanıcıları etiketlemektedir. Tablo "tblYorumlar", Facebook kullanıcılarının filtrelenmiş yorumlarını içermektedir. "tblKisiler" tablosunda, her kişi için juri elemanlarının cinsiyet, yaş ve eğitim düzeyi tahminleri tutulmaktadır.

Şekil 14. Etiketleme için hazırlanan web sitesinde kullanılan veritabanı tabloları

Geliştirlen web sitesi toplamda iki arayüzden oluşmaktadır. Birinci arayüz, giriş sayfası ve juri elemanlarının üye olmaları için bir formdan oluşmaktadır.

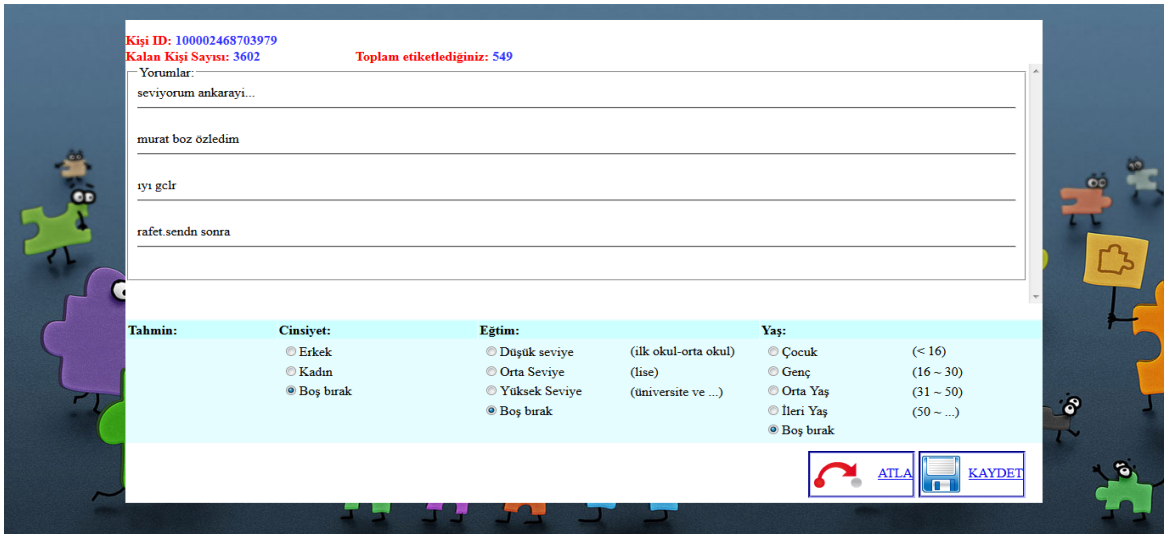
Birinci arayüz Şekil 15'te gösterilmiştir. Juri elemanı eğer daha önce üye olmuşsa kullanıcı adı ve şifresini girdikten sonra başla butonuna tıklayarak Facebook kullanıcılarını etiketleyeceği sayfaya (ikinci arayüz) yönlendiriliyor. Eğer daha önce üye olmamışsa yeni

üye kayıt formunu doldurduktan sonra üyeliğini tamamlar ve tekrar kullanıcı adı ve şifresini kullanarak Facebook kullanıcılarını etiketlemeye başlayabilir. Toplamda 20 juri elemanı bu web sitesine kayıt olmuştur ve Facebook kullanıcılarını etiketlemiştir. Bu çalışma Türkçe Facebook sayfalarında yapıldığından tüm juri elemanları, ana dili Türkçe olan kişilerden seçilmiştir. Böylece juri elemanlarının, yorumları anlayarak tahminlerini kaydettiklerinden emin olunmuştur.

İkinci arayüz Şekil 16'da gösterilmiştir. Juri elemanları, ikinci arayüzde random olarak seçilen bir kullanıcının yorumlarını "Yorumlar" panelinde okuduktan sonra, tahminlerini Cinsiyet kategorisi için {Erkek, Kadın}, Eğitim kategorisi için {düşük seviye (ilk okul, orta okul), orta seviye (lise), yüksek seviye (üniversite ve daha fazla)} ve Yaş kategorisi için {<16 (çocuk), ≥16 ve <30 (genç), ≥30 ve <50 (orta yaş), ≥50 (ileri yaş)} seçeneklerinden seçip, tahmin edemedikleri kategorileri ise "boş bırak" seçeneğini işaretleyerek kaydetmişlerdir. Hiçbir kategori için tahmin yürütemedikleri durumlarda "Atla" butonuna tıklayarak diğer Facebook kullanıcısına geçmişlerdir. Juri elemanları iki ay boyunca etiketleme işlemine devam etmişlerdir.

Şekil 15. Etiketleme web sitesine giriş arayüzü

Her üç gruptaki (Cinsiyet, Eğitim düzeyi ve Yaş) sınıf etiketinin kesinleştirilmesi için, tüm Facebook kullanıcılarının yorumları, minimum 3 juri elemanı tarafından en az bir kategori için tahmin edilmelidir. 4151 Facebook kullanıcısının yorumları, tüm juri



Şekil 16. Etiketleme için geliştirilen web sitesinin ikinci arayüzü

elemanlarının karşılarına random olarak çıkıp, okudukları yorum üzerine tahminlerini kaydetmesi durumunda, bir kullanıcının en az 3 juri elemanı tarafından etiketlenme olasılığı çok düşük bir rakam olacaktır. Kısa zaman içerisinde tüm Facebook kullanıcıları etiketlenemez. Dolayısıyla juride en aktif olan elemanı 'lider juri' olarak tanımlıyoruz. Lider juri, 4151 kullanıcıdan seçilen random yorumlar için, tahminlerini diğerlerinden önce kaydetmektedir. Öbür juri elemanları ise 4151 kullanıcıyı etiketlemek yerine, random olarak önüne çıkan ve sadece lider jurinin daha önce kaydetmiş olduğu yorumları okuyup tahminlerini işaretlemişlerdir. Bu yöntem ile etiketleme işlemi 2 ay sonra sona erdiğinde, 550 Facebook kullanıcısı en az 3 juri elemanı tarafından etiketlenmiş ve çalışma sadece bu 550 kişinin yorumları üzerinde devam etmiştir. Bahsedilen 550 kişiye ait yorum sayısı ise 9354'tür.

Facebook Kullanıcılarının etiketlenmesinin veritabanında kaydolma şekli için bir örnek Tablo 10'da gösterilmiştir. Bu örnekte görüldüğü üzere bir Facebook kullanıcısı, 7 juri üyesi tarafından etiketlenmiştir. Bu kullanıcıyı, Cinsiyet için 7 juri elemanından 6'sı Erkek olarak işaretlemiştir. Eğitim seviyesi için, işaretleme yapan 3 juri elemanından ikisi Düşük seviye olarak, Yaş kategorisindeyse tahmin belirten 4 juri elemanından 3'ü genç olarak işaretlemiştir.

Bu veriler etiketleme bittiğinde tekrar lokal veritabanına aktarılmıştır. Lokal veritabanına aktarılırken, juri elemanları tarafından yapılan tahminler her kişi için

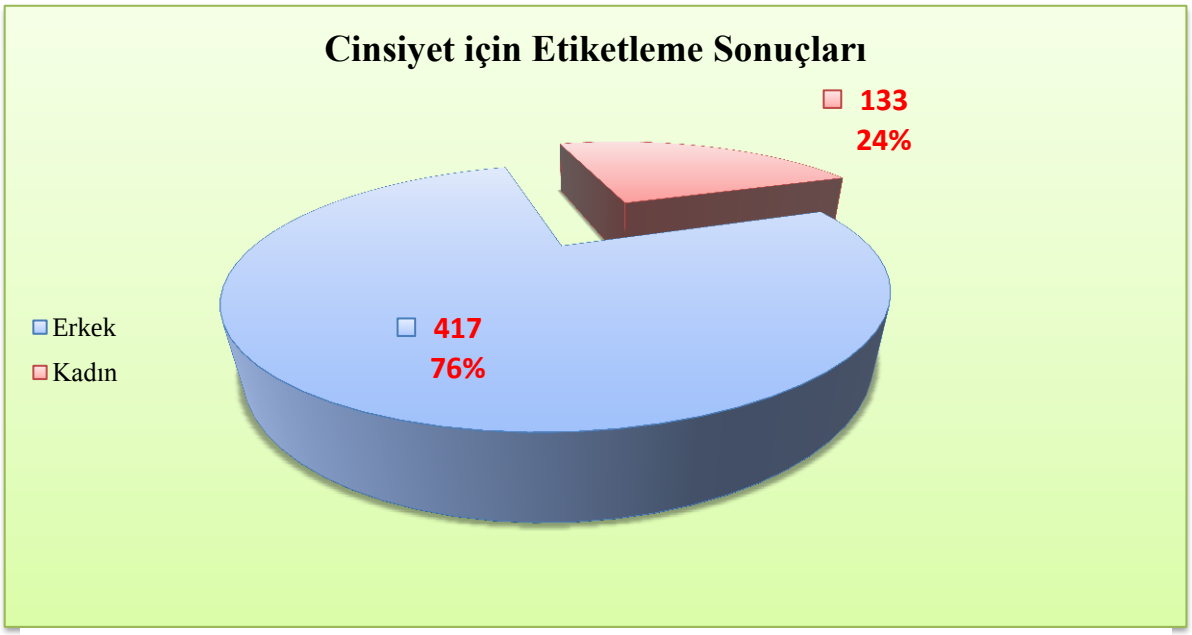
oylanarak yeni bir tabloya eklenmiştir. Yeni tabloda artık her Facebook kullanıcısı için, her kategoride oylama sonucundan çıkan sınıf etiketleri kaydolmuştur.

Tablo 10. Veri etiketlemenin veritabanına kaydolma şekli

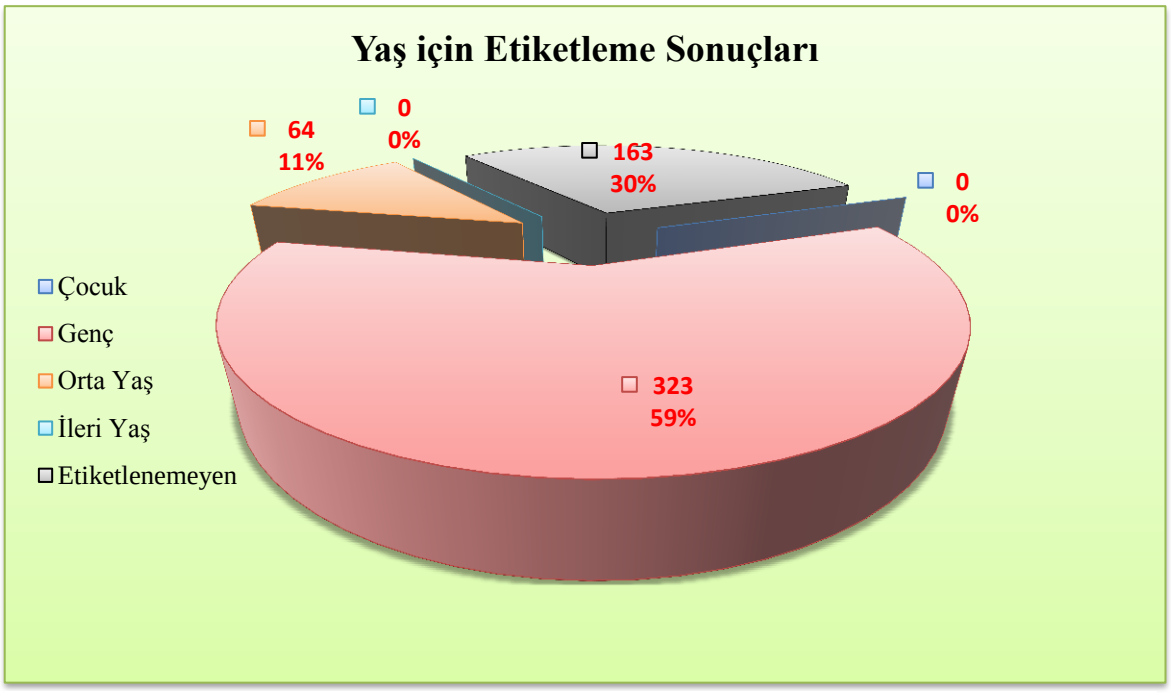
JuriID	KisiID	Cinsiyet	Egitim	Yas
21	563484083	Kadın	Düşük seviye	Çocuk
10	563484083	Erkek	Düşük seviye	Genç
6	563484083	Erkek		
8	563484083	Erkek		
4	563484083	Erkek	Yüksek Seviye	Genç
3	563484083	Erkek		Genç

Juri tarafından etiketleme sonuçları Cinsiyet için Şekil 17'de, Yaş için Şekil 18'de ve Eğitim düzeyi için ise Şekil 19'da gösterilmiştir. Ayrıca kişilerin Cinsiyetinin, yaşa ve Eğitime bağlı olarak dağılımı sırasıyla Şekil 20'de ve Şekil 21'de gösterilmiştir.

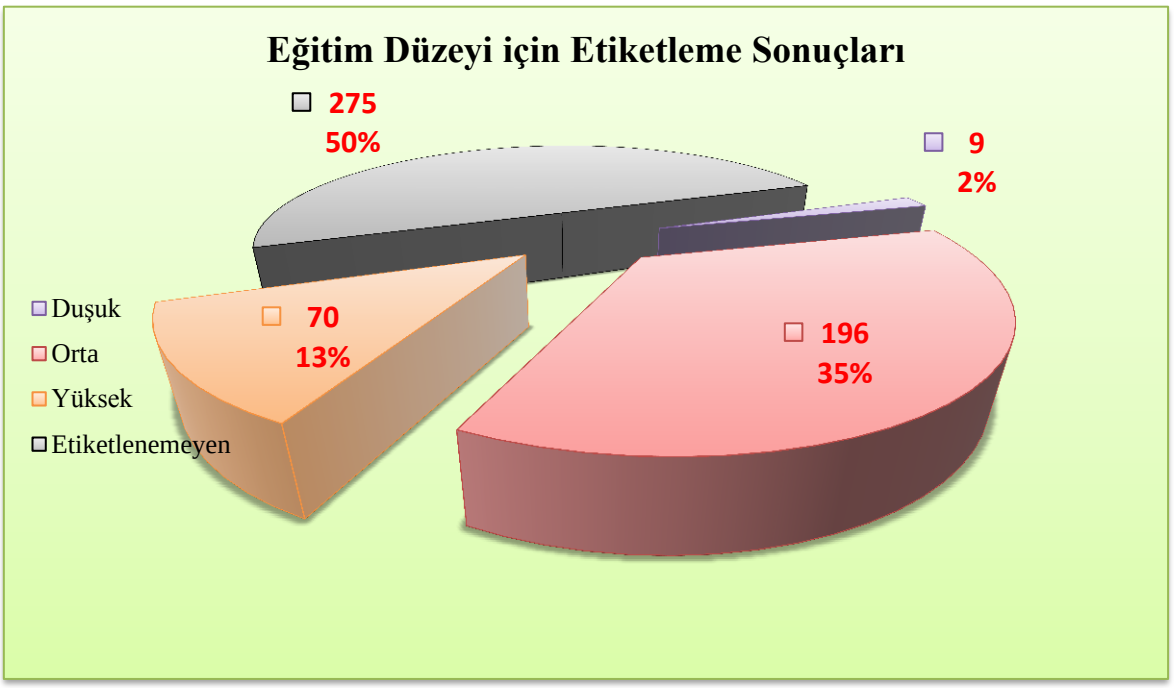
Şekil 17'de görüldüğü gibi kullanıcıların yaklaşık %76'ı erkektir, Şekil 18'de görüldüğü üzere çocuk ve ileri yaş için herhangi bir Facebook kullanıcısı etiketlenmemiştir. Şekil 19'da ise kullanıcıların %50'si eğitim için etiketlenememiştir.



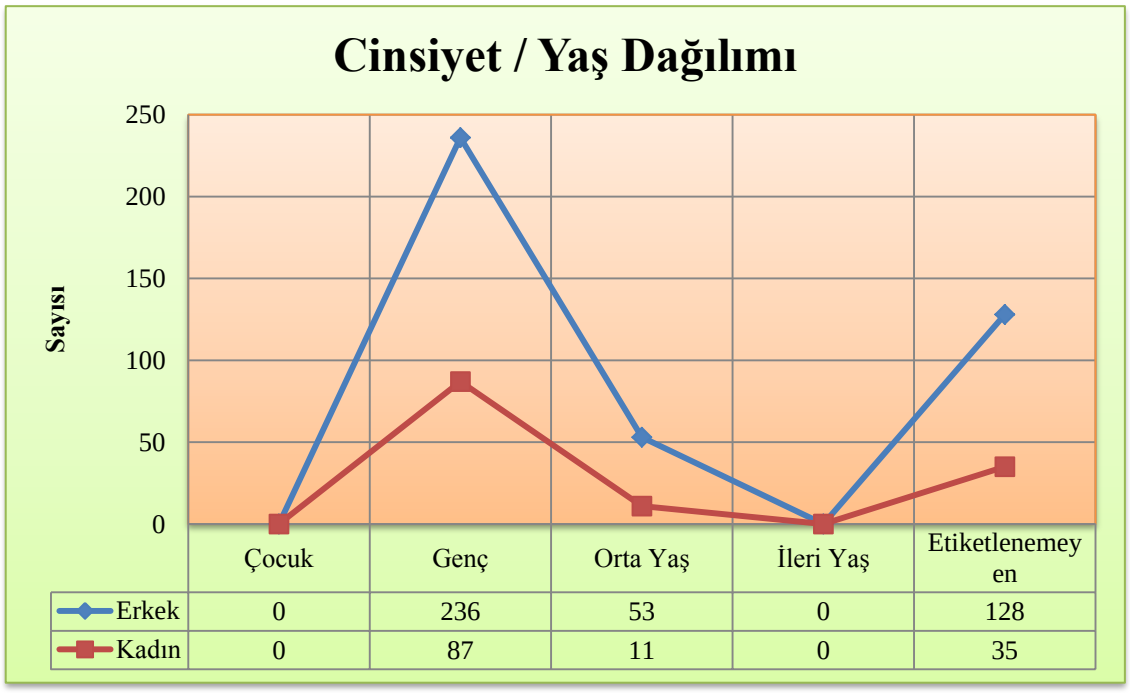
Şekil 17. Cinsiyet için etiketleme sonuçları



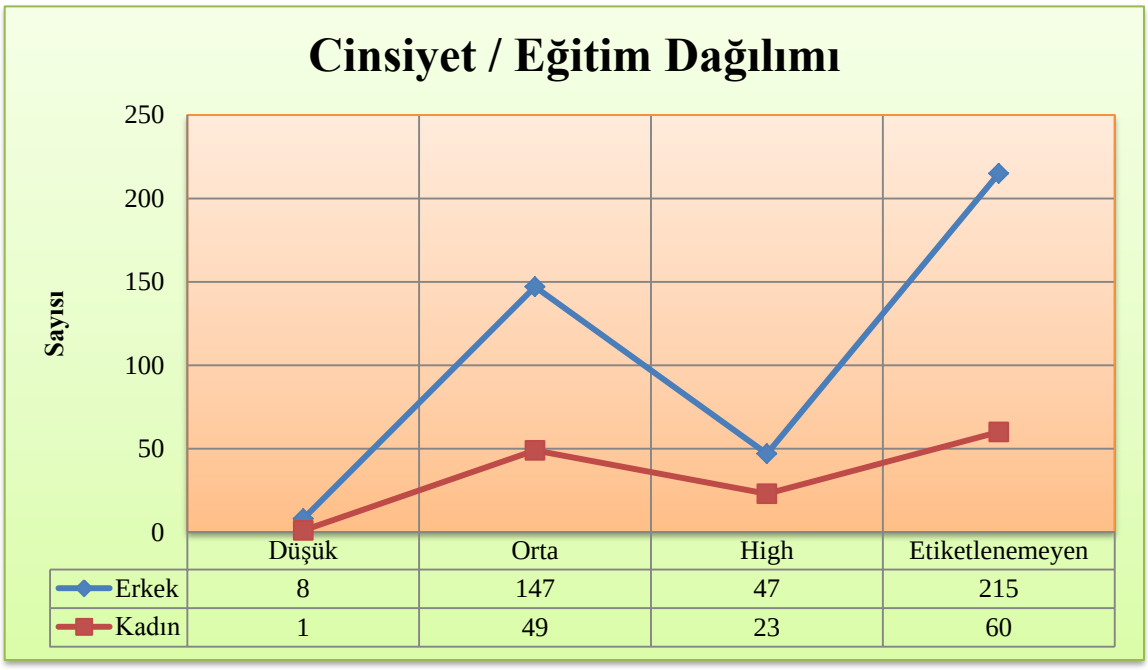
Şekil 18. Yaş için etiketleme sonuçları



Şekil 19. Eğitim düzeyi için etiketleme sonuçları



Şekil 20. Cinsiyet /Yaş dağılımı



Şekil 21. Cinsiyet / Eğitim dağılımı

Şekil 20'de görüldüğü üzere genç olarak etiketlenen kullanıcılardan 236 kişi erkek ve 87 kişi kadın olarak etiketlenmiştir. Ayrıca orta yaş olarak etiketlenenlerin sadece 11 tanesi kadındır.

Şekil 21'de görüldüğü üzere eğitim durumu düşük seviye olarak etiketlenen Facebook kullanıcıları arasından sadece 1 kişi kadındır. Ayrıca etiketlenemeyenlerin %86'sı erkektir.

3.4. Veri Önişleme

Metin halindeki veriyi yapılandırmak için metindeki terimler ayrıştırılmalıdır. Terimler, terimlerin tekrarlanma sayısı, terimin her dökümanda[2] geçme sayısı ve terimin yorumdaki indeksini kaydetmek için bir veritabanı oluşturulmuştur. Veritabanı iki tablodan oluşmaktadır. Tablo 11 ve Tablo 12'de bu iki tablo için birer örnek verilmiştir.

Tablo 11. Terimler tablosu

TermID	Term	Frequency
1	:)	1532
2	Slm	1250

Tablo 12. Kişiler ve terimleri kullanma tablosu

docID	termID	yorumID	Index
11205181	1	510	5
150849994	2	1053	0

Tablo 11'de ayrıştırılan her terim birer ID ve tüm dökümanlarda geçme sayısıyla kaydedilir. Tablo 12'de ise 'x' ID'li terimin, hangi kişi tarafından, hangi yorumda kullanıldığı bilgisi ve terimin yorumdaki indeksi kaydediliyor.

[2] Bu çalışmada, her kullanıcının tüm yorumları tek bir döküman olarak düşünülmüştür.

Yorumlardaki terimler çıkarıldığında önişleme ve temizleme yapılmadan, toplam 19888 terim bulunmaktadır. Ancak terimleri veritabanına eklemeden önce imla kontrolü, kök bulma ve işaretlerden arındırma gibi bazı metin madenciliği ön işleme adımları uygulanmalıdır.

3.4.1. Türkçenin Sohbet (*Chat*) Dilinde Karşılaşılan Sorunları Giderme

3.4.1.1. Alternatif Karakterler

Türkçenin chat versiyonundaki problemleri gidermek için birinci aşamada 2.5.1.3'te bahsedilen alternatif karakterlerden 'w' ve 'q' hariç, diğer tüm karakterler orjinal eşdeğerleriyle değiştirildi. Böylece alternatif harf kullanılan kelimeler, orjinal karakter kullanılanlarla aynı sayılmış oldu. 'w' ve 'q' karakterlerinin bazı kategoriler için bilgi verebileceği ve önemli nitelik olabilme ihtimali düşünüldüğü için bu karakterler incelenmek üzere değiştirilmemiştir, bu konuda daha detaylı bilgi, 3.4.3.2 bölümünde yer almıştır.

3.4.1.2.Türkçeye Özgü Harfler Yerine İngilizce Harfler Kullanmak

Yorumların analiziyle beraber Türkçe karakterler yerine İngilizce eşdeğerinin aşırı derecede kullanıldığı görülmektedir, bunları Türkçe eşdeğerlerine çevirmek ise semantik analiz yapmadan imkansızdır. Bu problemi basit bir şekilde ve semantik analize girmeden, düşük maliyetle çözmenin en mantıklı yolu, geriye kalan Türkçe karakterleri de ingilizce eşdeğerlerine dönüştürmektir. Böylece farklı karakterlerle kullanılan aynı sözcükler, veritabanına eklenirken aynısı olarak sayılacaktır.

Bu işlemin az da olsa bazı hataları oluşturmaktadır, örneğin "Söyle" ve "Şöyle" kelimleri ayrı ayrı kelimeler olmasına rağmen bu işlem yapıldıktan sonra ikisi de "Soyle" olarak kaydedilecektir. Ancak bu işlemin yanlış olarak değiştirdiği kelimelerin sayısı, bu işlemi uygulamadan önceki duruma karşın (aynı kelimelerin ayrı ayrı terim olarak

kullanılması) büyük bir derecede azdır ve bu işlem daha iyi sonuçlar almak için vazgeçilemez işlemlerden biridir.

3.4.1.3. Alternatif Yüz İfadeleri

Tüm alternatif yüz ifadeleri tek bir formata dönüştürülmüştür. Tekrarlanan son karakterler ise sınıflandırmada etkili olacağı düşünüldüğünden, en fazla iki tekrara kadar indirilmiştir. Örneğin eğer yüz ifadesi ":DDDDDDD" kullanılmışsa, ilk harf ve ikinci tekrar kalacak şekilde, geriye kalan tekrarlar silinir. Yani ":DD" olarak kaydedilmiştir. Böylece farklı sayılardaki tekrarlamalar tek bir formda veritabanına kaydedilmiştir.

3.4.1.4. Bazı Harflerin Tekrarı

"Allah", "İnşallah" vb. gibi kelimelerde görüldüğü üzre aynı harf tekrarlı biçimde kullanılmıştır ancak diğer kelimelerde tekrarlama sayısı 2'den farklı olduğu için aynı anlamı taşıyan kelimeler veritabanına farklı olarak kayedilebilir. Bu nedenle sadece iki kereden fazla tekrarlanan karakterler silinmiştir. Örneğin "Arkadaşlarrrrrr" kelimesi "Arkadaşlarr" olarak kaydolmuştur.

3.4.1.5.Bir Kelime İçin Farklı Formlar

Bir kelime için farklı formlardan kurtulmak amacıyla fiillerin kök bulma yöntemine başvurulmuştur. Diğer kelimeler için ise eklerden arındırmakla bu problemle başa çıkılmıştır.

Türkçe sondan eklemeli bir dil olduğu için [32] kök bulma ve eklerden arındırma işlemi, basit bir ekleri silme yöntemiyle uygulanmıştır. Bunun için iki ekler listesi düzenlenmiştir. Birinci ekler listesi çoğul eki gibi tüm kelimlere getirilebilecek eklerden oluşmaktadır. Bu liste EK 1'de verilmiştir. İkinci ekler listesi ise fiillere eklenebilecek,

genel olarak zaman eklerinin farklı formatlarından oluşmaktadır. İkinci liste Ek 2'de verilmiştir.

Kelimeler daha önce bahsedilen problemlerden arındırıldıktan sonra, ilk olarak eğer kelimenin sonunda birinci ekler listesinden bir ek var ise kelimenin sonundan çıkarılmıştır; daha sonra kelimenin sonunda ikinci ekler listesinden bir ek var ise, bu ek kelimenin sonundan silinerek yerine "mk" eki eklenmiştir. Örneğin "Gidiom" kelimesinden "iom" eki silinerek ve "mk" eklenerek "gidmk" olarak değiştirilmiştir.

3.4.1.6. Sesli Harfler

Kullanıcıların büyük bir bölümü, bazı sesli harfleri silerek kısaltma yapmıştır, bu da aynı anlama gelen kelimelerin birkaç farklı şekilde veri kümesinde mevcut olmasına sebebiyet vermiştir. Bu farklılıkları gidermek için tüm kelimelerde sesli harfler silinerek sadece sessiz harfler kalmıştır.

Kelimeleri sesli harflerden arındırmak, bazı farklı anlama gelen kelimelerin, aynı kelimeymiş gibi veritabanına kaydolmasına sebep olmuştur. Örneğin "Geç" ve "Güç" kelimeleri "Gç" olarak kaydolmuştur. Ancak bu duruma maruz kalan kelimlerin sayısı bu işlem yapılmadan önceki farklılıkların sayısının yanında önemsenemeyecek kadar azdır.

3.4.2. Değersiz Kelimelerden Arındırma

Bazı kelimler genel ve ortak olarak kullanılan kelimelerdir ve tekrarlanma sayısı genelde çok fazladır. Ancak bu kelimeler sınıf etiketi hakkında herhangi bir bilgi vermemektedir, örneğin "ve", "veya", "ne" vb. kelimeler. Bu kelimlerin listesi Ek 3'te verilmiştir. Değersiz kelimeler yorumlardan çıkarılarak, nitelik sayısı daha az tutulmaya çalışılmıştır.

3.4.3. Nitelikler

İmla kontrol ve düzeltme, değersiz kelimelerin silinmesi [33], fiiller için kök bulma işlemleri [32, 34, 35] ve önceki bölümde bahsedilen problemler giderildikten sonra farklı terimlerin sayısı 9565'e inmiştir. Bu terimlerin her biri nitelik olarak kullanılacaktır. Veri kümesinde en sık kullanılan 20 terim Tablo 13'te verilmiştir. Daha doğru sonuçlar üretilmesi için n-gramlar ve bazı yazım tarzı nitelikler de nitelikler kümesine eklenmiştir. Aşağıda eklenen n-gramlar ve yazım tarzı nitelikler hakkında bilgi verilmiştir.

Tablo 13. Veri kümesinde en sık kullanılan 20 terim

Sıra	Terim	Kullanma sıklığı
1	:)	1520
2	ben	1140
3	sen	621
4	:D	588
5	selam	514
6	bir	423
7	Ankara	418
8	kardelen	406
9	çok	403
10	iyi	395
11	Olmak	359
12	Abla	346
13	gelmek	278
14	olsun	257
15	güzel	243
16	arkadaş	234
17	demek	224
18	benim	222
19	:(	195
20	beni	189

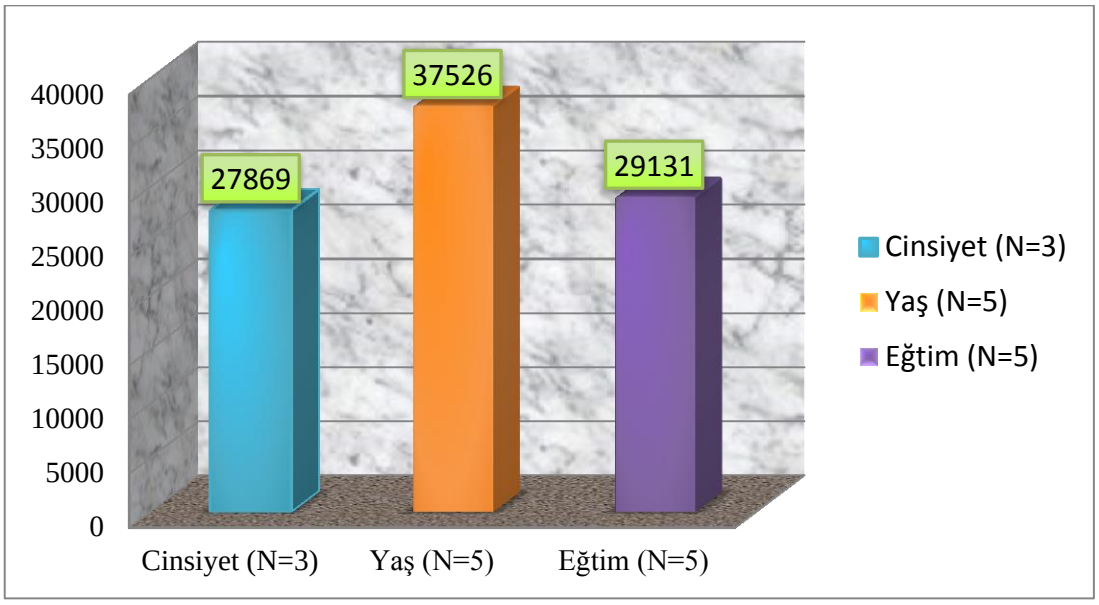
3.4.3.1. Terim N-Gram'ları

Bu çalışmada sadece terim n-gram'ları kullanılmıştır ve n sayısı da çıkan sonuçların, hesaplama karmaşıklığına değip değmemesine göre seçilmiştir. Başka bir değişle elde

edilen sonuçların, hesaplamalardaki karmaşıklığın oranına kıyasla kabul edilebilir derecede olup olmadığına bağlı olarak, n sayısı deneysel biçimde belirlenmiştir. Ayrıca N=n, n-gramlarını üretmek için, n=n-1, n=n-2 ... n=2, n-gramları da üretilmiştir. Örneğin eğer n=5, n-gramlar (5-gram) üretilecekse, n=4 (4-gram), n=3 (trigram) ve n=2 (bigram)'da hesaplanarak niteliklere eklenmiştir.

Şekil 22'de her kategori için seçilen n sayısıyla üretilen n-gramların eklenmesiyle toplam nitelik sayısı gösterilmiştir.

Bu çalışmada sadece terim n-gram'ları kullanılmıştır ve n sayısı sonuçlardaki etki ve hesaplama karmaşıklığına göre seçilmiştir başka bir değişle hesaplamadaki karmaşıklığın yükselmesi, çıkan sonuca değip değmediğine bağlı olarak n sayısı deneysel seçilmiştir.



Şekil 22. Cinsiyet, yaş ve eğitim kategorileri için üretilen n-gram'ların eklenmesiyle her kategori için nitelik sayısı

3.4.3.2. Yazı Stiline Bağlı Nitelikler

Facebook kullanıcılarının yazım tarzlarının, her kategoride sınıf etiketindeki etkisini analiz etmek için 6 yazım stili niteliği hesaplanarak nitelikler kümesine eklenmiştir. Bu yazım stilleri ve hesaplama teknikleri aşağıda verilmiştir.

- Ortalama Terim Uzunuğu (OTU)

Bir kullanıcının yorumlarda yazdığı terimlerin karakter bazındaki ortalama uzunluğunu gösterir. Bu nitelik, yorumlarında çok aşırı kısaltma yapan kullanıcıların sınıf etiketine bir etkisi olup olmadığını ölçmek için hesaplanmıştır. Bu niteliği hesaplamak için Formül 14 kullanılır.

$$OTU = \frac{Terimlerin\ Uzunluklarının\ Toplamı}{Toplam\ Terim\ Sayısı} \quad (14)$$

Formül 14'te de görüldüğü üzere, bir kullanıcının kullandığı tüm terimlerin uzunluklarının toplamının, kullandığı toplam terim sayısına oranı, ortalama terim uzunluğunu üretmektedir.

Bu niteliğin sınıf etiketine etkileri, cinsiyet için Şekil 23'te, yaş için Şekil 24'te ve eğitim düzeyi için Şekil 25'te gösterilmiştir.

Bu şekiller ve bu bölümdeki diğer şekiller bir "histogram" grafiği ile bir ya da birden fazla "çizge" grafiğinden oluşmaktadırlar. Histogram (çubuk) grafiği belirli bir niteliğin dağılımını göstermektedir. Yatay eksen ise değer aralıkları kümesini göstermektedir. Sol taraftaki dikey eksen, histogram grafiğine uygulanarak Facebook kullanıcılarının her hangi bir değer aralığındaki yüzdesini göstermektedir. Sağ taraftaki dikey eksen ise çizge graflara uygulanarak, Facebook kullanıcılarının yüzde kaçının ilgili özelliğe sahip olarak değerlendirileceğini göstermektedir.

Cinsiyet ve yaş kategorilerinde sınıf etiketi ikili olduğundan[3] sadece daha az örneğe sahip olan sınıf etiketi için çizge grafiği çizilmiştir, bunun simetrisi de diğer sınıf etiketinin çizge grafiğidir.

Yatay eksendeki değerler, min-max normalleştirme fomülüyle (Formül 15) 0 ile 1 aralığına normalleştirilmiştir.

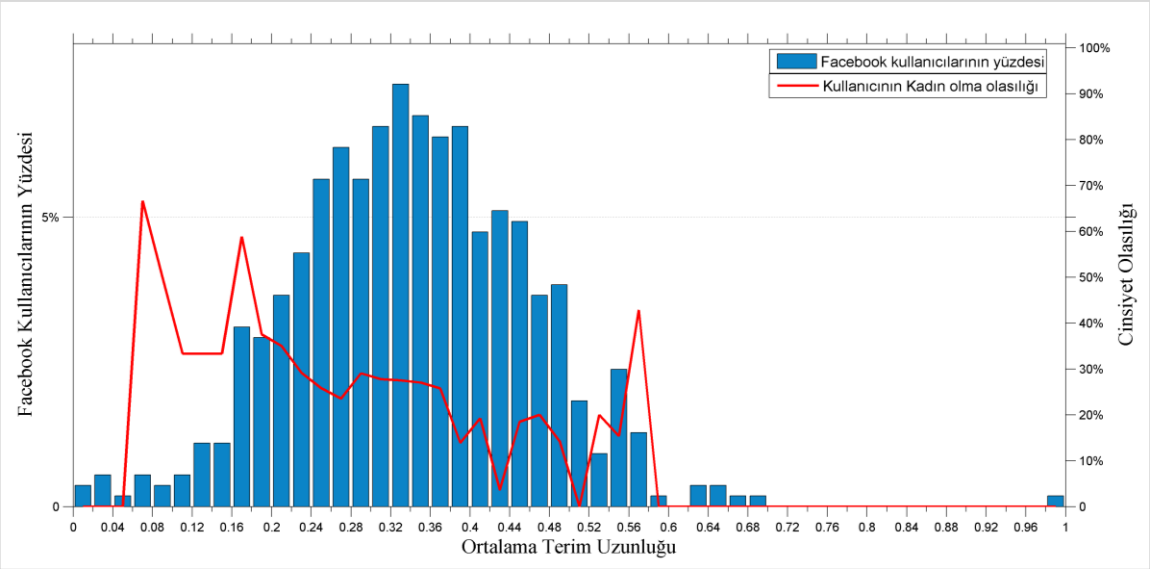
[3] Yaş kategorisinde çocuk ve ileri yaş için herhangi bir örnek juri tarafından etiketlenemediği için bu kategoride sadece genç ve orta yaş sınıf etiketleri kalmıştır ve bu nedenle bu kategoride cinsiyet gibi ikili sınıf etiketine sahiptir.

$$normal\ değer = \frac{x - min}{\max - min} \quad (15)$$

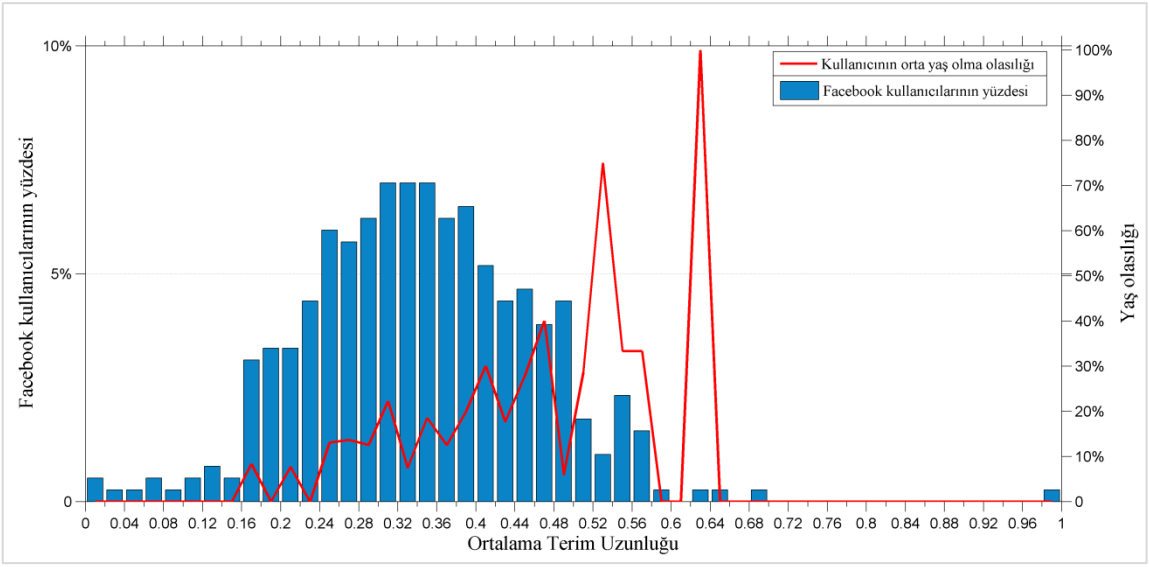
Yukarıdaki Formülde x, niteliğin gerçek değerini; min ve max ise aynı nitelikteki sırasıyla en küçük ve en büyük değerleri göstermektedir.

Şekil 23'te ortalama terim uzunluğu, cinsiyete göre gösterilmiştir. Burada yorumları, belirli bir ortalama terim uzunluğuna sahip olan facebook kullanıcılarının, yüzde kaçının kadın olma olasılığını kırmızı çizge grafiği göstermektedir. Örneğin ortalama terim uzunlukları [0.06 – 0.08] aralığındaolan tüm Facebook kullanıcılarının yaklaşık %65'i kadındır.

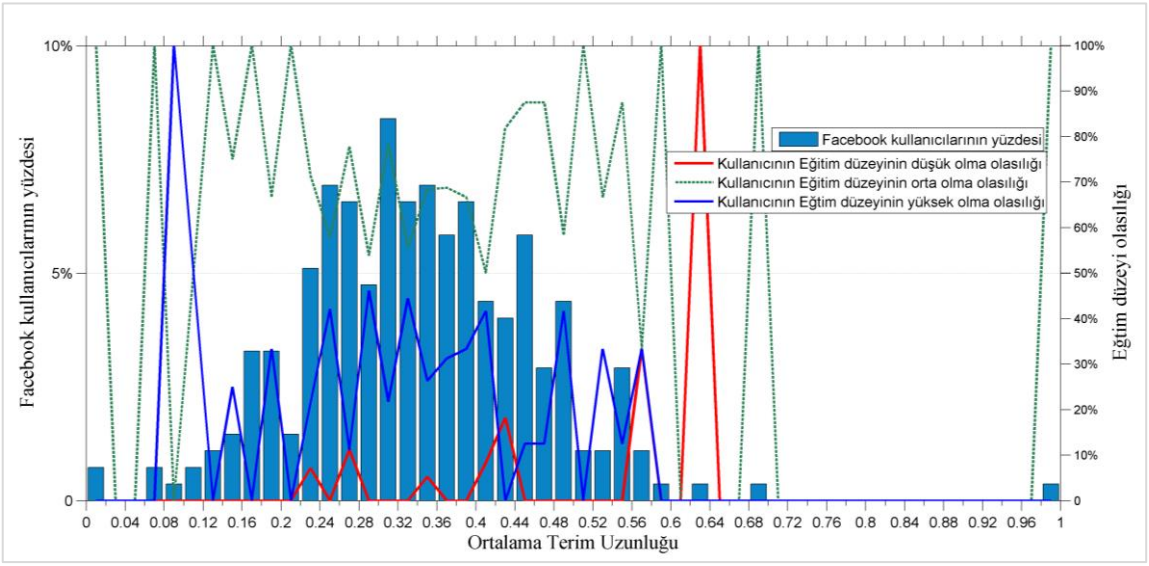
Şekil 24'te ortalama terim uzunluğu, kullanıcının yaşına göre gösterilmiştir ve kırmızı çizge grafiği aralıklar kümesindeki kullanıcıların, yüzde kaçının orta yaş olma olasılığını belirtmektedir. Örneğin ortalama terim uzunlukları [0.62 – 0.64] aralığında olan tüm Facebook kullanıcıları orta yaşlıdır.



Şekil 23. Ortalama terim uzunluğunun, kullanıcının kadın olma olasılığıyla bağlantısı



Şekil 24. Ortalama terim uzunluğunun, kullanıcının orta yaş olma olasılığıyla bağlantısı



Şekil 25. Ortalama terim uzunluğunun, kullanıcının eğitim düzeyi olasılığıyla bağlantısı

Şekil 25'te ise ortalama terim uzunluğu, kullanıcının eğitim düzeyine göre çizilmiştir. Bu grafikte önceki grafiklerin aksine 3 adet çizge grafiği bulunmaktadır. Kırmızı çizge grafiği, ortalama terim uzunluğu yatay eksende belirtilen aralıklarda olan kullanıcıların eğitim seviyesinin 'düşük' olma olasılığını göstermektedir. Kesik çizgili yeşil çizge grafiği 'orta' düzeyde ve

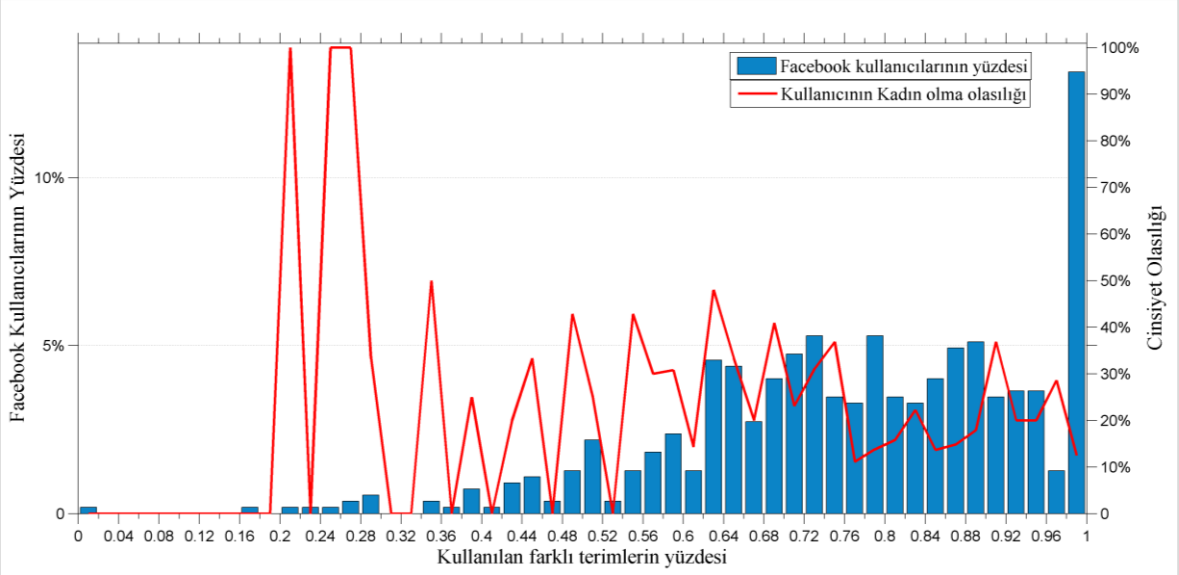
mavi çizge grafiği ise ‘yüksek’ seviyede olma olasılığını göstermektedir. Bu şekilde görülen bir örneği belirtmek gerekirse, ortalama terim uzunluğu [0.62 – 0.64] aralığında olan kullanıcıların tamamı düşük seviyede bir eğitim durumuna sahiptirler

- Kullanılan Farklı Terimlerin Yüzdesi

  Bir kullanıcının yorumlarında kullandığı farklı terimlerin, toplam terim sayısana bölümünden elde edilir ve bu ölçüme sözcük zenginliği denilir. Aşağıdaki formülden elde edilir:

$$sözcük\ zenginliği = \frac{Farklı\ terim\ sayısı}{Toplam\ terim\ sayısı} \quad (16)$$

Bu niteliğin sınıf etiketine etkisi, cinsiyet için Şekil 26’da gösterilmiştir. Bu grafikte mavi çubuk grafiği Facebook kullanıcılarının sözcük zenginliğine göre dağılımını göstermektedir. Kırmızı çizge grafikte kullanıcının kadın olma olasılığı, sözcük zenginliğinin belirli aralıklarında gösterilmektedir.



Şekil 26. Sözcük zenginliğinin, kullanıcının kadın olma olasılığıyla bağlantısı

Kırmızı çizgenin [0.2 – 0.28] aralığında olduğu durumda, kullanıcının %100 kadın olduğu söylenebilir.

Şekil 27'de sözcük zenginliği ile yaş olasılığı arasındaki ilişki gösterilmiştir. Bu grafikte kırmızı çizge grafiği kullanıcının orta yaş olma olasılığını göstermektedir. Görüldüğü üzere sözcük zenginliğinin yükselmesiyle birlikte kullanıcının orta yaş olma olasılığı düşmektedir, örnek sayısının düşük olması grafikteki gürültüye sebep olduğu gözlemlenebilir.

Şekil 28'de Sözcük zenginliğinin Eğitim düzeyine ilişkisi gösterilmektedir. Bu grafikte görüldüğü üzere maalesef örnek sayısı çok az olduğundan, kullanıcının düşük eğitim düzeyinde olma olasılığıyla ilgili pek fazla bilgi vermemektedir.

Şekil 28'de görüldüğü üzere sözcük zenginlikleri [0 – 0.02] ve [0.46 – 0.48] aralıklarında olan kullanıcıların yüksek eğitim düzeyinde olma olasılıkları %100'dür.

- Büyük Karakter Kullanımı

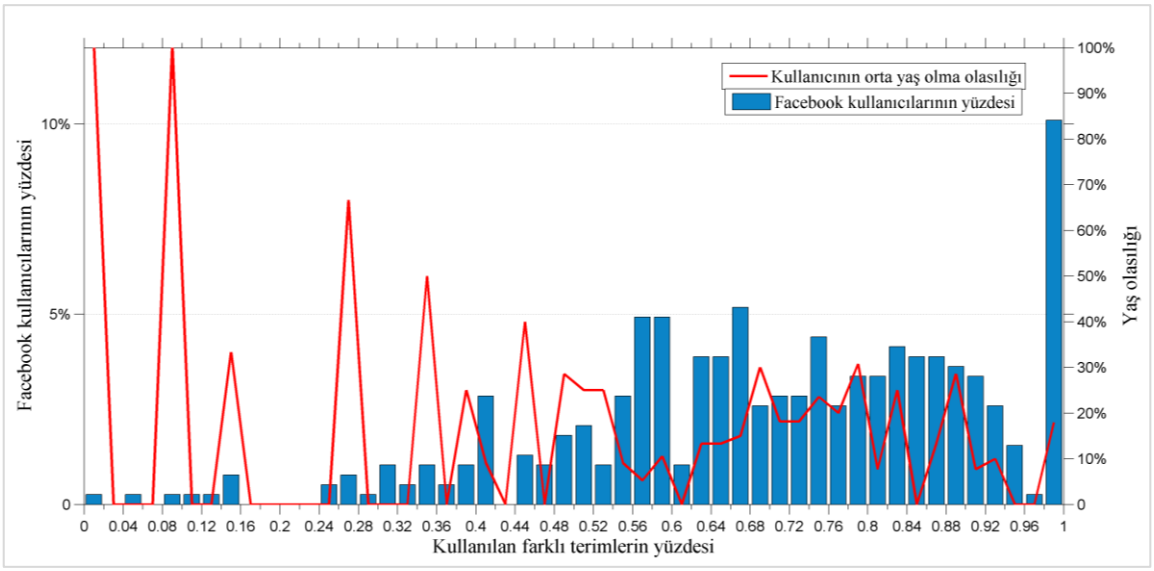
Kullanıcıların bazıları, yazdığı yorumlarda BKK'yi (Büyük Karakter Kullanımı) tercih etmektedirler, bu özelliğin sınıf etiketleriyle ilgili herhangi bir bilgi verip vermediğini öğrenmek için BKK yüzdesi, Formül 17'ye göre hesaplanmıştır.

$$BKK = \frac{Büyük\ Karakterlerin\ Sayısı}{Toplam\ yorumların\ uzunluğu} \quad (17)$$

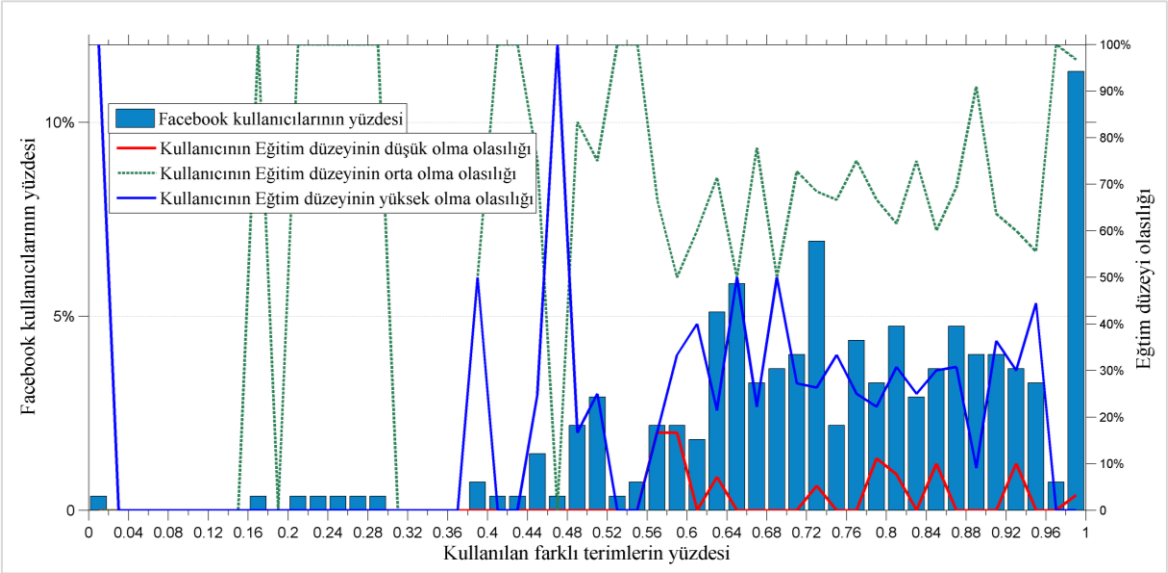
Formül 17'degörüldüğü üzere büyük karakterlerin sayısı kullanıcının toplam yorumlarının karakter bazındaki uzunluğuna bölünmüştür.

Şekil 29'da BKK'nın cinsiyet etiketiyle ilişkisi gösterilmiştir. Bu grafikte kırmızı çizge grafiği kullanıcının kadın olma olasılığını göstermektedir. [0.4 – 0.42] ve [0.88 – 0.90] aralıklarında kullanıcıların tamamı kadındır.

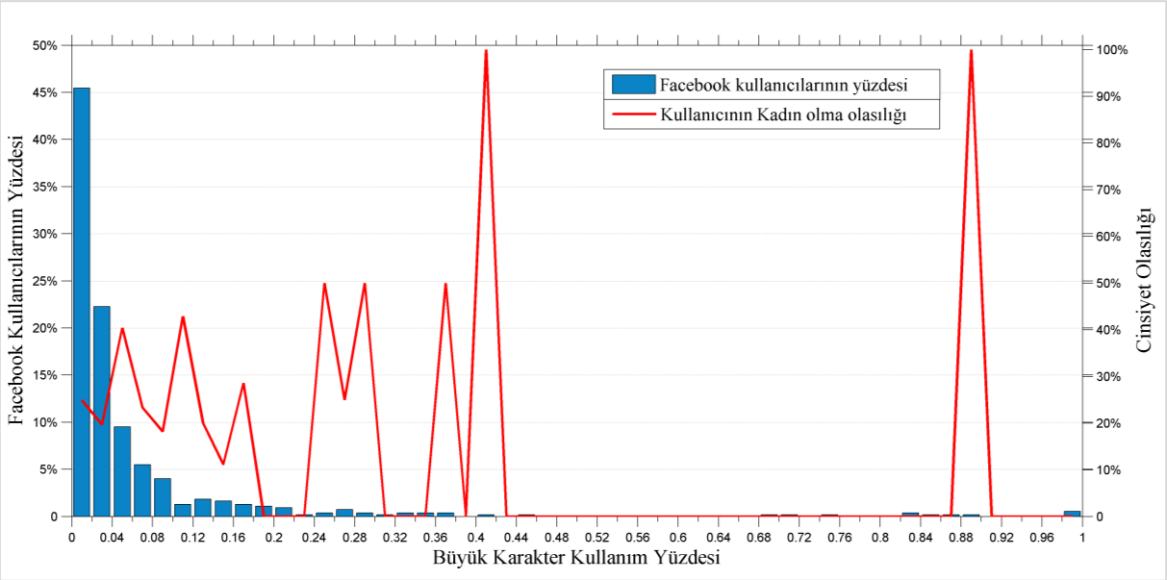
Şekil 30'da kullanıcının BKK'nın yaş etiketiyle ilişkisi gösterilmiştir. Kırmızı çizge grafiği kullanıcının orta yaş olma olasılığını göstermektedir.



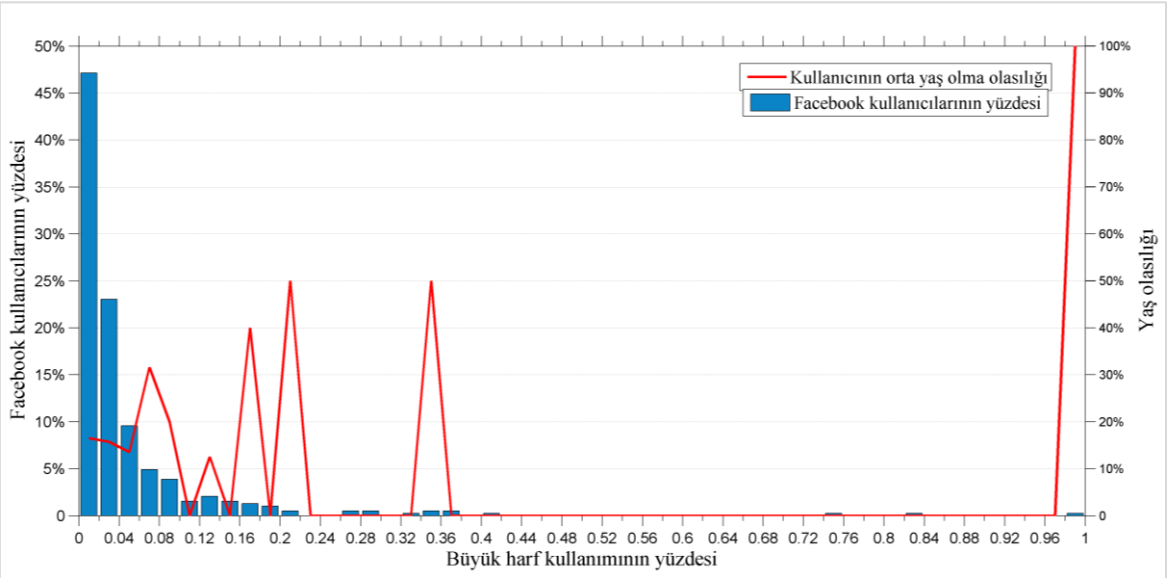
Şekil 27. Sözcük zenginliğinin, kullanıcının orta yaş olma olasılığıyla bağlantısı



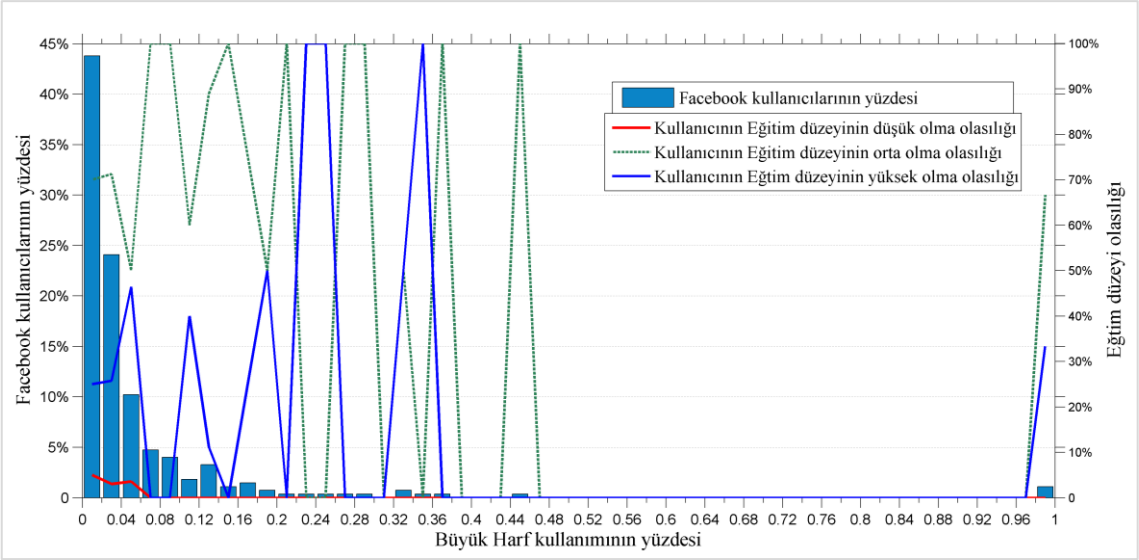
Şekil 28. Sözcük zenginliğinin, kullanıcının eğitim düzeyinin olasılığıyla bağlantısı



Şekil 29. BKK'nın, kullanıcının kadın olma olasılığıyla bağlantısı



Şekil 30. BKK'nın, kullanıcının orta yaş olma olasılığıyla bağlantısı



Şekil 31. BKK'nın, kullanıcının eğitim düzeyi olasılığıyla bağlantısı

Şekil 31'de kullanıcının BKK ve eğitim düzeyi olasılığı gösterilmektedir. Bu grafiğe baktığımızda düşük eğitim düzeyinde örnek sayısı çok az olduğundan dolayı maalesef bu nitelik düşük eğitim düzeyi hakkında bilgi vermemektedir (Kırmızı çizge grafik). Kullanıcının yüksek eğitim düzeyinde olma olasılığı mavi çizge grafikle gösterilmiştir. Bu çizge grafiği büyük harf kullanım değerlerinin [0.22 – 0.26] ve [0.34 – 0.36] aralıklarında olduğu zaman, kullanıcının yüksek eğitim düzeyinde olma olasılığını %100 olarak göstermektedir.

- Ortalama Yorum Uzunluğu

  Kullanıcıların yorum uzunluklarının cinsiyet, yaş ve eğitim düzeyi hakkında bir bilgi verip veremeyeceğini anlamak için OYU (Ortalama Yorum Uzunluğu) karakter bazında hesaplanmıştır. Bu nitelik Formül 18'le hesaplanır.

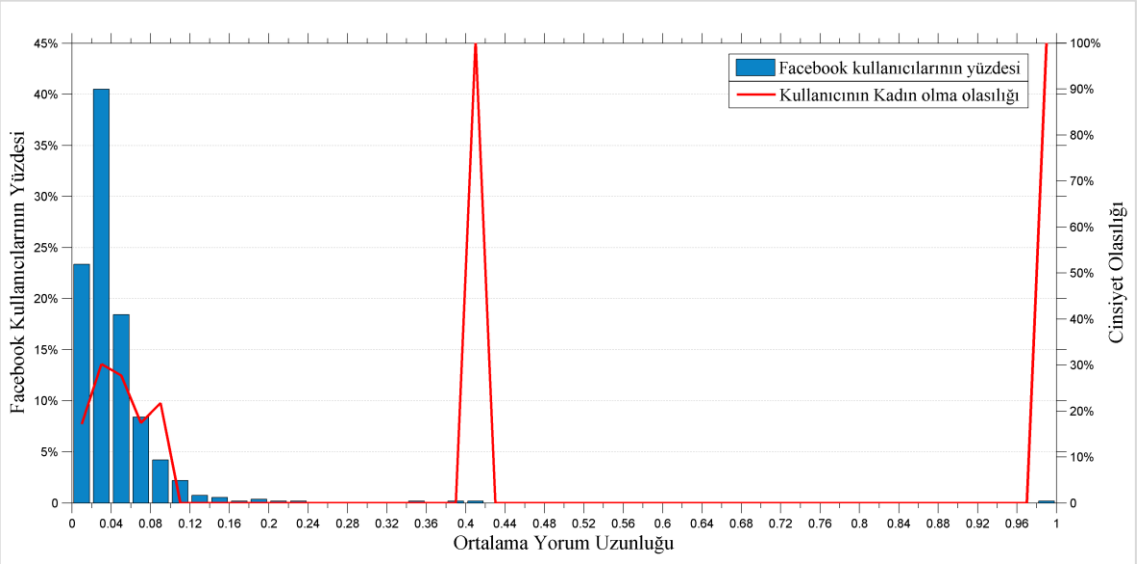
$$OYU = \frac{Yorumların\ Uzunluklarının\ Toplamı}{Toplam\ Yorum\ Sayısı} \qquad (18)$$

Formül 18'e göre; kullanıcının, tüm yorumlarının karakter bazındaki uzunluklarının toplamının, toplam yorum sayısına bölünmesiyle OYU elde edilir.

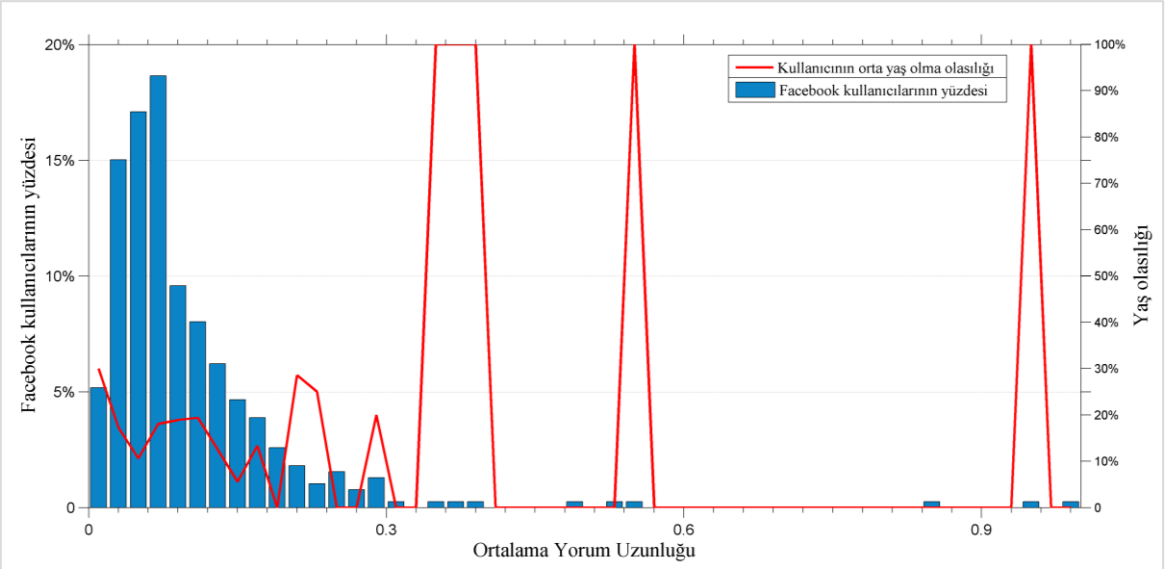
Bu niteliğin cinsiyet sınıf etiketine etkisi Şekil 32'de gösterilmiştir. Kırmızı çizge grafiği kullanıcının ortalama yorum uzunluğuyla bağlantılı olarak kadın olma olasılığını göstermektedir.

Şekil 33'te OYU'nun yaş etiketine etkisi gösterilmiştir. Kırmızı çizge grafiği OYU değerlerine karşın, kullanıcının orta yaş olma olasılığını göstermektedir. Bu grafikte görüldüğü üzere [0.34 – 0.4], [0.54 – 0.56] ve [0.94 – 0.96] aralıklarında OYU olan tüm kullanıcıların orta yaş olma olasılığı %100'dür.

Şekil 34'te OYU'nun, kullanıcının eğitim düzeyi olasılığıyla ilişkisi gösterilmiştir. Grafiğe bakıldığında [0.4 – 0.62] ve [0.98 – 1] aralıklarında OYU'su olan tüm kullanıcılarının %100 yüksek eğitim düzeyinde olma olasılığı vardır (mavi çizge grafik). Ancak düşük eğitim düzeyi hakkında bu grafik pek fazla bilgi sunmamaktadır.



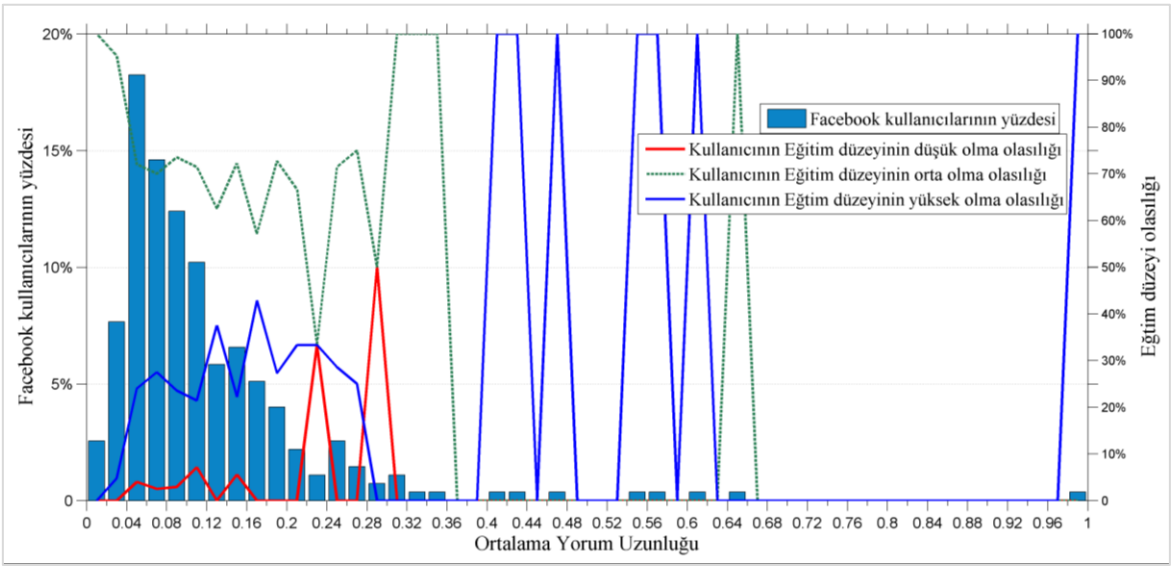
Şekil 32. OYU'nun, kullanıcının kadın olma olasılığıyla bağlantısı



Şekil 33. OYU'nun, kullanıcının orta yaş olma olasılığıyla bağlantısı

- 'q' Karakterinin Kullanılması

  Yorumların incelenmesiyle q karakterinin 'k' ve 'g' karakterleri yerine kullanıldığı dikkat çekmektedir. Bu harfi kullananların, sınıf etiketleriyle ilgili herhangi bir bilgi verip vermeyeceğini öğrenmek için bu karakter, yazım stili özelliklerinden biri olarak nitelikler kümesine eklenmiştir.

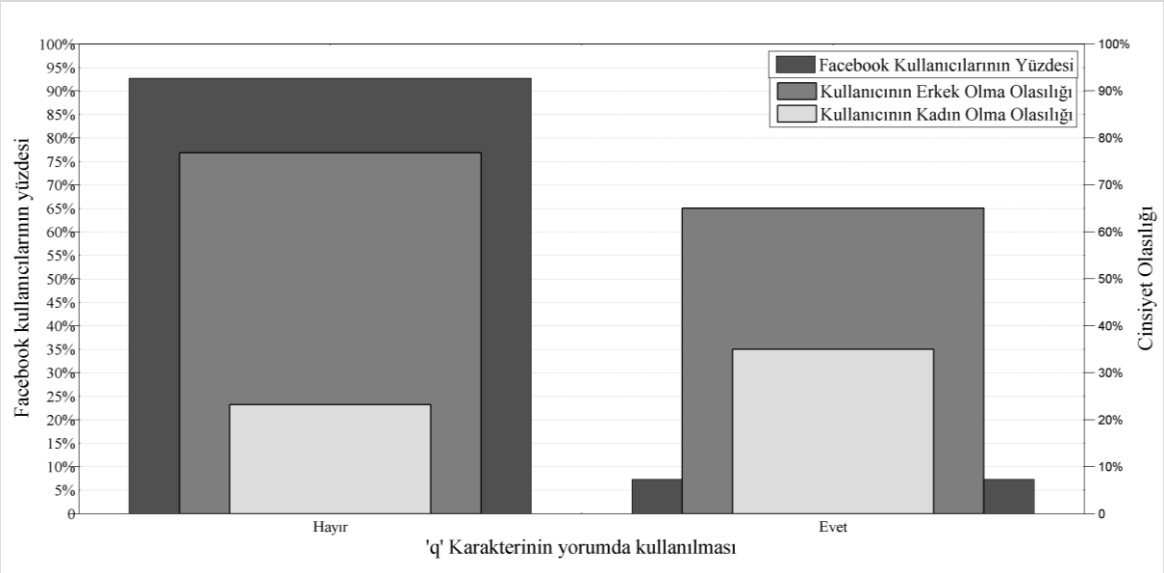


Şekil 34. OYU'nun, kullanıcının eğitim düzeyi olasılığıyla bağlantısı

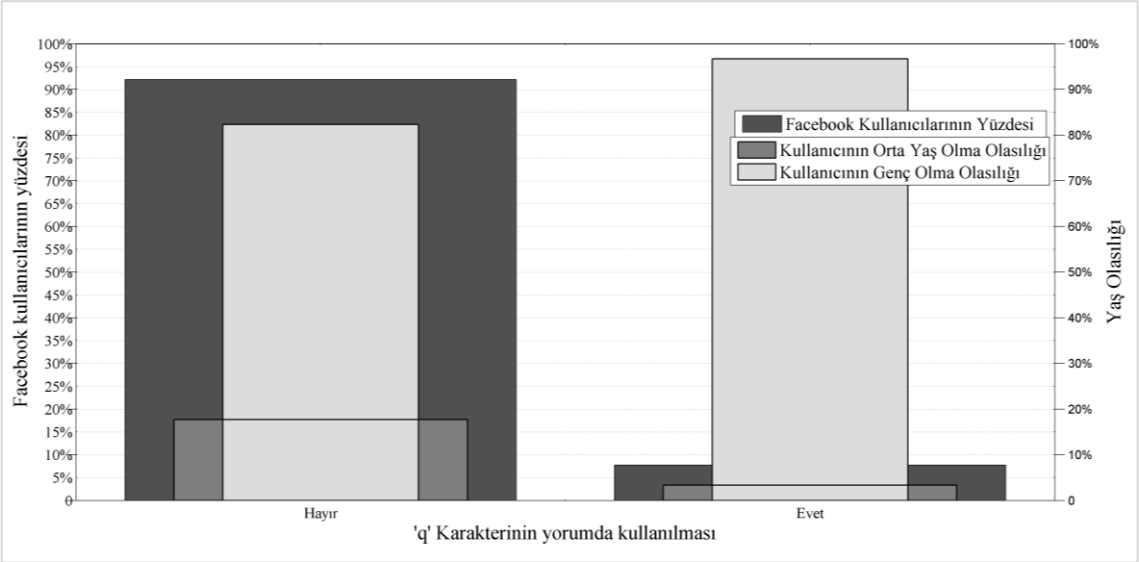
Şekil 35’te bu karakterin cinsiyet kategorisindeki etkisi gösterilmiştir ve bu bölümdeki bundan sonraki grafiklerde de 3 veya 4 histogram grafiği kullanılmıştır. Bu histogram grafiklerinin en geniş olanı (diğerlerinin arkasında kalan histogram), Facebook kullanıcılarının belirli karakteri (‘q’ veya ‘w’) kullanıp kullanmadıklarına göre olan dağılımı gösterendir. Grafikteki yatay eksen iki değerden oluşmaktadır (eğer belirli karakter kullanılmış ise ‘evet’ ve eğer kullanılmamış ise ‘hayır’). Sol taraftaki dikey eksen, Facebook kullanıcılarının dağılımını gösteren histogram grafiğine uygulanmıştır. Sağ taraftaki ise diğer histogramlara uygulanarak belirli bir sınıf etiketinin olma olasılığını göstermektedir.

Şekil 35’te orta gri tona sahip histogramda kullanıcının Erkek olma olasılığı, açık gri tonda ise kullanıcının Kadın olma olasılığı gösterilmektedir. Görüldüğü gibi ‘q’ karakterini kullananların yaklaşık %65’i erkektir.

Şekil 36’da ‘q’ karakteri kullanımının, kullanıcının yaş olasılığıyla ilişkisi gösterilmiştir. Bu grafikte, orta ton gri rengindeki alanda, kullanıcının orta yaş olma olasılığı ve açık gride ise kullanıcının genç olma olasılığı gösterilmektedir. Grafikte görüldüğü üzere ‘q’ karakterini kullananlarının yaklaşık %98’i gençtir.



Şekil 35. ‘q’ karakteri kullanımının, kullanıcının cinsiyet olasılığıyla bağlantısı



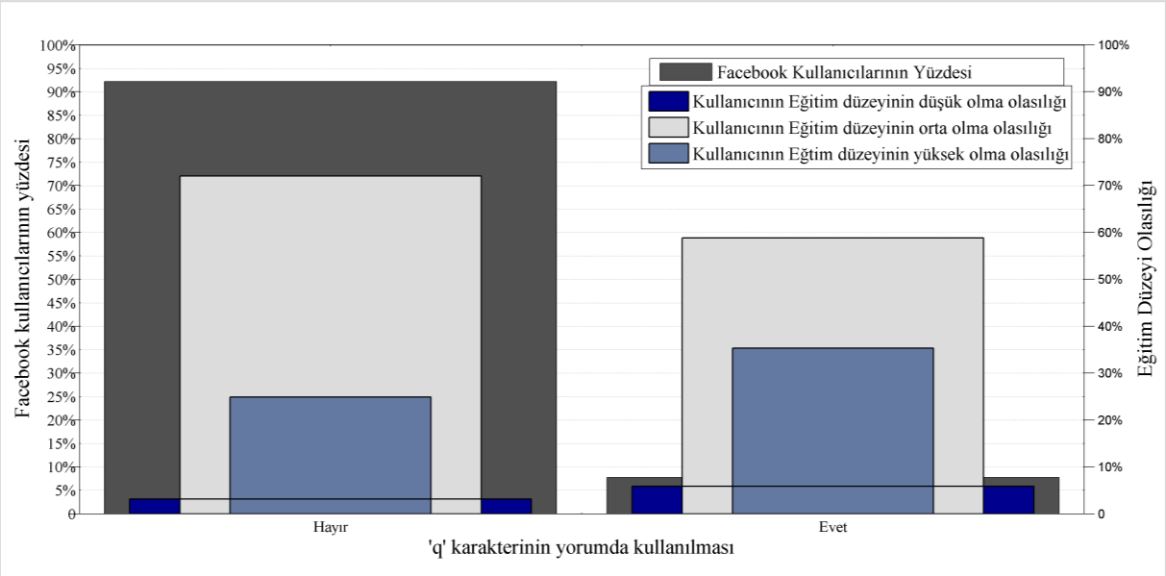
Şekil 36. ‘q’ karakteri kullanımının, kullanıcının yaş olasılığıyla bağlantısı

Şekil 37, ‘q’ harfi kullanımının, kullanıcının eğitim düzeyi olasılığıyla ilişkisini göstermektedir, koyu mavi renkli histogram grafiği, kullanıcının eğitim düzeyinin düşük olma olasılığını, açık mavi histogram, kullanıcının yüksek eğitim düzeyinde olma olasılığını ve açık gri histogram ise kullanıcının orta eğitim düzeyinde olma olasılığını göstermektedir. Bu grafikte ‘q’ karakterini kullananların yaklaşık %60’ı orta, yaklaşık %35’i yüksek ve yaklaşık %5’i düşük eğitim düzeyine sahiptirler.

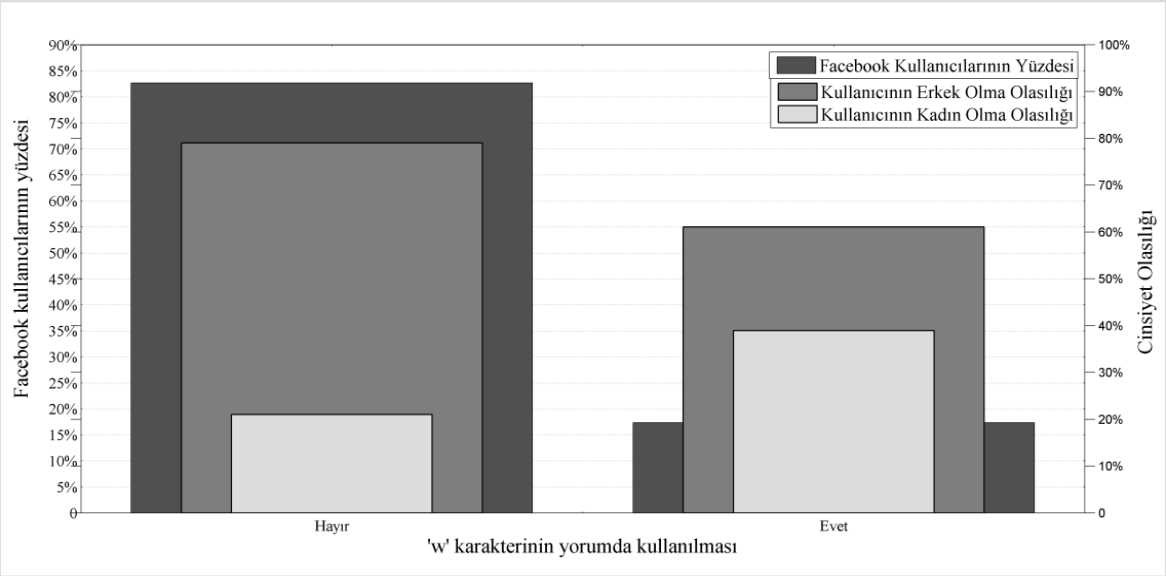
- ‘w’ karakterinin kullanılması

  ‘q’ harfinin kullanım sıklığı gibi ‘w’ karakterinin de yorumların büyük bir bölümünde ‘v’ yerine kullanıldığı görünmektedir. ‘w’ karakteri kullanımının ne derecede sınıf etiketine etkisi olduğunu gözlemlemek için, bu karakterin kullanıp kullanılmadığı bilgisi de nitelikler kümesine eklenmiştir.

  Şekil 38’de ‘w’ karakteri kullanımının, kullanıcının cinsiyetinin olasılığıyla ilişkisi gösterilmiştir. Bu grafikte görüldüğü gibi ‘w’ karakterini kullananların %60’ı erkek ve %40’ı ise kadındır. Halbu ki kullanmayanların %80’i erkek ve sadece %20’si kadındır. Bu nitelik, erkek ve kadın sayısına bakıldığında (Şekil 17); cinsiyet sınıflandırma hakkında çok iyi bilgi sunmaktadır. Öyle ki Tablo 16’da da görüldüğü gibi cinsiyet belirlemesi için bu nitelik 5. en iyi nitelik olarak seçilmiştir.



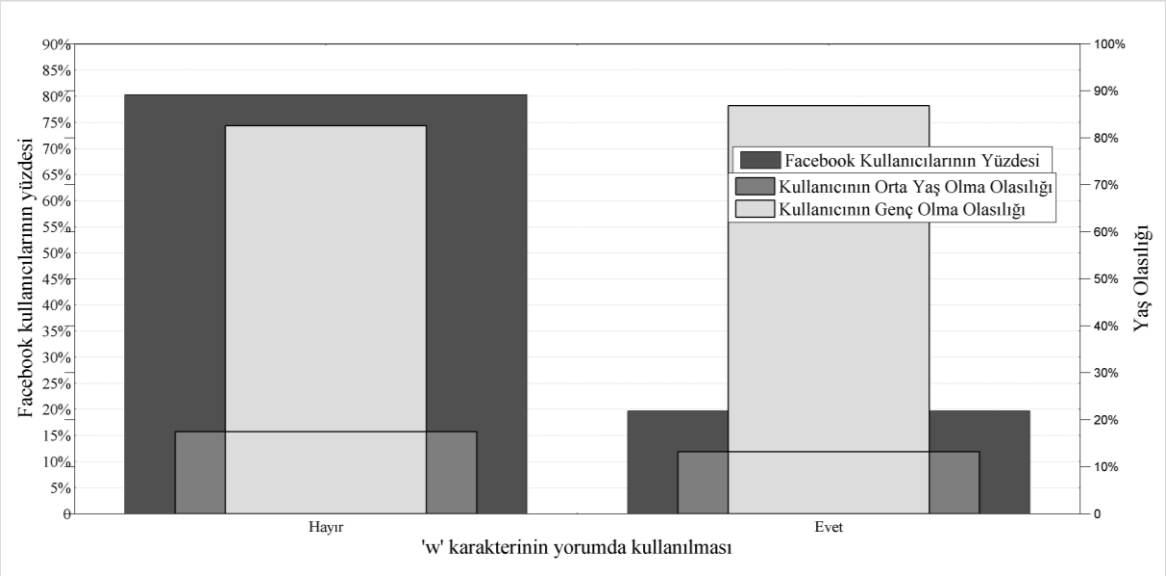
Şekil 37. ‘q’ karakteri kullanımının, kullanıcının eğitim düzeyi olasılığıyla bağlantısı



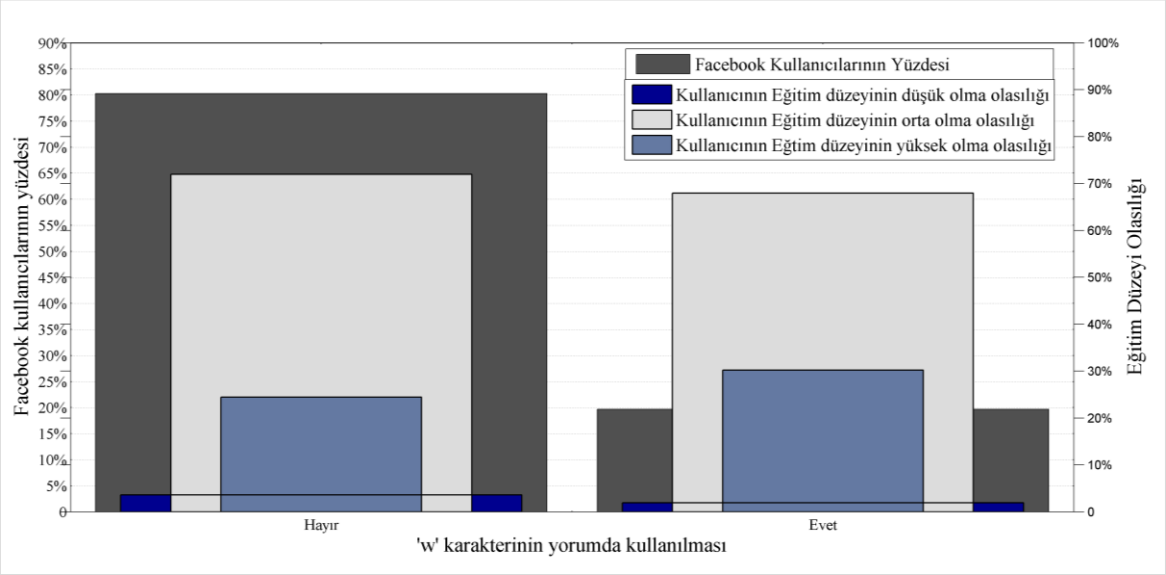
Şekil 38. ‘w’ karakteri kullanımının, kullanıcının cinsiyeti olasılığıyla bağlantısı

Şekil 39’da ‘w’ karakterinin kullanımının, kullanıcının Yaş olasılığıyla ilişkisi gösterilmiştir. Bu grafikte görüldüğü gibi ‘w’ karakterini kullananların yaklaşık %85’i gençtir, ancak kullanmayanların da yaklaşık %80’inin genç olması nedeniyle bu nitelik çok ayırt edici bir nitelik değildir.

Şekil 40'da 'w' karakterinin kullanımının, kullanıcının Eğitim düzeyi olasılığıyla ilişkisi gösterilmiştir. Bu grafikte 'w' karakterini kullanan ve kullanmayan kullanıcılar, her eğitim düzeyinde çok yakın bi yüzdeye sahip oldukları için, bu nitelik ayrıştırıcı bir nitelik değildir.



Şekil 39. 'w' karakteri kullanımının, kullanıcının Yaş olasılığıyla bağlantısı



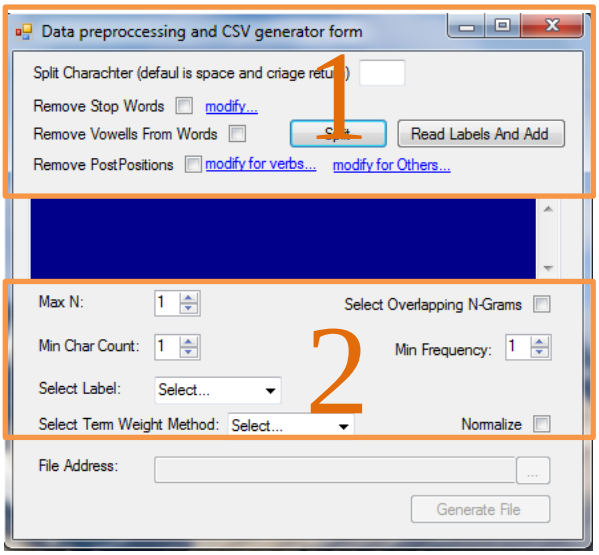
Şekil 40. 'w' karakteri kullanımının, kullanıcının Eğitim düzeyi olasılığıyla bağlantısı

3.5. Vektör Uzay Modeli Oluşturma

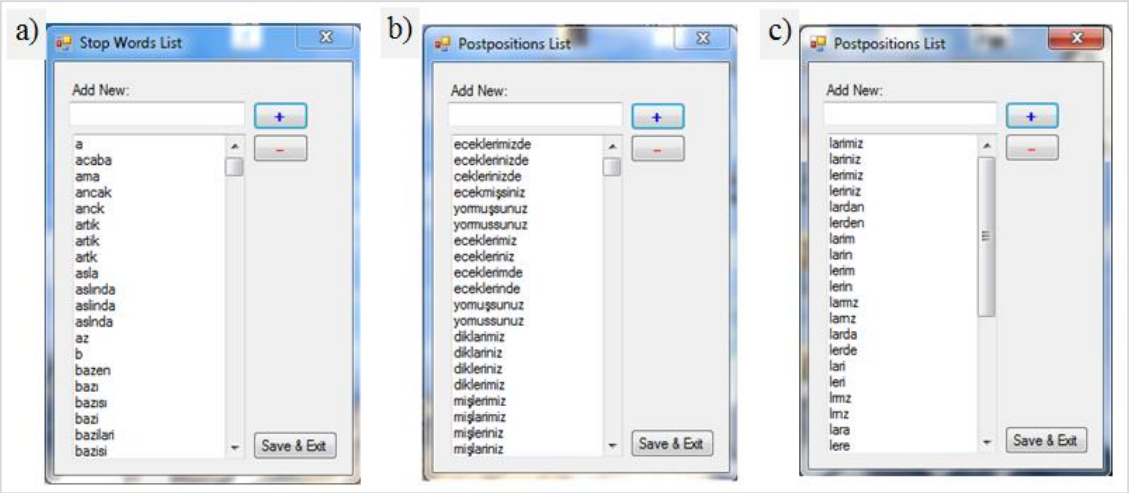
3.5.1. Kullanıcı Ara Yüzü

Veri önişleme bölümünde bahsedilen bütün işlemlerden, üretilen ve hesaplanan niteliklerden, tüm Facebook kullanıcıları için bir vektör uzay modeli oluşturmak üzre c#.net programlama dilinde bir arayüz geliştirilmiştir. Bu arayüz Şekil 41'de gösterilmiştir.

Şekil 41'de gösterilen arayüzün 1 numaralı kısmında, yorumlardaki terimler ayrıştırılmaktadır. Değersiz kelimelerin elenip, eklerin ve sesli karakterlerin silinmesiyle birlikte; belirtilen seçeneklerden seçilen şıklara göre elde edilen terimler, veri kümesine ekleniyor. Bu bölümde ayrıca değersiz kelimelerin listesinin (Şekil 42.a), fiiller için ekler listesinin (Şekil 42.b) ve diğer ekler listesinin (Şekil 42.c) yönetilmesi gibi modüller mevcuttur ve ilgili link tıklanarak listeler güncellenebilir. Bunların dışında bölüm 0'te bahsi geçen, juri tarafından etiketlenen Facebook kullanıcılarından kesin sınıf etiketinin üretilebilmesi ve bu etiketlerin veritabanına eklenebilmesi için bir fonksiyon mevcuttur ("Read Labels and ADD" butonu).



Şekil 41. Veri önişleme uygulaması ara yüzü



Şekil 42. a) Değersiz kelimeler modülü b) Fiil ekler modülü c) Diğer ekler modülü

Ara yüzün 1 numaralı kısmında gösterilen "split" butonu tıklandıktan sonra veritabanında tüm terimler, önişleme işlemleri yapılmış şekilde bulundurulur. Ayrıca bu bölümde, bölüm 3.4.3.2'de bahsedilen yazı stiline bağlı özellikler de hesaplanarak veritabanına eklenmiştir.

Ara yüzün 2.bölümündeki "Max N" alanında, üretilecek n-gramlar için maksimum n sayısının hangi değeri alacağı belirlenir. "Select overlapping n-grams" butonu işaretlenerek, üretilen n-gramların örtüşen olması belirtilir. Terimler vektörünü oluştururken, terimin vektöre dahil olması için en az kaç karaktere sahip olması gerektiğini belirtmek üzre "Min char count" kullanılmaktadır. Terimlerin ve n-gram'ların, vektör uzay modeline dahil olabilmeleri için minimum tekrarlanma sayısının kaç olması gerektiğini belirtmek üzre "min freq" kullanılır. Cinsiyet, Yaş ve Eğitim düzeyi kategorilerinin her birine ayrı bir vektör uzay modeli oluşturulması gerekmektedir. Vektör uzay modelini oluşturmak için ilgili kategori, "Select Label" listesinden seçildikten sonra, 2.3.3 bölümünde bahsedilen terim ağırlıklandırma yöntemlerinden (mantıksal, terim sıklığı ve TF-IDF) biri seçilir. Eğer terim ağırlıklandırma yöntemi TF-IDF seçilmişse, normalleştirmek için "normalize" seçeneği seçilir. "normalize" seçeneği veri değerlerini, min-max normalleştirme yöntemiyle 0 ile 1 aralığına normalleştirir ve bu seçenek diğer iki ağırlıklandırma yönteminde herhangi bir etki yapmaz.

Tüm bu ayarlamalar yapıldıktan sonra, dosya adı seçilerek "generate" tuşuna tıklanıldığında veri madenciliği araçlarının ortak kullandığı CSV (*Comma Seperated*

Vector) veya diğer adıyla "*Comma Delimited Vector*" dosyası ve ayrıca nitelik isimlerinin başka yerde tutulması için aynı isimde bir meta dosya oluşturulur.

CSV dosyasında, birinci satırda ',' işaretiyle ayrılmış niteliklerin ismi yer almaktadır, satırın sonunda ise sınıf etiketinin ismi (örneğin 'cinsiyet') mevcuttur. Daha sonraki satırlarda; her satır, bir dökümanın terim vektörünü temsil etmektedir. Nitelik değerleri ile nitelik isimleri aynı sıralamada olacak şekilde yazılmıştır, son değer ise sınıf etiketidir. Tablo 14'te CSV dosyası için küçük bir örnek gösterilmiştir. Bu tabloda terim vektörü 5 nitelik, doküman ID ve sınıf etiketinden oluşmaktadır. Terim vektörünün ağırlıklandırması için mantıksal yöntem uygulanmıştır.

Tablo 14. CSV dosyası için bir örnek

docID,attr1,attr2,attr3,attr4,attr5,Label
15155, 0, 1, 1, 0, 1,Erkek
12548, 1, 1, 0, 1, 0,Kadın

Meta dosyasında ise "id", "Nitelik Kodu" ve "Nitelik Adı" yer almaktadır. Tablo 14'te CSV için meta dosyası, Tablo15'te gösterilmiştir. Meta dosyasının kullanılma sebeplerinden en önemlisi CSV dosyasının bellek kullanımını en aza düşürmektir. Örneğin N=5 olarak seçilmişse, bir çok nitelik 5 terimden oluşacaktır ve bunların hepsinin CSV dosyasına yazılması, bellek kullanımını önemli derecede yükseltecektir. Ancak bir meta dosyası kullanarak 5 terim uzunluklu bir niteliğin ismini "attr" gibi kısa bir kodla göstermek bu problemi kolaylıkla çözmüş olacaktır. Meta dosyasının, bir başka kullanım nedeni ise nitelik isimlerinde kullanılan özel karakterlerdir. Oluşturulan CSV dosyası, herhangi bir veri madenciliği aracında kullanılabilmesi için nitelik isimlerinde özel karakterler (';', '.', vs.) kullanılmamalıdır.

Tablo15. CSV dosyası için bir meta dosyası örneği

İd	Nitelik Kodu	Nitelik Adı
1	attr1	merhaba
2	attr2	Ankara
3	attr3	merhaba ankara
4	attr4	:)
5	attr5	SA

3.6. Rapidminer ile Veri Madenciliği

Oluşturulan CSV dosyası, herhangi bir veri madenciliği aracında açılarak kullanılabilir. Bu çalışmada Rapidminer veri madenciliği aracı kullanılmıştır.

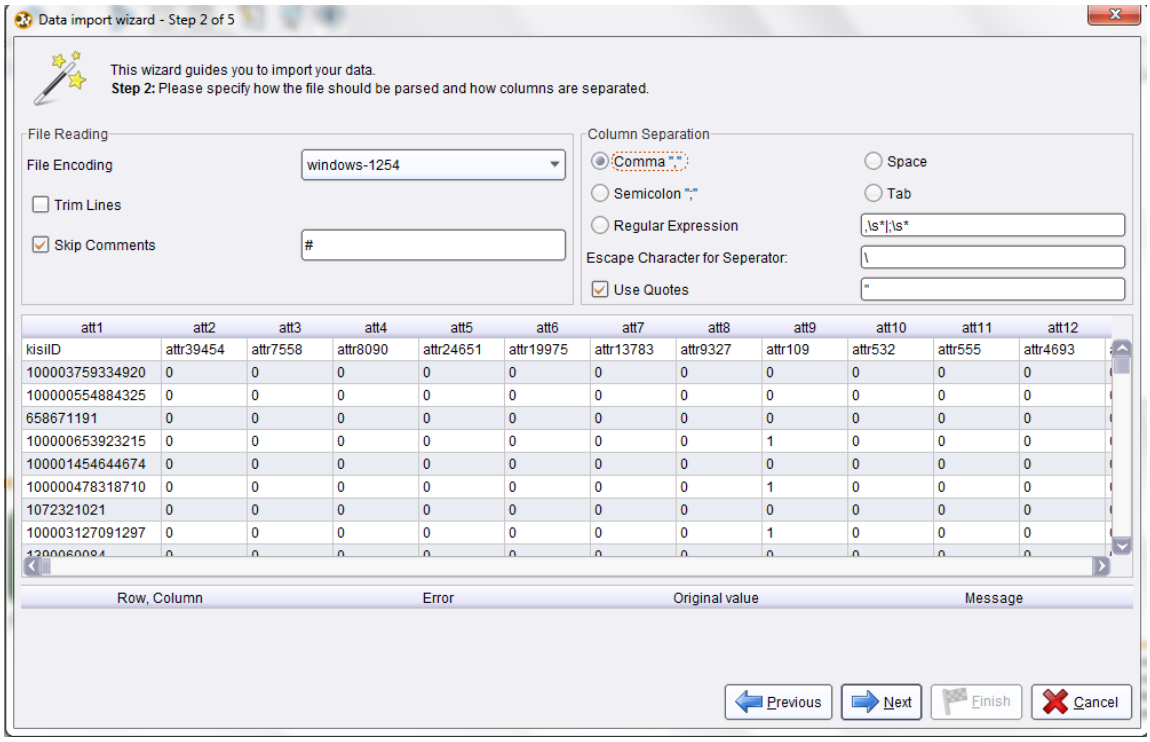
CSV dosyasını, Rapidminer'da açmak için "file → Import Data→ Import CSV file" adımları izlenir. CSV dosyası içeri aktarılmak üzre seçildikten sonra ikinci aşamada Şekil 43'te görüldüğü gibi ayrıştırıcı karakter olarak ',' işaretlenir. Üçüncü aşamada nitelik isimlerinin olduğu birinci satır belirlenir, dördüncü aşamada Rapidminer nitelik değerlerini inceleyerek bunların veri türünü tahmin etmeye çalışır. Bu aşamada doküman ID ve sınıf etiketinin hangi sutunda olacağı kullanıcı tarafından belirlenmelidir. Ayrıca Rapidminer, veri türünün tahmininde başarısız olmuşsa, kullanıcı manuel olarak veri türlerini her nitelik için belirleyebilir. Doküman ID, "id", nitelikler "attribute" ve sınıf etiketi "label" olarak seçilmelidir. En son aşamada veri, Rapidminer formatına dönüştürülüp, daha sonra kullanılmak üzre kaydedilir.

3.6.1. Rapidminer ile Veri Sınıflandırma

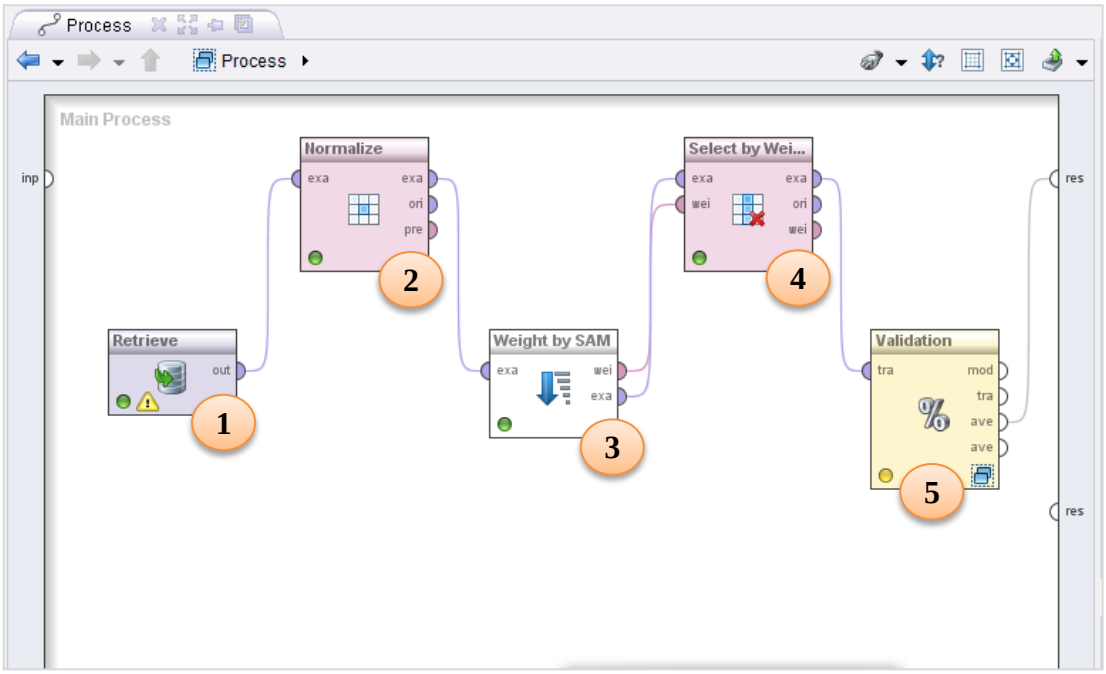
Rapidminer'de herhangi bir veri madenciliği işlemi yapmak için bir "Process" oluşturulur. Her "process" bir kaç işlem modülünden oluşur. Şekil 44 bu çalışma için

oluşturulan “process”i göstermektedir. Bu “process” 5 modülden oluşmaktadır, aşağıda bu 5 aşama açıklanmıştır:

1. Verilerin alınması: Veri depolarında daha önce dönüştürülerek kaydedilen CSV dosyası, veri madenciliği işlemi için hazırlanmaktadır.
2. Normalleştirme: CSV dosyasını oluştururken “normalize” seçeneği seçildiğinde, tüm terim nitelikleri ‘min-max’ yöntemiyle normalleştirildiğinden, bu seçenek sadece yazı stilindeki niteliklere uygulanmaktadır. Bu aşamada normalleşmemiş nitelik bırakılmamaktadır.
3. SAM ile nitelik ağırlıklandırma: Daha önce 2.3.4.1’de de bahsedildiği gibi bu çalışmada nitelik ağırlıklandırmak için SAM kullanılmıştır. Bu modül, SAM tekniğiyle nitelikleri ağırlıklandırarak bir sonraki aşamaya hazırlıyor.

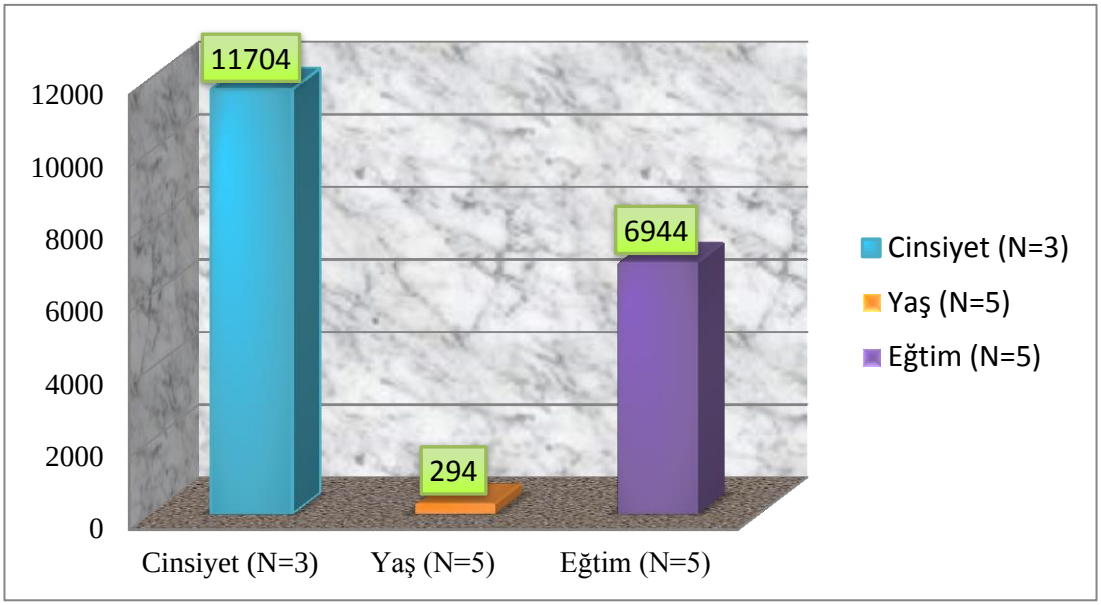


Şekil 43. Rapidminer'de csv dosyasının içeri aktarılması



Şekil 44. Rapidminer "process"i ve işlem modüllerinin bağlantısı

4. Ağırlığa göre nitelik seçme: Bu modülde, SAM ile ağırlıklandırılmış olan niteliklerin belirli bir sınırın altında olanlarını eleyip, kalan nitelikleri 5. modüle gönderiyor. SAM ile nitelik ağırlıklandırma ve boyut azaltma işleminden sonra, her kategori için kalan nitelik sayısı Şekil 45'te gösterilmiştir. Tablo 16'da her kategori için en yüksek ağırlıklandırılan ve önemli 10 nitelik listelenmiştir. Tüm kategorilerde ağırlıklar min-max normalleştirme tekniğiyle normalleştirilmiştir. Niteliklerin her sınıfta kullanılma sıklığı ve yüzdesi aynı tabloda gösterilmiştir.
5. K-Kat Çapraz Doğrulama (*K-Fold Cross Validation*): Bu Modül iki kısımdan oluşmaktadır. Birinci kısım "training" veya öğrenme, ikinci kısım "testing" veya sınama olarak adlandırılır. Şekil 46'da içeriği gösterilen bu modül için K değeri 10 olarak seçilmiştir. Bu modül 3 ayrı alt modülden oluşmuştur, bunların ilki "training" bölümünde yer alır (***a***) ve model oluşturmak için kullanılır. "testing" bölümündeyse modeli deneme ve performansını ölçmek için diğer 2 modül (***b*** ve ***c***) yer almaktadır. Aşağıda bu 3 modülün detayları anlatılmıştır.

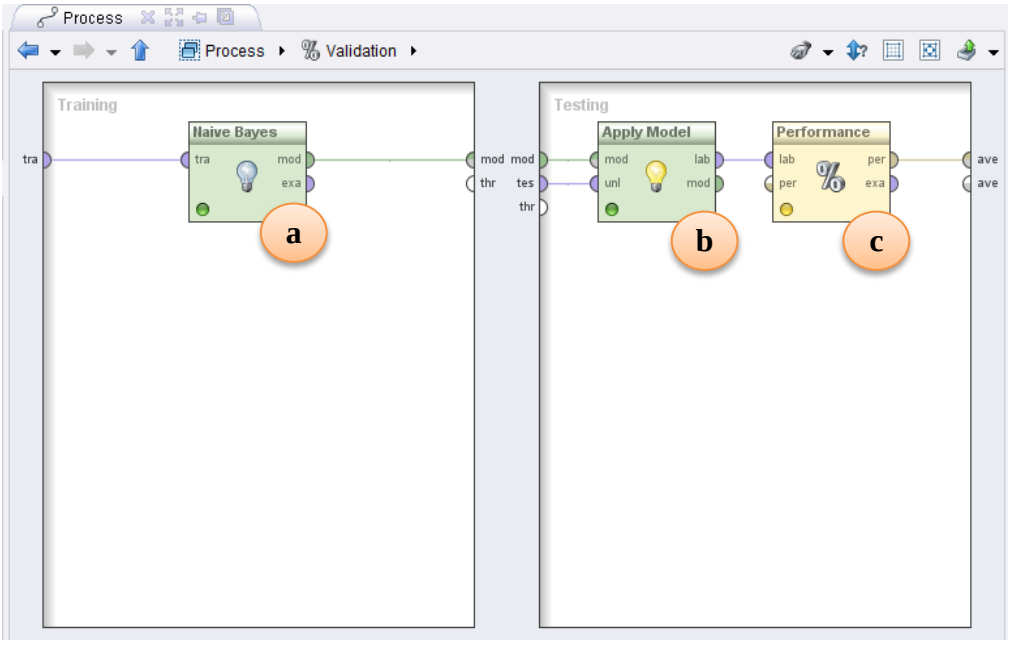


Şekil 45. SAM ile nitelik ağırlıklandırma ve boyut azaltma işleminden sonra her kategoride kalan nitelik sayısı

a. Model oluşturmak için gerekli veri madenciliği algoritması: Bu çalışmada Naive Bayes, KEYK ve DVM algoritmaları kullanılmıştır. Dolayısıyla bu bölümde, belirtilen algoritmalardan birinin modülü yerleştirilebilir.

b. Model deneme modülü: Sınama kümesindeki veriler, öğrenme bölümünde oluşturulan modele uygulanarak sonuçlar elde edilir. Sınama kümesi, sınıf etiketi olmadan modele uygulanır ve model sınıf etiketini sınama kümesi için tahmin eder.

c. Performans değerlendirme: Sınama kümesinin gerçek sınıf etiketleri ve model denemede üretilen sınıf etiketleri bu modülde karşılaştırılır. Doğru ve yanlış sınıflandırılan örnekler hesaplanarak sonuçlar tablosu veya karmaşıklık matrisi elde edilir. Bu karmaşıklık matrisi üzerinden kesinlik, anma ve f-ölçütü hesaplanabilir. Şekil 47'de cinsiyet için elde edilen karmaşıklık matrisi gösterilmiştir. Tüm bulgular ve tartışılacak detaylar SONUÇLAR bölümünde verilmiştir.



Şekil 46. K-Kat Çapraz Doğrulama modülünün içeriği

accuracy: 90.85% +/- 4.29% (mikro: 90.88%)			
	true erkek	true kadin	class precision
pred. erkek	387	21	94.85%
pred. kadin	29	111	79.29%
class recall	93.03%	84.09%	

Şekil 47. Cinsiyet sınıflandırması için karmaşıklık matrisi

Tablo 16. En önemli ve yüksek ağırlıklandırılmış 10 nitelik

SIRA	Cinsiyet				Eğitim Düzeyi					Yaş			
	Nitelik	Ağırlık	Sıklık		Nitelik	Ağırlık	Sıklık			Nitelik	Ağırlık	Sıklık	
			Erkek:416	Kadın:132			Düşük:9	Orta:195	Yüksek:70			Genç:322	Orta:64
1	çok	1.000	(106):**25.48%**	(58):**43.93%**	**avgCommLen**	1.000	Mean: 8.47, Mod: 7.0, Median: 6.88			:d	1.000	(99):**30.75%**	(3):**4.69%**
2	:)	0.899	(180):**43.27%**	(80):**60.61%**	yazmak	0.987	(1):**11.11%**	(11):**5.64%**	(8):**11.43%**	olsun	0.835	(66):**20.50%**	(27):**42.19%**
3	güzel	0.884	(74):**17.79%**	(44):**33.33%**	olmak	0.827	(0):**0%**	(15):**7.69%**	(11): **15.71%**	ankara	0.660	(95):**29.50%**	(30):**46.87%**
4	:(	0.867	(37):**8.89%**	(30):**22.73%**	sakal	0.761	(0):**0%**	(2):**1.03%**	(5): **7.14%**	:)	0.575	(175):**54.35%**	(25):**39.06%**
5	**W_Used**	0.831	(58):**13.94%**	(37):**28.03%**	selam	0.717	(1):**11.11%**	(73):**37.44%**	(18): **25.71%**	yaş	0.477	(27):**8.39%**	(13):**20.31%**
6	ay	0.784	(16):**3.84%**	(20):**15.15%**	orasi	0.673	(1):**11.11%**	(3):**1.54%**	(4): **5.71%**	akşam	0.448	(34):**10.56%**	(14):**21.88%**
7	seni	0.759	(46):**11.06%**	(31):**23.48%**	ders	0.658	(0):**0%**	(2): **1.03%**	(5): **7.14%**	hanım	0.447	(15):**4.56%**	(10):**15.63%**
8	♥	0.754	(15):**3.60%**	(19):**14.39%**	bilmek	0.658	(0):**0%**	(8):**4.10%**	(8): **11.42%**	benim	0.440	(59):**18.32%**	(19):**29.69%**
9	ben	0.729	(184):**44.23%**	(77):**58.33%**	24	0.642	(0):**0%**	(2): **1.03%**	(4): **5.71%**	dur	0.415	(27):**8.39%**	(12):**18.75%**
10	lan	0.631	(57):**13.70%**	(3):**2.27%**	görüşürüz	0.623	(1):**11.11%**	(3): **1.54%**	(4): **5.71%**	allah	0.400	(33):**10.25%**	(13):**20.31%**

4. SONUÇLAR

Bu bölümde Rapidminer kullanılarak her sınıflandırıcı için elde edilen sonuçlar verilmiştir. DVM, KEYK ve Naive Bayes sınıflandırıcılarından ve ayrıca bu üç sınıflandırma algoritmaları Mantıksal, Terim sıklığı ve TF-IDF terim ağırlıklandırma teknikleriyle çalıştırılıp, çıkan sonuçlar karışılaştırılmıştır.

Sınıflandırıcıları doğrulamak için 10-kat çapraz doğrulama yöntemi, tabakalı örneklemeyle kullanılmıştır.

4.1. Terim Sıklığı Ağırlıklandırma Yöntemiyle Sınıflandırma Sonuçları

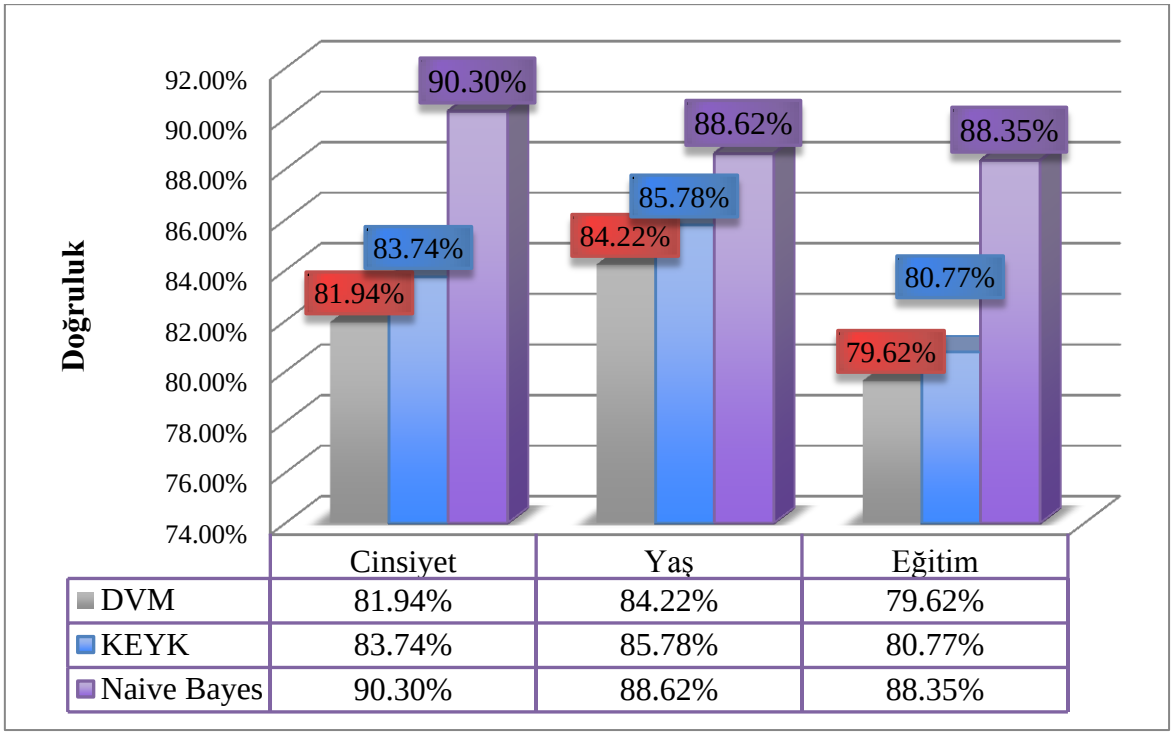
Terim sıklığı ağırlıklandırma yöntemiyle oluşturulan vektör uzay modelinin DVM, KEYK ve Naive Bayes sınıflandırıcı algoritmalarıyla her kategoride ürettiği sonuçlar Şekil 48'de verilmiştir.

Şekil 48'de görüldüğü gibi Naive Bayes sınıflandırma algoritması, terim sıklığına göre üretilen vektör uzay modelinde, her üç kategori için daha doğru sonuçlar üretmiştir.

4.2. Mantıksal Ağırlıklandırma Yöntemiyle Sınıflandırma Sonuçları

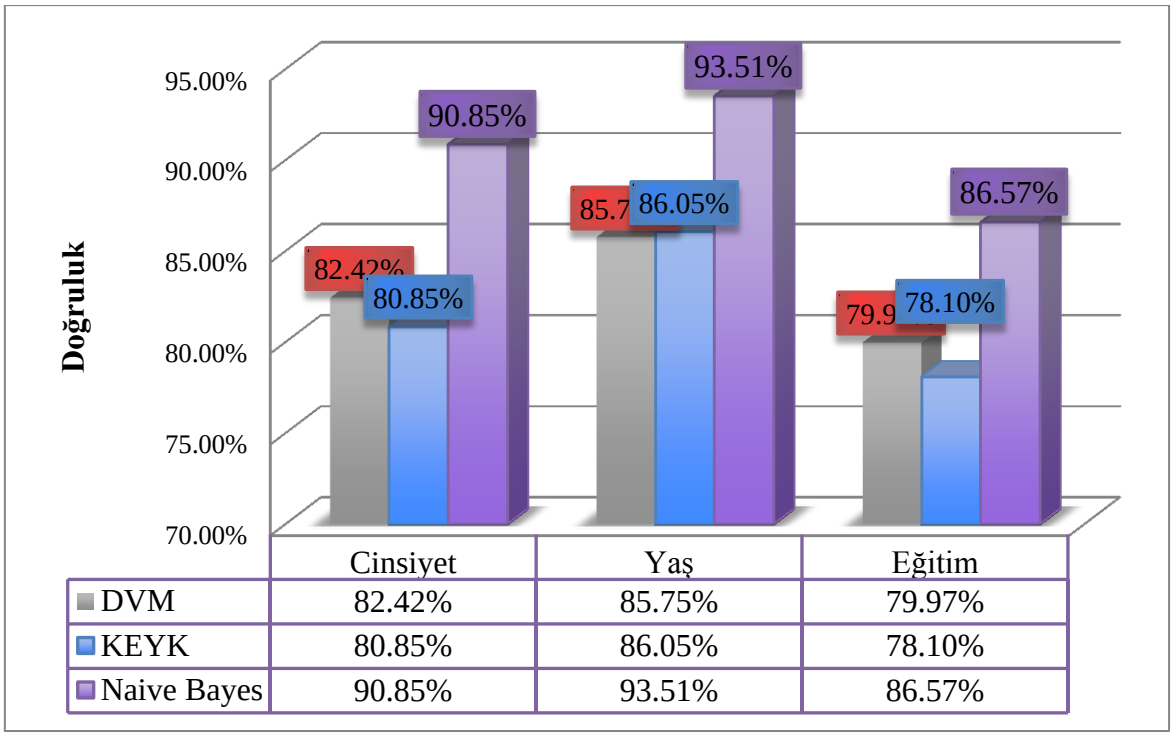
Bu bölümde, vektör uzay modelinin oluşturulmasında Mantıksal ağırlıklandırma kullanılmıştır, basit bir şekilde eğer kullanıcı terimi kullanmış ise 1, kullanmamış ise 0 değeri ağırlık olarak atanmıştır. Mantıksal terim ağırlıklandırma kullanılarak DVM, KEYK ve Naive Bayes sınıflandırıcılarıyla elde edilen sonuçlar, cinsiyet, yaş ve eğitim düzeyi için Şekil 49'da gösterilmiştir.

Şekil 49'da görüldüğü üzere Naive Bayes sınıflandırıcısı her üç kategoride de, terim sıklığı ağırlıklandırma yönteminde olduğu gibi mantıksal ağırlıklandırma yönteminde de daha iyi sonuçlar üretmiştir.



	Cinsiyet	Yaş	Eğitim
DVM	81.94%	84.22%	79.62%
KEYK	83.74%	85.78%	80.77%
Naive Bayes	90.30%	88.62%	88.35%

Şekil 48. Terim sıklığı ağırlıklandırma yöntemiyle oluşturulan vektör uzay modelinin DVM, KEYK ve Naive Bayes ile sınıflandırma sonuçları

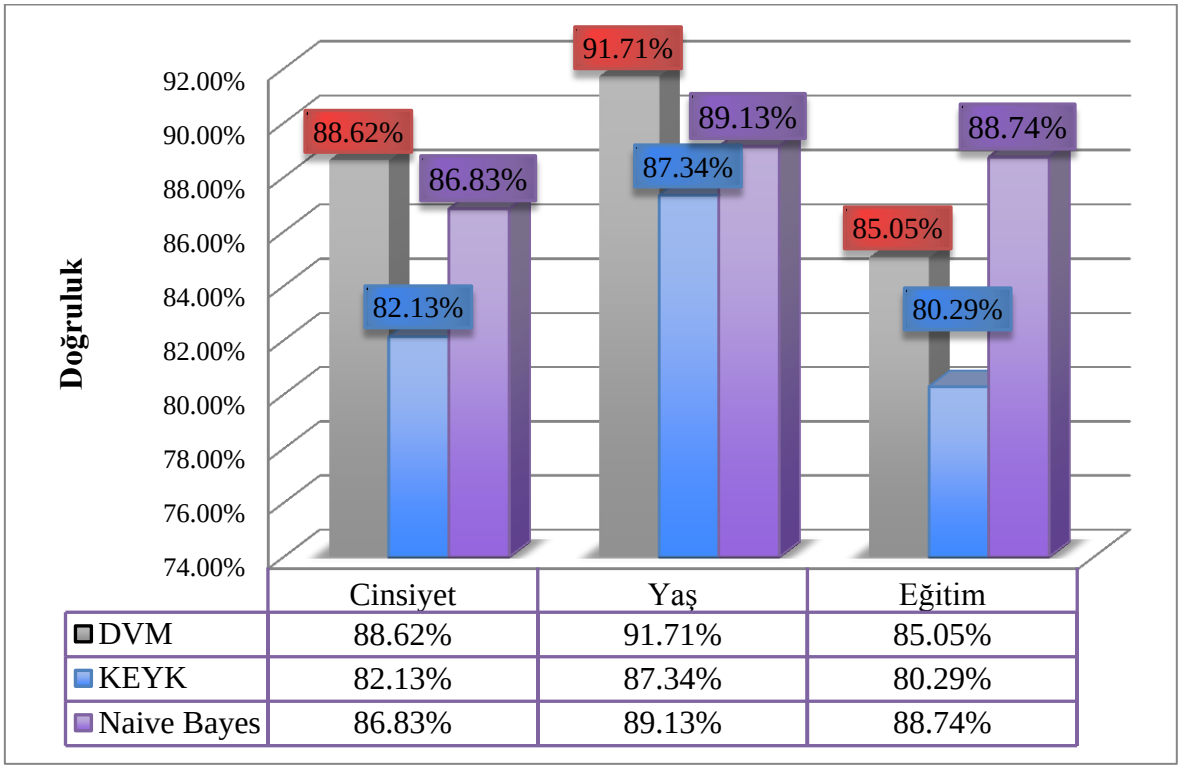


	Cinsiyet	Yaş	Eğitim
DVM	82.42%	85.75%	79.97%
KEYK	80.85%	86.05%	78.10%
Naive Bayes	90.85%	93.51%	86.57%

Şekil 49. Mantıksal ağırlıklandırma yöntemiyle oluşturulan vektör uzay modelinin DVM, KEYK ve Naive Bayes ile sınıflandırma sonuçları

4.3. TF-IDF Ağırlıklandırma Yöntemiyle Sınıflandırma Sonuçları

Bu bölümde, Vektör uzay modelinin oluşturulmasında TF-IDF terim ağırlıklandırma yöntemi kullanılmıştır. TF-IDF terim ağırlıklandırma kullanılarak DVM, KEYK ve Naive Bayes sınıflandırıcılarıyla elde edilen sınıflandırma sonuçları cinsiyet, yaş ve eğitim düzeyi için Şekil 50'de gösterilmiştir.



	Cinsiyet	Yaş	Eğitim
DVM	88.62%	91.71%	85.05%
KEYK	82.13%	87.34%	80.29%
Naive Bayes	86.83%	89.13%	88.74%

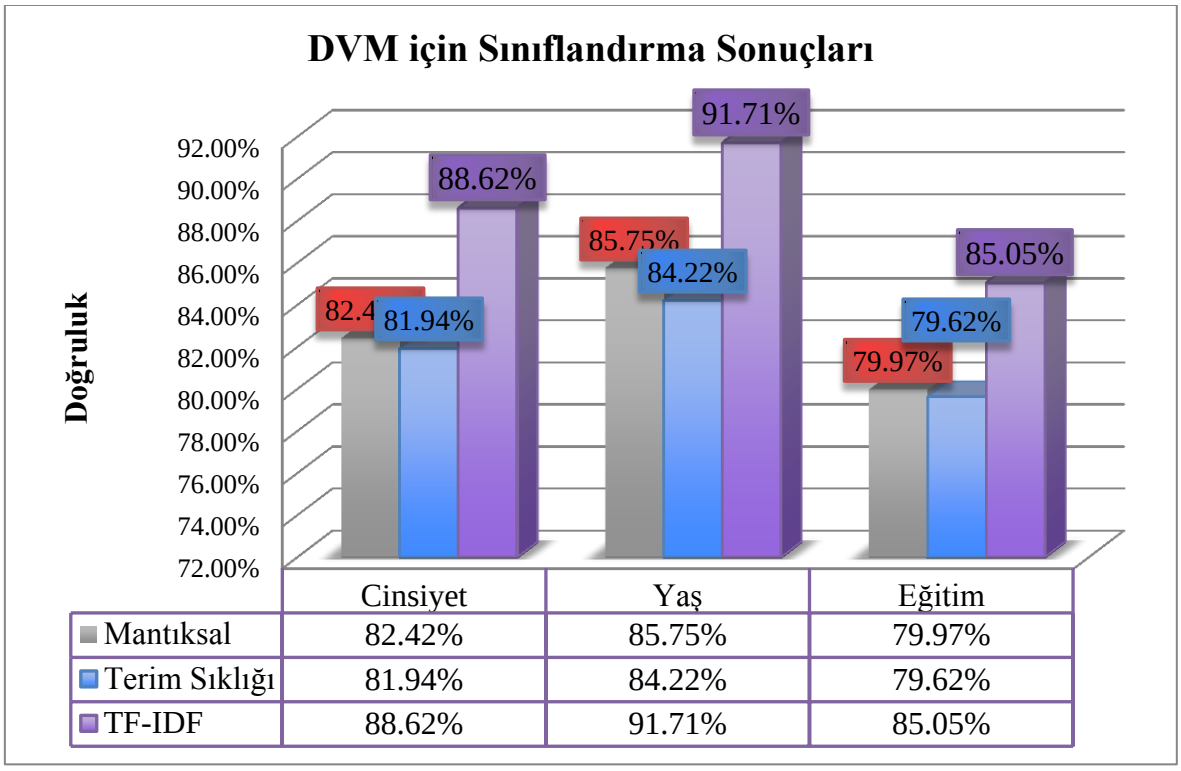
Şekil 50. TF-IDF ağırlıklandırma yöntemiyle oluşturulan vektör uzay modelinin DVM, KEYK ve Naive Bayes ile sınıflandırma sonuçları

Şekil 50'de görüldüğü üzere TF-IDF terim ağırlıklandırma kullanımında, Cinsiyet ve Yaş kategorileri için DVM sınıflandırıcısı; Eğitim düzeyi için ise Naive Bayes sınıflandırıcısı daha iyi sonuçlar üretmiştir.

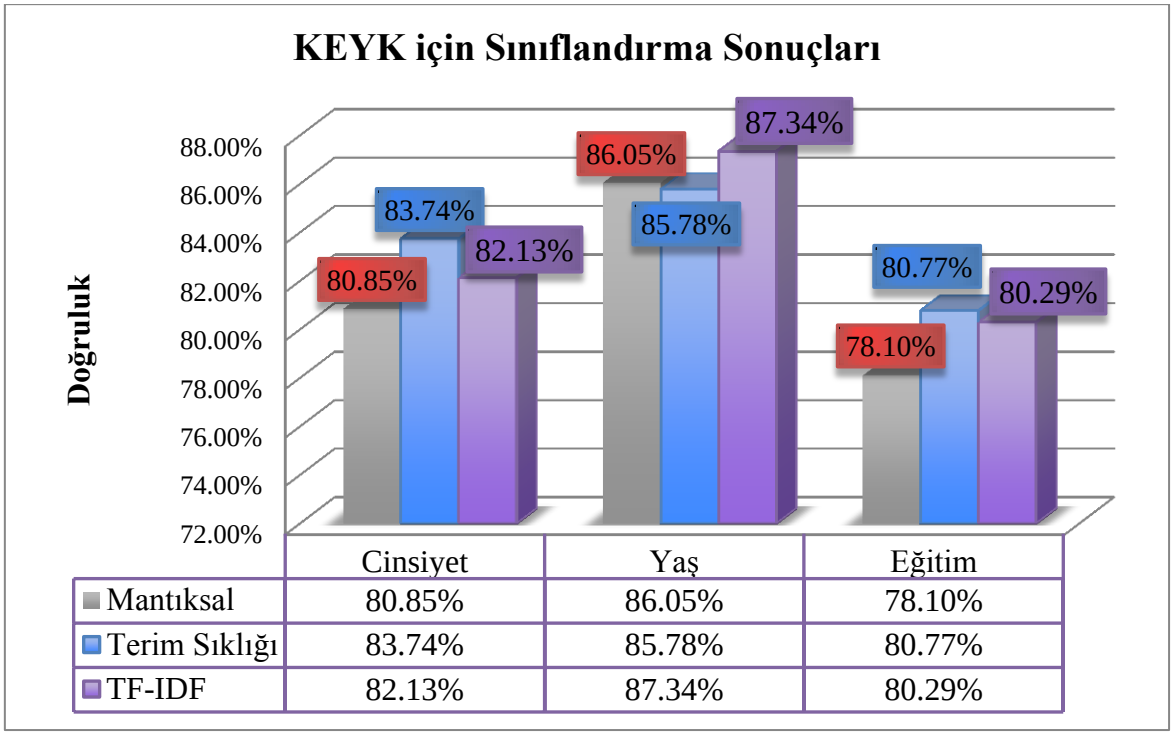
4.4. Terim Ağırlıklarının, Sınıflandırma Algoritmalarının Sonuçlarındaki Etkisi

Şekil 51'de DVM sınıflandırıcısının, farklı terim ağırlıklandırma yöntemleriyle sınıflandırma performansı karşılaştırılmıştır. Bu şekle göre DVM, tüm kategorilerde TF-IDF terim ağırlıklandırma yöntemiyle diğer iki ağırlıklandırma yönteminden daha iyi sonuçlar üretmiştir.

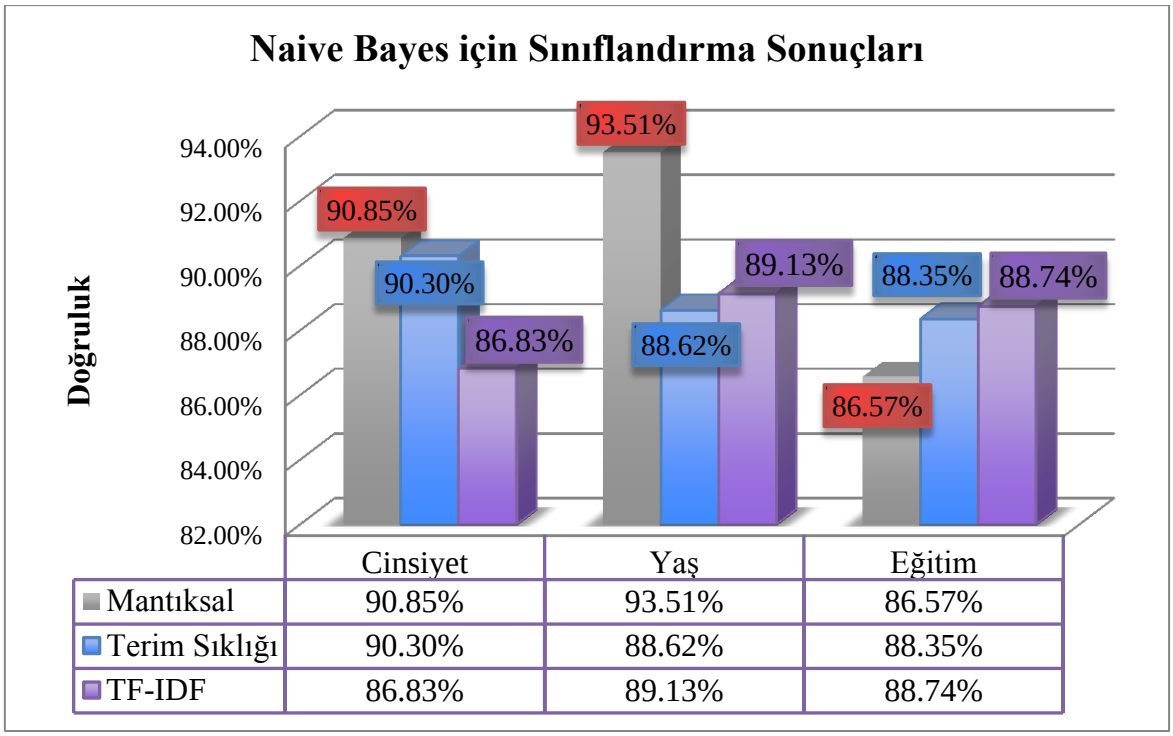
Şekil 52'de KEYK sınıflandırıcısının farklı terim ağırlıklandırma yöntemleriyle sınıflandırma performansı karşılaştırılmıştır. Bu şekilde gördüğümüz üzere KEYK, cinsiyet ve eğitim düzeyi için terim sıklığıyla, yaş için ise TF-IDF'le daha iyi sonuçlar üretmiştir.



Şekil 51. DVM sınıflandırıcısı için farklı terim ağırlıklandırma yöntemleriyle sınıflandırma sonuçları



Şekil 52. KEYK sınıflandırıcısı için farklı terim ağırlıklandırma yöntemleriyle sınıflandırma sonuçları



Şekil 53. Naive Bayes sınıflandırıcısı için farklı terim ağırlıklandırma yöntemleriyle sınıflandırma sonuçları

Şekil 53'te Naive Bayes sınıflandırıcısının farklı terim ağırlıklandırma yöntemleriyle sınıflandırma performansı karışılaştırılmıştır. Bu şekle göre Naive Bayes, cinsiyet ve yaş sınıflandırmasında mantıksal ağırlıklarla daha iyi sonuçlar üretmiştir, Eğitim için ise TF-IDF terim ağırlıkları yöntemiyle en iyi sonucu üretmiştir.

Kullanılan sınıflandırma algoritmalarının koşma sürelerini karşılaştırmak için Intel i3 işlemcili ve 3 GB ana bellekli bir bilgisayarda, cinsiyet kategorisinin mantıksal terim ağırlıklandırma yöntemiyle oluşturulan vektör uzay modeli, Naive Bayes, DVM ve KEYK sınıflandırma algoritmalarında sırasıyla çalıştırılmıştır ve sonuçlar Tablo 17'de gösterilmiştir.

Tablo 17. Naive Bayes, DVM ve KEYK sınıflandırma algorimaları için sınıflandırma süresi

Sınıflandırma Algoritması	Sınıflandırma Süresi (Saniye)
Naive Bayes	32
DVM	151
KEYK	85

Tablo 17'de de görüldüğü gibi Naive Bayes sınıflandırıcısı diğer iki sınıflandırma algoritmalardan daha hızlı bir şekilde sınıflandırma yapmaktadır.

Sonuç olarak Naive Bayes sınıflandırma yöntemi, Cinsiyet ve Yaş için Mantıksal terim ağırlıkları kullanarak sırasıyla %90,85 ve %93,51 doğruluk elde etmiştir. Aynı sınıflandırıcı, TF-IDF terim ağırlıkları kullanarak %88,74 doğrulukla eğitim kategorisinde de en iyi sonucu üretmiştir. Ayrıca Naive Bayes sınıflandırıcısı diğer iki sınıflandırıcıdan daha kısa bir sürede sonuçları üretmektedir.

5. ÖNERİLER

Bu çalışmada, Facebook kullanıcılarının yorumlarını analiz ederek yazarın cinsiyetini, yaşını ve eğitim düzeyini belirlemek üzre bir sistem geliştirilmiştir. Kullanıcıların sadece yorumları analize dahil edilmiş ve herhangi bir profil bilgisi, Ad, Soyad, Cinsiyet, doğum tarihi vs. kullanılmamıştır.

Bu çalışmayı daha da geliştirmek ve daha iyi sonuçlar elde etmek için aşağıdaki konular göz önünde tutulmalıdır.

- Türkçenin Sohbet (*Chat*) diline uygun bir imla kontrol algoritması veya aracı geliştirilmelidir.
- Türkçenin Sohbet (*Chat*) dilin uygun bir kök bulma algoritması veya aracı geliştirilmelidir.
- Bu çalışmada semantik analiz yapılmamıştır, dolayısıyla yorumların semantik analizi yapılarak daha iyi sonuçlar üretilip üretilemeyeceği incelenebilir.
- Bu tezin yazımı esnasında Facebook, yorumlar için yeni bir özellik geliştirmiştir. Bu özelliğe göre, artık yorumlarda belirli bir kişinin yorumuna cevap olarak yorum yazılabilmektedir. Bu konuyu göz önünde tutarak ve veri toplama böceğini yeni özellikle uyumlu çalışacak şekilde güncelleyerek yorum cevapları da analize tabi tutulabilir.
- Son olarak bir yorum hangi paylaşımın reaksiyonu olarak yazılmışsa, o paylaşımın kendi özellikleri de araştırmaya dahil edilerek, farklı sınıflardaki insanların farklı olaylardaki reaksiyonları da incelenebilir.

6. KAYNAKLAR

1. http://www.alexa.com/topsites/ Alexa, The Web Information Company, Alexa top 500 global sites. 22 Ocak 2013.

2. Anonim, Annual report of internet crime complaint center, Internet Crime Complaint Center (IC3), 2011.

3. Thelwall, M., Wilkinson, D. ve Uppal, S., Data mining emotion in social network communication: Gender differences in MySpace. Journal of the American Society for Information Science and Technology, 61, 1 (2009) 190-199.

4. Cheng, N., Chandramouli, R. ve Subbalakshmi, K., Author gender identification from text. Digital Investigation, 8, 1 (2011) 78-88.

5. Ellison, N.B., Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13, 1 (2007) 210-230.

6. Lenhart, A. ve Madden, M., Social networking websites and teens: An overview, Pew/Internet, 2007.

7. Wolak, J., Mitchell, K. ve Finkelhor, D., Online victimization of youth: Five years later, National Center for Missing & Exploited Children, 2006.

8. Ryan, A., Under the Hood: Hadoop Distributed Filesystem reliability with Namenode and Avatarnode. http://www.facebook.com/notes/facebook-engineering/under-the-hood-hadoop-distributed-filesystem-reliability-with-namenode-and-avata/10150888759153920, 9 Mayıs 2013.

9. Han, J., Kamber, M. ve Pei, J., Data mining: concepts and techniques, Morgan kaufmann, 2006.

10. Tusher, V.G., Tibshirani, R. ve Chu, G., Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences, 98, 9 (2001) 5116-5121.

11. http://rapid-i.com/content/view/181/190/ Rapid-i, RapidMiner. 10 Mayıs 2013.

12. http://www.kdnuggets.com/polls/2009/data-mining-tools-used.htm KDNuggets, Data Mining Tools Used Poll. 1 Mayıs 2013.

13. http://www.kdnuggets.com/polls/2011/tools-analytics-data-mining.html KDNuggets, Data Mining / Analytic Tools Used. 01 Mayıs 2013.

14. http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html KDNuggets, Data Mining / Analytic Tools Used Poll. 01 Mayıs 2013.

15. http://www.cs.waikato.ac.nz/ml/weka/ The University of Waikato, Weka 3: Data Mining Software in Java. 10 Mayıs 2013.

16. Damerau, F.J., Markov models and linguistic theory, Mouton, 1971.

17. Köse, C., Özyurt, Ö. ve Amanmyradov, G., Mining chat conversations for sex identification. Emerging Technologies in Knowledge Discovery and Data Mining, (2007) 45-55.

18. Köse, C., Özyurt, Ö. ve İkibaş, C., A comparison of textual data mining methods for sex identification in chat conversations. Information Retrieval Technology, (2008) 638-643.

19. Hariharan, S., Gender Prediction in chat based medium's using text mining. International Journal of Research and Reviews in Information Sciences (IJRRIS), 1, 1 (2011) 18-22.

20. Kobayashi, D., Matsumura, N. ve Ishizuka, M., Automatic Estimation of Bloggers' Gender, Proceedings of International Conference on Weblogs and Social Media; March 2007, Boulder, Colorado, U.S.A.

21. Mukherjee, A. ve Liu, B., Improving gender classification of blog authors, Proceedings of the 2010 conference on Empirical Methods in natural Language Processing; October 2010, Cambridge, MA, Association for Computational Linguistics, 207-217.

22. Prasath, R., Learning age and gender using co-occurrence of non-dictionary words from stylistic variations, Rough Sets and Current Trends in Computing; January 2010, Berlin Heidelberg, Springer, 544-550.

23. Rosenthal, S. ve McKeown, K., Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. Proc of ACL-11, (2011) 763-772.

24. Rustagi, M., Prasath, R., Goswami, S. ve Sarkar, S., Learning age and gender of blogger from stylistic variation. Pattern Recognition and Machine Intelligence, (2009) 205-212.

25. Yan, X. ve Yan, L., Gender classification of weblog authors, AAAI Spring Symposium Series Computational Approaches to Analyzing Weblogs; March 2006, Palo Alto, California, American Association for Artificial Intelligence 228-230.

26. Technical Report, Predicting gender from blog posts, University of Massachusetts, Amherst, USA, 2010.

27. Peersman, C., Daelemans, W. ve Van Vaerenbergh, L., Predicting age and gender in online social networks, Proceedings of the 3rd international workshop on Search and mining user-generated contents; October 2011, Glasgow, Scotland, UK, ACM, 37-44.

28. Burger, J.D., Henderson, J., Kim, G. ve Zarrella, G., Discriminating gender on Twitter, Proceedings of the Conference on Empirical Methods in Natural Language Processing; July 2011, Edinburgh, United Kingdom, Association for Computational Linguistics, 1301-1309.

29. Fink, C., Kopecky, J. ve Morawskib, M., Inferring Gender from the Content of Tweets: A Region Specific Example, Sixth International AAAI Conference on Weblogs and Social Media; May 2012, Dublin, Ireland.

30. Bamman, D., Eisenstein, J. ve Schnoebelen, T., Gender in Twitter: Styles, stances, and social networks. arXiv preprint arXiv:12104567, (2012) February

31. Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., Hu, W., Garzia, F., Tirocchi, N., Scarpiniti, M. ve Cusani, R., Gender Identification on Twitter Using the Modified Balanced Winnow. Communications and Network, 4, 3 (2012) 189-195.

32. Oflazer, K., Two-level description of Turkish morphology. Literary and linguistic computing, 9, 2 (1994) 137-148.

33. http://nlp.ceng.fatih.edu.tr/blog/?p=101 Fatih University, The Natural Language Processing Group, Turkish Stop Word List 1.1. 10 Ocak 2012.

34. Eryigit, G. ve Adali, E., An Affix Stripping Morphological Analyzer For Turkish, Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, 2004, Innsbruck, Austria, 299–304.

35. Hengirmen, M., Türkçe dilbilgisi, Engin Yayınevi, 2002.

7. EKLER

7.1. EK 1. Birinci Ekler Listesi

No	Ek	No	Ek
1	larimiz	18	lrnz
2	lariniz	19	lara
3	lerimiz	20	lere
4	leriniz	21	larr
5	lardan	22	lerr
6	lerden	23	daki
7	larim	24	deki
8	larin	25	lar
9	lerim	26	ler
10	lerin	27	lrm
11	larmz	28	lrn
12	larnz	29	lri
13	larda	30	lr
14	lerde	31	da
15	lari	32	de
16	leri	33	ta
17	lrmz	34	te

7.2. Ek 2. İkinci Ekler Listesi

eceklerimizde, eceklerinizde, ceklerinizde, ecekmişsiniz, yormuşsunuz, yormussunuz, eceklerimiz, ecekleriniz, eceklerimde, eceklerinde, yomuşsunuz, yomussunuz, diklarimiz, diklariniz, dikleriniz, diklerimiz, mişlerimiz, mişlarimiz, mişleriniz, mişlariniz, ceklerimde, ceklerinde, ceklermzde, ecekmişsin, ecekmişler, ecekmişsnz, yormuşsun, yormussun, yormuşlar, yormuslar, eceksiniz, eceklerim, eceklerin, ceklermde, ceklernde, ecekmişim, ecekmişiz, cekmişler, ecekmşsnz, ecekmişsn, ecekmişlr, ecektiniz, ecektiler, yorsunuz, yormuşum, yormusum, yrmuşsun, yomuşsun, yomussun, yormussn, yormuşuz, yormusuz, yormuşlr, yormuslr, yomuşsnz, yomussnz, yordular, yordunuz, diklarim, diklerim, diklerin, diklarin, dklrnzdn, dklrmzdn, miştiniz, miştiler, miştilar, mişsiniz, mişlerim, mişlarim, mişlerin, mişlarin, mişlermz, mişlarmz, mişlernz, mişlarnz, ecekleri, ceklermz, ceklernz, cklermde, cklernde, cklrmzde, cekmşsnz, ecekmşsn, ecekmişm, ecekmişz, cektiniz, cektiler, yorsunz, yrsunuz, yosunuz, yormuşm, yormusm, yomuşum, yomusum, yormşsn, yomuşsn, yomussn, yomuşuz, yrmuşuz, yomusuz, yomşlar, yomuslr, yrmşsnz, yrmssnz, yodunuz, yordunz, yordulr, diğimde, digimde, diğimda, digimda, diklari, dklarmz, dklarnz, dklernz, dklermz, dklrndn, dklrmdn, mişleri, mişlari, mşlermz, mşlarmz, mşlernz, mşlarnz, eceksin, ecekler, ceklerm, ceklern, cklermz, cklernz, cklrmde, cklrnde, cekmşsn, ecekmiş, ckmşsnz, cekmşlr, ecekmşz, ecekmşm, ecektim, ecektin, ecektik, ecektnz, ecektlr, yorsun, yorsnz, yrsnuz, yosunz, yosnuz, yorlar, yormşm, yrmuşm, yormsm, yrmşsn, yomşun, yrmssn, yomuşz, yormsz, yrmşlr, yrmslr, yomslr, yomşlr, ymşsnz, ymssnz, yordum, yorduk, yordun, yordnz, yrdunz, yodunz, yodulr, yrdulr, ydular, dklrmz, dklrnz, dklarm, dklarn, dklerm, dklern, dkleri, dklari, diktan, dikten, dgmzdn, dğmzdn, dğnzdn, dgnzdn, miştim, miştin, miştik, miştnz, mişsin, mişsnz, mişler, mişlar, mşlrmz, mşlrnz, eceğim, eceğiz, ecegiz, ecegim, cklerm, cklern, cklrmz, cklrnz, ckleri, cekler, cekmşm, ckmşsn, cekmşz, ckmşlr, ecekmş, ecekti, cektim, cektin, cektik, cektnz, cektlr, ecektm, ecektn, ecektk, ecktnz, ecktlr, yorum, yorsn, yrsun, yoruz, yrsnz, yonuz, yorlr, yolar, yorsa, yrmşm, yrmsm, yomşm, yomsm, yrmşz, yomşz, yomsz, yordu, yordm, yordn, yodun, yodum, yordk, yoduk, yrdnz, yodlr, yrdlr, diniz, diğim, diler, digim, dklrm, dklrn, dgmde, dğmde, dgmda, dğmda, dklri, dğmdn, dgmdn, dgndn, dğndn, miştm, miştn, mişti, miştk, mştnz, mştlr, mişim, mişiz, mşsnz, mşlrm, mşlrn, mşlri, ceniz, cklrm, cklrn, cklri, ckmşm, cekmş, madan, cekti, cektm, cektn, cektk, cktnz, cktlr, ecktm, ecktn,

Ek2.– Devamı

eckti, ecktk, yorm, yrum, yrsn, yosn, yorz, yruz, yonz, ynuz, yrlr, yolr, yosa, ymsz, yrdu, yodu, yrdm, yodm, yrdn, yodn, yrdk, yodk, ydnz, ydlr, dinz, digi, diği, dktn, mştm, mştn, mşti, mştk, mşsn, mşlr, ecek, ecez, cenz, ckmş, cktm, cktn, ckti, cktk, yor, yrm, ysn, yrn, yrz, yoz, ynz, ydm, ydn, ydk, ydu, din, dim, dik, dnz, dgm, dğm, dlr, dgi, dği, mşm, miş, mşz, cek, cem, cen, cnz, yr, yo, ym, yz, di, dn, dm, dk, mş, ck, cm, cn, cz

7.3. Ek 3. Değersiz Kelimeler Listesi

a, acaba, ama, ancak, anck, artık, artik, artk, asla, aslında, aslinda, aslnda, az, b, bazen, bazı, bazısı, bazi, bazilari, bazisi, belki, bezisi, bi, bile, birkac, birkaç, birsey, birseyi, birşey, bisey, biseyler, bisi, bişey, bişi, ble, blki, bnda, bndn, bole, bolece, boyle, boylece, böyle, böylece, bsey, bseyi, bseylr, bsy, bsyi, btn, bu, buna, bunda, bundan, bunu, burada, burda, butun, bzi, bzlri, bzn, bzsi, cnku, cogu, cogunuz, cunku, cünkü, da, daha, dan, de, degil, değil, den, diger, digeri, digerleri, diğer, dii, diil, diye, dlyi, dnr, dolayi, e, elbette, en, fakat, falan, felan, filan, fkt, fln, gbi, gene, gibi, gne, gore, gre, hala, hangi, hangileri, hangisine, hani, hatta, hc, hcbri, hcbrnz, hem, hepsi, hepsine, hepsini, her, herbir, herbiri, herkes, herkese, herkesi, herkez, hic, hicbiri, hiç, hiçbiri, hm, hngi, hnglri, hngsne, hpsi, hpsne, hr, hrks, hrkz, icin, icinde, icn, için, ile, ise, iste, işte, kac, kaç, kadar, kc, kdar, kdr, ki, kim, kime, kimin, kimisi, km, kmn, le, m, madem, mdm, mı, mısınız, mi, misin, misiniz, mu, musun, musunuz, mü, müsün, müsünüz, nasıl, nasil, ncn, ndir, ndn, ndr, ne, neden, nedir, nerde, nerede, nereden, nereye, nese, nesi, neyse, nicin, niçin, niye, nrde, nrdn, nrye, nsi, nsl, nye, o, oburu, ole, once, ondan, ondn, onlar, onlardan, onlr, onlrdn, orada, orda, otru, oturu, oyle, oysa, oysaki, oyski, ö, öbürü, öle, ön, önce, ötürü, öyle, ragmen, ramen, sa, sayet, se, sey, seyden, seye, seyi, seyler, si, simdi, smdi, sna, snda, sndn, snlr, snlri, snra, snu, sole, sonra, su, suna, sunda, sundan, sunlar, sunu, sunun, sy, syle, syt, şayet, şey, şimdi, şöyle, şu, şunda, şundan, şunlar, şunu, şunun, tabi, tamam, tan, tbi, ten, tm, tmm, tmu, tum, tumu, tüm, tümü, uzere, uzre, üzere, var, ve, veya, veyahut, vr, ya, yada, yan, yani, yerine, yine, yk, yksa, yne, yni, yok, yoksa, yri, yrne, zaten, zira, zra, ztn

ÖZGEÇMİŞ

Masoud (Mesut) TALEBİ, 1982 yılında Tebriz'de doğdu. Liseyi Ayetollah Talegani Lisesi'nde okudu. 2005 yılında Azad Shabestar Üniversitesi Bilgisayar Mühendisliği Bölümü'nü bitirdikten sonra, 2010 yılında Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim dalında yüksek lisans yapmaya başladı. Anadili Azerbaycan Türkçesi olmakla beraber, Türkçe, İngilizce ve Farsça dillerini çok iyi derecede bilmektedir.

1. Talebi, M. ve Köse, C., Facebook Yorumlarının Analiziyle Cinsiyet, Yaş Ve Eğitim Düzeyi Belirleme, 21. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı, Nisan 2013, Lefkoşa, KKTC.