T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TWITTER VERİLERİNİ ANLAMSAL SINIFLANDIRMA

MERİÇ MERAL

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN
DOÇ. DR. BANU DİRİ

İSTANBUL, 2014

ÖNSÖZ

Bu çalışma doğal dil işleme alanında Twitter verileri üzerinde yapılmıştır. Çalışmanın amacı, Twitter verilerini anlamsal olarak olumlu, olumsuz ve nötr sınıflandırabilen başarılı bir sınıflandırıcı oluşturulmasıdır. Bu amacı gerçekleştirebilmek için sözlük tabanlı ve n-gram olmak üzere 2 farklı doğal dil işleme yöntemi izlenmiştir. İzlenen 2 yöntem için farklı özellik vektörleri oluşturularak sınıflandırıcılar 9 farklı alandan toplanmış veriler ile eğitilmiştir. Bu çalışma ile farklı alanlardan toplanan veriler ve izlenen yöntemler sayesinde alandan bağımsız başarılı bir sınıflandırıcı oluşturulmuştur.

Çalışmamın her aşamasında bilgisi ve yol göstericiliği ile bana yardımcı olan ilgisini ve desteğini hiç eksik etmeyen danışmanım Sayın Doç. Dr. Banu Diri'ye sonsuz teşekkürlerimi sunarım.

Çalışmamın ilerlemesinde çok desteği olan arkadaşım Sadun Sözer'e ve bana yol gösteren Harun Kürşat Kofoğlu'na yardımları ve güveni için teşekkür ederim.

Çalışmamda bana her zaman büyük ilgi, destek ve anlayış gösteren aileme ve bütün arkadaşlarıma gönülden teşekkür ederim.

Mayıs, 2014

Meriç MERAL

İÇİNDEKİLER

Sayfa

KISALTMA LİSTESİ

API	Application Programming Interface
CFS	Correlation-based Feature Selection
DDİ	Doğal Dil İşleme
DVM	Destek Vektör Makinesi
JSON	JavaScript Object Notation
NB	Naive Bayes
NoSQL	No Structured Query Language
RO	Rastgele Orman
ROC	Receiver Operating Characteristic

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

TWITTER VERİLERİNİ ANLAMSAL SINIFLANDIRMA

Meriç MERAL

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Doç. Dr. Banu DİRİ

İnternetin büyümesi ve insanların kişisel fikirlerini paylaşmaları ile sosyal medya önemli bir bilgi kaynağı haline gelmiştir. Sosyal medya verileri ham haliyle çok kirli olduğu için üzerinde işlem yaparak gerçeğe uygun sonuçlar çıkarmak mümkün olamaz. Bu çalışmada, Twitter'dan toplanan veriler üzerinde algı analizi yapılmıştır. Bu analiz gerçekleştirilirken doğal dil işleme ve Naïve Bayes, Rastgele Orman ve Destek Vektör Makinesi gibi makine öğrenmesi yöntemleri kullanılarak, akıllı bir sistem oluşturulmuş ve elde edilen sonuçlar karşılaştırmalı olarak verilmiştir.

Anahtar Kelimeler: Sosyal medya, sınıflandırma, n-gram, sözlük tabanlı, kök bulma, Destek Vektör Makinesi, Rastgele Orman, Naive Bayes

ABSTRACT

SEMANTIC CLASSIFICATION OF TWITTER DATA

Meriç MERAL

Department of Computer Engineering

MSc. Thesis

Advisor: Assoc. Prof. Dr. Banu DİRİ

Social media has become an important information source with the expanding internet and ideas shared by the people. The social media raw data which is quite disordered and messy can not be processed as it is and obtain adequate results. In this study, sentiment analysis has been performed by collecting data from the Twitter. To perform this analysis an intelligent system has been created by using machine learning methods such as Naïve Bayes, Random Forest, Support Vector Machine and the compared results has been given.

Keywords: Social media, classification, n-gram, Word based, finding root, Support Vector Machine, Random Forest, Naive Bayes

BÖLÜM 1

GİRİŞ

İnternetin büyümesi ile insanların bu platformdaki paylaşımları, kişisel fikirleri ve bu fikirlerin alışverişi önemli bilgi kaynakları haline gelmiştir. Bu büyük bilgi kaynağı sadece araştırmacıların değil büyük ve orta büyüklükteki şirketlerin hatta hükümetlerinde ilgisini çekmektedir. Şirketler insanların satın aldığı ürün ve hizmetler hakkındaki fikirlerine ilgi duyarken, hükümetler belirli olaylar karşısında insanların algısıyla ilgilenmektedir. Böylece yapılan algı analizi ile geri bildirimlerden beslenerek aksiyon alınabilecek toplum tarafından olumsuz karşılanan bir durum düzeltilebilecektir. Aynı şekilde toplumsal algının olumlu olması durumunda doğru gerçekleştirilen aksiyonlar belirlenebilecek ve bu aksiyonlar desteklenmiş olacaktır. Gerçek dünyadaki bu tip uygulamaların dışında pek çok akademik çalışma sosyal medya verilerini kullanmaktadır. Sosyal medyadaki verilerin çeşitliliği (kişi, konu, lokasyon, veri tipi) akademisyenlerin ve araştırmacıların üzerinde her dalda çalışma yapmasını mümkün kılmaktadır.

Günümüz dünyasında rekabet halindeki firmalar, geleneksel pazarlama yöntemleri ile müşteriye ulaşmanın yeterli olmadığını görmüşler ve sahip olduğu potansiyel nedeni ile gündemlerine sosyal medyayı almışlardır. İnsanlar artık eskisi kadar televizyon seyretmemekte ya da radyo dinlememektedir. İnternetin ve sosyal medyanın büyümesi ile her geçen gün bunları günlük rutinleri arasına daha fazla yerleştirmekte olan insanlar, firmaların da sosyal medya ortamlarında daha fazla yer almasına neden olmaktadırlar. Artık neredeyse bütün büyük ve orta ölçekli firmalar sosyal medyayı kullanarak pazarlama yapmakta, müşteri görüşlerini öğrenmekte ve ürünlerini bu fikirler yönünde geliştirmektedir. Bu açıdan sosyal medya çok önemli ve büyük bir veri

kaynağıdır ve buradan elde edilen bilgi birçok sektöre yön vermektedir. İnsanların bir ürün ya da hizmet hakkındaki olumlu ve olumsuz görüşleri sosyal medyadaki ağ aracılığı ile bütün dünyaya yayılmaktadır. Sosyal medyada yayılan fikirler, ürün ve hizmetlerin ne kadar başarılı olduğunu yansıtmakta ve dolayısı ile satışları ve ekonomiyi etkilemektedir. Oluşan ekonomik etki zamanla firmaların sosyal medyadaki algıyı keşfetmeleri ihtiyacını doğurmuştur ve sosyal medyada algı analizi çalışmalarına ön ayak olmuştur.

Twitter verileri çoğunlukla ticari kaygılar doğrultusunda işlenmesinin yanında insanlığın yararı içinde kullanılabilmektedir. Bir doğal afet alanındaki durumun ve oluşan ihtiyaçların analizi buna en iyi örnektir [6]. Böylece Twitter verileri analiz edilerek doğal afetlerin türüne göre acil ihtiyaçların öncelikleri belirlenebilir ve bir sonraki doğal afette bu önceliğe göre aksiyon alınabilir. Twitter verilerinin bir diğer faydalı çalışma alanı olarak ilaç endüstrisidir. İlaç firmaları sosyal medya verileri üzerinde yaptıkları veri madenciliği çalışmaları ile ilaçların bilinmeyen yan etkilerini ve sıklıklarını keşfetmektedirler [3]. Bu çalışmalar her ne kadar kazanç kaygısı ile yapılsa da insan sağlığı ve yaşam kalitesini arttırdığı için iyi amaçlara hizmet ettiği söylenebilir. Ayrıca, sosyal medyanın gerçek zamanlı olarak izlenmesi pek çok çalışma konusunun önünü açmaktadır. Şikayet yönetim sistemleri bu konunun hedef uygulamalarının başında gelir. Ticari şirketlerden, halk sağlığı ve can güvenliği hizmetlerine kadar birçok alanda sosyal medya verileri gerçek zamanlı izlenebilir, buradan elde edilen bilgiler ışığında hızlı aksiyonlar alınarak insan sağlığı, güvenliği ve memnuniyeti arttırılabilir.

Açık fikirlerin sürekli sosyal medyada dile getirilmesi bu ortamlardaki verilerde büyük çeşitlilik oluşturmaktadır ve yukarıda ifade edilen nedenlerden dolayı anlamlandırılması ihtiyacını gündeme getirmiştir. Bunu sağlamak için verilerin bir ön işlemden geçirilerek süzülmesi, dönüştürülmesi gereklidir ve işlenen veriler ile sınıflandırma algoritmaları eğitilerek akıllı bir sistem oluşturulmalıdır. Bu çalışmada sosyal medya verilerini işleyerek sınıflandırabilen başarılı bir sistem geliştirilmesi hedeflenmektedir.

1.1 Literatür Özeti

Sosyal medya içerdiği verinin çeşitliliği ve büyüklüğü nedeni ile günümüzde oldukça popüler bir çalışma alanıdır. Ayrıca kullanım kolaylığı, özellikle fikirlerin ifade edildiği bir

ortam oluşu ve veriye erişimde sağladığı kolaylıklar nedeni ile Twitter diğer sosyal medya sitelerinden ayrılmaktadır. Bizde çalışmamızda algı analizini Twitter verileri üzerinde yaptığımız için, araştırmamızı veri kaynağı olarak Twitter'ı kullanan çalışmalar üzerinde yaptık. Twitter ile yapılan çalışmalar içerisinde doğal dil işleme ve veri madenciliği yöntemlerinin oldukça fazla olduğu görülmektedir. Genellikle çok büyük miktarda verilerin işlendiği bu çalışmalarda verilerin işlenmesinde yaşanan zorluklar ve bu zorlukların aşılmasında izlenen yöntemler yol gösterici niteliktedir.

"Naive Bayes and Unsupervised Artificial Neural Nets for Caneun Tourism - Social Media Data Analysis"[1] isimli çalışmada, 70 milyon Twitter verisi üzerinde veri madenciliği yapılarak Meksika'da yer alan Cancun adındaki tatil yerine giden turistlerin bu tatil yeri hakkındaki algısı analiz edilmektedir. Doğal dil işleme ile fiillerin köklerine indirgenmesi, üçseferden az geçen kelimelerin veri setinden çıkarılması, ağırlıklı frekans matrisi oluşturulması işlem adımları yapılmıştır. Naive Bayes algoritması ve ikili seçim anahtar sözcük(binary choice keyword) yöntemleri birleştirilerek tivitler üzerindeki algının ölçülmesi için melez bir metodoloji izlenmektedir. Algıyı modellemek için veri içinde grupların açıkça ayırt edilebildiği yöntem olan Kendini Örgütleyen Haritaları (Self Organizing Maps) ve turizm alan bilgisi kullanılmaktadır. Görselleştirilen veriler incelenerek birbirleri ile ilişkili kelimeler çıkarılmakta ve pozitif, negatif algı tespiti yapılmaktadır. Bu çalışma ile turist ve yolcuların sıkıntıları, ilgi duydukları konular keşfedilmekte ve hedef kitlesi turist ve yolcular olan turizm şirketleri için pratik bilgiler sağlanmaktadır.

"Analysis of the Relation Between Turkish Twitter Messages and Stock Market Index" [2] isimli çalışmada, 45 günlük zaman diliminde toplanmış 1.9 milyon Türkçe tivit kullanılarak borsadaki değişimin kişilerin ekonomi ile ilişkili attıkları tvitler ile bir ilişkisinin olup olmadığını belirlemek üzere bir araştırma yapılmıştır. Tivitler üzerinden sekiz farklı duyguya bağlı özellikler çıkarılmış ve veri seti mutlu ve mutsuz olmak üzere iki sınıflı olarak etiketlenmiştir. Çalışma sonucunda borsadaki değişimlerin tvitler ile %45 ilişkili olduğu bilgisi elde edilmiştir.

"Towards Large-scale Twitter Mining for Drug-related Adverse Events"[3] isimli çalışmada, toplanan 2 milyar tivit içerisinden ilaç kullanımı ile ilgili olanlar seçilerek

işlenmiş, ilaçlardan kaynaklanan yan etkiler ve buna bağlı diğer sorunların saptandığı bir sistem geliştirilmiştir. İlaçlardan kaynaklanan yan etkiler hastalar için büyük risk oluşturmaktadır. Bu nedenle yan etkilerin önceden tespit edilmesi hayati önem taşımaktadır. Günümüze kadar hastaların problemlerini kendi kendine raporlaması ve sonrasında bunları bir geri bildirim ile ilaç firmalarına sunması istenmekteydi. Bu çalışma ile raporlama ortamının yerini sosyal medya almakta ve burada paylaşılan bilgiler doğal dil işleme ve veri madenciliği teknikleri ile işlenmektedir. Alanda uzman kişiler tarafından etiketlenen veriler öncelikle bir doğal dil işleme sürecinden geçirilmiş, sonrasında Destek Vektör Makinesi ile sınıflandırılmıştır. Sistem içerisinde iki sınıflandırıcı barındırmaktadır: veriler öncelikle ilaç kullanıcılarının tespit edildiği bir sınıflandırıcıdan geçirilmekte ve sonrasında ikinci sınıflandırıcı ile ilaçların neden olduğu yan etkiler tespit edilmektedir. Sınıflandırmanın 1000 kez tekrarlandığı çalışmada %74 başarı elde edilmiştir.

"Predicting Collective Sentiment Dynamics from Time-series Social Media" [4] isimli çalışmada, beş aylık zaman dilimi içerisinde toplanan 12 milyon tivit içerisinden İngilizce olan 7 milyon tivit analiz edilerek geçmiş verilerde algının zamanla değişimine bağlı olarak, gelecek zaman penceresi içindeki algıyı tahminleyen bir sistem geliştirilmiştir. Sosyal medya dinamiğinin modellenmesi için çıkarılan 80 özellik ile seçilen veri madenciliği algoritmaları çalıştırılmıştır. Sınıflandırma modeli olarak Destek Vektör Makinesi, lojistik regresyon ve karar ağaçlarının kullanıldığı çalışmada Destek Vektör Makinesi %85'in üzerinde başarı ile diğer yöntemleri geride bırakmaktadır.

"Automated Twitter Data Collecting Tool for Data Mining in Social Network"[5] isimli çalışmada, Twitter API'sinin saatte 350 çağrı limiti nedeni ile otomatik veri toplayan Java tabanlı bir araç geliştirilmiştir. Geliştirilen araç kullanılarak 30 Ocak – 15 Şubat tarihleri arasında toplanan 5.1 milyon tivit ile yapılan çalışmada, 2012 yılında ABD'de düzenlenen Amerikan futbolu ligi şampiyonluk maçı olan Super Bowl sırasında yayınlanan reklamların araç üreticileri hakkında konuşulmaya etkisi analiz edilmektedir. Analiz edilen tivitler içerisinden anahtar sözcük olarak otomobil üreticileri, reklamlarda oynayan aktörlerin isimleri ve araç modelleri seçilmekte ve WEKA[29] ile tivitler görselleştirilmiştir. Görselleştirilen iki haftalık veride, Super Bowl sırasında araç markalarıyla ilgili tivit sayısında artış olduğu gösterilmiştir.

“Tweeting about the Tsunami - Mining Twitter for Informantion on the Tohoku Earthquake and Tsunami”[6] isimli çalışmada, Tohoku depreminden önce toplanmış 280 bin tivit ile depremden sonra toplanmış 1 milyondan fazla tivit analiz edilmektedir. Sosyal medyadan toplanan veriler deprem felaketi sırasında insanların nelere ihtiyaç duyduğu ve gereksinimlerin çıkartılması konusunda önemli bilgiler ihtiva etmektedir. Çalışmada bilginin keşfi için TAKMI adında bir veri madenciliği aracı geliştirilmiştir. Bu araç ile bilinmeyen örüntüler keşfedilebilmekte ve kelimeler arasında korelasyon hesaplanabilmektedir. Yapılan çalışmanın sonucunda geliştirilen araç ile korelasyon hesaplanarak tedarik sıkıntısı çekilen ihtiyaçlar tespit edilmekte, depremden sonra tedarik ve sıkıntı anlamları içeren sözcüklerin birlikte artan kullanım eğiliminde olduğu çıkarılmaktadır.

“A machine learning approach to Twitter user classification” [7] isimli çalışmada, Twitter kullanıcılarının davranış, iletişim ağı yapısı, kullanıcının tivitlerindeki içerik özellikleri kullanılarak sınıflandırma yapılmaktadır. Öğrenme algoritması olarak içerisinde birden fazla karar ağacı içeren Gradyan Artırımlı Karar Ağacı kullanılmaktadır. Bu algoritma daha hızlı çözümleme zamanı ve sahip olduğu daha küçük bir model yapısı ile çalışarak Destek Vektör Makinesi gibi algoritmalarla yarışmaktadır. Politik yönden Twitter kullanıcılarının sınıflandırıldığı deneysel çalışmanın sonucunda belirli sınıflarda %64 - %98 arasında başarı elde edilmiştir.

“Analysis and Classification of Twitter Messages” [8] isimli çalışmada, öğreticili makine öğrenmesi yaklaşımı kullanılarak destek vektör makinesini temel alan bir sınıflandırıcı geliştirilmiştir. Destek Vektör Makinesi algoritması 4.800 tivit ile eğitilerek %80 başarının elde edildiği çalışmada, n-gram gösterimi, eğitim verisi setinin büyüklüğü ve kelimelerin köke dönüşümünün sınıflandırma başarısına etkisi incelenmiştir. Yapılan deneylerin sonucunda n-gram gösteriminin ve kelimelerin köke dönüşümü yönteminin sınıflandırma başarısını arttırmadığı görülmüştür.

“Sentiment Analysis on Social Media” [9] isimli çalışmada, 1.000 adet Facebook iletisi analiz edilerek kamu yayın hizmeti veren Rai ile La7 isimli özel şirket hakkındaki algı karşılaştırılmaktadır. Çalışma kapsamında sistem içerisinde veri toplayıcı, anlam yorumlayıcı motor, doğal dil destekli arama motoru ve verileri kümeleyen sınıflandırma

motoru geliştirilmiştir. Anlamsal analiz çalışmasında cümle içindeki kelimeler anlamına (etken, hedef, neresi, neden vb.) ve kullanım yerine (isim, fiil, sıfat vb.) göre %96 doğrulukla etiketlenmiştir. Deneysel çalışmanın sonucunda anlamsal analiz %93 başarı göstermiştir.

"A Novel Data-Mining Approach Leveraging Social Media to Monitor and Respond to Outcomes of Diabetes Drugs and Treatment" [10] isimli çalışmada, şeker hastalığı konusuna adanmış "Diabetes Daily" internet forumundan toplanan iletiler ile şeker hastalarının kullandığı tıbbi araç ve ilaçlar ile ilgili deneyimler ölçülmektedir. Hastaların tıbbi malzemeler hakkındaki fikirlerini daha iyi anlamak için forum iletilerini sayısal olarak analiz etme konusunda kendiliğinden organize haritalar kullanılmıştır. "Rapidminer" isimli veri madenciliği aracı ile uzay vektör modeli kullanarak tıbbi ilaç ve cihaz ile ilgili her ileti olumlu, olumsuz ve nötr olarak etiketlenmektedir. Modelden çıktı olarak elde edilen olumlu, olumsuz ve nötr kelime listeleri ile veriler kendiliğinden organize haritalar kullanılarak görselleştirilmiştir. Elde edilen sonuçlara göre tıbbi malzemeler ile ilgili olumlu kelime kullanımının forumun çoğunluğuna hakim olduğu görülmüştür.

"Kemik Doğal Dil İşleme Grubu" [33] Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği Bölümü bünyesinde çalışan bir çalışma grubudur. Metin sınıflandırma/kümeleme, yazar tanıma, otomatik soru cevaplama, semantik, konuşma işleme, hayat bilgisi veritabanları, metin sıkıştırma, metin özetleme, duygu analizi alanlarında çalışan Kemik grubunun "CSdb" adında Türkçe hayat bilgisi veritabanı oluşturma, "Text2Arff" adında Türkçe metinlerin çeşitli özelliklerini otomatik çıkarılması ve "AutoDMOZ" adında DMOZ web ontolojisi için otomatik sınıflandırma ve anlamsal tanım çıkartma yapan projeleri vardır.

1.2 Tezin Amacı

Milyonlarca insan her gün sosyal medyada satın aldığı ürünler, kullandığı hizmetler ve markalar, kurumlar, ekonomi, siyaset, spor vb. konular hakkında olumlu ve olumsuz fikirlerini paylaşmaktadır. Sosyal medya siteleride (Facebook, Twitter, Linkedin ve Google+ vb.) insanların yayınladığı bütün bu içeriği sağladıkları servisler aracılığı ile paylaşmaktadır. Ancak, paylaşılan bu ham veride anlamsal bir etiket bilgisi (konu,

olumluluk, olumsuzluk vb.) bulunmamaktadır. Dolayısı ile bir konu hakkındaki algının ne olduğu bilgisi sosyal medyada yığın halindeki verinin içerisinde gizlidir. Bu çalışmada kullanılan doğal dil işleme ve veri madenciliği teknikleri ile Twitter verileri işlenerek bu bilginin çıkarımı sağlanacak ve veriler algısal olarak sınıflandırılacaktır.

Sosyal medya verileri işlenmemiş haliyle üzerinde çalışılması oldukça zordur ve anlamlı sonuçların elde edilebilmesi için birçok ön işlemden geçirilmesi gereklidir. Veriler incelendiğinde çoğunluğunun hatalı yazılmış kelimeler, kısaltmalar ve günlük konuşma dilinde kullanılmayan sosyal medyaya özgü jargon sözcüklerden oluştuğu gözlemlenmiştir. Bu büyük miktardaki veri işlenmemiş ham yapısıyla tek tek incelenmeye çalışıldığında insan algısıyla bile anlaşılması güçtür. Bu nedenle öncelikle amaca yönelik olarak sosyal medya verilerinin süzülmesi, işlenmesi ve belirli kalıplara sokulması gerekir. Verileri bu şekilde ön işlemden geçirirken doğal dil işleme yöntemlerinden faydalanılır. Bu işlemler çalışmamızın ilerleyen bölümlerinde detaylandırılacak, deneysel sonuçlar paylaşılacaktır.

Twitter, sosyal medyada en büyük ve en hızlı büyüyen ortamlardan birisidir. Sağladığı halka açık API sayesinde araştırmacıların ve veri analizi yapanların ilgi odağı haline gelmiştir. Ayrıca, 140 karakter sınırlaması yapılan paylaşımların içeriğini etkileyerek Twitter'ı bilginin doğrudan, kısa ve net ifade edildiği bir ortama dönüştürmüştür. Twitter'dan toplanan veriler ile doğal dil işleme ve veri madenciliği ile pek çok araştırma yapılmakta ve endüstride projeler gerçekleştirilmektedir. Bunlara örnek olarak salgınları önceden tahminleyen [23], ilaçların bilinmeyen yan etkilerini keşfeden [3], algının zamanla değişimini tahminleyen [4], turistik bir beldeye gelen turistlerin tivitleri üzerinde algı analizi yapan [1] çalışmalar verilebilir. Bu çalışma Twitter'ın sağladığı API ile verilere erişimin kolay, iletilerin içeriğinin kısa ve açık ifadeler bulundurması nedeniyle Twitter verileri üzerinde yapılmıştır.

Bu çalışmanın amacı, Twitter API'si kullanılarak 9 farklı alanda atılmış tivitlerin içerisinden süzülerek çıkarılmış etiketlenmiş veriyi baz alarak 'olumlu', 'olumsuz', 'nötr' sınıflandırma yapabilen başarılı bir sınıflandırıcının oluşturulmasıdır. 9 farklı alandan belirli anahtar sözcükler aratılarak oluşturulan veri seti ile sınıflandırıcının alana bağımlılığı azaltılmıştır. Her alan için toplanan veri seti kendi içinde alan bağımlıdır

ancak 9 farklı alandan toplanan veri setlerinin birleşimi alan bağımsız veri seti olarak kabul edilmektedir. Böylece alandan bağımsız olarak sınıflandırma yapabilen, dayanıklılığı yüksek başarılı bir sınıflandırıcı hedeflenmektedir.

Bunun yanı sıra yukarıda belirtilen 9 alanın her biri için oluşturulan veri setleri ile de testler yapılmıştır. Etiketlenen veriler ile aynı sınıflandırma yöntemleri kullanılarak her bir alanın verileri ile ayrı eğitilen ve eğitildiği alanda başarılı sınıflandırıcı geliştirilmesi de hedeflenmektedir. Sonuç olarak hem alan bağımlı hem de alan bağımsız veriler ile izlenen doğal dil işleme yönteminin başarısı karşılaştırılacak ve izlenen yöntemlerin hangi durumda daha başarılı olduğunu gösteren nesnel bir sonuç elde edilecektir.

İlerleyen bölümlerde başarılı bir sınıflandırıcının oluşturulması yolunda veriler üzerinde yapılan doğal dil işleme teknikleri, karşılaşılan zorluklar ve bu zorlukların aşılması için yapılan deneysel çalışmalar anlatılacaktır. Ayrıca, deneysel çalışmalarda kullanılan yöntemlerin 'Naive Bayes', 'Destek Vektör Makinesi' ve 'Rastgele Orman' sınıflandırma algoritmalarının başarısına etkileri gösterilecek ve karşılaştırılacaktır.

1.3 Hipotez

Günümüze kadar makine öğrenmesi teknikleri ile markette birlikte satılan ürünleri tespit ederek varlıkların birlikteliklerini bulan, bir dokümanın internetten benzerlerini öneren, bir dizi özelliklere sahip verileri kümeleyen, banka müşterilerinin kredi kartı kullanım sıklıklarına göre kredi skoru çıkaran, zaman içinde gerçekleşen olayları analiz ederek sıralı örüntüler bulan uygulamalar gibi pek çok çalışma yapılmıştır. Bu çalışmaların bazılarında üzerinde çalışılan veriler doğası gereği dağınık ve özellikleri çıkarılmamış iken, bazılarında ise oldukça düzenli ve işlenmeye hazırdır. Üzerinde çalışılan sosyal medya verileri ise çoğunluğu hatalı yazılmış kelimeler, kısaltmalar ve sosyal medya jargonunda ifade edilen kelimelerden oluştuğu için makine öğrenmesi teknikleri ile anlamlandırılabilmesi oldukça güçtür.

Sosyal medyanın genel özellikleri Twitter verileri üzerinde de görülmekle beraber yukarıda ifade edilen problemler ortamın yarattığı kısıtlamalardan dolayı daha yoğun bir şekilde görülmektedir. Bu durum Twitter verilerinin anlamlandırılmasının doğal dil işleme yapmadan başarılı olamayacağını göstermektedir. Bu nedenle verilerin

anlamlandırılarak başarılı bir sınıflandırıcının geliştirilebilmesi için çalışmada bir dizi doğal dil işleme yöntemi uygulanmıştır.

Doğal dil işleme sürecinde işlenecek veriler metinsel olduğu için biri karakter diğeri kelime düzeyinde olmak üzere iki farklı yol izlenerek daha başarılı bir sistemin oluşturulmasında en iyi sonucu veren yöntem tercih edilmiştir.

Sosyal medya sitelerinde iletiler incelendiğinde hatalı yazılan kelimelerin oldukça sık olduğu görülür. Sözlük tabanlı işleme sürecinde eğer bu hatalı kelimeler düzeltilmez ise her kelime için birden fazla özellik oluşur ve bu durum öğrenme algoritmasının başarısını düşürür. Geliştirilecek bir kelime öneri sistemi ile hatalı yazılan kelimeler düzeltilerek her kelime için bir özellik çıkarılacaktır. Böylece sınıflandırma başarısının yükselmesi hedeflenmektedir.

Aynı şekilde, Türkçe kelimelerin aldığı eklerden dolayı da her kelime için birden fazla özellik oluşur ve bu durum öğrenme algoritmasının başarısını düşürür. Oluşan çeşitlilikten dolayı sınıflandırma başarısındaki düşmenin önüne geçebilmek için kök elde etme işlemi her kelime için gerçekleştirilmelidir.

Sınıflandırma algoritmaları işlenen veriler ile eğitilerek en iyi sonucu veren ve eğitim veri setine göre en kararlı yöntem seçilmelidir. Bunun için seçilen ‘Naive Bayes’, ‘Destek Vektör Makinesi’ ve ‘Rastgele Orman’ algoritmaları kullanılarak sistem eğitilmeli ve başarıları belirlenen ölçüm değerleri ile karşılaştırılmalıdır.

BÖLÜM 2

TWITTER VERİLERİ

Twitter, kullanıcılarının “tivit” adı verilen en fazla 140 karakterden oluşan mesajlar göndermesine ve okumasına izin veren sosyal ağ ve mikroblog sitesidir. 2006 yılında Jack Dorsey tarafından kurulan Twitter, günümüzde en çok ziyaret edilen 10 internet sitesinden biridir.

Bloglarda olduğu gibi Twitter’daki mesajlar korumalı hesaplar hariç herkese açıktır. Twitter kişilere hesaplarından gönderdikleri mesajları onu takip etmeyenlere göstermeme seçeneği sunmaktadır. Bu koruma dışında Twitter üzerinde paylaşılan bütün mesajlar halka açıktır. Twitter, paylaşılan mesajlar üzerinde sorgu çalıştırılabilmeyi mümkün kılan gelişmiş bir arama motoruna sahiptir. Bu arama motoruna kelime, kullanıcı (kişi), yer, tarih, durum gibi kriterler verilerek ihtiyaca göre sonuçlar elde edilebilmektedir.

Twitter’ın en bilinen ve belirgin özelliği gönderilen mesajların 140 karakter oluşudur. Bunun sebebi cep telefonlarının mesajlarının 160 karakter ile sınırlı olmasıdır. Çünkü Twitter ilk zamanlarında özellikle mobil mesajlaşma için kullanılıyordu ve günümüzde de bu geçerliliğini korumaktadır. Bundan yola çıkarak Jack DorseyTwitter’ı kurarken mesajlaşma uzunluğunun mevcut kısa mesaj uzunluğu sınırında kalmasına karar vermiştir. Twitter 20 karakteri kullanıcı adı için ayırır ve diğer 140 karakteri ise mesajın kendisi için kullanır. Böylece Twitter mesajları cep telefonlarından da kolaylıkla gönderilebilir. Aynı zamanda 140 karakter ile mesajlaşmak, mesajların bilgiyi doğrudan vermesini sağlayarak kolay okunması ve yazılmasına olanak sağlar.

Twitter'ın bir özelliği de belirlenen kişileri takip edilebilmesidir. Böylece sadece takip edilen kişilerin mesajları ana sayfa üzerinde görüntülenir. Karşılıklı takip zorunluluğu olmadığı için, takip edilen kişi tarafından tanınıyor olmaya gerek yoktur. Twitter'ın önemli diğer iki özelliği etiketleme ve bahsetme özellikleri jargonu ilgilendirdiği için bölüm 2.1'de anlatılmaktadır.

2.1 Twitter Jargonu

Twitter'ın 140 karakter sınırlamasından dolayı kullanıcılar yazdıkları tivitlere mümkün olabildiğince çok bilgi sığdırma stratejileri geliştirmişlerdir [8]. Bu stratejilerin başında diyez(#) karakteri gelir. Kullanıcıların kullandığı ve Twitter arama motorunun da desteklediği diyez karakteri(#) yazılan belirli bir başlıkla tivitler etiketlenebilir. Bu etiketleme sistemi ile kullanıcılar tivitin konusunu belirleyebilmekte ve kullanıcı arabiriminden etiketlenen başlık seçildiğinde, Twitter çalıştırdığı sorgu ile o etikete sahip bütün tivitleri listeleyebilmektedir. Örnek olarak etiketlenen bir tivit aşağıdaki gibi görünür:

'Savaşa hayır! Yaşasın #dunyabarisi'

işaretinden sonra anahtar kelime veya kelime gruplarının gelmesiyle oluşan etiketler insanların bir duyguyu, düşünceyi ya da eylemi ifade etmede en çok kullandığı yöntemlerden birisidir. Bu çalışmada toplanan veriler içerisinde en popüler etiketler belirlenerek özellik listesine eklenmiştir. Çalışmanın detayları bölüm 3.1'de anlatılmaktadır.

140 karakter sınırlamasının Twitter jargonuna bir etkisi de çok fazla kısaltmanın varlığıdır. Kullanıcılar gönderdikleri iletilerde daha fazla bilgiye yer verebilmek için okuyanın anlayabileceği şekilde kelimeleri kısaltma eğilimi göstermektedir. Bunlara 'slm' (selam), 'hrksn' (karikasın), 'lol' (kahkaha) gibi örnekler verilebilir. Bu tip kısaltmaların işlenme yöntemleri bölüm 3.1'de detaylı olarak anlatılmaktadır.

Twitter kullanıcıları diğer sosyal medya sitelerinde olduğu gibi duygularını ifade etmek için karakterler kullanmaktadır. Birden fazla karakteri bir araya getirerek ':)' (mutlu), ':(' (mutsuz), '<3' (sevgi) gibi duygu ifadeleri oluşturulabilmektedir. Bu ifadeler ile ilgili çalışma yine bölüm 3.1'de detaylandırılmaktadır.

2.1.1 Twitter İstatistikleri

2014 yılı 1 Ocak'ta Twitter hakkında alınan istatistik bilgileri [11] şöyledir:

- Twitter'da toplam aktif kullanıcı sayısı 645.750.000'dur ve her gün 135.000 yeni kullanıcı kayıt olmaktadır.
- Her ay 190 milyon tekil kullanıcı Twitter'ı ziyaret etmektedir ve 115 milyonu aktif kullanıcıdır.
- Günde ortalama 58 milyon tivit atılmakta, Twitter arama motorunda 2.1 milyon sorgu çalıştırılmaktadır.
- Twitter'ın reklam gelirleri 2011 yılında 139 milyon dolar, 2012 yılında 259 milyon dolar ve 2013 yılında 405 milyon 500 bin dolar olmak üzere sürekli artma yönündedir.

2.2 Verilerin Toplanması

Bu çalışma kapsamında verilerin toplanması için geliştirilen sosyal medya arama katmanı, Twitter API'sine erişerek belirlenen anahtar sözcükler ile sorgulama yapmakta ve sorgulama sonucunda dönen verileri ayrıştırarak anlaşılır hale getirmektedir. Arama katmanı Java ortamında geliştirilmiştir ve Twitter API'sine erişiminde Twitter4J [22] java kütüphanesi kullanılmıştır. Bu bölümde verilerin toplanması için çalıştırılan sorgular, Twitter API'den gelen veriler ve bu verilerin anlaşılır hale getirilerek saklanması aşamaları anlatılmaktadır.

2.2.1 Twitter API

Çalışmamızda Twitter API[21] 1.1 sürümü kullanılmıştır. Twitter API'nin tivit arama servisini(search/tweets) kullandığımız bu sürümde çalıştırılan sorguların sonucunda dönen veriler json formatındadır. Json formatındaki veriler içerisinde bir tivit ile ilgili birçok bilgi gelmektedir: ileti, yazar, retivit sayısı, coğrafi bilgi, dil, takipçi sayısı. Json formatındaki bu bilgiler daha sonra sosyal medya arama katmanı ile ayrıştırılmakta ve uygulamanın anlayabileceği varlık yapısına dönüştürülmektedir.

Twitter API'sine yapılan sorgularda operatörler dahil en fazla 1.000 karakter sınırlaması vardır. Çalıştırdığımız sorgularda ortalama karakter sayısı oldukça düşük olduğu için bu

sınırlama yapılan çalışmada bir engel oluşturmamıştır. Twitter API'sine yapılan sorgulamalarda tarih aralığı belirtmek için 'since' ve 'until', konum belirtmek için 'locale' ve anahtar sözcükleri aramak için de 'q' en sık kullanılan kriterlerdir.

2.2.2 Çalıştırılan Sorgular

Çalışmamızda veriler,9 farklı alan için belirlenen anahtar sözcükler Twitter API ile 07.06.2013 – 13.07.2013 tarihleri arasında sorgulanarak toplanmıştır. Alanlar ve kullanılan sorgular hakkında detaylı bilgiler Çizelge 2.1'de verilmektedir:

Çizelge 2. 1 Twitter'da ilgili alana ait verileri toplamak için sorgulanan anahtar sözcükler

Alan ID	Alan Adı	Anahtar Sözcükler
1	Telekomünikasyon	vodafone, turkcell, ttnet, gprs, port, proxy, cep telefonu, bilgisayar, adsl, iletisim baskanligi, elektronik, TIB, AR-GE, 3G, tablet bilgisayar, genpa
2	Sağlık	doktor, eczane, ilaç, reçete, acil müdahale, hemsire, steteskop, bas agrisi, bas dönmesi, mide bulantisi, kan tahlili, röntgen, asi, kanser, kolestrol, ameliyat, acil servis
3	Finans	banka, imkb, ihlas finans, yapi kredi, aa finans, finans sektörü, is bankasi, akbank, ziraat bankasi, ekonomi, girisimcilik, merkez bankasi, borsa, finans masasi, döviz, para, dolar, euro, bütçe, finans kaynagi
4	Spor	fenerbahçe, galatasaray, besiktas, trabzonspor, maç, gol, halisaha, galibiyet, malubiyet, futbol, voleybol, basketbol, tenis, real madrid, barcelona, hakem, sampiyonlar ligi, avrupa ligi, fifa, uefa
5	Sigorta	poliçe, araba sigortasi, saglik sigortasi, ferdi kaza sigortasi, ana teminat, anadolu sigorta, günes sigorta, axa sigorta, kasko, sigortalar kurumu, trafik sigortasi, sigorta sirketi, sigortalar birligi, sigorta primi, sigorta sektörü, ücretsiz sigorta, sigorta emeklileri, özel sigorta
6	Gıda	market, abur cubur, saglikli ürünler, hormonlu gidalar, hormonsuz gidalar, konserve, helal gida, gida bankasi, gida yardimi, gida zehirlenmesi, gida mühendisligi, gida boyasi, gida güvenligi
7	Otomotiv	mercedes, bmw, audi, otomobil, araba galerisi, motor, antifriz, hava yastigi, turbo, vites, süspansiyon, egzos, ford, citroen, volkswagen, porsche, sanziman, beygir gücü, ferrari, yakit tüketimi, araç muayene, rot balans

Çizelge 2. 1 Twitter'da ilgili alana ait verileri toplamak için sorgulanan anahtar sözcükler (Devam)

8	Gayrimenkul	toki, satilik ev, kira, konut, kiralik ev, gecekondu, satilik daire, kiralik daire, haciz, ipotek, tapu, arazi, tapu kadastro, gayrimenkul yatirim ortakligi, gayrimenkul sahibi, gayrimenkul fiyatlari, GYO, ali agaoglu, emlak, konut vergisi, soyak, gayrimenkul
9	Siyaset	siyaset, siyasi, anayasa, parti, seçim, hükümet, bakan, muhalefet, cumhurbaskani, vali, komisyon, tayyip erdogan, devlet bahçeli, milletvekili, devlet, veto, kemal kiliçdaroglu, referandum, basbakan, meclis, divan, gensoru

Çizelge 2.1'de belirtilen anahtar kelimeler ile yapılan sorgulamalara bir örnek verilecek olursa:

https://api.twitter.com/1.1/search/tweets.json?count=100&locale=tr&since:2013-06-25&until:2013-06-26&q=otomotiv

Bu sorguda 2013-06-25 ve 2013-06-26 tarihleri arasında, konumu Türkiye olan ve içinde otomotiv kelimesi geçen 100 tivit isteği Twitter API'sine gönderilmektedir.

Çalıştırılan sorgular ile arama sonucunda "json" formatında dosyalar dönmektedir. Bu dosyalar geliştirilen sosyal medya arama katmanı tarafından işlenerek anlaşılır hale getirilmektedir. Verilerin işlenmesinde öncelikle "json" formatındaki veriler ayrıştırılır. Ayrıştırma sonucunda her tivit için 'yazar', 'ileti', 'gönderilme tarihi' gibi birçok özellik elde edilir. Sosyal medya arama katmanı bu özellikleri birleştirerek ileti adı verilen varlıklara dönüştürür. Son aşamada 'json' formatından ileti varlığına dönüşen tivitleri saklamak için Solr veritabanına yazılmak üzere veritabanı yönetimi katmanına gönderilir.

2.3 Verilerin Saklanması

Büyük miktarda verilerin saklanıp üzerinde hızlı arama yapabilmek için NoSQL veritabanları araştırılarak bunların içerisinden Solr seçilmiştir ve bu çalışmanın içerisinde kullanılmıştır. Yapılan geliştirme sonucunda NoSQL kullanan ve bir arama motoru görevi gören kararlı bir veri tabanı yönetim katmanı ortaya çıkarılmıştır.

Solr sağladığı indeksleme yapısı ile büyük veriler üzerinde hızlı sorgular çalıştırılmasını mümkün kılmaktadır. Kurulan Solr veritabanında aynı şema yapısına sahip 9 koleksiyon yaratılmıştır. Koleksiyonlar kendilerine ait indeksleri olan belirli bir şema yapısına sahip veri saklama alanlarıdır. Solr'ın şema yapısı verinin nasıl saklanması gerektiğini ifade eder ve indekslenmesi gereken alanlar ve bu alanların veri tipi bilgisi şemada yer alır. Bu çalışma kapsamında oluşturulan Solr şema yapısı Şekil 2.1'deki gibidir.

```
<schema name="SearchItem" version="1.4">
    <types>
        <fieldType name="string" class="solr.StrField" sortMissingLast="true" omitNorms="true"/>
        <fieldType name="int" class="solr.TrieIntField" precisionStep="0" omitNorms="true" positionIncrementGap="0"/>
        <fieldType name="long" class="solr.TrieLongField" precisionStep="0" omitNorms="true" positionIncrementGap="0"/>
        <fieldType name="date" class="solr.TrieDateField" omitNorms="true" precisionStep="0" positionIncrementGap="0"/>
        <fieldType name="text_general" class="solr.TextField" positionIncrementGap="100">
        <analyzer type="index">
            <charFilter class="solr.MappingCharFilterFactory" mapping="mapping-ISOLatin1Accent.txt"/>
            <tokenizer class="solr.WhitespaceTokenizerFactory"/>
            <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords_tr.txt" enablePositionIncrements="true"/>
            <filter class="solr.TurkishLowerCaseFilterFactory"/>
            <filter class="solr.SnowballPorterFilterFactory"/>
        </analyzer>
        <analyzer type="query">
            <charFilter class="solr.MappingCharFilterFactory" mapping="mapping-ISOLatin1Accent.txt"/>
            <tokenizer class="solr.LetterTokenizerFactory"/>
            <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords_tr.txt" enablePositionIncrements="true"/>
            <filter class="solr.TurkishLowerCaseFilterFactory"/>
            <filter class="solr.SnowballPorterFilterFactory"/>
        </analyzer>
        </fieldType>
    </types>
    <fields>
        <field name="entryId" type="string" indexed="true" stored="true" required="true"/>
        <field name="sourceId" type="long" indexed="true" stored="true"/>
        <field name="searchKriteriaId" type="long" indexed="true" stored="true"/>
        <field name="searchId" type="long" indexed="true" stored="true"/>
        <field name="statusId" type="long" indexed="true" stored="true"/>
        <field name="title" type="text_general" indexed="true" stored="true"/>
        <field name="link" type="string" indexed="true" stored="true"/>
        <field name="entryDate" type="date" indexed="true" stored="true"/>
        <field name="languageCode" type="string" indexed="true" stored="true"/>
        <field name="authorName" type="text_general" indexed="true" stored="true"/>
        <field name="authorLink" type="string" indexed="true" stored="true"/>
        <field name="searchDate" type="date" indexed="true" stored="true"/>
        <field name="tagList" type="string" indexed="true" stored="true" multiValued="true"/>
        <field name="extUserId" type="string" indexed="true" stored="true"/>
        <field name="packageId" type="long" indexed="true" stored="true"/>
        <field name="readFlag" type="int" indexed="true" stored="true" default="0"/>
        <field name="friendsCount" type="int" indexed="true" stored="true" default="0"/>
        <field name="followersCount" type="int" indexed="true" stored="true" default="0"/>
    </fields>
    <uniqueKey>entryId</uniqueKey>
    <defaultSearchField>title</defaultSearchField>
</schema>
```

Şekil 2. 1 Solr şema yapısı

Solr veritabanında oluşturulan 9 adet koleksiyonun içinde alana göre belirlenmiş anahtar sözcüklerin sorgulanmasıyla elde edilen 2 milyonun üzerinde veri tutulmaktadır. Her koleksiyonda ilgili alan için tanımlanmış anahtar sözcüklerin

Twitter'da aratılmasıyla dönen tivitler saklanmaktadır. Koleksiyonlarda yer alan her bir kayıt şema yapısına uygun oluşturulur ve kaydedilir. Tivitlerin içeriği bu kayıtlarda yer alan bir alana yazılır ve bu alan veritabanına yazma sırasında indekslenir. Böylece verilen bir anahtar kelime ile tivitin içeriğinin tutulduğu alan üzerinde sorgu çalıştırılarak ihtiyaca uygun veriler elde edilebilir. Solr veritabanında çalıştırılan sorgular ilişkisel bir veritabanına göre daha hızlıdır. Çünkü sorgu tek bir şema üzerinde çalıştırılır ve başka bir şema ile ilişkisi olmadığı için veriler başka bir yere referans olmaksızın tek bir indeks üzerinden elde edilebilir. Bu çalışmada gerektiğinde milyonlarca tivit üzerinde sorgu çalıştırılacağı ve gelecek çalışmalarda tivit sayısının daha da artması ihtimaline karşı ilişkisel olmayan bir veritabanı tercih edilmiştir.

2.4 Verilerin Etiketlenmesi

9 farklı alandan toplanan ve insan eli ile sınıflandırılan veriler hakkında detaylı bilgiler Çizelge 2.2'de verilmiştir:

Çizelge 2. 2 Alan bazlı etiketli tivit sayıları

Alan ID	Alan Adı	Olumlu	Olumsuz	Nötr
1	Telekomünikasyon	161	685	172
2	Sağlık	92	314	299
3	Finans	115	498	376
4	Spor	196	417	566
5	Sigorta	468	686	690
6	Gıda	30	102	299
7	Otomotiv	35	139	185
8	Gayrimenkul	39	83	211
9	Siyaset	208	794	461

Tivitler işlenmeden ham halleriyle insan eli ile etiketlenmiştir. Etiketleme işlemi 8.321 adet tivit için yapılmıştır. Bu tivitler kullanıcıların karşısına Solr veritabanındaki 2 milyon tivit içerisinden rastgele seçilerek listelenmiştir ve etiketlemeleri sağlanmıştır. Etiketleme işlemi 4 kişi tarafından yapılmıştır ve yapılan işlemin kişisel algıya göre farklılık göstermesi mümkün olacağından aynı tivit en az 2 kişi tarafından etiketlenmiştir. Bu da bir kişi tarafından etiketlenmiş tivitlerin sistem tarafından seçilerek tekrar farklı bir kullanıcıya görüntülenmesi ile sağlanmıştır. Böylece verinin sınıf doğruluğu birden fazla kişinin onayı ile doğrulanmıştır. Verilerin farklı alanlardan

toplanmasındaki amaç alana bağımlılığı azaltarak genel anlamda sosyal medya verileri üzerinde başarılı ve alan değişimine dayanıklı bir sınıflandırıcı oluşturmaktır.

Veri setleri etiket sayısı dikkate alınarak homojen hale getirilmiştir. Her alan için eleman sayısı en az sınıf baz alınarak sayıları indirgenmiştir. Sınıf eleman sayıları her alanın kendi içinde homojen hale getirilmiştir ve toplam 4.032 adet tivitin yer aldığı veri seti oluşturulmuştur. Bölüm 5'te yapılan deneysel çalışmalar hem homojen hem de homojen olmayan veri setleri üzerinde yapılmıştır.

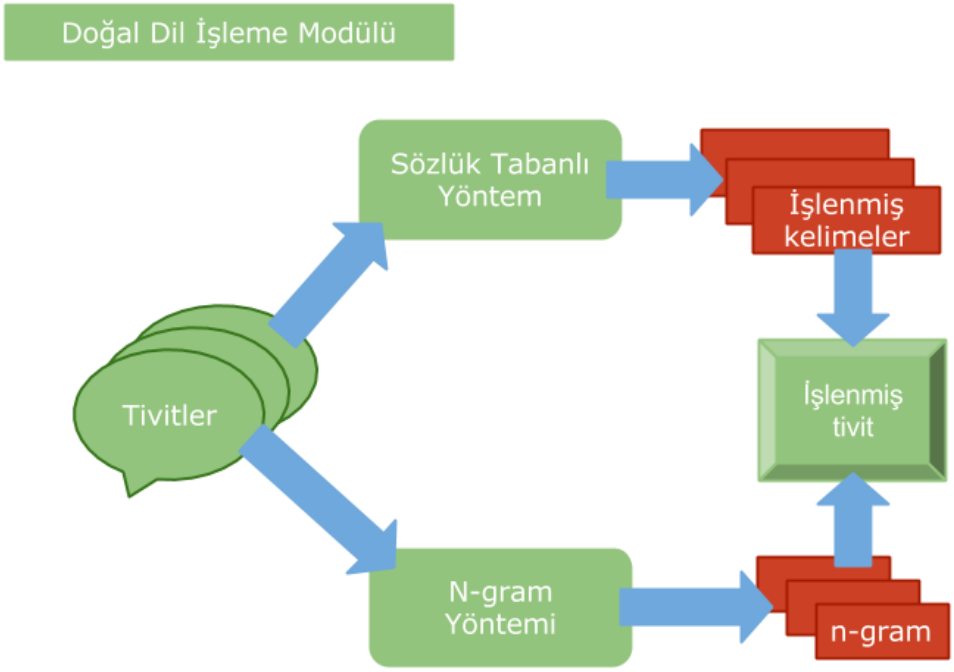
Çizelge 2. 3 Alan bazlı etiketli homojen tivit sayıları

Alan ID	Alan Adı	Olumlu	Olumsuz	Nötr
1	Telekomünikasyon	161	161	161
2	Sağlık	92	92	92
3	Finans	115	115	115
4	Spor	196	196	196
5	Sigorta	468	468	468
6	Gıda	30	30	30
7	Otomotiv	35	35	35
8	Gayrimenkul	39	39	39
9	Siyaset	208	208	208

BÖLÜM 3

VERİLERİN İŞLENMESİ

Twitter verilerinin sınıflandırılmasında 2 farklı yöntem uygulanmıştır. Birinci yöntemde tivitler kelimelere ayrıştırılarak işlenirken, ikinci yöntemde karakter n-gramlar oluşturulmuştur. Bu iki yöntemi birbirinden ayıran en temel fark birinin karakter diğerinin ise kelime düzeyinde uygulanmasıdır. N-gram oluşturma işlemi basitçe karakterlerin ardışık olarak ayrıştırılması iken, sözlük tabanlı yöntemde ise verilerin birçok ön işlemden geçirildiği karmaşık bir sürece sahiptir. Doğal dil işleme modülünün genel veri işleme yapısı Şekil 3.1'de verilmektedir.



Şekil 3. 1 Doğal dil işleme modülü veri akış şeması

3.1 Sözlük Tabanlı Yöntem

Türkçe bir tivit incelendiğinde içerdiği kelimeler açısından oldukça çeşitliliğe sahiptir. Kullanılan bir kelimenin değişik çekimlere sahip halleri, farklı yapım ekleri ile farklı anlamlara kavuşması, sosyal medyada kullanılan jargonun Türkçedeki kelimelerin yerini alması bu çeşitliliğe verilebilecek örneklerden birkaçıdır. Ayrıca, tivitler oldukça gürültülü yapıdadır. Anlam ifade etmeyen kelimeler, yanlış yazılmış kelimeler, Twitter'a ait özel ifadeler bunlara örnek verilebilir.

Twitter verileri yukarıda sayılan özelliklerinden dolayı üzerinde çalışılması oldukça zor, insan gözüyle bile anlamının anlaşılması kolay olmayan bir yapıdadır. Bu anlaşılmazlığın ortadan kaldırılabilmesi ve verilerin temizlenmesi, düzeltilmesi için birçok önişlemden geçirilmesi gerekmektedir. Bu önişlemler veriyi sınıflandırma algoritmasının ayrımını yapabileceği özelliklere kavuşturarak başarılı bir sınıflandırıcının yaratılmasını sağlayacaktır.

Türkçe sondan eklemeli bir dildir. Bir kelime 10'dan fazla ek alabilir ve herhangi bir fiil veya isim üzerinden istenildiği kadar sözcük türetilebilir. Türkçedeki bu zenginlik verilerin sınıflandırılması önünde bir engel teşkil etmektedir. Herhangi bir kök veya gövde aldığı birden fazla çekim ekiyle onlarca türevi oluşturulabilir ve bu durum bir sınıflandırma işleminde her türev için yeni bir özellik çıkarılması anlamına gelir. Özelliklerin bu derece artması algoritmanın başarısını düşürmektedir. Bunun nedeni bir kelimeye anlamını veren kısmının gövdesi oluşudur. Türkçede olumsuz eki ("-ma", "-me") dışında diğer çekim ekleri olumluluk ya da olumsuzluğu belirleyecek bir özelliğe sahip değildir. Bu duruma örnek olarak, yaptığımız çalışmada geçmiş zaman çekiminin başarıyı nasıl etkileyebileceği üzerine bir deneme yapılarak başarıyı etkilemediği hatta başarıyı düşürdüğü görülmüştür. Bir sınıflandırma algoritmasında kelimenin gövdesini kullanmak yerine onun birden fazla türevini özellik olarak almak o özelliğin sınıflandırmada belirleyiciliğini azaltmaktadır.

Bu nedenden dolayı Türkçe tivitlerin doğru sınıflandırılabilmesi için kelimeler kök, gövde ve eklerine ayrıştırılmıştır. Kelimelerin ayrıştırılması için Türkçede oldukça başarılı olan açık kaynak kodlu Zemberek [25] kütüphanesi kullanılmıştır. Zemberek sahip olduğu Türkçe sözlüğü sayesinde kelimelerin çözümlemesini yaparak kök ve

eklerine ayrıştırabilmektedir. Doğru yazılmış Türkçe kelimelerin büyük bir çoğunluğu çözümlenerek gövdeleri elde edilmiş ve olumsuzluk eki dışındaki çoğul, zaman çekim, hal, iyelik, ilgi gibi çekim ekleri temizlenmiştir. Kelimenin çözümlenerek kökünün eklerden ayrıştırılması ve olumsuzluk eklerinin belirlenmesi aşağıdaki örneğin işlenmesiyle sağlanmaktadır.

"sevmiyorum" kelimesinin çözümlemesi:

[Kok: ***sev****, FIIL] Ekler:* ***FIIL_OLUMSUZLUK_ME*** *+ FIIL_SIMDIKIZAMAN_IYOR + FIIL_KISI_BEN*

Yukarıdaki çözümlemede kök "sev" ve ekler içerisinde fiil olumsuzluk eki "-me" tespit edilmiştir. Olumsuzluk ekleri kelimenin anlamını değiştirdiği için kelimelerin olumsuz halleri de özellik olarak kabul edilmiştir ve bir dönüşüm uygulanarak olumsuzluk ifadesi kelimenin başına "not_" eklenerek sağlanmıştır. Çizelge 3.1'deki iki özellik bu duruma örnektir.

Çizelge 3. 1 "sev" özelliğinin olumlu ve olumsuz 2 durumu

Özellik No	Özellik Adı
1	sev(mek)
2	not_sev(mek)

Zemberek doğru yazılmış kelimeleri çözümleyebilirken, yanlış yazılmış kelimeleri kök ve eklerine ayrıştıramamaktadır. Bu durumun aşılabilmesi için kütüphanenin önerme fonksiyonu kullanılmıştır. Önerme fonksiyonu yanlış yazılmış kelimelere içindeki uzaklık algoritmasını kullanarak sözlükteki benzer kelimeleri öneri olarak sunmaktadır. Öneriler çoğunlukla birden fazla karakterin yanlış yazıldığı kelimelerde çok doğru olmamaktadır. Bu fonksiyonun kullanımında birden fazla önerinin dönmesi durumuna çok sık rastlanmıştır ve bunlardan hangisinin doğru olduğunu bulmak oldukça zor bir işlemdir. Bu nedenle işlenen verilerin %1'inden daha azını düzeltebilmiştir. Burada Zemberek'in öneri fonksiyonuna ek olarak Apache Lucene'in içinde yer alan Levenshtein uzaklık algoritması kullanılmıştır. Sözlük olarak Zemberek ile aynı veri setini verdiğimiz bu algoritma yaptığı tekil öneriler ile sosyal medya verilerinde daha başarılı olmuştur ve başarısından dolayı kelimelerin düzeltilmesinde bu yöntem tercih edilmiştir.

Twitter getirdiği 140 karakter sınırlaması ile tivitlerin içeriğini etkilemiş ve kendine özgü bir jargonun oluşmasına neden olmuştur. Karakter sayısındaki sınırlama insanların duygu ve düşüncelerini ifade etmek için yetersiz kalmış ve bu durum kullanılan kelimelerin farklılaşmasına neden olmuştur. Böylece çoğunlukla ünlü harflerin yutulduğu hem Türkçe hem de İngilizce ifadeler içeren bir sosyal medya jargonu oluşmuştur. Duygusal ifadeleri de karakterlerle ifade edebilen bu jargona çeşitli örnekler verilebilir:

Olumlu jargon: lol, ^.^, :), :D, tşkkr, tbrk, yuppi, cnm...

Olumsuz jargon: :(, :|, cıks, :S...

Jargonların bir durumu ifade etmeyi kolaylaştırdığı için çok sıklıkla kullanıldığı gözlemlenmiştir. Bu nedenle sosyal medya verilerinin sınıflandırılması için önemli bir etken olarak kabul edilmiştir. Veriler üzerinde yapılan gözlemler ile jargon sözlüğü oluşturulmuş ve sözlükteki ifadelerin herbiri sınıflandırıcı için bir özellik olarak kabul edilmiştir. Oluşturulan sözlük sınıflandırılamayan birçok Türkçe tivitin özelliklerini içerdiği için sınıflandırma başarısını arttırmıştır.

Twitter'da mükerrer veriler oldukça fazla miktardadır. Bunun nedeni birçok kişinin başka kişilerin tivitini yani paylaşımını kendi sayfasında yayımlamasıdır. "Retivit" (retweet) olarak ifade edilen paylaşımların tekrar yayınlanması durumu Twitter'da yeniden yayınlanan içeriğin başına "RT" ifadesinin getirilmesiyle gerçekleşmektedir. Yapılan her retivit, Twitter'da yeniden mükerrer bir verinin oluşmasına neden olur. Verilerin tekrar tekrar işlenmemesi ve sınıflandırıcıya verilen eğitim verisinde tekrar eden verilerin engellenmesi için oluşan mükerrer kayıtların toplanan Twitter verileri içerisinden temizlenmesi gereklidir. Bu iş için düzenli ifadeler (regular expressions) kullanılmıştır. Yazılan düzenli ifade sayesinde "RT" ile başlayan tivitler toplanan Twitter veri setinin içerisinden temizlenmiştir.

Twitter'da tekrar eden veriler sadece retivitler değildir. Birinden bahsetme, söz etme durumları için kullanılan menşınlar(mention) da oldukça çok tekrar eden verilerdir. Bu ifadeler "@" karakterinden sonra bahsedilmek istenen kişinin Twitter kullanıcı adı belirtilerek kullanılır. "@kullanıcıadı" şeklinde kullanılan menşınlar çok sık tekrar etmelerinin yanında tivitin anlamında bir olumluluk ya da olumsuzluk ifade

etmemektedir. Bu nedenle retivitler gibi bu ifadelerin de verilerin içerisinden temizlenmesi gerekli görülmüştür. Tekrar düzenli ifadeler kullanılarak "@" karakteri ile başlayan kelimeler tivitlerin içerisinden temizlenmiştir.

Her dilde olduğu gibi Türkçede de bir cümlenin anlamını olumluluk ya da olumsuzluk açısından etkilemeyen durak sözcükler(stop words) vardır. Tivitleri işlemeden önce bu çok sık tekrar eden kelimeler oluşturulan bir sözlük aracılığıyla temizlenmiştir. Durak sözcük sözlüğünün içerdiği kelime tiplerine örnek olarak aşağıdakiler verilebilir:

- Sayılar(dokuz, yüzelli, bin vb.)
- Zamirler(ben, sen, o, şu, orası, nereye, kim vb.)
- Edatlar(göre, diye, sanki vb.)
- Bazı isimler ve fiiller(et(-mek), ol(-mak), şey vb.)

Yukarıda anlatılan bütün yöntemlerin izlenme sırası ve yapılan işlemler sırasıyla aşağıdaki gibi özetlenebilir:

1. Tekrarlardan ve "RT"lerden temizlenen veriler tivitler halinde okunur.

2. Her cümle için

2.1 Kelimelere ayrılır

3. Her kelime için

3.1 Zembereğin denetle metoduyla kelimenin doğru yazılıp yazılmadığı test edilir.

3.2 Lucene'in Levenshtein uzaklık algoritması ile bulunan sonuç varsa bu öneri, yoksa zemberek kütüphanesinin öner metoduyla herhangi bir öneri olup olmadığı değerlendirilir.

4. 3.1 ve 3.2'de kelime denetleme işlemi başarısız olmuş ise veya herhangi bir öneri yapılamamış ise önerilmeyenler şeklinde farklı bir dosyaya kaydedilmek üzere saklanır.

5. 3.2'den bir öneri geldiği durumda 6. adımdaki işlemler yapılır.

6. if'in else kolunda kelime için Lucene'in Levenshtein uzaklık algoritması ile bulunan en iyi öneri alınır. Lucene ile öneri üretilemez ise zembereköner (kelime) metodundan gelen ilk öneri alınır.

7. Bu öneri kelimesi Zemberek kütüphanesinin çözümleme metoduyla kelime dilbilgisi çözümlemesi yapılır.

8. Eğer çözümlemenin içerisinde 'FIIL' geçiyorsa

8.1 Bu fiilin kökü alınır.

8.2 Eğer çözümlemede "FIIL_OLUMSUZLUK_ME", "FIIL_OLUMSUZLUK_DEN" ya da "FIIL_OLUMSUZLUK_SIZIN" eklerinden biri geçiyorsa ve "FIIL_ISTEK_E" ve "FIIL_DONUSUM_ILI" geçmiyorsa

Fiil "not_" öneki alınarak geçerli cümleye eklenir.

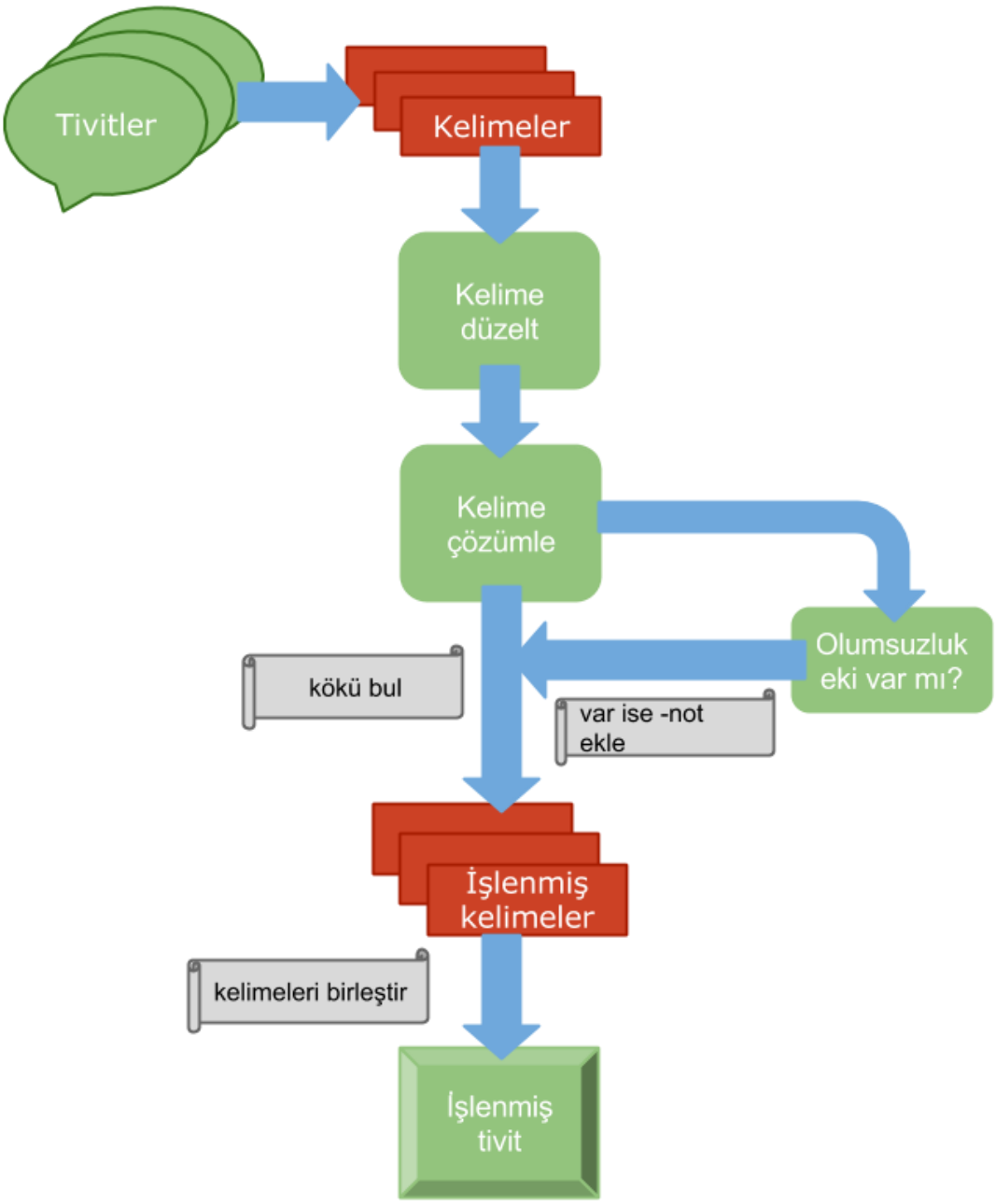
Aynı zamanda özellik listesine eklenmek üzere not_"fiil" haliyle bir dosyaya yazılır.

8.3 8.2'deki koşul sağlanmadığı durumda fiil olduğu gibi mevcut cümleye eklenir.

9. Eğer 8'deki koşul sağlanmaz ise ve fiil değil ise kök hali mevcut cümleye eklenir.

10. Bu işlemler 2'den 9'a kadar tüm cümleler için tekrarlanır.

Sonuç olarak köklerine ayrılmış çözümlenmiş kelimeleri içeren tüm sonuçlar elde edilmiş olur. Yukarıda anlatılan adımlar akış olarak özetle Şekil 3.2'deki gibidir.



Şekil 3. 2 Sözlük tabanlı yöntem için izlenen akış

Eğitim verisi olması amacıyla toplanan tivitler üzerinde yapılan incelemeler sonucunda içinde jargonların, kısaltmaların, olumlu ve olumsuz kelime köklerinin bulunduğu 589 özellik çıkarılmıştır. Bu 589 özellik kullanılarak ön işlemden geçirilmiş veriler bir özellik vektörüne dönüştürülerek sınıflandırılacaktır.

3.2 N-gram'lar

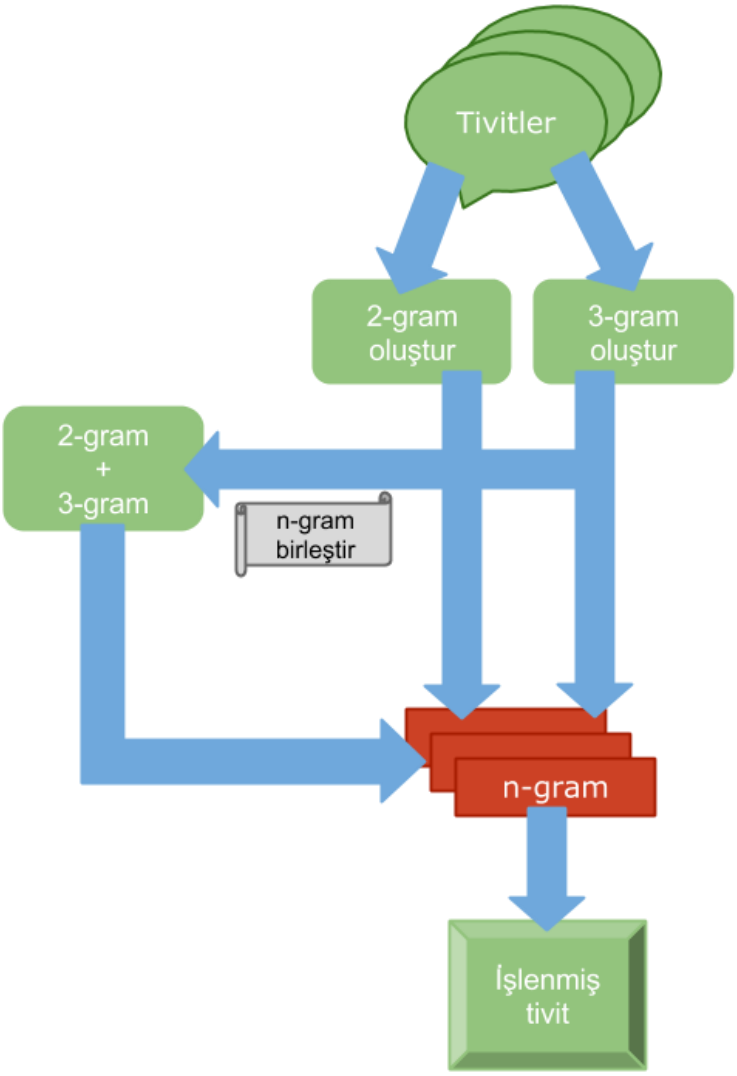
N-gram'lar bir dizi verinin ardışık sıralı n elemanlı alt kümelerinin bulunduğu yöntemdir. Bu yöntemle dizi halindeki veriler üzerinde gözle görülemeyen gizli örüntüler keşfedilebilmektedir. Dizi halindeki veriler bir olay dizisi, protein ya da DNA dizisi, metinsel verideki bir karakter ya da kelime dizisi olabilir. Metinsel veriler için n-gram'ların dil algılama, konuşma tanıma, gibi oldukça fazla kullanım alanı vardır. Biz de çalışmamızda bu yöntemi sosyal medya verilerinin anlamlandırılmasında kullandık.

Çalışmamızda n-gram birimlerimiz olarak karakter ya da kelime olmak üzere 2 yöntem izlenebilirdi. Birim olarak karakter seçilmesi durumunda Twitter jargonunda kullanılan çeşitli kısaltma ve yansıma ifadelerin yakalanması oldukça mümkün görülmektedir. Ayrıca, sık kullanılan olumlu ve olumsuz ifadelerin içinde tekrarlanan karakter dizilerinin keşfedilmesinde de oldukça faydalı olacağı düşünülmüştür. Birimler kelime olarak seçildiğinde ise verilerden kaynaklı örüntülerin bozulduğu gözlemlenmiştir. Kelimelerin birlikte kullanımı ve sırası içerdiği kelimelerin farklı ya da hatalı yazımından kaynaklı başarılı n-gram dizileri çok az sayıdadır. Bu nedenle n-gramla izlediğimiz yöntemde karakter dizilerini kullandık.

İzlediğimiz n-gram yönteminde birden fazla yaklaşım kullanılmıştır. Tivitler, bigram, trigram olmak üzere iki farklı şekilde sıralı karakter dizileri haline getirildi. Bu yaklaşımlardan bigramda, tivit içerisinde geçen kelimeler ikili karakterler halinde ayrıştırıldı. Tivitler önce kelimelere bölündü, daha sonra karakter bazında diziler elde edildi. Eğitim verisindeki bütün karakter dizileri elde edildikten sonra mükerrer kayıtlar tespit edilerek tekilleştirme işlemi gerçekleştirildi. Böylece elde ettiğimiz karakter ikilileri modellenecek sınıflandırıcı için birer özellik haline getirildi. 8.321 adet tivitin ayrıştırılmasıyla 1.604 adet bigram elde edilmiştir.

Tivitler, bigram yönteminden sonra trigramlar için işleme sokulmuştur. Bu yaklaşımda tivit içerisinde geçen kelimeler öncelikle kendi arasında, daha sonra kelime içinde üçlü karakterler halinde ayrıştırıldılar. Tivitlerin ayrıştırılması sonucu 10.067 adet trigram elde edilmiştir. Bu kadar çok sayıda olmasının nedeni hem bigramlara göre daha çok varyasyona sahip olması, hem de Türkçe tivitlerdeki karakter dizilerinin çok farklılık göstermesidir.

Tivitlerde geçen bigram ve trigramları oluşturulması ile elde edilen özellik listesi birleştirilerek de 2-gram ve 3-gram'ların birleşiminden oluşan bir özelik listesi elde edilmiştir. Doğal dil işleme modülünde n-gramların oluşturulması için izlenen akış Şekil 3.3'deki gibidir.



Şekil 3. 3 N-gram yöntemi için izlenen akış

Sonuç olarak hem sözlük tabanlı, hem de n-gram yöntemleri kullanarak özellik listeleri elde edilmiştir. Oluşturulan özellik listeleri hakkında bilgiler Çizelge 3.2'de yer almaktadır. Doğal dil işleme modülünde elde edilen özellik listeleri bir sonraki aşamada kullanılarak eğitim verisi özellik vektörlerine dönüştürülmektedir. Özellik vektörleri de kullanıldığı yöntemin adına göre birleştirilerek Weka[29] kütüphanesinin işleyebileceği

düzendeki arff dosyaları oluşturulmuştur. Bu aşamadan sonra oluşturulan arff dosyaları sınıflandırma modülünde Bölüm 5'te anlatıldığı gibi test edilmektedir.

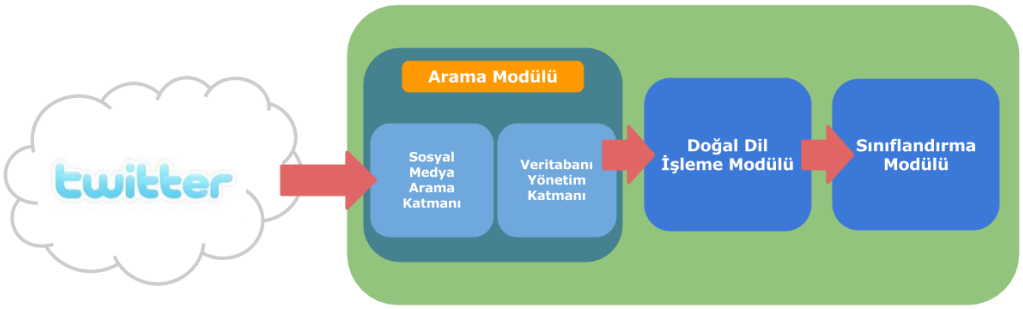
Çizelge 3. 2 Özellik Listeleri

Yöntem	**Özellik Yapısı**	**Özellik Sayısı**
Sözlük tabanlı	Kelime kök ve gövdeleri	589
2-gram	2-gram karakterler	1604
3-gram	3-gram karakterler	10067
2-gram + 3-gram	2-gram ve 3-gram karakterler	11671

BÖLÜM 4

GELİŞTİRİLEN SİSTEM

Bu çalışmada Şekil 4.1'de görülen sosyal medya erişim katmanı, veritabanı yönetim katmanı, doğal dil işleme modülü, sınıflandırma modülü olmak üzere 4 uygulama bileşeninden oluşan bir sistem geliştirilmiştir. Geliştirilen sistem bir sunucu uygulaması olarak çalışmaktadır. Uygulama geliştirme sürecinde Apache Tomcat ve Oracle Weblogic sunucuları üzerinde istihdam edilmiştir ve testleri bu ortamlarda gerçekleştirilmiştir. Sistem her iki sunucuda da ek bir konfigürasyon gerektirmeden sorunsuz çalışmıştır. Bu bölümde geliştirilen sistemin bileşenleri tanıtılmakta, birbirleri arasında ve dış sistemler ile nasıl iletişim kurdukları anlatılmaktadır.



Şekil 4. 1 Geliştirilen Sistem

4.1 Sosyal Medya Erişim Katmanı

Sosyal medya erişim katmanı projenin içerisinde kütüphane olarak uygulamayla entegre çalışmaktadır. Bu modülün işlevi, sosyal medya kanallarıyla veri alış verişini

sağlamaktır. Bu iş için sosyal medya kanallarının özel uygulama parçalarını kullanmaktadır.

Twitter'a erişim için java ortamında geliştirilmiş Twitter4J kütüphanesi kullanılmıştır. Bu kütüphane sağladığı uygulama programlama arayüzü (API) ile anahtar sözcük bazlı arama yapma, mesaj gönderme/silme, uygulamaya kaydolma, direkt mesaj gönderme, arkadaş listesine ekleme gibi fonksiyonlar sağlamaktadır. Bu modülde bölüm 2'de anlatılan önceden belirlenmiş sözcükler Twitter4J kütüphanesinin sağladığı arayüz ile aratılarak Twitter verileri toplanmaktadır. Bu arama işlemi yazılan bir zamanlayıcı uygulama parçası ile periyodik hale getirilmiştir ve uygulama kullanıcısının belirlediği periyotlara göre sürekli arama yapmaktadır.

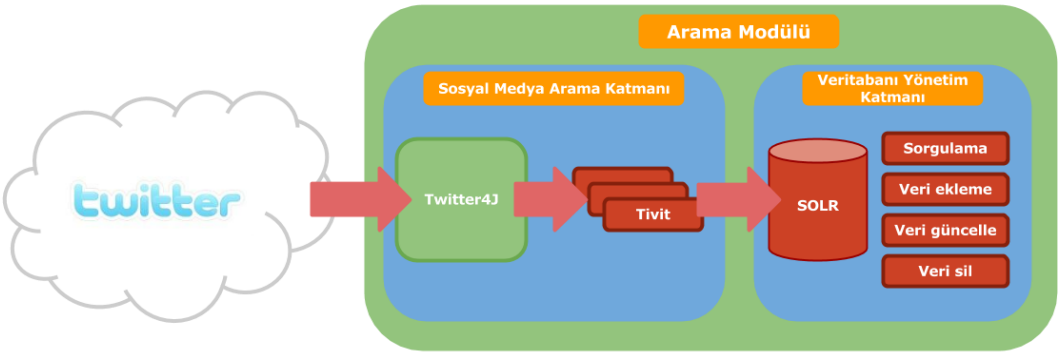
Zamanlayıcı uygulama parçası yüklenmesi ve operasyonu açısından zorluk yaşanmaması için mevcut uygulamanın paketlenmiş halinin farklı bir kopyasında diğer modüllerden bağımsız çalışabilmekte ve birden fazla farklı sunuculara kurulabilmektedir.

4.2 Veritabanı Yönetim Katmanı

Veri tabanı yönetim katmanı Solr veritabanı üzerindeki verilerin eklenmesi, güncellenmesi, silinmesi ve arama işlemlerini yöneten modüldür. Solr, ilişkisel olmayan (NoSQL) verilerin saklandığı bir veritabanıdır. Hızlı sorgulama yapma ve büyük miktarda verileri yüksek performansla yazma ve okuma özelliğinden dolayı tercih edilmiştir. Solr verileri kendi özel formatıyla dosya sistemi üzerinde tutmaktadır. Veritabanı yönetimi katmanı aynı zamanda arama motoru modülü olma özelliği ile bu veriler üzerindeki sorgu çalıştırma işlemlerini yönetmektedir.

Sosyal medya verileri öncelikle Sosyal medya erişim katmanı ile toplanmakta ve Solr veritabanına yazılması için veritabanı yönetim katmanına verilmektedir. Bu veri akışı Twitter'dan Solr veritabanına kadar Şekil 4.2'de görülmektedir. Bu katmana erişim diğer modüllerin kullanabileceği bir programlama arayüzü ile gerçekleşmektedir. Böylece diğer katmanlar verilerin nasıl yazıldığı ya da okunduğu ile ilgilenmeksizin ilgili işleri bu katmana yaptırmaktadır. Katmanda programlama arayüzleri kullanıldığı için verilerin yazıldığı veritabanı da ihtiyaçlar doğrultusunda farklı bir veritabanı ile kolayca

değiştirilebilir. Bunun için yeni kurulacak veritabanı için gerekli arayüzlerin gerçekleştirimini sağlamak yeterlidir.

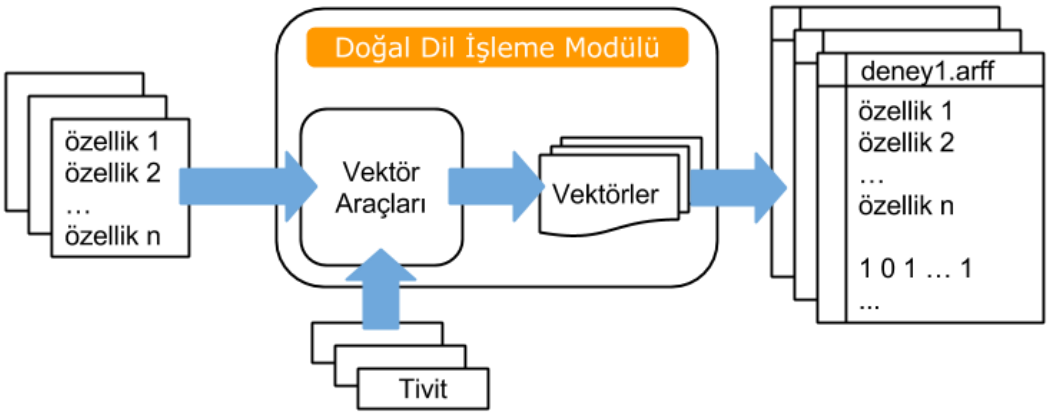


Şekil 4. 2 Arama Modülü

Bu katmandan erişimi sağlanan Solr veritabanı geliştirilen sistem gibi kendi başına bir sunucu uygulamasıdır ve Apache Tomcat, Oracle Weblogic sunucuları üzerinde yapılan testlerde başarılı bir şekilde kurulduğu ve kararlı çalıştığı görülmüştür.

4.3 Doğal Dil İşleme Modülü

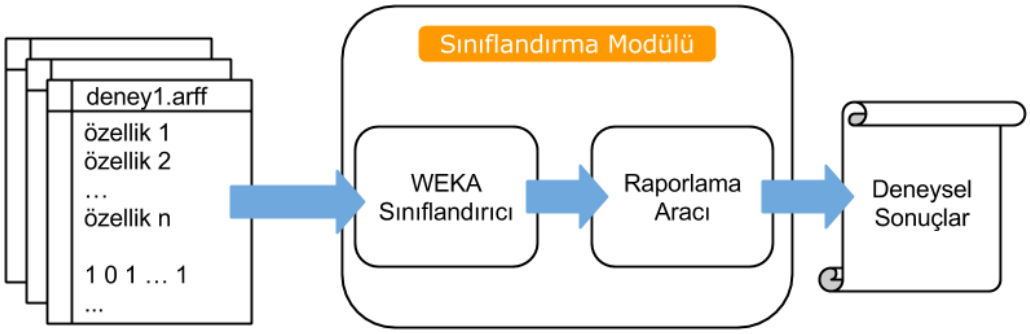
Doğal dil işleme(DDİ) modülü Bölüm 3'te anlatılan süreçlerin yürütüldüğü uygulama bileşenidir. Bu modülde bütün işlemler araç sınıflar ve bunları kullanan bir yönetici sınıf vasıtası ile gerçekleştirilmektedir. Veritabanı yönetim katmanı ile okunan veriler bu modül ile işlendikten sonra Weka[29] kütüphanesinin işleyebileceği 'arff' doküman formatına dönüştürülerek dosya sistemine kaydedilir. Arff dosyalarını oluşturabilmek için dosya sistemindeki özellik listelerinden okuyan modül her bir veri kaydını bu özellikleri kullanarak özellik vektörüne dönüştürür. Bu özellik listeleri önceden doğal dil işleme modülünde sözlük tabanlı ve n-gram yöntemleri ile oluşturulmuştur. Sözlük tabanlı, 2-gram, 3-gram ve bunların birleşiminden oluşan özellik listeleri ile vektörlere dönüştürülen tivitler 'arff' dosyalarında saklanarak sınıflandırma modülünde yer alan Weka[29] kütüphanesi ile sınıflandırılmaktadır. Bu sürecin özet hali Şekil 4.3'de yer almaktadır.



Şekil 4. 3 Doğal Dil İşleme Modülü

4.4 Sınıflandırma Modülü

Sınıflandırma modülü belirli yöntemler izlenerek işlenmiş veriler ile sınıflandırma algoritmalarının eğitildiği ve test edildiği modüldür. Weka'nın[29] 3.6.3 sürümünün kullanıldığı bu modül geliştirilen araç sınıflar ile 'arff' dosya formatındaki veri setini açarak sistemi eğitmektedir. 10 parçalı çapraz doğrulama ile eğitilen sistem sınıflandırma modülünün sağladığı programlama arayüzü ile dış bir sistem uygulamaya erişerek etiketsiz test verisini sınıflandırabilmektedir. Dışarıdan erişen bir sistem sınıflandırma modülündeki programlama arayüzüne test verisi, sınıflandırma algoritması ve izlenen yöntemini parametre olarak verdikten sonra sonuç alabilmektedir. Bu sürecin özet hali Şekil 4.4'de yer almaktadır.



Şekil 4. 4 Sınıflandırma Modülü

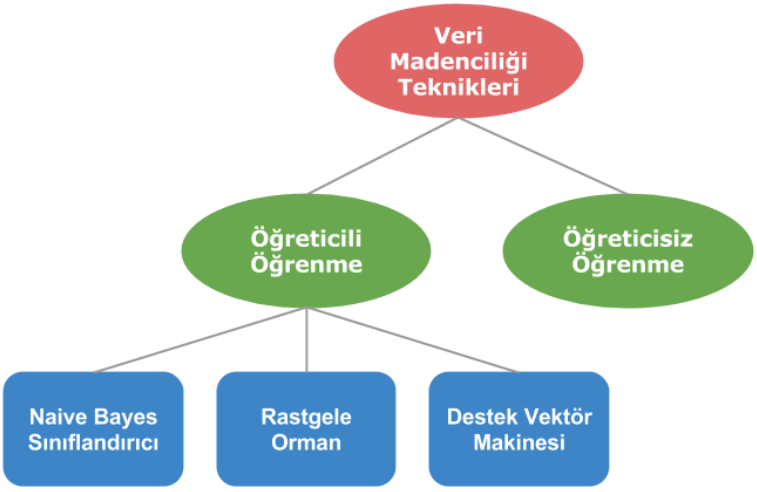
BÖLÜM 5

DENEYSEL SONUÇLAR

Bu bölümde Bölüm 3'de anlatılan yöntemlerle işlenen veriler ile sınıflandırma algoritmaları eğitilerek geliştirilen sistemin başarısı ölçülecektir. İzlenen yöntemlerin başarısını ölçmek için Naïve Bayes, Rastgele Orman ve Destek Vektör Makinesi sınıflandırıcıları kullanılmıştır. Seçilen bu algoritmalar ile izlenen yöntemlerin tutarlığını ve başarısını karşılaştırırken çoğunlukla f-ölçüm [25] ve bazı yerlerde alıcı işletim karakteristiği (orijinal adıyla; Receiver Operating Characteristic - ROC) [24] değerleri kullanılmıştır. F1 skoru olarakda bilinen f-ölçüm değeri kesinlik (precision) ve duyarlılık (recall) değerleri ile hesaplanır ve ikili sınıflandırma probleminin testinde doğruluğu gösterir. ROC eğrisi, sınıflandırma probleminde 2 sınıf için ayrım eşik değerinin farklılık gösterdiği durumda hassasiyetin kesinliğe olan oranıdır. ROC daha basit anlamıyla doğru pozitif sınıflandırılmış verilerin sayısının, yanlış pozitif sınıflandırılmış olanların sayısına oranıdır.

5.1 Sınıflandırma Algoritmaları

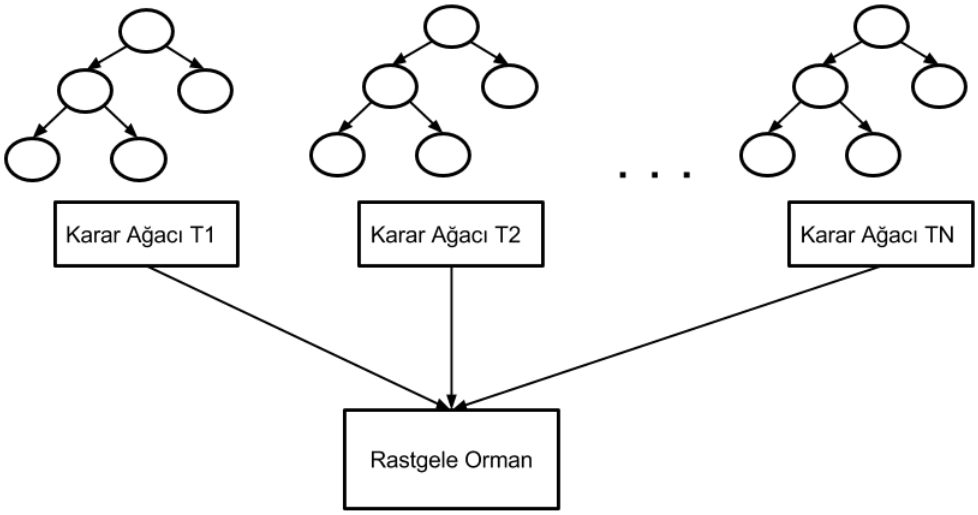
Bu bölümde deneysel çalışmalarda kullanılan ve Şekil 5.1'de de gösterilen öğreticili öğrenme algoritmaları anlatılmaktadır. Sırasıyla Rastgele Orman, Naive Bayes ve Destek Vektör Makineleri ile ilgili çalışma mantığı anlatılacak ve ilgili formüller paylaşılacaktır.



Şekil 5. 1 Sınıflandırma Yöntemleri

5.1.1 Rastgele Orman

Deneysel çalışmalarda kullandığımız ilk algoritma olan Rastgele Orman algoritması toplu sınıflandırma yöntemleri arasında yer alır. Toplu sınıflandırma yöntemleri, birden çok sınıflandırıcı üreterek onların tahminlerinden alınan sonuçlar ile sınıflandırma sonucuna karar veren algoritmalardır. En bilinen toplu sınıflandırıcılardan torbalama algoritmasında orijinal eğitim veri setinden birden çok önyüklemeli eğitim veri setleri oluşturulur ve her bir önyüklemeli eğitim veri seti için bir ağaç üretilir. Ardışık gelen ağaçlar bir öncekinden bağımsızdır ve tahmin için en büyük oy baz alınır. Torbalama yöntemine rastgele özellik seçimi eklenerek Rastgele orman geliştirilmiştir. Breiman ve Cutler tarafından sunulmuş olan Rastgele Orman tüm değişkenler arasından en iyi dalı kullanarak her bir düğümü dallara ayırmak yerine, her bir düğümde rastgele olarak seçilen değişkenler arasından en iyisini kullanarak her bir düğümü dallara ayırır. Her bir veri seti orijinal veri setinden yer değiştirmeli olarak üretilir. Sonra rastgele özellik seçimi kullanılarak ağaçlar geliştirilir [26][27].



Şekil 5. 2 Rastgele Orman yöntemi

Şekil 5.2'de yukarıda anlatılan her bir veri seti için elde edilen karar ağaçları ve bunları birleşimi ile Rastgele Orman sınıflandırıcısının oluşumu görülmektedir. WEKA[29] ile yapılan deneylerde Rastgele Orman yöntemi ile rastgele sayıda seçilen özellik ile 10 ağaç oluşturularak sistem eğitilmiş ve testler yapılmıştır.

5.1.2 Naive Bayes

Diğer kullandığımız sınıflandırma algoritması olan Naïve Bayes, Bayes teoremini temel alan bir yöntemdir. İstatistiksel bir yöntem olan Bayes Teoremi bir rassal değişken için olasılık dağılımı içinde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi açıklar. Naïve Bayes sınıflandırıcı Bayes teoreminin bağımsızlık önermesiyle basitleştirilmiş halidir. Bu önerme örüntü tanımada kullanılacak her özelliğin istatistiksel açıdan bağımsız olması ve aynı derecede önemli olması gerekliliğini ortaya atar. Bayes kuralı (1)'deki gibi ifade edilebilir.

$$P(C_i|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|C_i)P(C_i)}{\mathrm{P}(\mathbf{X})} \quad (1)$$

- $P(x|C_i)$: Sınıf i'den bir örneğin x olma olasılığı
- $P(C_i)$: Sınıf i'nin ilk olasılığı
- P(x) : Herhangi bir örneğin x olma olasılığı
- $P(C_i|x)$: x olan bir örneğin sınıf i'den olma olasılığı (son olasılık)

Naïve Bayes sınıflandırıcıda son olasılığı en büyük olan durum *maks(P(C_i|X))*aranmaktadır. Böylece en büyük olasılığı veren durumda test verisi o sınıfa dahil edilir. P(x) olasılığı bütün sınıflar için sabit olduğu için sadece (2)'deki olasılık için en büyük değer aranır.

$$P(C_i|\boldsymbol{X}) = P(\boldsymbol{X}|C_i)P(C_i) \quad (2)$$

Naive Bayes yönteminde bilinmeyen X örneğini sınıflandırırken her C_i sınıfı için $P(\boldsymbol{X}|C_i)P(C_i)$ hesaplanır. X örneği en yüksek değere sahip S_i sınıfı olarak etiketlenir. Karşılaştırma için basitleştirilen ifade bütün özelliklerin birbirinden bağımsız olması durumunda P(x|Si) (3)'deki olarak ifade edilebilir

$$P(x|C_i) \approx \prod_{k=1}^{L} P(x_k|C_i) \quad (3)$$

Böylece Bayes karar teoremi (4)'deki halini alır. Bayes karar teorisine göre X örneği eğer (4)'deki büyüklük ifadesindeki gibiC_i sınıfında olma olasılığı diğerinden yüksek ise sınıf C_i'ye aittir.

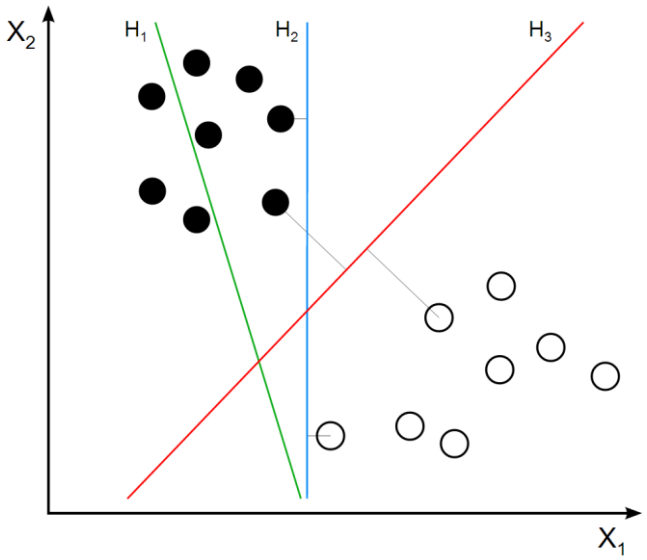
$$P(C_i)\prod_{k=1}^{L} P(x_k|C_i) > P(C_j)\prod_{k=1}^{L} P(x_k|C_j) \quad (4)$$

5.1.3 Destek Vektör Makineleri

Kullandığımız üçüncü yöntem olan Destek Vektör Makineleri sınıflandırma, regresyon ve aykırı değer belirleme için kullanılabilen bir eğiticili öğrenme yöntemidir. İstatistiksel öğrenme teorisini ve yapısal risk minimizasyonuna dayanır. Destek Vektör Makineleri sinyal işleme, yapay zeka ve veri madenciliği alanlarındaki birçok çalışmada elde ettiği yüksek başarısından dolayı bu çalışmada kullanılmıştır. Yöntemin yüksek doğruluğunun yanında karmaşık karar sınırlarını modelleyebilme, yüksek sayıda birbirinden bağımsız değişken ile çalışabilme, hem doğrusal hem de doğrusal olarak ayrılamayan verilere uygulanabilme gibi avantajları vardır.

Temelde iki sınıflı problemleri doğrusal olarak ayırabilen DVM yönteminde verileri sınıflandırırken ayırabilecek sonsuz sayıdaki doğru içerisinden marjini en yüksek olan doğru seçilir. Bu sonsuz sayıdaki doğrudan 3 farklı doğrultudaki örneğinin görüldüğü Şekil 5.3'de H_1doğrusu sınıfları başarılı bir şekilde ayıramazken H_2'de en yüksek marjini

sağlayamamıştır. H_3 ise hem sınıfları başarılı bir şekilde ayırabilmekte hem de en yüksek marjini sağlamaktadır.



Şekil 5. 3 Destek Vektör Makinesi için çizilen marjinler(H_1, H_2, H_3)

Şekil 5.4'de taranan bölgeye marjin denir ve oluşan hiperdüzlem eşitlik (5) ile ifade edilir. Eşitliğin 0'a eşit olduğu doğruda sınıfları ikiye ayıran doğrudur. Şekilde yer alan iki sınıfın marjin üzerindeki örnekleri ise destek vektörleri olarak adlandırılmaktadır.

$$w \cdot x - b = 1 \quad (5)$$

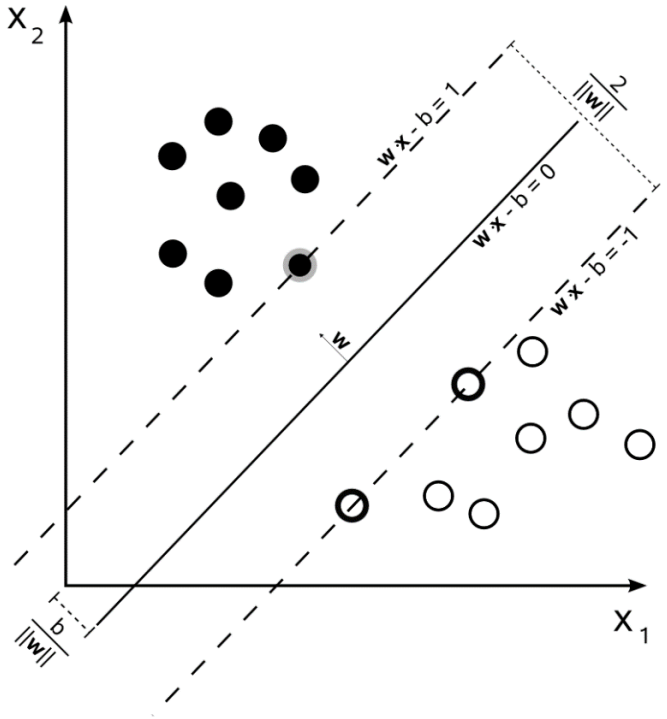
$$w \cdot x - b = -1$$

Bu iki hiperdüzlem arasındaki uzaklık $\frac{2}{|w|}$ değerindedir ve bu nedenle DVM yönteminde amaç |w| değerini minimize etmektir. |w| değeri ne kadar küçükse marjin genişliği o kadar artar. Düzlemdeki belirli x_i noktaları ve y çıktı verileri kullanarak eğitilmesi ile en uygun hiperdüzlem (6)'daki formül ile ifade edilir;

$y = +1$ ise $w.x_i+b >=1$ (6)

$y = -1$ ise $w.x_i+b <1$

tüm i'ler için $y_i(w.x_i + b) >= 1$'dir. Bir test X örneği sınıflandırılırken önceden belirlenmiş hiperdüzlemin hangi tarafına düştüğü bulunur. Hiperdüzlemin bir tarafı negatif sınıf, diğer tarafı pozitif sınıftır ve X örneği hangi tarafa düşer ise o sınıf olarak etiketlenir.



Şekil 5. 4 Destek Vektör Makinesi için çizilen marjin ve hiperdüzlem

Çok sınıflı Destek Vektör Makinelerinde çoklu sınıf probleminin çözümünde en sık kullanılan yaklaşım problemin ikili sınıflandırıcı problemine indirgenmesidir. İkili Destek Vektör Makineleri ile eğitilen sistemde 3 farklı sınıf olduğu için olumlu/olumsuz, olumlu/nötr ve olumsuz/nötr olmak üzere 3 sınıflandırıcı modeli oluşturulmuştur. 3 modelin sınıfları farklı olduğu için kullanılan özelliklerinde ağırlıkları her model için farklılık göstermektedir ve sınıflandırma sonucu 3 sınıflandırıcıdan elde edilen sonuçların birleşimidir. Bu sonucun birleşimi oylama stratejisi ile belirlenir ve sınıflandırıcılardan en yüksek oyu alan sınıf kazanan olarak seçilir.

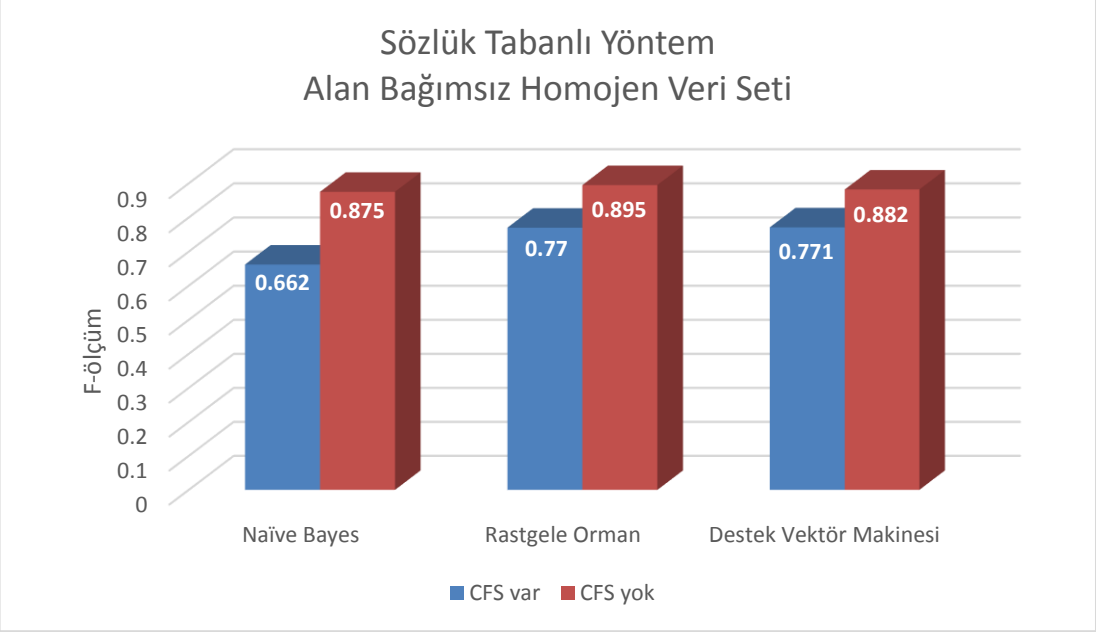
5.2 Özellik Seçimi

Sınıflandırma işleminde bazı durumlarda özellik uzayının boyutunun artması sınıflandırıcının başarısını azaltıcı etki yapar. Bu durumda sınıflandırma gücü olmayan ve gereksiz işlem yükü getiren özelliklerin ayıklanması gerekliliği ortaya çıkar. Ayrıca, ayırt edici nitelikte olmayan özelliklerin kullanıldığı özellik vektörleri sınıflandırma başarısını düşürebilir. Bu nedenlerden dolayı çalışmamızda özellik seçimi yapılmıştır ve yöntem olarak korelasyon tabanlı özellik seçimi (CFS) kullanılmıştır.

CFS, korelasyona dayanan bir özellik alt kümesi oluşturma yöntemidir. CFS çıktı sınıfı ile yüksek korelasyona sahip özellikleri seçen ve sınıflandırma başarısına etkisi olmayan hatta başarıyı düşüren gereksiz özellikleri eleyen bir yöntemdir. Bu çalışmada özellik alt kümeleri Weka'da [29] yer alan CFS uygulaması ile oluşturulmuştur.

Çalışmamızda yaptığımız deneylerde CFS yanında ki-kare dağılımı (Chi Square) ve bilgi kazancı (Information Gain) yöntemleri de test edilmiştir. Yapılan deneyler sonucunda en iyi sonuç CFS yöntemi ile elde edilmiştir ve sınıflandırma algoritmalarının karşılaştırılmasında CFS ile elde edilen özellik uzayı kullanılmıştır.

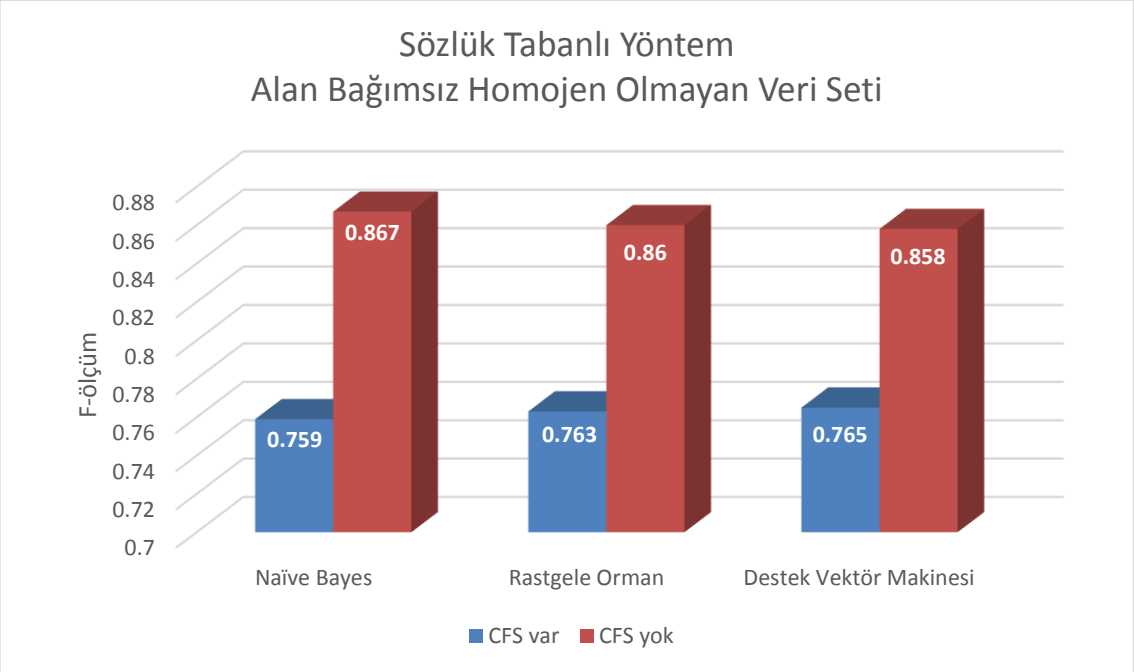
Çalışmamızda Naïve Bayes, Rastgele Orman ve Destek Vektör Makinesi sınıflandırıcıları hem özellik seçimi yapılarak hem de özellik seçimi yapılmaksızın eğitilmiştir. Korelasyon tabanlı özellik seçimi diğer yöntemlere göre daha iyi sonuçlar vermesine rağmen alan bağımsız veri setinde başarının düşmesine neden olmuştur. Şekil 5.5'de sözlük tabanlı yöntem ve homojen alan bağımsız veriler ile eğitilen sistemde Naïve Bayes, Rastgele Orman ve Destek Vektör Makinesi algoritmalarının CFS kullanarak ve kullanmadan elde edilen f-ölçüm değerleri gösterilmektedir.



Şekil 5. 5 Alan bağımsız homojen veri setinde kullanılan CFS'nin sözlük tabanlı yöntem ile eğitilen sınıflandırıcıların başarısına etkisi

Homojen veri setinde CFS kullanımında elde edilen düşük başarı aynı şekilde homojen olmayan veri setinde de görülmektedir. Şekil 5.6'da sözlük tabanlı yöntem ve homojen

olmayan alan bağımsız veriler ile eğitilen sistemde Naïve Bayes, Rastgele Orman ve Destek Vektör Makinesi algoritmalarının CFS kullanarak ve kullanmadan elde edilen f-ölçüm değerleri gösterilmektedir.

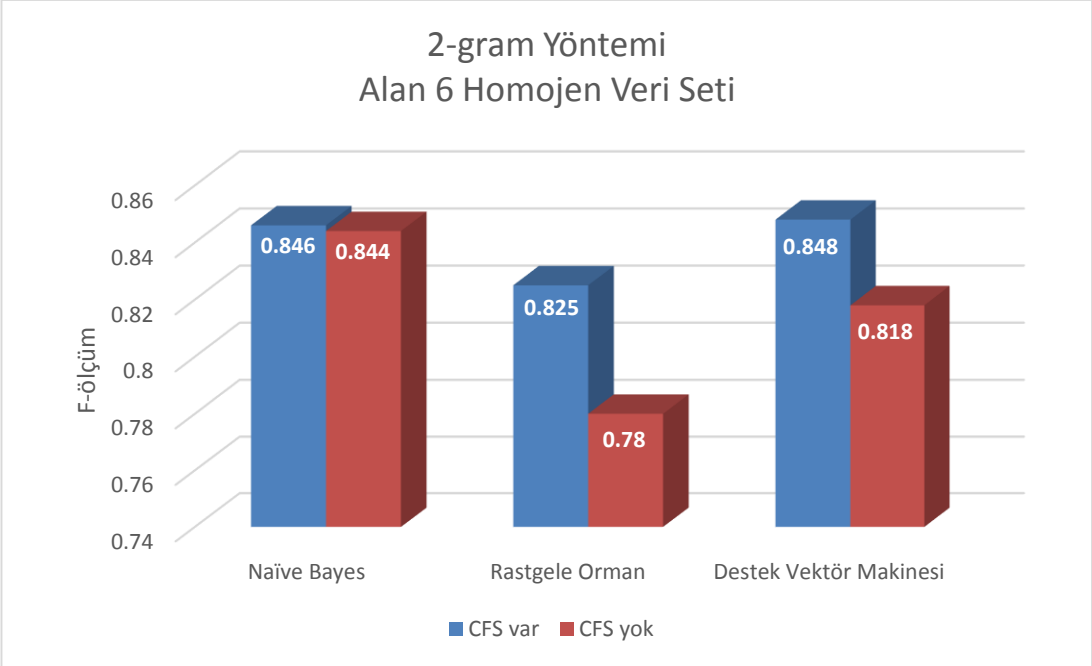


Şekil 5. 6 Alan bağımsız homojen olmayan veri setinde kullanılan CFS'nin sözlük tabanlı yöntem ile eğitilen sınıflandırıcıların başarısına etkisi

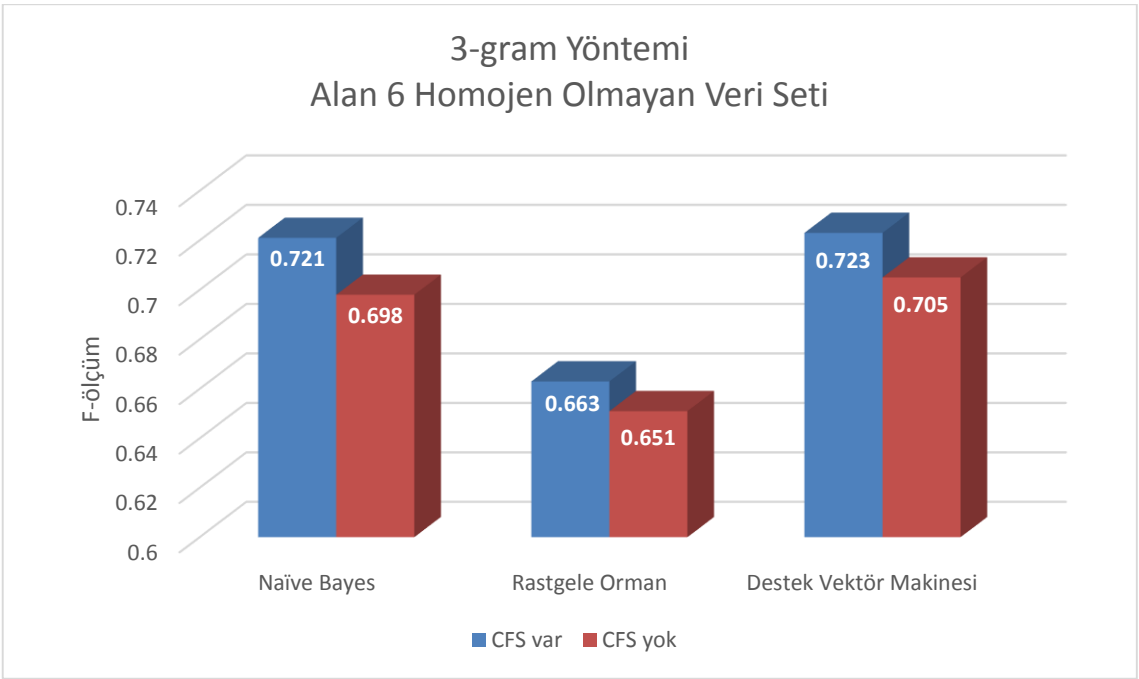
Sözlük tabanlı yöntem ile eğitilen alan bağımsız veri setinde başarının düşmesine neden olan korelasyon tabanlı özellik seçimi alan bağımlı veri setinde n-gram yöntemi ile eğitildiğinde başarının artmasını sağlamıştır. Özellik seçimi kullanılmaması durumunda 0.895 f-ölçüm değeri ile Rastgele Orman algoritması en yüksek başarıyı vermektedir. Özellik seçimi kullanıldığında ise 0.125 kadar f-ölçüm değerini düşürmekte ve 0.77 elde edilmektedir. Korelasyon tabanlı özellik seçimi yönteminin burada en büyük etkisinin Naive Bayes sınıflandırma yönteminde olduğu görülmektedir. Bu durumda Naive Bayes sınıflandırıcısında 0.213 kadar başarıda düşüş gözlenmektedir. Özellik seçiminden en az etkilenen sınıflandırıcı ise 0.111 başarı kayıp oranı ile Destek Vektör Makinesidir.

Şekil 5.7 ve Şekil 5.8'de 6 numaralı alan olan gıda alanındaki homojen verilerle 2-gram yöntemi ile eğitilen sistemde Naïve Bayes, Rastgele Orman ve Destek Vektör Makinesi algoritmalarının CFS kullanarak ve kullanılmadan elde edilen f-ölçüm değerleri

gösterilmektedir. Her iki şekilde de görüldüğü gibi özellik seçiminin uygulanması halinde özellikle Rastgele Orman ve Destek Vektör Makinesi algoritmalarının başarısında gözle görülür bir artış sağlanmıştır.

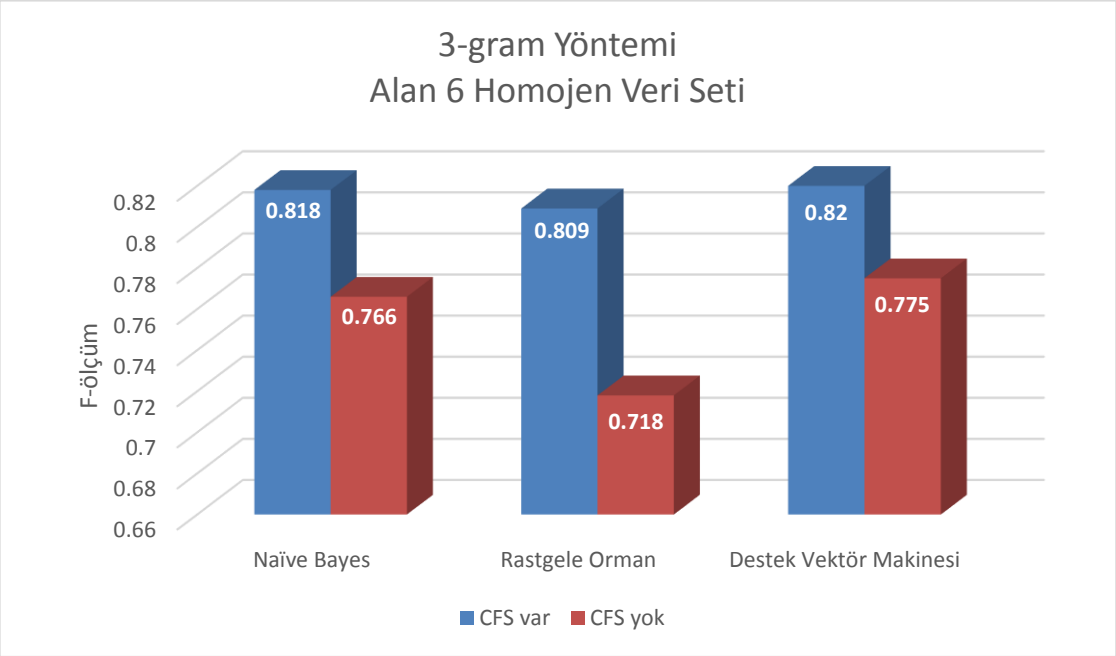


Şekil 5. 7 Alan 6 homojen veri setinde kullanılan CFS'nin 2-gram ile eğitilen sınıflandırıcıların başarısına etkisi

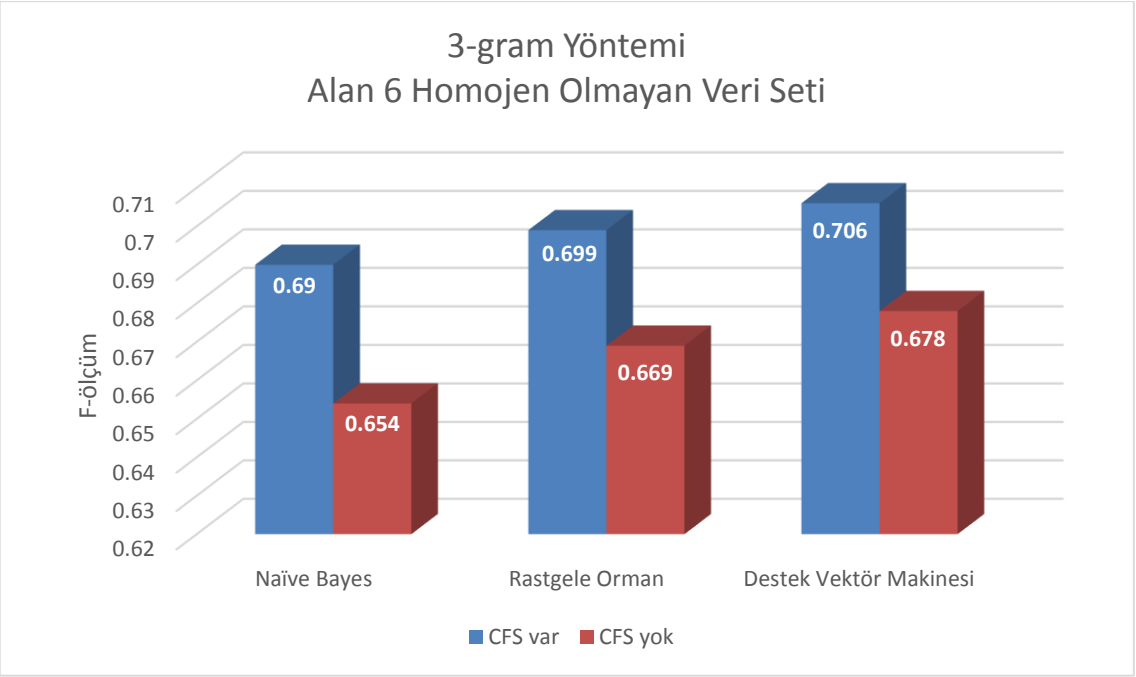


Şekil 5. 8 Alan 6 homojen olmayan veri setinde kullanılan CFS'nin 2-gram ile eğitilen sınıflandırıcıların başarısına etkisi

Alan 6 verilerinde 2-gram ile yapılan deneyler Şekil 5.9'de görüldüğü gibi 3-gram için tekrarlandığında benzer durum burada da görülmüştür. CFS ile özellik seçimi yapıldığında en büyük kazanç Rastgele Orman yöntemi olmaktadır. Başarı 0.78'den 0.825'e yükseldiği Rastgele Orman sınıflandırıcısının yanında diğer yöntemlerde de aynı şekilde yükseldiği görülmektedir. Aynı şekilde homojen olmayan veri seti ile deneyler yapıldığında Şekil 5.10'da görüldüğü gibi CFS ile özellik seçimi yapıldığında bütün sınıflandırıcılarda başarının yükseldiği görülmektedir.

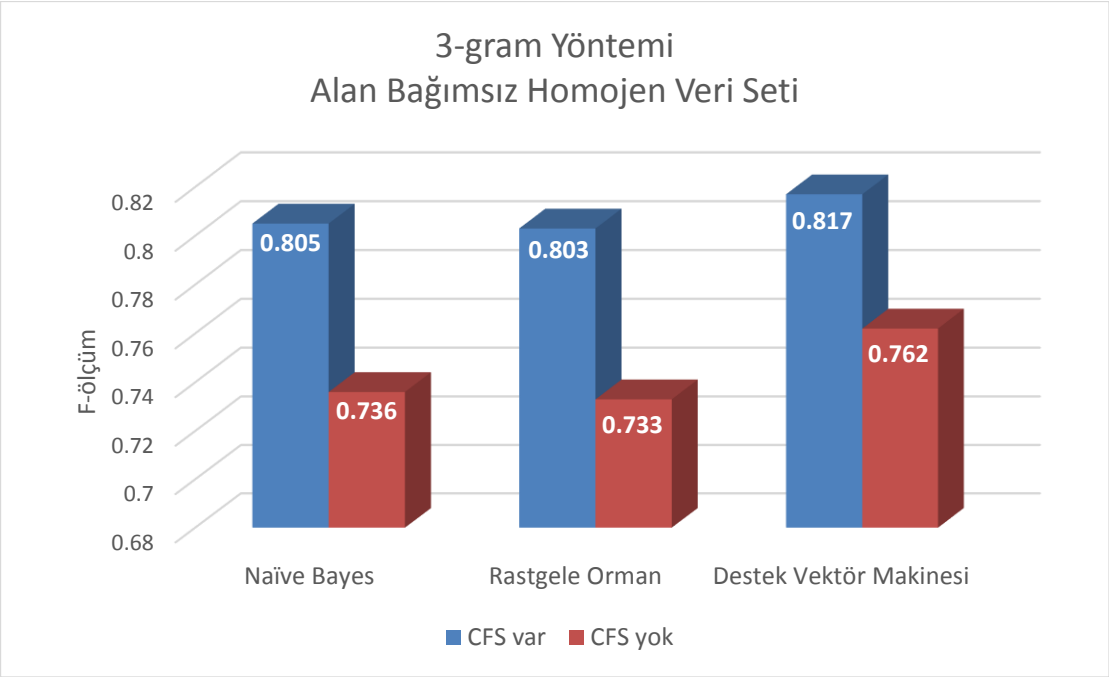


Şekil 5. 9 Alan 6 homojen veri setinde kullanılan CFS'nin 3-gram ile eğitilen sınıflandırıcıların başarısına etkisi



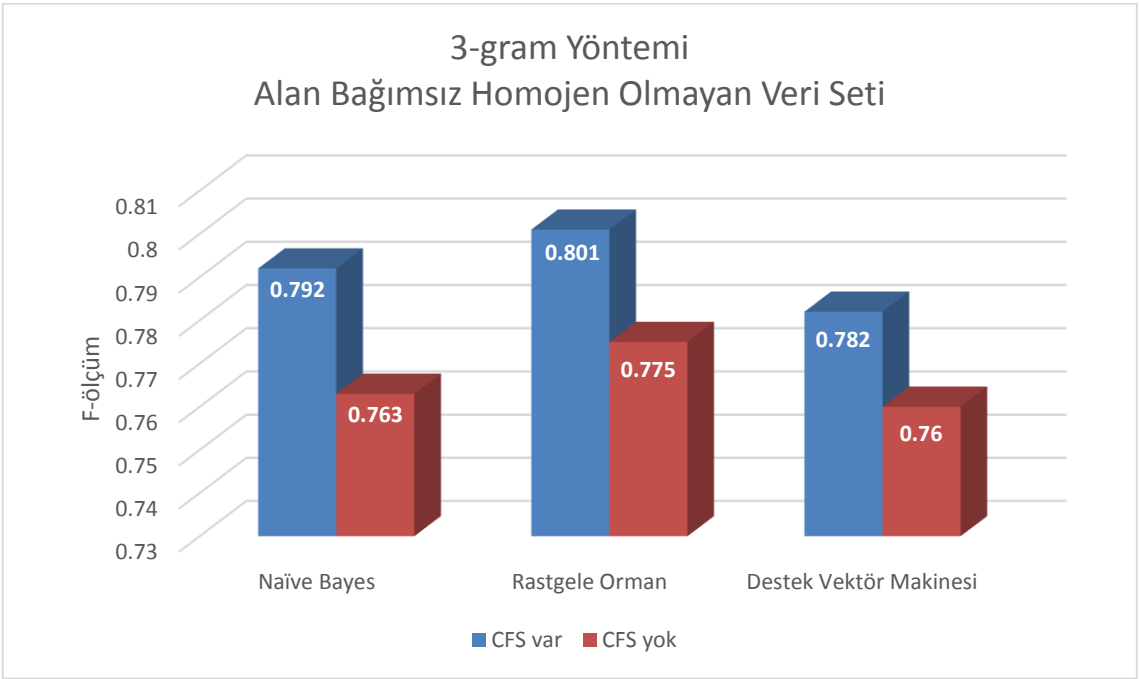
Şekil 5. 10 Alan 6 homojen olmayan veri setinde kullanılan CFS'nin 3-gram ile eğitilen sınıflandırıcıların başarısına etkisi

Şekil 5.11'de ise bütün alanları içinde barındıran alan bağımsız veri seti ile 3-gram yöntemi test edilmiştir. CFS ile özellik seçimi yapılması durumunda alan bağımlı verilerde olduğu gibi alan bağımsız verilerde de başarı artmaktadır. CFS'nin 0.7 f-ölçüm değer farkı ile en çok Rastgele Orman yöntemine etkisi olduğu görülmektedir. N-gram ile yapılan deneylerde CFS'nin kesinlikle başarıyı yükselttiği ancak, sözlük tabanlı yöntemde aynı durumun söz konusu olmadığı görülmektedir. Bu nedenle hedeflenen başarılı sınıflandırıcının seçilmesi için yapılan karşılaştırmalarda n-gram yönteminde CFS uygulanarak, sözlük tabanlı yöntemde ise CFS uygulanmadan elde edilen sonuçlar kullanılmıştır.



Şekil 5. 11 Alan bağımsız homojen veri setinde kullanılan CFS'nin 3-gram ve alan bağımsız veriler ile eğitilen sınıflandırıcıların başarısına etkisi

Homojen veri setinde CFS kullanımında sağlanan başarı artışı aynı şekilde homojen olmayan veri setinde de görülmektedir. Şekil 5.12'de sözlük tabanlı yöntem ve homojen olmayan alan bağımsız veriler ile eğitilen sistemde Naïve Bayes, Rastgele Orman ve Destek Vektör Makinesi algoritmalarının CFS kullanarak ve kullanmadan elde edilen f-ölçüm değerleri gösterilmektedir.



Şekil 5. 12 Alan bağımsız homojen olmayan veri setinde kullanılan CFS'nin 3-gram ve alan bağımsız veriler ile eğitilen sınıflandırıcıların başarısına etkisi

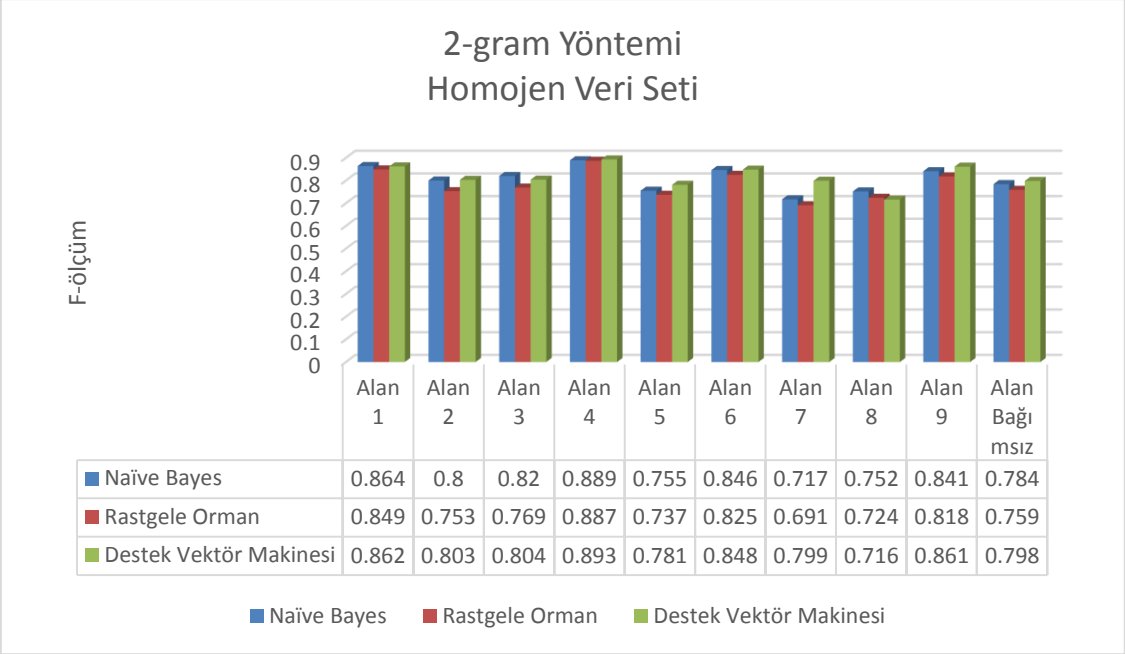
Korelasyon tabanlı özellik seçimi uygulandığında oluşan yeni yeni özellik listelerinin boyutları çizelge 5.1'dedir. Özellik sayısının en çok n-gram yönteminde azaldığı görülmektedir ve hemen hemen bütün veri setlerinde %99'un üzerinde bir oranla küçülmüştür. Sözlük tabanlı yöntemde ise küçülme oranı veri setine bağlı olarak %67 ile %99.5 arasında değişmektedir.

Çizelge 5.1: CFS ile oluşan özellik listesi boyutları

Veri Seti	Sözlük tabanlı	2-gram	3-gram	2-gram + 3-gram
Alan 1	63	56	51	63
Alan 2	50	60	62	60
Alan 3	86	54	31	60
Alan 4	86	32	31	58
Alan 5	85	60	34	65
Alan 6	76	69	57	77
Alan 7	34	48	30	53
Alan 8	44	50	49	50
Alan 9	76	89	63	87
Alan Bağımsız	190	90	115	96

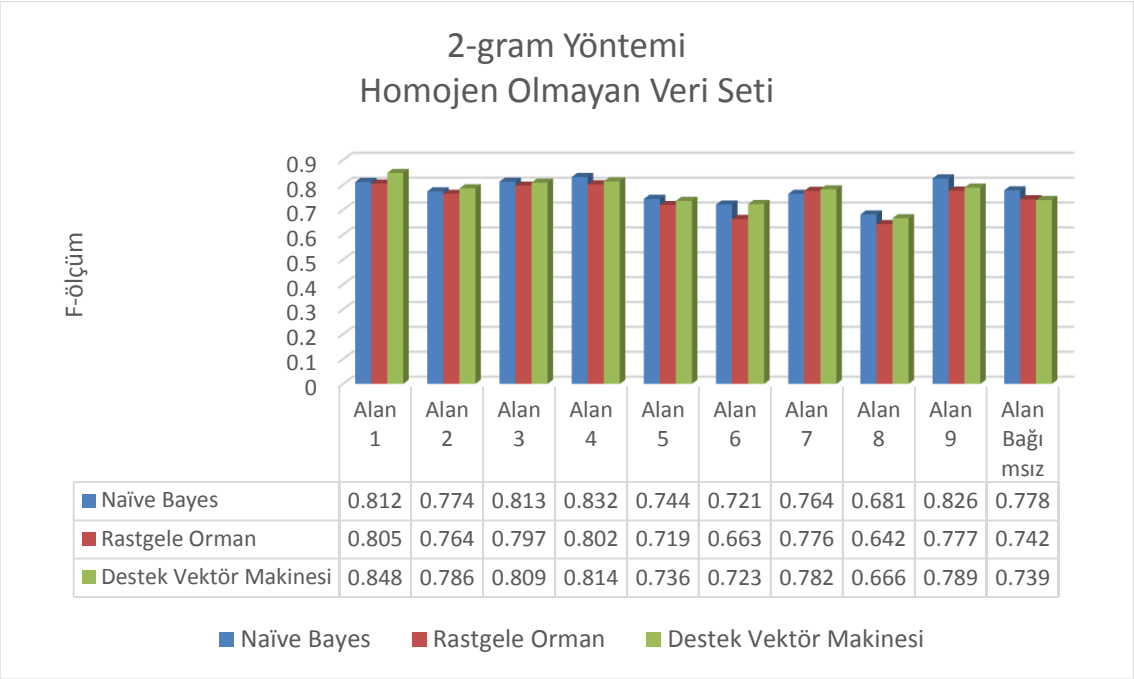
5.3 Sınıflandırma Algoritmalarının Karşılaştırılması

Sınıflandırıcıların karşılaştırmasında 2-gram, 3-gram, 2-gram + 3-gram ve sözlük tabanlı yöntemlerinin tümü için deneyler yapılmıştır. Her yöntem için alan 1-9 ve tüm alanların birleştirildiği alan bağımsız veri seti kullanılmıştır.

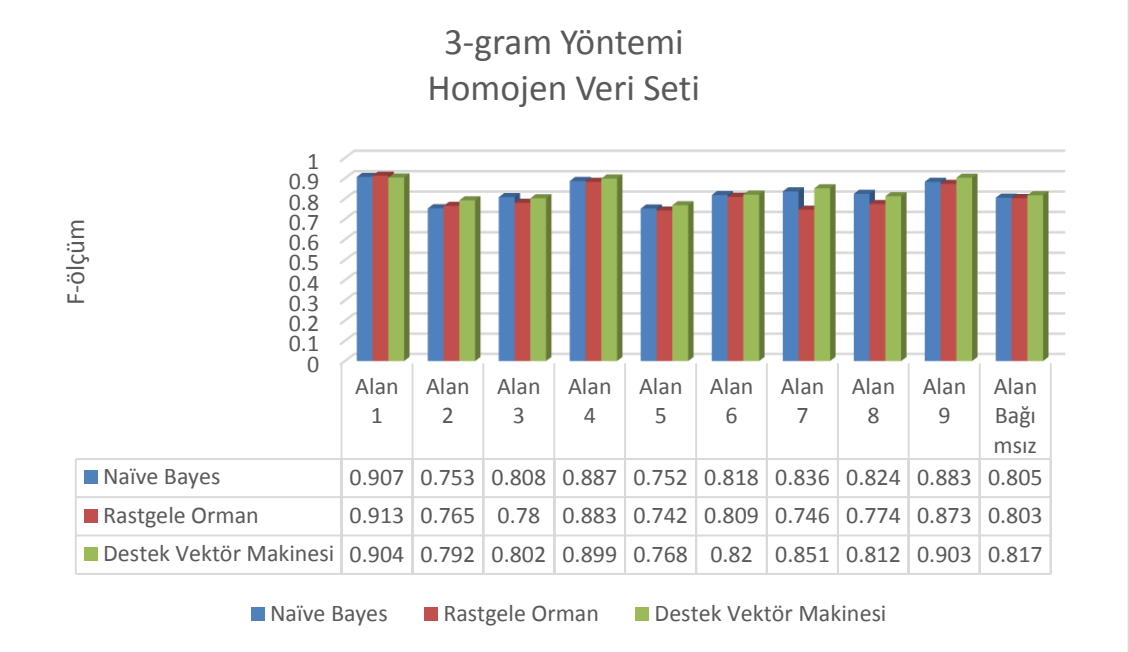


Şekil 5. 13 Homojen veri seti ile 2-gram yöntemi sınıflandırma başarısı

2-gram yöntemi oluşturulan özellik vektörü ile eğitilen sistemde 9 farklı alan ve bütün alanları içeren veri setleri kullanılmıştır. Destek Vektör Makinesinin alanların çoğunda ve alan bağımsız veri setinde en iyi sonuçları verdiği görülmüştür. Bazı alanlarda Naive Bayes en başarılıdır, Rastgele Orman ise diğer yöntemlerden daha az başarılıdır (Şekil 5.13). Aynı yöntem kullanılarak 9 farklı alan ve bütün alanları içeren homojen olmayan veri setleri kullanıldığında Şekil 5.14'deki sonuçlar elde edilmektedir. Homojen veri setinde elde edilen sonuçlarda olduğu gibi homojen olmayan veri setinde de bazı alanlarda Naive Bayes en başarılıdır, Rastgele Orman ise genellikle diğer yöntemlerden daha az başarı göstermiştir. Veri setleri karşılaştırıldığında 2-gram yönteminin genelde homojen veri setinden daha başarılı olduğu görülmektedir.



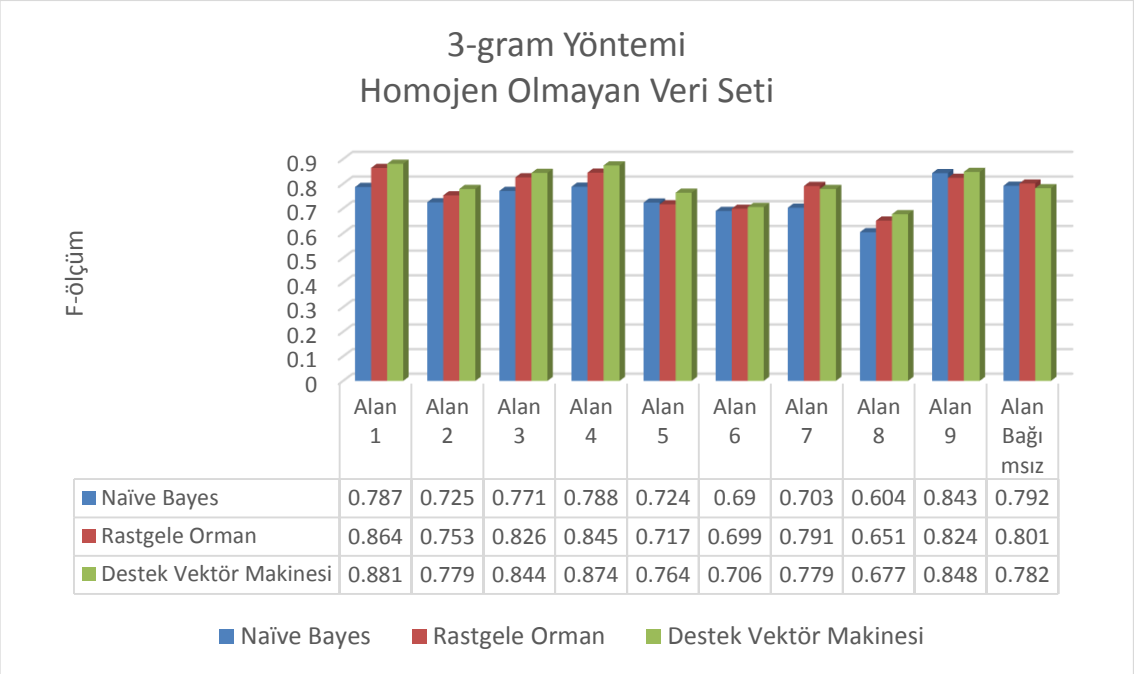
Şekil 5. 14 Homojen olmayan veri seti ile 2-gram yöntemi sınıflandırma başarısı



Şekil 5. 15 Homojen veri seti ile 3-gram yöntemi sınıflandırma başarısı

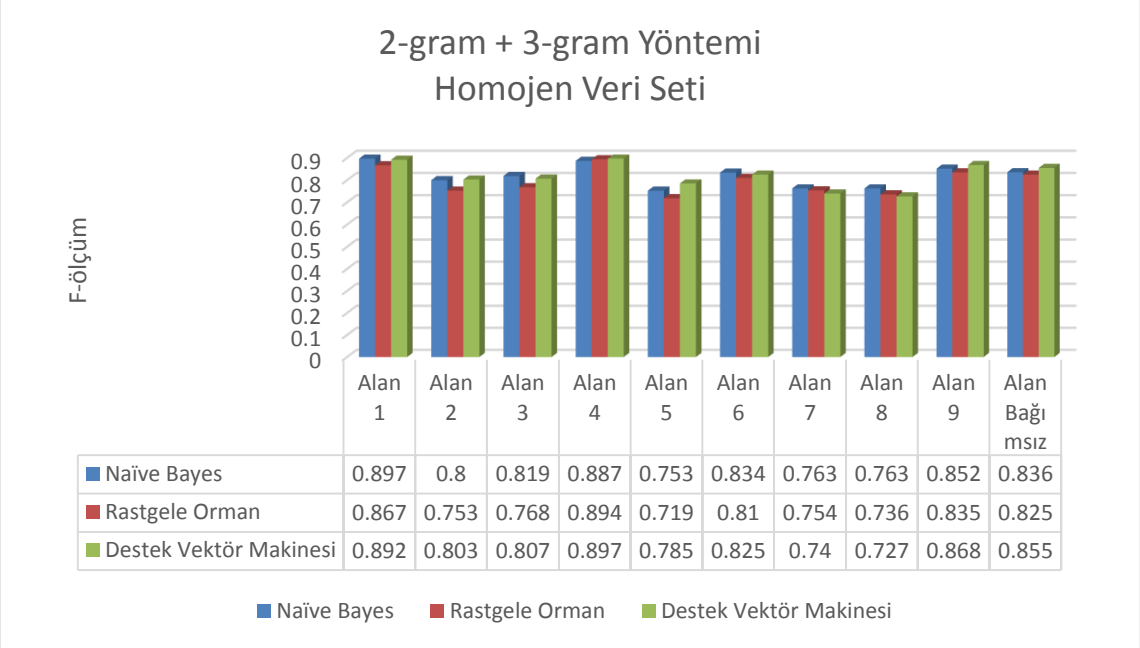
3-gram yönteminde Destek Vektör Makinesinin çoğunlukla en iyi sonuçları verdiği şekil 5.15 ve şekil 5.16'da görülmektedir. Homojen olan ve homojen olmayan veri setleri karşılaştırıldığında ise homojen veri seti ile daha iyi sonuçlar elde edildiği görülmektedir. Elde edilen başarılar büyük oranda veriye bağlı olduğu için çalıştırılan algoritmalar aynı veri setinde birbirine çok yakın sonuçlar çıkmaktadır. Şekil 5.17 ve

şekil 18'de 2-gram ve 3-gram özellik vektörünü birleştirdiğimiz yöntemde de hem homojen hem de homojen olmayan veri seti ile Destek Vektör Makinesinin en iyi sonuçları verdiği görülmektedir. Bu yöntemde hem alan bağımsız hem de alan verilerinde homojen veri setinin homojen olmayan veri setinden daha başarılı olduğu söylenebilir.



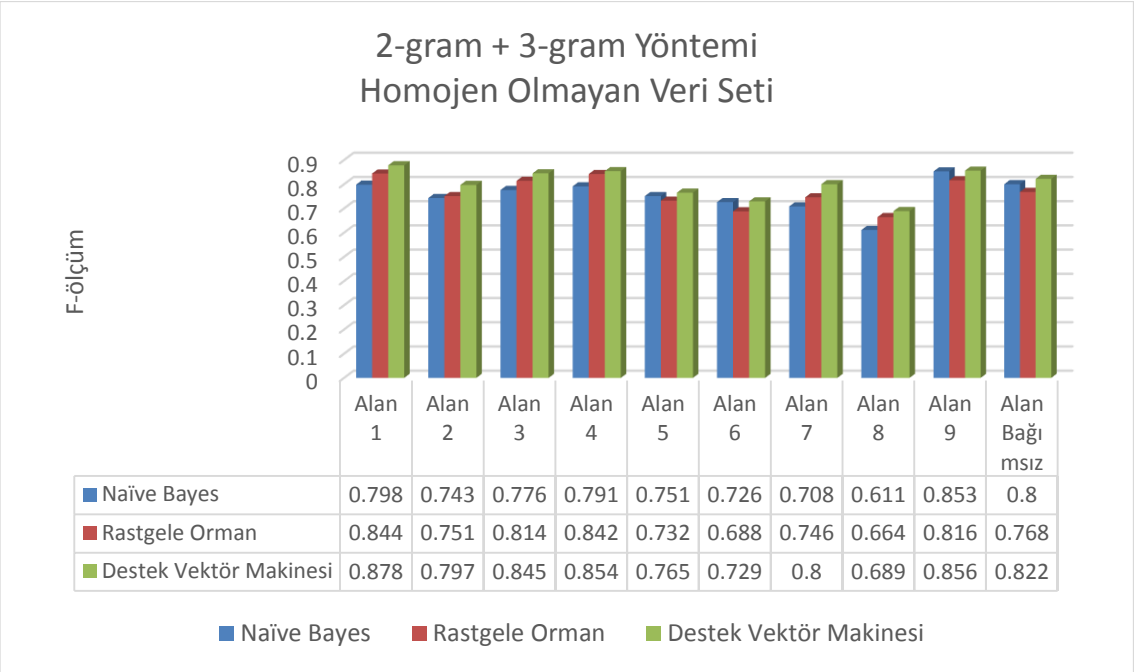
	Alan 1	Alan 2	Alan 3	Alan 4	Alan 5	Alan 6	Alan 7	Alan 8	Alan 9	Alan Bağımsız
Naïve Bayes	0.787	0.725	0.771	0.788	0.724	0.69	0.703	0.604	0.843	0.792
Rastgele Orman	0.864	0.753	0.826	0.845	0.717	0.699	0.791	0.651	0.824	0.801
Destek Vektör Makinesi	0.881	0.779	0.844	0.874	0.764	0.706	0.779	0.677	0.848	0.782

Şekil 5. 16 Homojen olmayan veri seti ile 3-gram yöntemi sınıflandırma başarısı



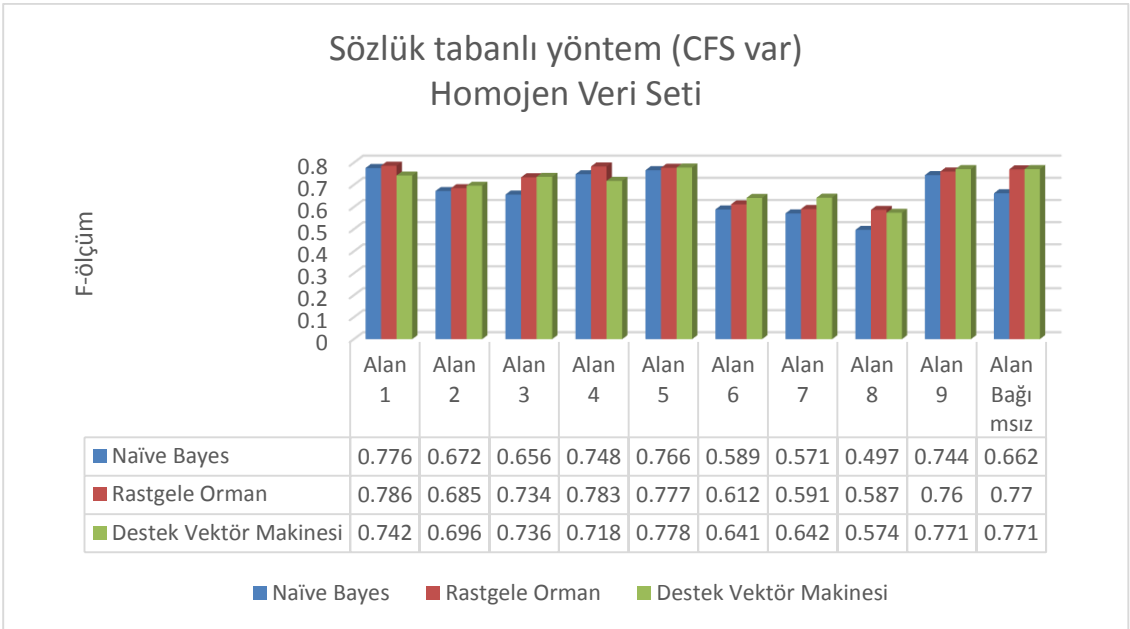
	Alan 1	Alan 2	Alan 3	Alan 4	Alan 5	Alan 6	Alan 7	Alan 8	Alan 9	Alan Bağımsız
Naïve Bayes	0.897	0.8	0.819	0.887	0.753	0.834	0.763	0.763	0.852	0.836
Rastgele Orman	0.867	0.753	0.768	0.894	0.719	0.81	0.754	0.736	0.835	0.825
Destek Vektör Makinesi	0.892	0.803	0.807	0.897	0.785	0.825	0.74	0.727	0.868	0.855

Şekil 5. 17 Homojen veri seti ile 2-gram + 3-gram yöntemi sınıflandırma başarısı

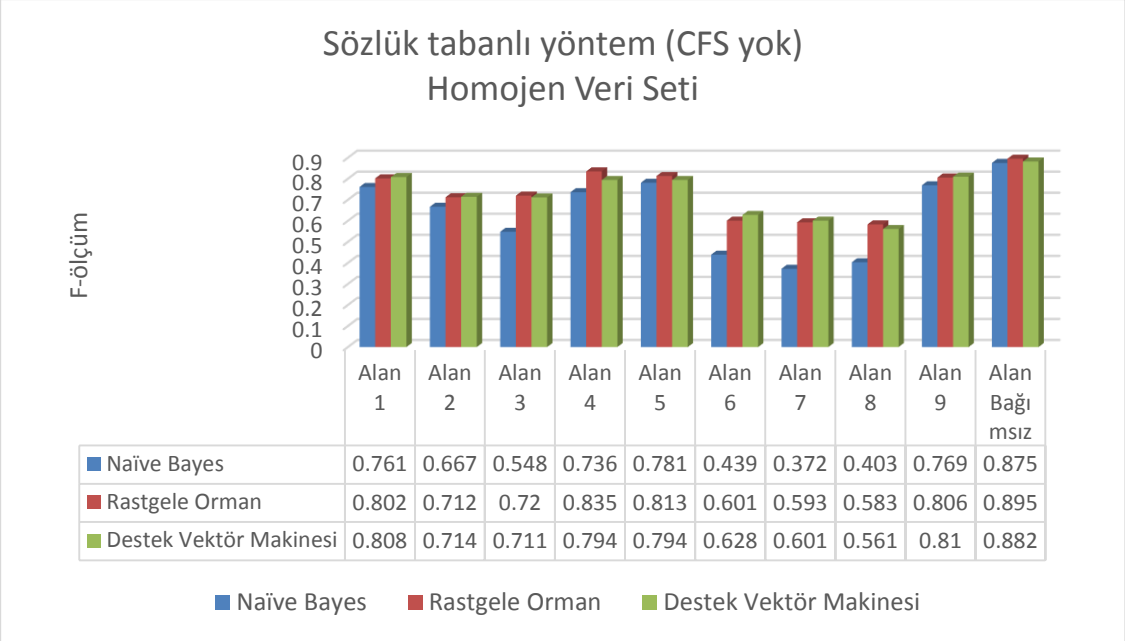


Şekil 5. 18 Homojen olmayan veri seti ile 2-gram + 3-gram yöntemi sınıflandırma başarısı

Bu bölümde sözlük tabanlı yöntem ile algoritmaları karşılaştırırken n-gram yönteminde de olduğu gibi korelasyon tabanlı özellik seçimi kullanılmıştır. Sözlük tabanlı yöntemde başarıların genel olarak düşük olmasına rağmen n-gram yönteminde de en iyi f-ölçüm değerlerini veren Destek Vektör Makinesi sınıflandırıcısı diğer sınıflandırıcılardan daha başarılıdır (Şekil 5.19).

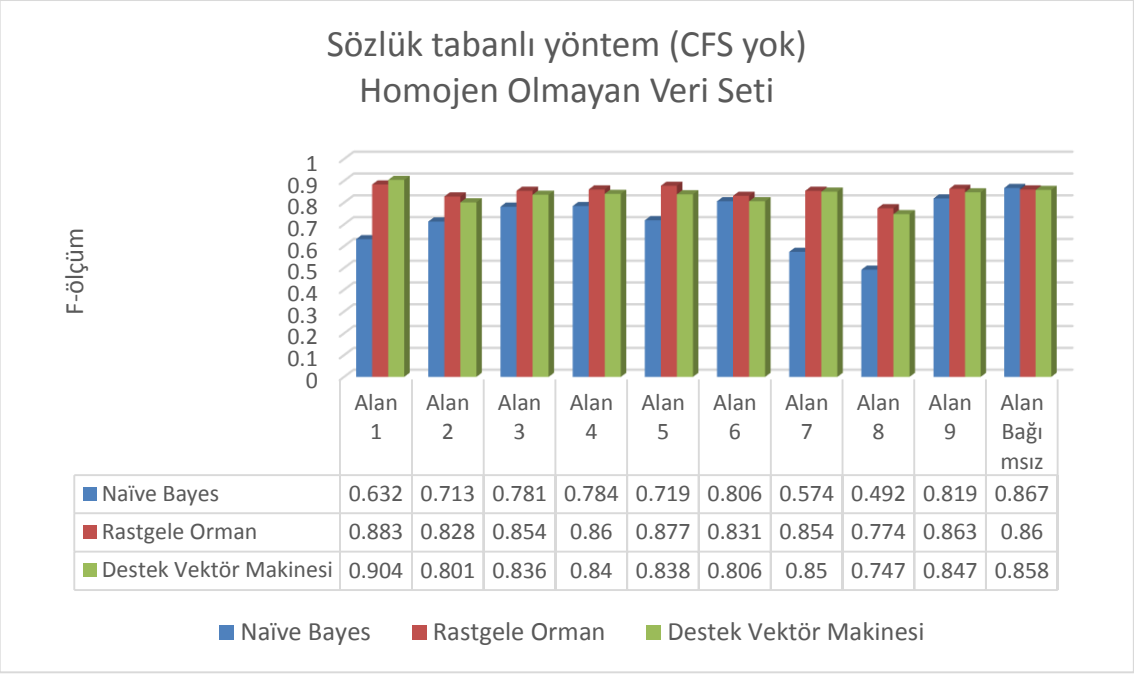


Şekil 5. 19 Homojen veri seti ile CFS kullanarak sözlük tabanlı yöntem başarısı



Şekil 5. 20 Homojen veri seti ile CFS kullanmadan sözlük tabanlı yöntem başarısı

Şimdiye kadar yapılan karşılaştırmaların tümünde korelasyon tabanlı özellik seçimi kullanılmış ve bir istisnai durum hariç Destek Vektör Makinesinin en başarılı sınıflandırıcı olduğu görülmüştür (Şekil 5.19). Bu istisnai durum özellik seçim yönteminin kullanılmadan sözlük tabanlı yöntemler ile elde edilen sonuçlardır. Rastgele Orman sınıflandırıcısı eğitildiği takdirde 0.895 f-ölçüm değeri ile alan bağımsız veriler içinde elde edilen en iyi sonucu vermiştir (Şekil 5.20).

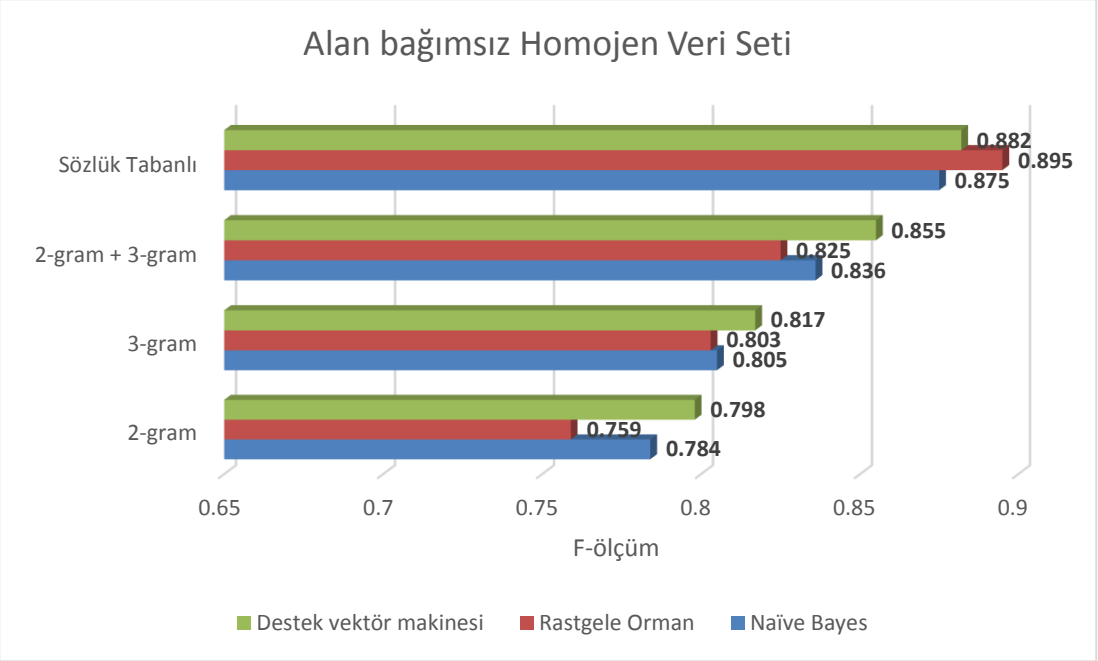


Şekil 5. 21 Homojen olmayan veri seti ile CFS kullanmadan sözlük tabanlı yöntem başarısı

Şekil 21'de homojen olmayan veri seti ile yapılan deneyler incelendiğinde homojen veri setinden farklı bir sonuç olduğu görülmektedir. Alan bağımsız verilerde homojen veri setine göre daha az başarı gösterirken alan bağımlı verilerde çok daha yüksek başarı göstermiştir. DVM'nin alan 1'de RO'nun da diğer alan verilerinde en yüksek başarıyı elde etmiştir. Alan bağımsız verilerde ise 0.867 f-ölçüm değeri ile NB en yüksek başarıyı göstermiştir.

5.4 Yöntemlerin Karşılaştırılması

Bölüm 3'te verilerin ön işlemden geçirilmesi aşamasında izlenen 2 farklı yöntem anlatılmıştır. Bu yöntemlerden biri kelime diğeri karakter seviyesinde işlemler izlediği için birbirinden farklılık göstermektedir. Verilerin anlamlandırılmasında kelimelerin etkisi göz ardı edilemeyecek kadar fazladır ancak karakter kullanımında normalde gözle farkedilemeyecek gizli örüntüler bulunabilir ve bu örüntüler ile başarılı bir sınıflandırıcı elde edilebilir. Bu bölümde sözlük tabanlı ve n-gram yöntemleri karşılaştırılacaktır.



Şekil 5. 22 Alan bağımsız homojen veri seti için yöntemlerin karşılaştırması

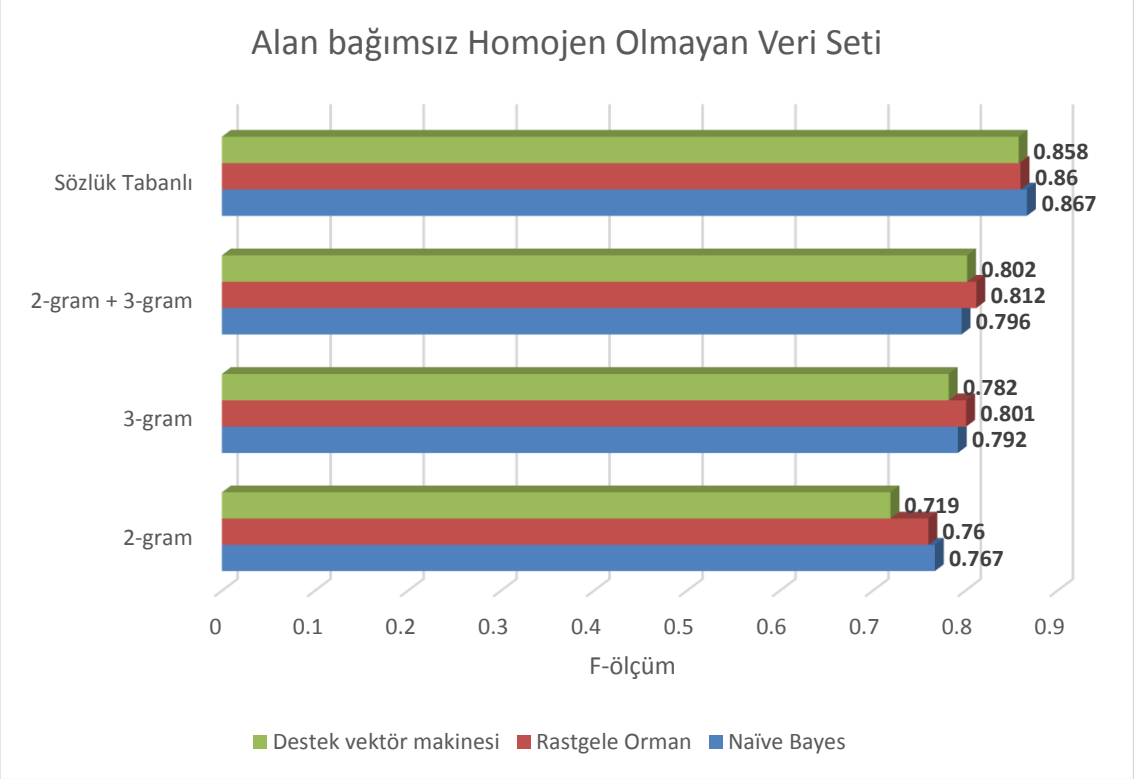
Şekil 5.22 incelendiğinde alan bağımsız veri setinde 2-gram ve 3-gram yöntemlerinin tivitleri sınıflandırmada seçici özelliğinin diğer yöntemlere göre daha az olduğu

görülmektedir. 2-gram yönteminde Rastgele Orman sınıflandırıcısı ile en düşük olan 0.759 f-ölçüm değeri elde edilebilmiştir. Burada en başarılı olan Destek Vektör Makinesi ise 3-gram yöntemi ile karşılaştırıldığında en düşük başarı elde eden Rastgele Orman sınıflandırıcısından da düşük bir sonuç elde etmiştir.

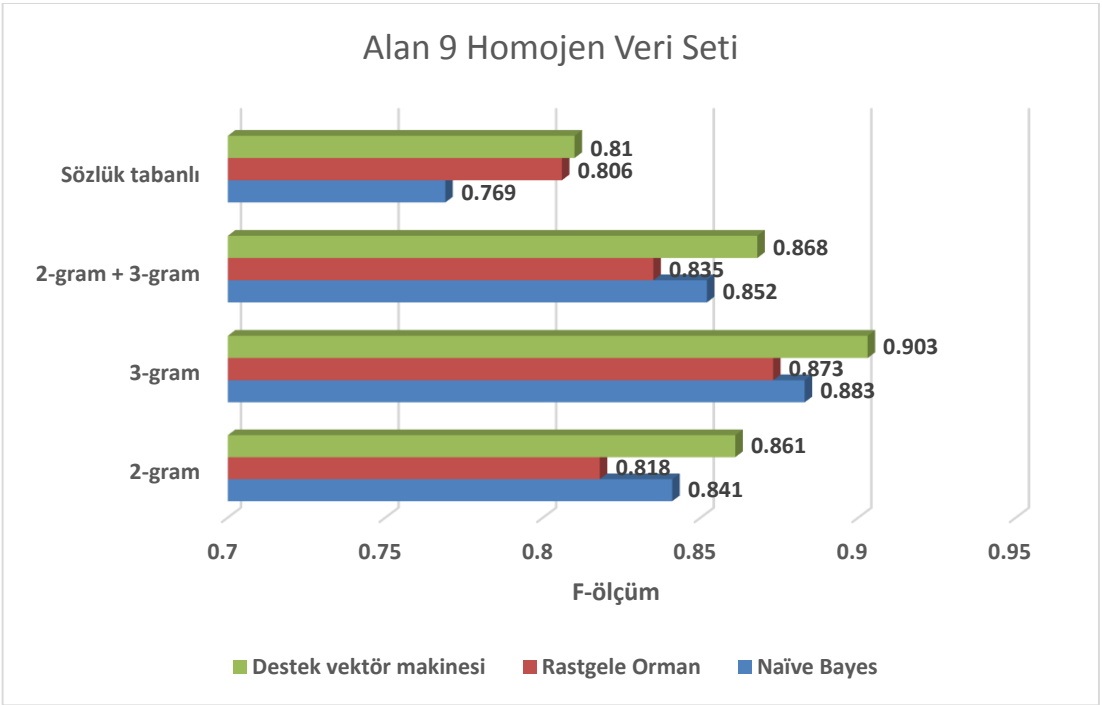
2-gram ve 3-gram özellik vektörünü birleştirdiğimiz yöntemin alan bağımsız verilerde diğer n-gram yöntemlerine göre daha fazla başarılı olduğu görülmektedir. Diğer n-gram yöntemlerindeki gibi burada da en iyi sonucu 0.855 f-ölçüm değeri ile Destek Vektör Makinesi sınıflandırıcısı vermektedir.

N-gram yönteminde özellik seçimi uygulandığında başarının arttığı ancak, sözlük tabanlı da aynı durumun söz konusu olmadığı bölüm 5.2'de belirtilmiştir. Bu nedenle sözlük tabanlı yöntemde algoritmaları karşılaştırırken korelasyon tabanlı özellik seçimi ile elde edilen özellik vektörü kullanılmamıştır. 589 adet özelliğin kullanılarak alan bağımsız verilerle eğitilen sınıflandırıcılar en yüksek başarıyı sözlük tabanlı yöntemde vermiştir. Bu yöntemde Rastgele Orman sınıflandırıcısı 0.895 f-ölçüm değeri ile diğer yöntemlerden daha başarılı olmuştur.

Homojen veri setinde oluşan başarı sıralaması aynı şekilde homojen olmayan veri setinde de görülmektedir. Şekil 5.23'de homojen olmayan alan bağımsız veriler ile eğitilen sistemde Naïve Bayes, Rastgele Orman ve Destek Vektör Makinesi algoritmalarından elde edilen f-ölçüm değerleri gösterilmektedir. Alan bağımsız veriler ile iki farklı veri setinde yapılan deneyler karşılaştırıldığında homojen veri setinin diğerinden daha başarılı olduğu görülmektedir. Sözlük tabanlı yöntemde homojen veri seti ile elde edilen en iyi değer 0.895 f-ölçüm değeri iken homojen olmayan veri setinde en iyi değer 0.867'dir. N-gram yöntemlerinden 2-gram ve 3-gram yönteminde homojen veri seti ile elde edilen en iyi değer 0.855 f-ölçüm değeri iken homojen olmayan veri setinde en iyi değer 0.812'dir. 3-gram yönteminde homojen veri seti ile elde edilen en iyi değer 0.817 f-ölçüm değeri iken homojen olmayan veri setinde en iyi değer 0.801'dir. 2-gram yönteminde homojen veri seti ile elde edilen en iyi değer 0.798 f-ölçüm değeri iken homojen olmayan veri setinde en iyi değer 0.767'dir.

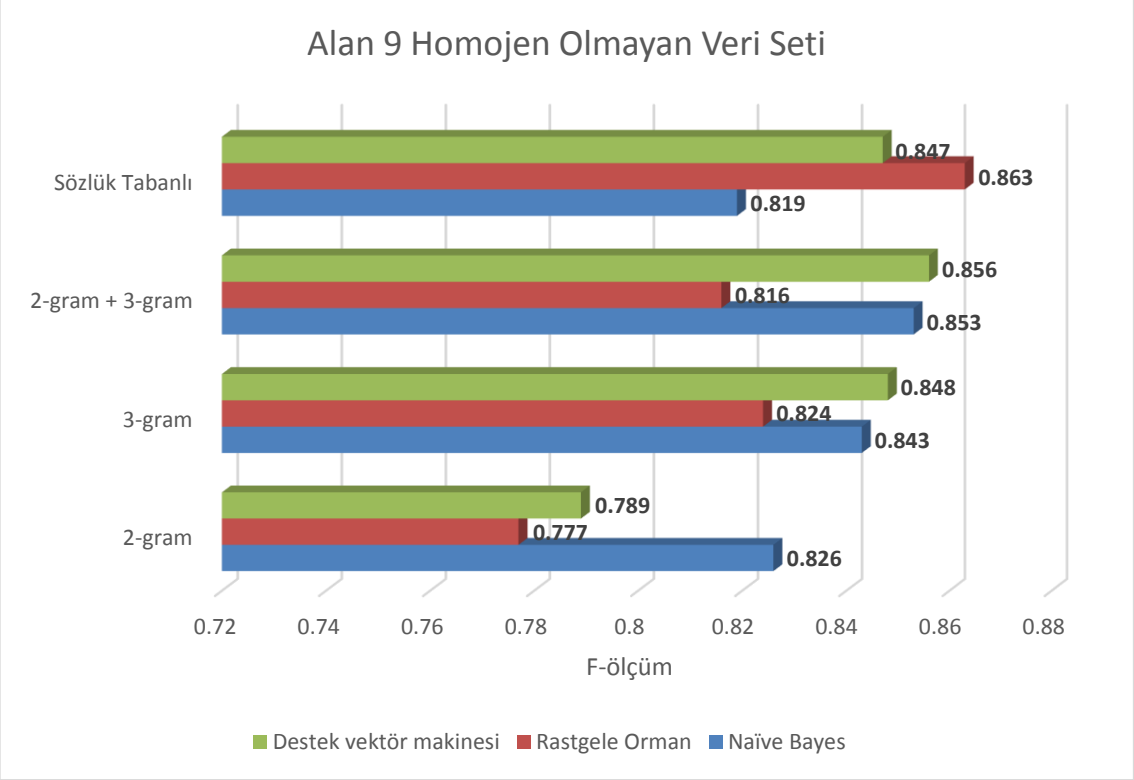


Şekil 5. 23 Alan bağımsız homojen olmayan veri seti için yöntemlerin karşılaştırması



Şekil 5. 24 Alan 9 homojen veri seti için yöntemlerin karşılaştırması

Alan bağımlı veri setleri ile yapılan deneyler göz önüne alındığında ise farklı bir durum ile karşılaşılmaktadır. Şekil 5.24'de görüldüğü gibi sözlük tabanlı yöntem diğer yöntemler kadar başarı göstermemiştir. Alan bağımlı verilerde daha başarılı olan n-gram yöntemi örnek olarak 9. alan için incelendiğinde %80 - %90 arasında değişen sınıflandırma başarısı ile oldukça iyi sonuçlar vermiştir. 3-gram özellik vektörü ile eğitilen DVM sınıflandırıcısı 0.903 f-ölçüm değeri elde etmiştir ve sözlük tabanlı yöntem ile aralarında %10'dan daha fazla fark olduğu görülmektedir. Şekil 5.25'de homojen veri seti ile aynı karşılaştırma yapıldığında sözlük tabanlı yöntemin rastgele orman sınıflandırıcısı ile n-gram yöntemlerinden daha başarılı olduğu görülmektedir. N-gram yöntemleri ise kendi içinde en başarılıdan başlayarak 2-gram, 3-gram birleşimi, 3-gram ve 2-gram olarak sıralanmaktadır.



Şekil 5. 25 Alan 9 homojen olmayan veri seti için yöntemlerin karşılaştırması

BÖLÜM 6

SONUÇ

Bu çalışmada 9 farklı alandan toplanan Twitter verileri olumlu, olumsuz ve nötr olarak sınıflandırılmıştır. TwitterAPI üzerinden toplanan tivitler insan eli ile etiketlenerek No-SQL veritabanı olan Solr'da saklanmıştır. Etiketlenen tivitler sözlük tabanlı ve n-gram tabanlı olmak üzere 2 farklı şekilde işlenmektedir. Çalışmanın amacı işlenen bu tivitler ile başarılı bir sınıflandırıcının oluşturulmasıdır. Bunun için sistem hem alanları tek tek kendi başına hem de bütün alanları bir arada içeren bir eğitim seti ile eğitilmiştir. Sınıflandırma yapılırken Naive Bayes, Rastgele Orman ve Destek Vektör Makinesi yöntemleri kullanılmıştır. Sınıflandırma başarıları için f-ölçüm değerleri dikkate alınmıştır ve bütün karşılaştırmalar bu değer üzerinden yapılmıştır.

Bu çalışmada bütün alanlardan toplanan verileri içeren alan bağımsız veri seti ile en büyük başarı sözlük tabanlı yöntem ile elde edilmiştir. Homojen veri seti ile homojen olmayan veri seti karşılaştırıldığında homojen olanın daha başarılı olduğu görülmüştür ve bu bölümde homojen veri setinden elde edilen değerler verilmektedir. Çizelge 6.1'de bütün alanlar için NB, RO ve DVM sınıflandırıcılarının CFS kullanmadan eğitilmesi ile elde edilen sonuçlar görülmektedir. Alan bağımsız veri setinde 0.895 f-ölçüm değeri ile mükemmele en çok yakınsayan başarı Rastgele Orman yöntemi ile elde edilmiştir. Rastgele Orman alan bağımlı veri setlerinde de 9'unun 4'ünde en yüksek başarıyı göstermiştir. Destek Vektör Makinesinin ise 5 alanda birden diğerlerine göre daha başarılı olduğu için alan bağımlı veri setlerinde daha başarılı olduğu söylenebilir.

n-gram yönteminde ise en yüksek başarı 2-gram ve 3-gram özellik vektörleri birleştirilerek elde edilmiştir. Özellikle alan bağımlı veri setlerinde daha başarılı olunan bu yöntemde sınıflandırıcılar daha yüksek başarı verdiği için CFS kullanılmıştır. Çizelge

6.2'de görüldüğü gibi Destek Vektör Makinesi alan bağımsız veri setinde en yüksek başarıyı 0.855 f-ölçüm değeri ile verirken alan bağımlı veri setlerinin 4'ünde başarılı olabilmiştir. Alan bağımlı homojen veri setlerinin 5'inde en yüksek başarı ise Naive Bayes yöntemi ile alınmıştır.

Çizelge 6.1: Sözlük tabanlı yöntem ile alınan f-ölçüm değerleri

Veri Seti	**NB**	**RO**	**DVM**
Alan 1	0.761	0.802	0.808
Alan 2	0.667	0.712	0.714
Alan 3	0.548	0.72	0.711
Alan 4	0.736	0.835	0.794
Alan 5	0.781	0.813	0.794
Alan 6	0.439	0.601	0.628
Alan 7	0.372	0.593	0.601
Alan 8	0.403	0.583	0.561
Alan 9	0.769	0.806	0.81
Alan Bağımsız	0.875	0.895	0.882

Çizelge 6.2: 2-gram ve 3-gram özellik vektörleri ile eğitilerek alınan f-ölçüm değerleri

Veri Seti	**NB**	**RO**	**DVM**
Alan 1	0.897	0.867	0.892
Alan 2	0.8	0.753	0.803
Alan 3	0.819	0.768	0.807
Alan 4	0.887	0.894	0.897
Alan 5	0.753	0.719	0.785
Alan 6	0.834	0.81	0.825
Alan 7	0.763	0.754	0.74
Alan 8	0.763	0.736	0.727
Alan 9	0.852	0.835	0.868
Alan Bağımsız	0.836	0.825	0.855

Bütün yöntemler incelendiğinde alan bağımsız veri setlerinde sözlük tabanlı yöntemin en yüksek başarıyı Rastgele Orman sınıflandırıcısı ile verdiği Çizelge 6.3'de görülmektedir. 2-gram ve 3-gram özellik vektörlerinin birleştirilmesi ise bütün sınıflandırıcılar için başarıyı yükseltmiştir. N-gram yöntemlerinde başarı oranları düşükten başlanırsa sırasıyla 2-gram, 3-gram, 2-gram + 3-gram özellik vektörleri ile elde edilmiştir.

Çizelge 6.3: Alan bağımsız veri setleri için yöntemlerin karşılaştırması

Sınıflandırıcı	2-gram	3-gram	2-gram+3-gram	Sözlük Tabanlı
NB	0.784	0.805	0.836	0.875
RO	0.759	0.803	0.825	0.895
DVM	0.798	0.817	0.855	0.882

Yapılan bu çalışmada çok kirli olduğu söylenebilecek Twitter verileri işlenerek makinenin anlamlandırılıp olumlu, olumsuz, nötr olarak sınıflandırabileceği bir sınıflandırıcı oluşturulmuştur. Yapılan deneylerde sözlük tabanlı yöntem ve Rastgele Orman sınıflandırıcısı kullanılarak 0.895 f-ölçüm değeri ile neredeyse %90'a yakın bir başarı elde edilmiştir. Literatürde metin sınıflandırma konusunda oldukça başarılı çalışmalar vardır ancak, söz konusu Twitter verileri için elde edilen bu sonuç oldukça tatminkardır.

Sosyal medya sürekli olarak gündeme göre içeriği ve jargonu değişen dinamik bir mecradır. Bu nedenle tivitlerin sürekli başarılı bir şekilde anlamlandırılabilmesi için dinamik bir şekilde kendi kendini eğiten bir sistemle sağlanabilir. Gelecekte bu mantığı baz alan ve bu çalışmada kullanılan yöntemleri birleştirerek sınıflandırıcılardan elde edilen sonuçları oylama mantığı ile seçen bir proje yapılabilir. Böylece sadece alandan değil gündemden de bağımsız başarılı bir sınıflandırıcı edilmiş olacaktır.

KAYNAKLAR

[1] Claster W. B., Dinh H. ve Cooper M.,(2010). "Naive Bayes and Unsupervised Artificial Neural Nets for Caneun Tourism - Social Media Data Analysis", Second World Congress on Nature and Biologically Inspired Computing

[2] Şimşek M. U., Özdemir S.,(2012). "Analysis of the Relation Between Turkish Twitter Messages and Stock Market Index", 978-1-4673-1740-5/12 IEEE

[3] Bian J., Topaloglu U. ve Yu F., (2012). "Towards Large-scale Twitter Mining for Drug-related Adverse Events", SHB'12

[4] Nguyen L. T., Wu P., Chan W. ve Peng W., (2012). "Predicting Collective Sentiment Dynamics from Time-series Social Media", WISDOM '12

[5] Byun C., Lee H. ve Kim Y., (2012). "Automated Twitter Data Collecting Tool for Data Mining in Social Network", RACS'12

[6] Murakami A., Nasukawa T., (2012). "Tweeting about the Tsunami - Mining Twitter for Informantion on the Tohoku Earthquake and Tsunami", WWW 2012 Companion

[7] Pennacchiotti M., Popescu A., (2011). "A machine learning approach to Twitter user classification", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media

[8] Horn C., (2010). "Analysis and Classification of Twitter Messages", Master's Thesis, Graz University of Technology

[9] Neri F., Aliprandi C., Capeci F., Cuadros M. ve By T., (2012). "Sentiment Analysis on Social Media", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

[10] Akay A., Dragomir A., (2013). "A Novel Data-Mining Approach Leveraging Social Media to Monitor and Respond to Outcomes of Diabetes Drugs and Treatment", 2013 IEEE Point-of-Care Healthcare Technologies (PHT)

[11] Twitter İstatistikleri, http://www.statisticbrain.com/twitter-statistics/, 4 Ocak 2014

[12] Phelan O., McCarthy K. ve Smyth B., (2009). "Using Twitter to Recommend Real-Time Topical News", RecSys'09

[13] Shamma D. A., Kennedy L. ve Churchill E. F., (2009). "Tweet the debates understanding community annotation of uncollected sources", WSM'09

[14] Degirmencioglu E. A., Uskudarli S., (2010). "Exploring Area-Specific Microblogging Social Networks", Web Science Conf.

[15] Aldahawi H. A., Allen S. M., (2013). "Twitter Mining in the Oil Business: A Sentiment Analysis Approach", IEEE Third International Conference on Cloud and Green Computing

[16] Hassan A., Abbasi A., Zeng D., (2013). "Twitter Sentiment Analysis: A Bootstrap Ensemble Framework", SocialCom/PASSAT/BigData/EconCom/BioMedCom

[17] Tang J., Liu H., (2012). "Unsupervised Feature Selection for Linked Social Media Data", KDD'12

[18] Katarzyna W., Bougueroua L. ve Dziczkowski G., (2011). "Social Media Analysis for e-Health and Medical", 2011 International Conference on Computational Aspects of Social Networks (CASoN)

[19] Gelernter J., Wu G., (2012). "High performance mining of social media data", XSEDE12

[20] Bachi G., Coscia M., Monreale A. ve Giannotti F., (2012). "Classifying Trust/Distrust Relationships in Online Social Networks", ASE/IEEE International Conference on Social Computing and ASE/IEEE International Conference on Privacy, Security, Risk and Trust

[21] Twitter API, https://dev.twitter.com/docs/api/1.1/get/search/tweets, 24 Ekim 2013

[22] Twitter4J: Twitter API erişimi için geliştirilmiş java kütüphasi, http://twitter4j.org/, 24 Ekim 2013

[23] Szomszor M., Kostkova P. ve Swineflu D. Q. E., (2009). "Twitter predicts swine flu outbreak in 2009" In Proceeding of 3rd International ICST Conference on Electronic Healthcare for the 21st century

[24] Alıcı işletim karakteristiği - ROC, http://tr.wikipedia.org/wiki/ROC, 22 Şubat 2014

[25] F-ölçüm değeri, http://en.wikipedia.org/wiki/F1_score, 22 Şubat 2014

[26] Archer K.J., (2008). "Emprical characterization of random forest variable ımportance measure", 52(4),2249-2260

[27] Breiman L.,Cutler A., Random forest, http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm, 5 Kasım 2013

[28] Zemberek Doğal Dil İşleme Kütüphanesi, https://code.google.com/p/zemberek/ , 1 Mart 2014

[29] Weka: Data Mining Software, http://www.cs.waikato.ac.nz/ml/weka/, 3 Mart 2014

[30] Eroğul U., (2009). "Sentiment analysis in turkish". Middle East Technical, Ms Thesis, Computer Engineering, 2009

[31] Vural A. G., Cambazoglu B. B., Senkul P. ve Tokgoz Z. O., (2013).“A framework for sentiment analysis in turkish: Application to polarity detection of movie reviews in turkish”, In Computer and Information Sciences III, pages 437–445

[32] Turkmenolgu C., Tantuğ A. C.,(2014). “Sentiment Analysis in Turkish Media”, International Conference on Machine Learning (ICML)

[33] Kemik Doğal Dil İşleme Grubu, http://www.kemik.yildiz.edu.tr/, 22 Haziran 2014

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Meriç MERAL

Doğum Tarihi ve Yeri : 5.11.1989 BURSA

Yabancı Dili : İngilizce

E-posta : mericmeral@gmail.com

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Y. Lisans	Bilgisayar Müh.	Yıldız Teknik Üniversitesi	-
Lisans	Bilgisayar Müh.	Ege Üniversitesi	2011
Lise	Fen Bilimleri	Atatürk Lisesi	2007

İŞ TECRÜBESİ

Yıl	Firma/Kurum	Görevi
2011-Devam	Etiya Bilgi Teknolojileri	Yazılım Uzmanı

YAYINLARI

Bildiri

"Twitter Üzerinde Duygu Analizi", Meriç Meral, Banu Diri, (2014). *SİU 2014(IEEE 22. Sinyal İşleme ve İletişim Uygulamaları Kurultayı*), 23-25 nisan 2014, Trabzon