DUYGU ANALİZİNDE ÖZNİTELİK SEÇME METRİKLERİNİN DEĞERLENDİRİLMESİ: TÜRKÇE FİLM ELEŞTİRİLERİ

Assessment of Feature Selection Metrics for Sentiment Analysis: Turkish Movie Reviews

Fırat AKBA

Doç. Dr. Ebru AKÇAPINAR SEZER

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin

Bilgisayar Mühendisliği Anabilim Dalı için Öngördüğü

YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2014

Fırat AKBA'nın hazırladığı "**DUYGU ANALİZİNDE ÖZNİTELİK SEÇME METRİKLERİNİN DEĞERLENDİRİLMESİ: TÜRKÇE FİLM ELEŞTİRİLERİ**" adlı bu çalışma aşağıdaki jüri tarafından **Bilgisayar Mühendisliği Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof. Dr.

Başkan — Hayri SEVER

Doç. Dr.

Danışman — Ebru AKÇAPINAR SEZER

Yrd. Doç. Dr.

Üye — Sevil ŞEN AKAGÜNDÜZ

Yrd. Doç. Dr.

Üye — Erhan MENGÜŞOĞLU

Öğretim Görevlisi Dr.

Üye — Fuat AKAL

Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından **YÜKSEK LİSANS TEZİ** olarak onaylanmıştır.

Prof. Dr. Fatma SEVİN DÜZ

Fen Bilimleri Enstitüsü Müdürü

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- Tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- Görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- Başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- Atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- Kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- Ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

Beyan ederim.

27/11/2014

FIRAT AKBA

ÖZET

DUYGU ANALİZİNDE ÖZNİTELİK SEÇME METRİKLERİNİN DEĞERLENDİRİLMESİ: TÜRKÇE FİLM ELEŞTİRİLERİ

Fırat AKBA

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Danışmanı: Doç. Dr. Ebru AKÇAPINAR SEZER

Kasım 2014, 76 sayfa

İnternet hizmetlerindeki gelişmeler ve İnternet kullanıcıları sayısındaki artış, günlük aktivitelerimizde İnternet kullanımını daha gelişmiş bir seviyeye taşımıştır. Günümüzde insanlar talep ettikleri bilgiye İnternet üzerinde yaptıkları basit bir arama ile kolayca ulaşabilmektedirler. Hatta İnternet kullanımındaki yeni gelişmeler, kullanıcılara İnternetteki bilgileri sorgulayabilmelerine olanak vermektedir. İnternetteki bilgilerin büyük bir kısmı geribildirim yapılmasına açıktır. Bu geri bildirimler; anketler ve forum web siteleri aracılığıyla ilgili kuruluşlar tarafından yeni fikirleri analiz edebilmek için toplanmaktadır. Geribildirimlerin kısa sure içerisinde insan gücü ile analizi çok zordur. Çünkü çok fazla İnternet kullanıcısı olmasından dolayı bu yorumların değerlendirilmeleri için uzun bir işlem süresine gereksinim duyulmaktadır. Duygu analizi kavramı da bu fikirlerin ne kadar olumlu ve olumsuz olduğunu sınıflama aşamasında ortaya çıkan problemlerin çözümü noktasında keşfedilmiştir. Bu tezde, duygu analizi yöntemlerinin başarı oranları karşılaştırılarak incelenmiştir. Uygulanan birtakım deney sonuçlarına göre, kısa sürede cevap verebilecek ve daha az insan gücüne ihtiyaç duyacak bir sistem

oluşturulmasına çalışılmıştır. Tezde kullanılan veriler Türkçe film yorumları web sitesi üzerindeki kullanıcılar tarafından yorumlanıp puanlandırılmıştır. Bu veriler 0,5 ile 5,0 aralığında puanlandırılmıştır. Öznitelik seçme metrikleri istatistik alanında yaygın olarak kullanılmaktadır. Veriler elde edildikten sonra, bu araştırmanın gereksinimlerine cevap verebilmek için Destekçi Vektör Makineleri (SVM) ile öznitelik seçme metriklerinin çeşitli kategorilere ait yorumlardaki ayrıştırıcı özelliği kullanılarak, SVM'nin başarısına nasıl bir katkı yaptığı tespit edilmiştir. Önerilen sistem tasarımında sadece olumlu ve olumsuz kategorileri sınıflarken %83,9 F_1 değeri elde edilmiştir. Olumlu, olumsuz ve nötr yorumları sınıflarken ise %63,3 F_1 başarı değerine ulaşmıştır. Literatür araştırmalarındaki bulgulara dayanarak; önerilen sistem tasarımı, öznitelik seçme metriklerinin duygu analizinde başarı ile kullanılabileceğini kanıtlamaktadır. Önerilen bu yeni sistem tasarımının duygu analizi konusuna yeni bir bakış açısı getireceğine inanmaktayız.

Anahtar Kelimeler: Duygu Analizi, Öznitelik Seçme, SVM, Naive Bayes, Türkçe Derlem

ABSTRACT

ASSESMENT OF FEATURE SELECTION METRICS FOR SENTIMENT ANALYSIS: TURKISH MOVIE REVIEWS

Fırat AKBA

Master of Science, Department of Computer Engineering

Supervisor: Assoc. Prof. Dr. Ebru AKÇAPINAR SEZER

November 2014, 76 pages

Achievements in the Internet services and the increase in the number of Internet users transformed our daily Internet activities to an upper level. People can easily access to their demand via a simple search on the Internet. Even these achievements enable users to query Internet information. Most part of the information presented on the Internet is open for feedbacks. User feedbacks have been captured from polls and forum web sites to be analyzed and produce new ideas. In fact it is hard to analyze them manually in a short time due to existence of huge amount of Internet users' reviews and so much processing time and effort required to evaluate these reviews. Sentiment Analysis concept has been discovered to solve problems occurred while classifying opinions by separating positive and negative reviews. Sentiment analysis concept was discovered at the point of assessing these reviews by means of classifying them into "positive" and "negative". In this thesis, sentiment analysis methods are investigated by considering their success rates. According to several experiment results, it is tried to develop a system that answers in a short time and needs less human efforts.

The data used in the thesis was commented and rated by users of Turkish movie reviews web site. This data was rated by gaps between ‘0.5’ and ‘5.0’ points. Feature Selection metrics have been used frequently in the field of statistical. Upon gathering the data, how the use of discrimination feature of SVM and feature selection metrics in comments of various categories has contributed to SVM’s success has been discovered. In the proposed system design, 83.9% F1 score is obtained while classifying only positive and negative reviews. While classifying both of positive, negative and neutral reviews, 63.3% F1 achievement score was reached. Based on findings in the literature review, proposed system design proves feature selection metrics can be used successfully in the sentiment analysis. We believe this proposed system design will bring a new perspective in the field of Sentiment analysis.

Keywords: Sentiment Analysis, Feature Selection, Support Vector Machine, Naïve Bayes, Turkish Corpus.

TEŞEKKÜR

Yüksek lisans tezimde ve eğitimimde emeği çok geçmiş ve yardımlarını hiç bir zaman esirgemeyen sayın anabilim dalı başkanımız Prof. Dr. Hayri SEVER ve tez danışmanım olan Sn. Doç. Dr. Ebru AKÇAPINAR SEZER'e minnet ve teşekkürlerimi sunuyorum. Tez çalışmam üzerindeki fikirlerinden ve iyi niyetli çabalarından dolayı değerli tez jürisi üyelerimiz olan Sn. Yrd. Doç. Dr. Sevil ŞEN AKAGÜNDÜZ, Sn. Yrd. Doç. Dr. Erhan MENGÜŞOĞLU ve Sn. Dr. Fuat AKAL hocalarımıza buradan ayrı ayrı teşekkür etmek istiyorum.

Ayrıca meslektaşım ve yakın arkadaşlarım olan Sn. İhsan Tolga MEDENİ, Sn. Behzad NADERALVOJOUD, Sn. Selman BOZKIR ve Sn. Alaettin UÇAN'a teşekkür ederim. Bu tezi maddi ve manevi olarak desteklerini hiçbir zaman üzerimden eksik etmeyen değerli aileme ve yakın zamanda kaybetmiş olduğum sevgili anneannem merhume Melehat ERTUĞRUL'a atfediyorum.

İÇİNDEKİLER

Sayfa

ŞEKİLLER

Sayfa

TABLOLAR

Sayfa

SİMGELER VE KISALTMALAR

Simgeler

Kısaltmalar

DAL	Dictionary of Affect in Language (Dil Etkileşim Sözlüğü)
IR	Information Retrieval (Bilgi Erişimi)
ML	Machine Learning (Makine Öğrenimi)
NB	Naive Bayesian
SA	Sentiment Analysis (Duygu Analizi)
SVM	Support Vector Machine (Destekçi Vektör Makinesi)
VSM	Vector Space Model (Vektör Uzay Model)

1. GİRİŞ

1.1. Problem Tanımı

Geçmişten günümüze teknolojinin gelişmesi ve akabinde bilgisayarların günlük hayatımıza girmesiyle beraber işlerimiz daha hızlı bir şekilde ilerleye başlamıştır. İşlerin daha hızlı ilerlemeye başlaması demek normalde harcanan sürede daha fazla iş yapabilme potansiyeli anlamına geldiğinden dolayı bilgisayarların önceden yapılan işlerin geçmişini depolayabilmesi için sahip oldukları depolama ünitelerinin günden güne geliştirildiğine şahit olmaktayız. Buradaki en iyi çözüm sadece bilgisayarların depolama birimlerini veya işlem kapasitesini yükseltmek değildi. Çünkü toplanan bu veriler ileride bir gün lazım olması için depolanmaktaydı. Bu veriler çeşitli türlerde olabilmektedir. Örneğin, fatura bilgileri, satış deneyimleri, personel sigorta bilgileri veya şikâyet, öneriler gibi duygu içeren veriler olabilmektedir. Bu tür verileri işleyebilmek kadar, bu verileri toparlama işlemi de zahmetli olmaktadır. Sayısal verilerin analiz edilebilmesi için muhasebe yazılımları geliştirilmiş, arşiv yöneticileri ve birçok buna benzer uygulamalar geliştirilmiştir. Böylelikle insan kontrolü yavaş yavaş yerini bilgisayar kontrolündeki otomasyon sistemlerine yerini bırakmaya başladı. Bu otomasyonlar o kadar etkili bir biçimde çalışıyordu ki çoğu durumda yaptığı analizler sayesinde insanların yerine karar veren ve Karar Destek Sistemleri (Decision Support Systems) [1] adı verilen uygulamalar geliştirilmeye başlandı. Buna yapay zekâ yazılımların katkısını da ekleyecek olursak, neredeyse her durum için akıllı uygulamalar geliştirebilir hale geldi. Teknolojinin bu denli hızlı ilerlemesi sayesinde makine öğrenimi yöntemlerinin başarısı artık göz ardı edilmemelidir.

İnternet kavramı da teknoloji ile beraber gündelik yaşamımıza girmiş ve şu an için dünyadaki en fazla veriyi kendi bünyesinde barındırmaktadır. Bu nedenle veri madenciliği ile ilgili yapılmış çalışmaların İnternet üzerinde yapılması kaçınılmazdır. İnternetin bu kadar yaygın kullanılması sayesinde insanlar birbirileriyle ucuz bir şekilde iletişim kurmakta, çevrim içi sanal mağazalarını oluşturmakta hatta bunlar aracılığıyla uluslararası ticaret gerçekleştirmektedirler. Çevrim içi alışverişlerin güvenliği günümüzde halen sorgulanmakta ve bu konuda

araştırmalar yapılmaktadır. Bu araştırmaların bir tanesinde Sultan ve ekibi [2] sormuş oldukları çevrim içi alışveriş yapan birinden “Çevrim içi alışverişlerde riski en asgari düzeye indirmek için önceki kullanıcıların geribildirimleri çok yardımcı oluyor.” diye bir cevap almışlardır ve tezlerinde bunun önemine dikkat çekmişlerdir. Yani bir insana güven duymak nasıl bir süreç içerisinde gerçekleşiyorsa, İnternet üzerinde de bu süreç ne kadar iyi geribildirimler aldığına bağlı olmaktadır. Bunun için İnternet üzerindeki bilgi paylaşan veya ticaret yapan web sitesi sahipleri geribildirimlerle gerçek insanların etkileşimde bulunduğu bu kişilerin memnuniyetlerini veya şikâyetlerini paylaşabilmesini sağlamışlardır. Bu hem web site sahibi tarafından görülen bir eksikliği giderme amaçlı, hem de kullanıcıların olumlu veya olumsuz duygularını dile getirmesini sağlayan bir yöntemdir. Bu işlem ticari olarak şirketlerin piyasa profilleri hakkında araştırma yapan anket kuruluşları tarafından günümüzde halen gerçekleştirilmektedir. Önceleri, insanların sahip olduğu duygu ve düşünceler, değerlendirilmesi çok uzun zaman alan bu tür yöntemlerle elde edilmeye çalışılmaktaydı. Bu işlem için her kapı tek tek dolaşılmakta ve anket için doğru kişileri bulana kadar durmaksızın devam edilmekteydi. Elde edilen bu yazılı anket bilgilerinin değerlendirilmesi içinse ayrıca bir o kadar süre ve iş gücü harcanması gerekmekteydi. Bu nedenle insan gücünden kaynaklı olarak işin maddi gideri de anket çalışmalarını yapan kurum ve kuruluşlara çok ağır bir yük teşkil etmekteydi. Teknolojinin ve İnternet’in kullanımının bu denli yaygınlaşması ile beraber kurum ve kuruluşlar anket çalışmalarına giden giderleri azaltmak için İnternet üzerinden anket bilgi yönetim sistemleri geliştirerek daha kolay bir şekilde insanlara ulaşmaya başlamıştır.

Duygular kişiden kişiye, konudan konuya değişkenlik göstermektedir. Bu değişkenlik olumlu, olumsuz veya hiç duygu içermiyor olarak sınıflandırılabilir. Örneğin İnternet kullanıcıları sanal mağaza üzerinden ürün satın aldıktan bir süre sonra o ürünü değerlendirmek için, ürünün eksik yanlarını eleştirebildiği gibi, iyi olan yanlarını da söylemek isteyebilir. Böylece sonradan aynı ürünü almak isteyen bir müşterinin ürünü satın alma işlemini daha bilinçli bir şekilde gerçekleştirmesine olanak sağlanabilir. Benzer şekilde, bir filmi izlemek isteyen kişinin IMDB ya da benzeri puanlama yapılan siteler üzerinden o filmi izleyen diğer kişilerin yazmış olduğu yorumlar vasıtasıyla filmi izleyip izlemeyeceğine karar vermesi sağlanabilir.

İnternet üzerindeki bilgi, kaynak ve doküman sayısının fazlalığı nedeniyle insanların tek tek bu düşünceleri okuyup, değerlendirmesi ve bunlardan duyguları ayrıştırması çok zaman almaktadır. Bu noktada anket bilgi yönetim sistemleri yorumların otomatik değerlendirilmesi konusunda halen insan müdahalesine muhtaçtır. Bu konuda bilgisayarların yüksek işlem gücü kabiliyetine başvurmak, insan gücüne olan ihtiyacı ortadan kaldırabilecektir. Bu safhada, günden güne daha da bir önem kazanmakta olan makine öğrenimi yöntemleri sınıflamalar konusunda çok güzel sonuçlar elde etmektedir. Binlerce dokümanı ayıklayıp sınıflamak insan eliyle yapıldığında çok zaman almaktadır. Oysa bunu bir makine öğrenimi yöntemiyle gerçekleştirmek hiç de fazla vakit almayacaktır. Sadece makine öğrenimi yöntemleri kullanılarak elde edilen başarılar yetersiz kalmaktadır. Bu noktada sonuçların daha tutarlı elde edilebilmesi için bir takım yöntemler kullanılmaktadır. Bu tezde yapılan çalışma, makine öğrenimi yöntemleri sonuçlarının iyileştirilmesi açısından kullanılan yöntemlerin başarılarına katkı sağlayacak bir çalışma içermektedir.

1.2. Çalışma Konusu ve Kapsamı

Bu tez genel olarak metinler üzerindeki duygu analizinin, IR (Bilgi Erişimi Sistemleri) ve ML (Makine Öğrenimi) çalışma alanlarının yardımıyla gerçekleştirilmesi üzerine dayanan bir sistematiği içermektedir. IR üzerindeki Öznitelik Seçme Metriklerinin (Feature Selection Metrics), Türkçe metinlerdeki ayrıştırıcı terimlerin tespitinde ve SVM üzerindeki başarım artırımı yöntemleri incelenmiş ve değerlendirilmiştir. Aynı zamanda ML yöntemlerindeki en yaygın ve başarılı olarak kullanılmakta olan gözetimli öğrenme (*Supervised Learning*) üzerinde yapılmış olan sınıflandırma (Classification) çalışmalarını ve uygulamalarını içermektedir.

1.2.1. Duygu Analizi (Sentiment Analysis)

Dünya üzerindeki doğal diller kullanılarak yazılmış olan metinlerin olumluluk veya olumsuzluk durumunun incelenmesine Duygu Analizi (SA) adı verilir. SA yapılabilmesi için incelenecek olan metinlerin belirli dilbilgisi kuralları içerisinde olması gerekir. Aksi takdirde metin içerisindeki anlam ve duygu vurgusunun tespiti mümkün olmayabilmektedir. SA klasik olarak el ile veya bilgisayar tabanlı bazı algoritmalar ve makine öğrenimi yöntemleri ile gerçekleştirilebilmektedir.

Metinlerin içermiş olduğu duyguların analizinin sağlıklı yapılabilmesi için öncelikle duygu içeren kelimelerin tespitinin başarılı bir şekilde gerçekleştirilmesi gerekir. Bunu sağlayabilmek için önceden belirlenmiş kelime listeleri kullanılır veya bu kelimeler Öznitelik Seçme (Feature Selection) gibi yöntemler ile belirlenir. Böylelikle cümlenin içermiş olduğu duygu doğru bir şekilde analiz edilebilir hale gelir.

Metin tabanlı duygu analizi gerçekleştirilmesi esnasında gerek insan gücünden kaynaklı maliyet olsun gerekse de geniş uygulanabilirlik alanı konusunda olsun, bilgisayar tabanlı duygu analizi yaklaşımları üzerine bilim dünyasında birçok araştırma ve geliştirme faaliyetleri gerçekleştirilmektedir. Bu faaliyetler literatür özeti kısmında tek tek incelenmiştir.

1.2.2. Öznitelik Seçme (Feature Selection)

Belirli sayıda ve sınıfta değişkenleri bazı matematiksel algoritmalar ışığında ağırlıklandıran öznitelik seçme metrikleri, uygulanan veriler içerisindeki en değerli veya ayrıştırıcı değişkenlerin tespiti konusunda çok yardımı olan bir uygulamadır. Öznitelik seçme, makine öğrenimi, veri madenciliği ve istatistik bilim alanları içerisinde çok yaygın olarak kullanılan yöntemlerden biridir. Özellikle verileri sınıflama ve eleme işlemleri için yaygın olarak kullanılır. Öznitelik seçme metriklerinin faydası olarak büyük boyutlu veriler üzerinde yapılan verilerin eliminasyonunda harcanan süre ve maliyeti düşürmesi gösterilebilir. Bunu gerçekleştirirken eğitim setinde sıkça kullanılan ve analiz sonucuna katkı

sağlamayan kelimelerin elenmesini sağlar. Böylelikle cümlelerde daha az geçen ve sonucu birebir değiştirme etkisi bulunan kelimelerin tespiti gerçekleştirilir.

Öznitelik seçme metrikleri bu işlemi matematiksel denklemler kullanır ve her bir kelime için ayrı ayrı puanlamalar yaparak önceden tanımlanmış kategoriler için en değerli kelimeleri belirler. Analizi yapılan kelimeleri ise sayısal değerlikler vererek hangisinin en değerli ve değersiz olduklarını ifade eder. Böylelikle cümleler en sağlıklı bir şekilde ve içerisindeki bilgi kirliliği azaltılmış olarak analiz yapılabilir hale getirilir.

Öznitelik seçme yöntemleri farklı denklemler kullanılarak gerçekleştirilebilmektedir. Bu denklemlerin başarıları genel olarak birbirlerine yakın oranlardadır. Bu denklemlerdeki farklılıklar daha az veya fazla değişkene bağlı hesaplamalar yapılmasından kaynaklanır. Öznitelik seçme metriklerinin etkililik ve tutarlılık oranı geçmişte yapılan akademik literatür de baz alındığında gayet başarılı, ve çok yaygın olarak halen kullanılmakta olduğu söylenebilir.

1.3. Amaç

Bilgi erişimi sistemlerinin amacı belirli bir derlem üzerinde var olan bilgileri istenilen biçimde sağlamak olduğundan, bu tezde yapılan çalışmanın amacı Türkçe film web siteleri üzerinde var olan yorumların içerisindeki olumlu, olumsuz veya duygu içermeyen (Nötr) yorumların en iyi başarı oranıyla tespit edilip, bu üç kategoride (Olumlu, Olumsuz, Nötr) istenen veri çıkarımının gerçekleştirilmesidir. Şu an mevcut olup, metinsel duyguların analizinde halen kullanılmakta olan iki yöntem mevcuttur; ilki "*öncül yön"* puan hesaplama (SentiWordNet) ve ikincisi de makine öğrenimi yöntemleridir. Tezde yapılan literatür çalışmalarıyla, bu yöntemlerin başarıları ve uygulanabilirlikleri incelenmiş ve bu yöntemlerden ML uygun görülmüştür. Veri çıkarımının doğru bir şekilde yapılabilmesi için IR yöntemlerinin literatür çalışmalarında var olan, fakat SA ve ML literatüründe daha önce rastlanmamış olan öznitelik seçme metriklerinin etkilerinin incelenmesi kararlaştırılıp, ML literatüründeki bazı başarılı yöntemlerin sonuçlarıyla karşılaştırılması amaçlanmıştır.

1.4. Motivasyon ve Özgün Değer

Ele alınan problemin önemi ve güncelliği, uygulanması planlanan yöntemlerin özgünlüğü ile ayrıca Türkçe üzerinde şimdiye kadar bu tür çalışma yapılmamış olması göz önünde bulundurulduğunda, bu tez sonuçları Türkçe dili için yapılacak başka çalışmalara referans olma niteliğindedir. Bu tezde bilgi erişimi ve veri madenciliği alanlarında kullanılan öznitelik çıkarımı yöntemleri kullanılarak Türkçe dilinde yazılmış olan film eleştirilerinin içerdiği duygular tespit edilmeye çalışılmıştır. Bu tezin Türkçe cümleleri en sağlıklı şekilde temsil edecek kelime sayısı hakkında deney ve çalışmaları içerisinde barındırmasıysa diğer çalışmalardan farklı olmasını sağlamaktadır. Deney sonuçları tamamlandıktan sonra Türkçe cümleler üzerinde duygu analizi yapılırken göz önünde bulundurulması faydalı olabilecek bir eşik değeri olan 375 adet kelime sayısına ulaşılmıştır. Şayet bir Türkçe film yorumları üzerinde yapılacak analizde öznitelik seçme işlemindeki puanı en yüksek ilk 375 adet terimin kullanılmasıyla başarı değeri en ideale yakın olacaktır.

Gerek metinler üzerindeki duygu analizinin veri madenciliği çalışma alanında uğraşılan en gözde konulardan biri olması, gerekse de bu tez çalışmasında belirtilen sistematiğin çalışma mantığıyla elde edilen sonuçların gelecekte yapılacak benzer çalışmalara ışık tutabilecek olması bu çalışmanın bir diğer önemli yanını oluşturmaktadır.

1.5. Tez Çalışmasında Kullanılan Başarım Ölçütü Terimleri

Yapılmış olan çalışmaların tez vasfını kazanabilmesi için elde edilen bulguların benzer ya da daha önceden yapılmış çalışmalara olan üstünlüklerinin veya zayıf kalan yanlarının tespit edilebilmesi gerekmektedir. Bu nedenle tez çalışmasında elde edilen deney sonuçlarının tüm bilim insanları tarafından kabul görmüş başarım hesaplama yöntemleri kullanılarak bu değerlerin ifade edilmesi gerekmektedir. Bu tez çalışmasındaki deney sonuçlarınca elde edilmiş olan

bulguların başarısını ifade etmek için kullanılmış olan başarım ölçütleri ve hesaplanma yolları alt başlıklar şeklinde açıklanmıştır.

1.5.1. Duyarlık Değeri (Precision)

Deneyler sonucunda yapılmış olan ölçümlerin birbirine ne derecede yakın olduğunu gösteren başarım ölçütü terimidir. Bu değer hesaplanırken doğru sınıflandırılmış pozitif örnek sayısının (TP), toplam pozitif örnek sayısına (TP+FP) bölünmesiyle bu değere ulaşılır (Eşitlik 1.1). Bu değer her zaman 0-1 aralığında olmaktadır.

$$Duyarlık = \frac{TP}{TP + FP}$$

Eşitlik 1.1 Duyarlık Değeri Eşitliği

1.5.2. Anma Değeri (Recall)

Anma değeri hedefi tutturma oranı olarak bilinmektedir. Yani gerçekte ulaşılması gereken bilgilere ne oranda ulaşılmış olduğunu gösteren bir değerdir. Bu değer hesaplanırken doğru sınıflandırılmış ilgili pozitif örnek sayısının (TP), toplam ilgili belgelerin sayısına (TP + FN) bölünmesiyle bu değere ulaşılır (Eşitlik 1.2).

$$Anma = \frac{TP}{TP + FN}$$

Eşitlik 1.2 Anma Değeri Eşitliği

1.5.3. Doğruluk (Accuracy)

Doğruluk değeri deneyde yapılan analizin gerçek değere ne kadar yakın olduğunu gösterir. Yani aynı şartlarda bu deney tekrarlanırsa yeni sonucun, önceki sonuca ne derecede benzer olacağını gösterir. Bu değer hesaplanırken doğru sınıflanmış pozitif ve negatif değerlerin sayısının toplamı (TP + TN), sınıflanan verilerin tümünün sayısına (TP + FP + TN + FN) bölünmesiyle elde edilir (Eşitlik 1.3).

$$Doğruluk = \frac{TP + TN}{TP + FP + TN + FN}$$

Eşitlik 1.3 Doğruluk Hesaplama Yöntemi

1.5.4. F1-Ölçümü (F1-Measure)

Önceki bölümlerde açıklanan Duyarlık ve Anma değerliklerinin harmonik ortalamaları (Eşitlik 1.4) hesaplanarak bu değer elde edilir. Bu değer 0 ile 1 değerleri arasında olur. Yapılmış olan testin doğruluğunu ifade eden bir değerdir. Veri madenciliği ve makine öğrenimi alanlarında yapılmış çalışmalar için en yaygın kullanılan başarı ölçütüdür.

$$F1 = 2 * \frac{Duyarlık * Anma}{Duyarlık + Anma}$$

Eşitlik 1.4. F1 Ölçümü Yöntemi

2. Literatür Özeti

Bu bölümde yapılan duygu analizi ile ilgili benzer çalışmalarda kullanılmış olan yöntemler ve varsa başarı oranları belirtilerek anlatılmıştır.

2.1. Sınıflama (Classification)

Herhangi bir verinin niteliğinin diğer verilerin niteliklerine göre kıyaslama yapılarak belirlenmesi işlemine sınıflama adı verilir. Veri sınıflamanın iki çeşit yöntemi vardır. Bunlar "Gözetimli" (Supervised) ve "Gözetimsiz" (Unsupervised) sınıflandırmalardır.

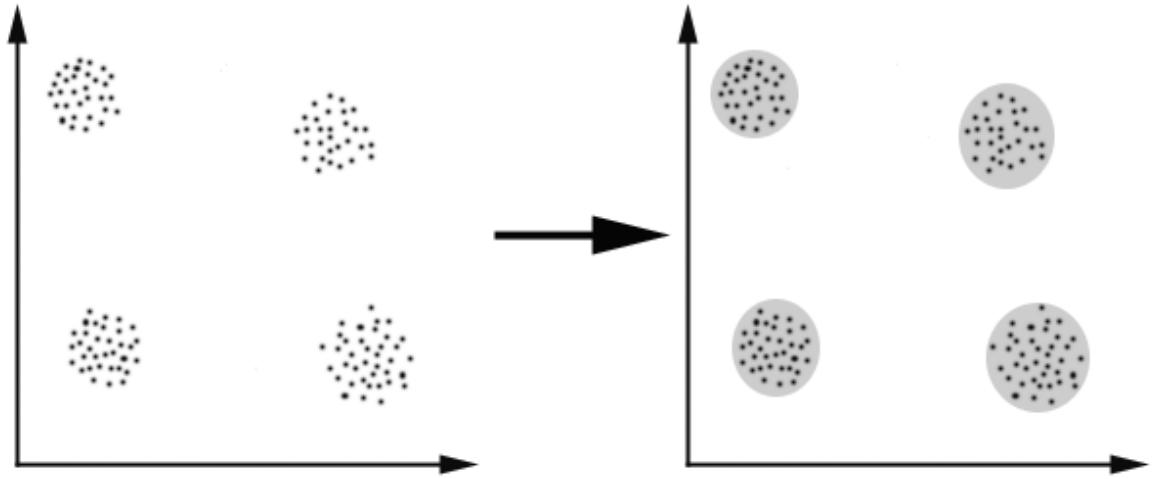
Gözetimli sınıflamada önceden değerleri ve sınıfları bilinen veriler, geliştirilen yöntemin testi veya analizinin başarısının tespit edilebilmesini sağlar. Gözetimsiz sınıflama ise etiketlenmemiş ve sınıflanmamış veriler üzerindeki bilinmeyen yapının tespiti için kullanılır. Bu uygulamaların daha iyi bilinen ve veri madenciliği alanında geçen isimleri aşağıda yer almaktadır.

i. Kümeleme (Clustering)
ii. İlişkilendirme (Association)

2.2. Kümeleme (Clustering)

Veri sınıflama yöntemlerinden biri olan gözetimsiz olarak, yani sınıf sayısı ve türü bilinmeyen verilerdeki yapılan sınıflamalar için kullanılır. Kümeleme işlemi, nesneler kümesi üzerinde belirli demetleme algoritmaları kullanılarak içerisindeki nesneleri gruplama işlemi olarak bilinir. Birbirine uzaklık olarak yakın noktalarda biriken benzer noktalara sınırlar verilerek verileri kümeleme işlemi gerçekleştirilir. Bu aradaki uzaklığı belirlerken *k-NN* algoritmasında olduğu gibi *Gaussian* ve benzeri uzaklık hesaplama yöntemleri kullanılır. Burada etiketleme işlemi kullanılmadığı için kategoriler oluşacak olan kümelerin sayısı kadar olacaktır.

Şekil 2.1'de [3] gösterilen iki ayrı görüntünün ilkinde dört farklı yerde kümelenen noktalar ikinci resimde sınırları belirlenerek kümelenmiştir.



Şekil 2.1 Kümeleme Yöntemi

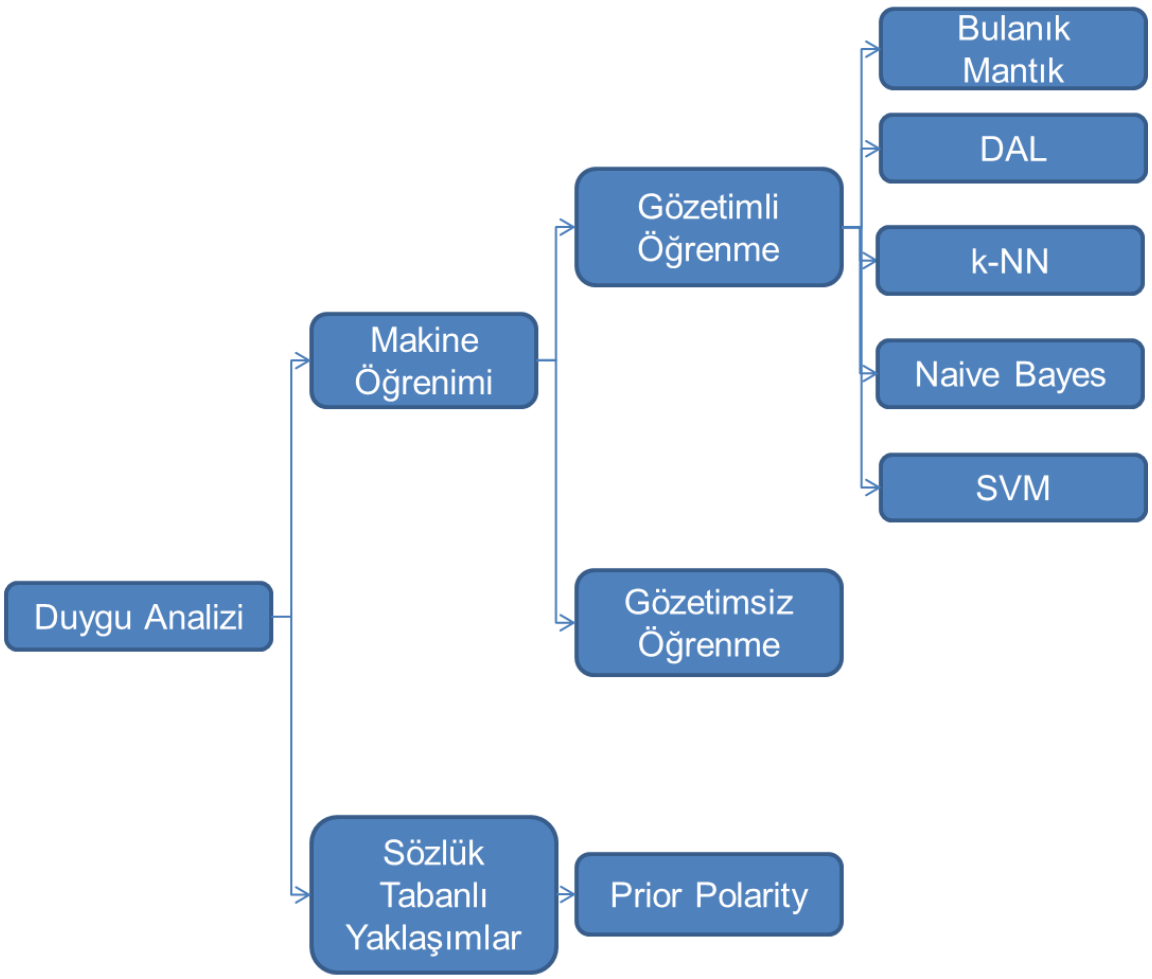
Bu kümeleme işleminin sağlıklı bir şekilde yapılabilmesi Han'ın [4] belirtmiş olduğu bazı özelliklere sahip olması gerekmektedir. İlk olarak ölçeklendirilebilir olması gerekmektedir. Yani terim sayısının artışına bağlı olarak uygulanan algoritmanın başarımı çok fazla değişmemelidir. Ayrıca farklı veri türlerini barındıran bir veri kümesi kullanarak kümelenecek verilerin siyah ve beyaz gibi birbirinden kolayca ayrışabilmesi gerekmektedir. Sonrasında gereken özellik ise gürültü (Noise) olan ortamlarda görevini başarılı bir şekilde devam ettirebilecek bir sistem tasarımına sahip olmasıdır. Son iki özellik olarak oluşturulan sistemin çok boyutlu veri tabanlarında uygulanabilir olması ve kümelemeye her nereden başlarsa başlasın sonuçların aynı tutarlılık derecesine sahip olması lazımdır.

2.3. İlişkilendirme (Association)

Veri sınıflama yöntemlerinden bir diğeri olan gözetimli olarak, yani sınıfların sayısı ve türleri bilinen veriler üzerinde yapılan sınıflamalar için kullanılan yöntemlerdir.

Genel olarak etiketli veriler ve önceden belirlenmiş sınıflar ile test ve eğitim kümesi arasında ilişkilendirilmeler yapılır. Buna örnek olarak süpermarket problemini verebiliriz. Salata ile ilişkilendirilen domates ve maydanozu beraber alan bir kişinin salatalık alması çok yüksek bir olasılık olması veya bu üç ürünü birlikte alan kişinin soğan alması olasılığının yüksek olması gibi. Bu şekil bağıntılar ile sistem eğitilerek bir kişinin alacağı bir sonraki ürünün tahmin edilmesi sağlanır. Bu türde yapılan veri sınıflama türüne ilişkilendirme adı verilir.

2.4. Duygu Analizi Literatüründe Kullanılmış Yöntemler



Şekil 2.2 Literatürde Kullanılmış Yöntemler

Bu bölümde geçmiş literatürlerde en sık kullanılmış yöntemlerden bahsedilmektedir. Aşağı kısımda anlatılan yöntemlerin bazılarının veri sınıflama türlerine göre hangi kategorilerde yer aldığı basitçe Şekil 2.2'de gösterilmektedir.

Duygu analizinin iki ayrı yaklaşım ile çözüldüğünden tezin giriş kısmındaki alt başlıklarda bahsedilmekteydi. Makine öğrenimi ve sözlük tabanlı yaklaşımlar ile bu analiz işlemi gerçekleştirilir. Şekil 2.2'de yöntem farklılıklarına göre dallara ayrılan yöntemlerin bütünü ve bunların başarısını arttırmaya yönelik kullanılan uygulamaların bütünü bu bölümdeki alt başlıklarda incelenmektedir.

2.4.1. Dil Etkileşim Sözlüğü (DAL)

İlk olarak 1989 yılında psikolojik olarak kelimelerin duygusal ifadelerinin puanlanması amaçlanarak oluşturulmuş ve duygu analizi alanında çok iyi bir başarı elde etmiş bir yöntemdir. İlk olarak İngilizce dili için geliştirilmiş olup ayrıca İspanyolca, Almanca ve Fransızca dilleri için başarılı bir şekilde uyarlanmıştır ve halen duygu analizinde kullanılan etkin yöntemler arasında sayılmaktadır.

Bu yöntem geliştirilirken 200 gönüllü kişinin yardımlarıyla İngilizce sözlük içerisinde yer alan toplam 8.742 adet duygu içeren kelimelerin her birine kendi elleriyle 3 ayrı kategoride 1 ile 5 puan aralığında sayısal değerler verilerek etiketlendirilmişlerdir. Her bir kelime hoşluk, aktivite ve tasvir yönleriyle ayrı ayrı puanlandırılmıştır. 1 ile 2 aralığı olumsuz duygu içeren, 4 ve 5 aralığı olumlu duyguları içeren kelimeler olarak, geri kalan ise duygu içermeyenler olarak sınıflandırılmıştır.

Bu işlemler gerçekleştirildikten sonra test olarak verilen cümleler bir kök eliminasyonu işlemine tabi tutulur. Test için verilmiş olan bu cümledeki tüm kök halindeki kelimeler eğer sistem içerisinde mevcut ise bulunan bu köklerin değerliklerinin toplanıp aritmetik ortalaması alınarak cümlenin nasıl bir duygu içerdiğinin tespitini yapmaktadır.

Kelime	Kelime Sınıfı	P puanı	A puanı	I puanı
amigo	isim	3	2,4	3
esperar	fiil	1,2	1	2,8
poder	fiil	2,8	2,8	2,2
terminar	fiil	2,2	3	2,8
prueba	isim	1,8	2,4	2,2
tiempo	isim	2	2	2,2
Ortalama Puan:		2,17	2,27	2,53

Tablo 2.1 İspanyolca DAL Puanlandırma Örneği

Tablo 2.1'de gösterilen örnekte İspanyolca olarak verilen bir cümlenin duygu puanının nasıl belirlendiği gösterilmektedir. Şekildeki "P puanı" hoşluk puanını, "A puanı" aktivite puanını ve "I puanı" tasvir puanını temsil etmektedir. Bu aritmetik ortalaması ise "Ortalama Puan" olarak gösterilmiştir.

İspanyolca için yapılmış olan çalışma olumlu ve olumsuz olarak verilen cümleleri bu yöntemle %62,33 oranıyla doğru sınıflayabilmiştir. İngilizce ve diğer diller içinde yapılmış olan çalışmalar yaklaşık olarak %62 civarlarında başarı oranı sağlamaktadırlar.

2.4.2. Naive Bayes (NB)

NB makine öğrenimi yöntemleri arasında en temel olarak kabul edilen bir sınıflandırma yöntemlerinden biridir. Dayanak noktası olarak Bayesian sınıflandırma yöntemi temellerine dayanmaktadır. Bu yöntemin en bilinir özelliği olayları olasılık hesaplamaları yaparak değerlendirmesidir. Bunun içinse önceden etiketlenmiş olarak sistemdeki eğitim kümesi verileri kullanmaktadır. Yöntem üzerinden test edilecek bir sorgulamada önceden verilmiş olan eğitim kümesini kullanarak bir olasılık değeri elde edip, bu skora göre de test verisinin hangi kategoriye benzediğini belirlemektedir.

Naive Bayesian ile Bayesian yöntemlerinin (Eşitlik 2.1) farkı ise yeni gelen test verisindeki bir değerin eğitim kümesi içerisinde yer almaması durumunda oluşacak

sıfır olasılık değerine olan bakış açılarıdır. NB bu durumu sıfır vermek yerine, belirlenmiş bir eşik değeri ekleyerek sonuçtaki hassaslık oranının daha da yükselmesini sağlamaktadır.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Eşitlik 2.1 Bayesian Yöntemi Denklemi

P(A|B) ; B olayı gerçekleştiği durumda A olayının meydana gelme olasılığıdır.

P(B|A) ; A olayı gerçekleştiği durumda B olayının meydana gelme olasılığıdır.

P(A) ve P(B) ; A ve B olaylarının öncelikli olasılıklarıdır.

NB $P(B \mid A)\, P(A)$ değerinin sıfır olması durumunda sıfır yerine uygun görülen bir eşik değeri konularak hesaba dahil edilir. Günümüzdeki literatür çalışmalarında en temel yöntem olarak kabul gören NB duygu analizi ve diğer türdeki sınıflamalar için yaygın olarak kullanılmaktadır.

2.4.3. k-Nearest Neighbors (k-NN)

Sınıflandırma problemlerinde etkin olarak kullanılmakta olan yöntemlerden bir tanesidir. Belirlenen “k” değerine göre verilen sorgunun “k” birim komşuluğunda yer alan vektörleri tespit ederek sınıflama işlemini gerçekleştirir. Bu yöntemin doğru uygulanabilmesi için iyi bir eğitim kümesi oluşturulması şarttır. Eğitim kümesi bu yöntemin başarısındaki en önemli faktördür. Uygulanmasının kolay olması nedeniyle *k-NN* sınıflama problemlerinde sık sık kullanılmaktadır. Bilgi kirliliğinin olduğu dokümanlarda sınıflama kabiliyeti güçlüdür.

k-NN algoritmasının çalışma mantığı her bir sorgunun ayrı ayrı hesaplanmasını gerektirdiğinden dolayı bu yöntemin hesaplanma maliyeti çok yüksektir. En yakın komşuluk bağıntısına dayandığı için vektör uzayında ifade edilen terimlerin birbirine olan uzaklıkları *Manhattan* yöntemi (Eşitlik 2.2), *Euclidean* yöntemi (Eşitlik 2.3) ya da Minkowski yöntemi (Eşitlik 2.4) yardımıyla hesaplanır. k-NN eşitlikleri içinde gösterilen “k” değeri komşuluk derecesini, “x” değeri kategorinin vektörünü “y” değeriyse sonucun vektörünü temsil etmektedir. Verilen sorgudaki terim benzerlik oranı “1” değerine en yakın olan kategoriye eklenir.

$$\sum_{i=1}^{k} |x_i - y_i|$$

Eşitlik 2.2 Manhattan Yöntemi

$$\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Eşitlik 2.3 Euclidean Yöntemi

$$(\sum_{i=1}^{k} (|x_i - y_i|)^q)^{1/q}$$

Eşitlik 2.4 Minkowski Yöntemi

2.4.4. Kelime Öbekleri (Bag-of-Words)

Doğal dil işleme ve bilgi erişimi sistemleri alanlarında yaygın olarak kullanılan bir başka yöntemdir. Türkçesi kelime öbekleri olarak bilinen bu yöntem uygulanırken,

cümle içerisinde geçen birden fazla kelimeyi tek bir kelimeyle ifade edermiş gibi değerlendirilmesini sağlar. Böylelikle isim veya sıfat tamlamalarının daha düzgün bir biçimde değerlendirilmelerini sağlar. Bunun gerçekleştirilebilmesi için değerlendirilmeye alınacak kelime öbeklerinin sisteme öğretilmesi gerekmektedir. Sisteme verilecek bir sorgu eğer önceden eğitilmiş olan veri kümesindeki kelime öbeklerini içeriyorsa o kelime öbekleri vektörler halinde test haznesine kaydedilirler. Daha sonradan geçme sıklıklarına göre veya önceden verilmiş olan ağırlıklarına göre gerekli hesaplamalar yapılarak sorgu sonucuna ulaşılır. Bu yöntem daha çok ML yöntemleri uygulanmadan önce sorgulardaki terim çıkarımı yapılırken tercih edilmektedir.

2.4.5. Konuşma Bölümü Etiketi & Duygu Etiketleri (PoS-Tag & Emotion Tags)

PoS-Tag dokümanlardan çıkarılan kelimelerin bir nevi dil yapısına göre etiketlendirilmesiyle oluşturulan veri setleridir. Bu veri seti kelimelerin isim, sıfat, zarf, zamir, edat ve bağlaç gibi durumlarına göre istenilen özelliklerde etiketlenir. Bu etiketlere göre belirli ağırlıklar verilerek sorgulanan cümlelerin hangi kategoride yer aldığının tespitinde rol alır. Konuşma dillerindeki özne-yüklem-tümleç sırasına göre cümledeki vurgulanmak istenen kelimenin ön plana çıkarılmasına da yarar. *Emotion Tags* ise işaret diliyle ifade edilen duyguları etiketleyebilmek için kullanılmaktadırlar. Özellikle internet ortamında yazılan cümlelerin içerisinde sık sık yer almaktadır. Semboller bazen cümlede direk olarak anlatılmak istenen duyguyu tespit etmekte çok başarılı olabilmektedirler. Bu nedenle çoğu terim çıkartma işleminde yaygın olarak kullanılmaktadır.

2.4.6. TF-IDF Ağırlıklandırma

IR ve metin madenciliği alanlarında veri setlerindeki bilgi kirliliğini azaltmak için sıkça kullanılan terim çıkartma yöntemlerinden bir tanesidir. Cümlelerde yer alan kelimelerin ne sıklıkla kullanıldığına yani frekansına ve diğer dokümanlarda geçme sıklıklarını birlikte hesaplayarak, kategoriler için en önemli kelimelerin tespitini

sağlar. Böylelikle test verisi üzerinde analiz edilecek kelimelerin sayısını azaltarak hem işlem süresinin kısalmasına hem de cümlelerin içerisinde çok sık kullanılan ve sistemin sonucuna katkıda bulunmayan zamir, bağlaç gibi istenmeyen kelime yapılarının sistemin sonucuna etki etmesini önleyecektir.

Term Frequency (TF) burada terimin sıklığını ifade etmektedir. Hesaplanırken dokümanda o terimin geçme sayısının (f), o terimin toplam dokümanlarda kaç kere geçmiş ise o sayıya (df) bölünmesiyle elde edilir. Inverse Document Frequency (IDF) ise analiz edilen kelimenin ne kadar eşşiz (Unique) olduğunu hesaplanması (Eşitlik 2.6) işlemidir.

IDF hesaplanırken "N" değeri toplam doküman sayısını temsil eder. "df" ise hesabı yapılan kelimenin kaç dokümanda yer aldığını temsil eder. Cümledeki *IDF* değerinin yüksek olması demek o kelimenin kullanılmasının, kategorinin belirlenmesi için çok değerli olduğunu gösterir. Örneğin "çok" kelimesi bir cümlede tek başına bir duygu ifadesi içermemektedir. Bu nedenle hem olumlu cümle hem de olumsuz cümlenin içerisinde sık olarak geçebilir. *TF-IDF* (Eşitlik 2.7) yöntemi tam bu esnada bu kelimenin ağırlığını düşük bularak önemsiz olduğunun tespitinde büyük rol oynar.

$$TF = \frac{f}{df}$$

Eşitlik 2.5 IDF Hesaplanması

$$IDF = \log\left(\frac{N}{df}\right)$$

Eşitlik 2.6 IDF Hesaplanması

$$TF - IDF = TF * IDF$$

Eşitlik 2.7 TF-IDF Hesaplanma Yöntemi

2.4.7. Bulanık Mantık (Fuzzy Logic)

Daha önceki bahsedilen yöntemler sorgulara her zaman iki seçenekli yani “Doğru” veya “Yanlış” şeklinde cevaplar vermek üzere tasarlanmıştı. Bulanık mantıkta ise verilen bir sorgunun “Çok doğru”, “Çok yanlış” veya “Az çok doğru” gibi cevapları olmaktadır. Bunun amacı da “Doğru” veya “Yanlış” olan sorgunun sonucunu hesaplarken yanılma riskinin olmasındandır. Bu yanılma riskinden dolayı Loutfi Zadeh tarafından “Bulanık Mantık” teorisi ortaya atılmıştır. Metin madenciliğinde kullanılmasının başlıca sebeplerinden bir tanesi sorgu sonuçlarının sözel olarak ifade edilebilen değerler olmasıdır. Bulanık mantık sayıların birbirine olan yakınlığına dayanır. Yani bir değer incelenirken diğer değere olan komşuluğu referans alınır. Yani eğitim kümesi içerisindeki belirtilen değer sorguda yer almadığı sürece komşusundaki değerlere bakıp bunun gerçekleşeceğini diğer istatiksel yöntemler gibi söylemeyecektir. Fakat bu komşu değerlere göre bunun olması ihtimalini veya tetiklenebileceğini kestirip doğru değere yakın bir sonucu vermeye çalışacaktır.

2.4.8. Tohum Kelimeler (Seed Words)

Tohum Kelimeler yöntemi bir eğitim verisi olmamakla beraber konuşma dillerinde sıkça kullanılan duygu ifadelerinin sonradan test edilen sistemin analizinin başarısını arttırması amacıyla sisteme dâhil edilen kelimelerdir. Bu yöntem daha çok hazır kurulu olan sistemlerin başarısını arttırması için, insan desteğiyle bunu gerçekleştirmektedir. Örneğin sistem eğitim setini öğrenme aşamasında analizi yapılacak o konu hakkında çok değerli bir kelimeyi öğrenememiş olsun. Bu durumda gelen test sonuçları değerli olan kelimeyi önemsemeyeceği için

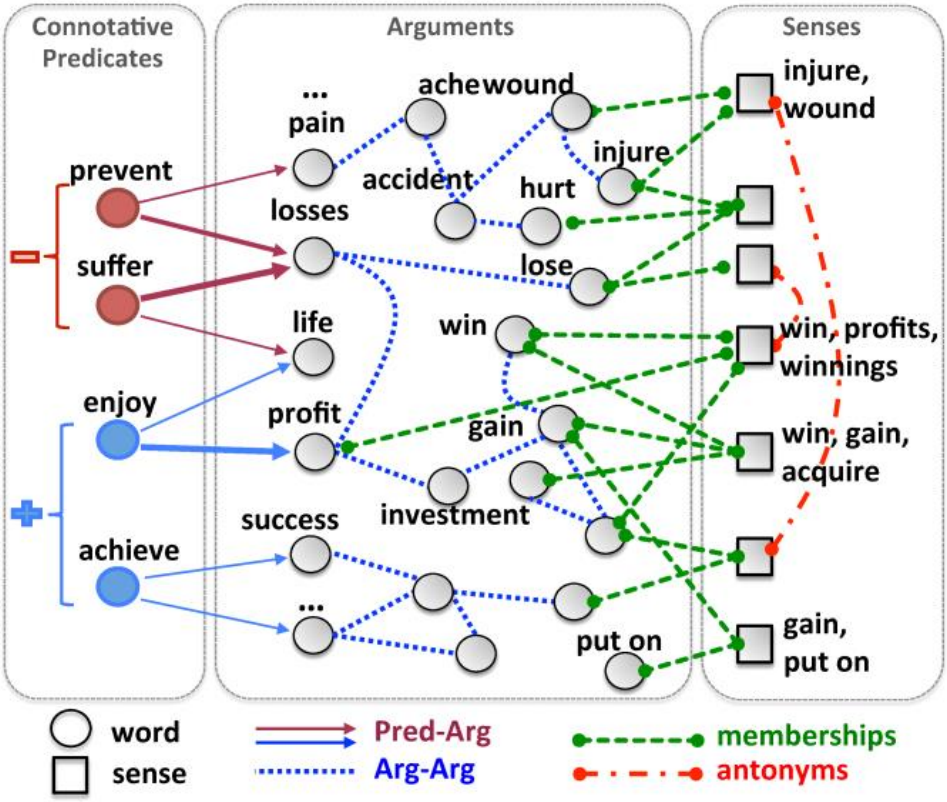
gerçekten olması gerektiği kategori yerine olmaması gereken bir kategoride olarak değerlendirilmesine neden olacaktır. Böylelikle sistemin başarımı düşük değerlerde seyredecektir. Tohum kelimeler kişinin eğitim setindeki verileri değiştirmesi yerine, değerlendirmeye girmesi gereken ve sistemde bulunmayan değişkeni dışarıdan müdahale ederek tohum gibi sisteme ekmesi işlemidir.

2.4.9. Öncül Yön Puanlandırma (Prior Polarity)

Şimdiye kadar literatür taraması bölümünde anlatılan çoğu yöntem ML yöntemlerinden veya ML yöntemlerinin başarısını arttırmaya yönelik uygulamalardı. Duygu analizi konusunda ML yöntemlerinin haricinde kullanılan diğer bir başarılı yöntemse öncül yön puanlandırma sistemidir. Bu sistemde önceki bölümlerde de anlatılan *Pos-Tag* yöntemiyle kelimeler dilbilgisi kuralları içerisindeki türüne (Fiil, isim, sıfat vb.) ve duygu içeriğine göre yüzlerce insana tek tek etiketlendirilip, ayrı ayrı puanlandırılmaktadır. Bu puanlara 'Öncül Yön' (*Prior*) puanlar denmektedir ve kelimenin dilbilgisi türüne göre bu puanlar farklılık gösterebilmektedir. Böylelikle kelimeler diğer kelimelerle ilişkilendirilip onlara ait duygu puanları oluşturulduktan sonra her cümledeki kelimelerin sahip olduğu puanlar hesaplanır. Bu hesaba göre de analiz edilen cümlenin hangi duyguyu içerdiği tespit edilmektedir. Bu yöntem ilk defa İngilizce için geliştirildiğinde "SentiWordNet" adı verilmiştir. Daha sonradan İngilizce ile aynı dil grubunda yer alan birkaç dil içinde aynı yöntem birebir kelimeler tercüme edilerek gerçekleştirilmiştir. Duygu analizi işini hem olumlu, olumsuz hem de duygu içermeyen cümleler için gerçekleştirebilmektedir. Eğer kullanılan dil içindeki kelimelerin çoğunu temsil edebilen bir puanlandırma cetveli oluşturulmuşsa bu yöntem çok başarılı bir şekilde çalışacaktır.

Duygu analizi alanında sıfırdan uygulanması en maliyetli olan yöntemlerden biridir. Bu puanlandırma için gereken insan sayısı ve zaman sayısı ML yöntemlerine göre oldukça çok fazladır. Türkçe için "SentiStrength" [8] adındaki yazılımın tercümesi yapılarak elde edilmiş bir sürümü mevcuttur. Fakat başarı oranlarıyla ilgili şimdiye kadar bir bilgiye rastlanılmamıştır. Şekil 2.3'te [6] yer alan şemada kelimelerin ilişkilendirilme işlemi tasvir edilmiştir. Görüldüğü üzere *Connotative Predicates*

(Çağrışım Fiilleri), *Arguments* (Konular) ve *Senses* (Hisler) olarak ayrılmış kümelerde olumsuz çağrışım fiilleri olarak belirtilmiş *prevent* (Önleme) ve *suffer* (Acı çekmek) yer almakta. Bunun yanında *enjoy* (Eğlenmek) ve *achieve* (Başarmak) fiilleri de olumlu çağrışımlar olarak tanımlanmıştır. *Arguments* (Konular) kısmında ise sözlükte yer alan ve değerlendirme için öneme sahip kelimeler bulunmaktadır. Böylelikle üç ayrı grupta ilişkilendirilen kelimeler vasıtasıyla yeni gelen cümlelerin analizi yapılabilmektedir.



Şekil 2.3 Öncül Yön Puanlandırma Kelime İlişkilendirilmesi

2.5. Benzer Çalışmalar

Yapılan literatür araştırmalarında, duygu analizinin birçok dilde uygulanmış olduğu fark edilmiştir. Bu çalışma alanında yapılmış en başarılı örnekler genellikle İngilizce için yapılmıştır. Diğer Avrupa dilleri için yapılmış olan çalışmalar, İngilizce

ile olan dil benzerliği nedeniyle hemen hemen aynı başarı oranlarına sahip olmuştur. Öte yandan Türkçe, morfolojik yapısı olarak sondan eklemeli bir dil olduğundan ve özne, yüklem, tümleç dizilişi Avrupa dillerinden farklı olduğundan dolayı Avrupa dilleri için uygulanmış olan bazı yöntemler, Türkçe için uygulandığında sınırlı düzeyde başarı sağlanmıştır.

Genel olarak literatür taramasında, SA uygulamalarının iki ayrı taraftan ilerlediği görülmüştür. Şu an var olan SA hakkında yapılmış var olan araştırmalar; ML yöntemleri yaklaşımıyla ve *Lexicon* (Sözlük) tabanlı yöntemler olmak üzere iki ayrı noktadan ilerlemektedirler. Sözlük tabanlı uygulamalar ilk olarak Esuli ve ekibinin yaptığı çalışmada [7] İngilizce için yapılmış olan DAL'ın (Dictionary of Affect in Language) [8] farklı dillere nasıl çevrilip hesaplanabileceği anlatılmaktadır. Bu yöntem ile Ghorbel ve ekibi [9] Fransızca için, Valdivia ve ekibi [10] İspanyolca için, Denecke [11] ise Almanca başta olmak üzere birçok dil için duygu analizi yapılabileceğini kanıtlamışlardır. Türkçe için uyarlanmış olan tek sözlük tabanlı uygulama olan SentiStrength, Vural ve ekibi [5] tarafından gerçekleştirilmiştir.

Duygu analizi yapmanın diğer bir yolu ise ML yöntemlerini kullanmaktır. Pang ve ekibi [12] geleneksel konu bazlı metin sınıflama yaklaşımını duygu analizi için uyarlamış ve Naive Bayes [13], Maximum Entropy Classification [14], Nearest Neighbor [15] ve Support Vector Machine [16] gibi klasik ML yöntemlerini kullanarak duygu sınıflandırması yapmışlardır. Genel olarak en başarılı sonuçlar SVM sınıflayıcıları ile alınmıştır.

Agarwal ve ekibi [17] kullanmış oldukları tek tek insan eliyle tek tek etiketlenmiş Twitter verileri üzerinde *Unigram*, *Senti-Feature* ve *Tree Kernel* modellerini test ederek bir takım sonuçlar elde etmişlerdir. Bunları gerçekleştirirken duygu sembolleri ve bazı kısaltmaları da deney içerisine entegre edecek uygulamalar kullanmışlardır. Böylelikle ikili (Olumlu, Olumsuz) sınıflamalar içerisinde en iyi sonuçları Senti-Feature ve Unigram modellerinin harmanlaması sonucu %75,39 doğrulukla hesaplamışlardır. Üçlü (Olumlu, Olumsuz, nötr) sınıflama içinse Senti-Feature ve Tree Kernel modellerinin harmanlanması sonucu doğruluk oranını %60,83 olarak hesaplamışlardır.

Bir başka çalışmada ise Becker ve ekibi [18], 2012 ve 2013 yılları içerisinde elde edilmiş toplam 475.000 adet etiketlenmemiş İngilizce SMS ve Tweeter verisi

üzerinde *Polarity Bag-of-Word*, duygu sembolleri ve *PoS Tag* yöntemlerini kullanarak üretmiş oldukları *Polarity Lexicon* (Kutupsal Sözlük) yöntemlerini SVM yardımıyla test etmişlerdir. İkili sınıflamalarda %88,9 ile %70,4 F_1 değerleri arasında başarılar almışlardır fakat nötr kategoriler için alınan sonuçlar %9 ile %31 F_1 değerleri aralığında kalmıştır.

Habernal ve ekibi [19] Çek sosyal medyası üzerinde dokuz farklı Facebook hayran sayfasından elde etmiş oldukları 10.000 adet yorumu toplayıp, her birini tek tek kendileri etiketlemeye çalışmışlardır. Bu yorumlar olumlu, olumsuz ve nötr olarak sınıflanmıştır. Hem olumlu hem de olumsuz ifadelerin geçtiği yorumlar ise *Bipolar* olarak adlandırılmış ve bu yorumlar hesaplamalarda kullanılmamışlardır. Daha sonra geriye kalan sınıflanmış yorumları N-gram [20], PoS features ve TFIDF yöntemleriyle sınıflamaya çalışmışlardır. İkili (Olumlu ve Olumsuz) sınıflamada en iyi %90 F_1, Üçlü (Olumlu, Olumsuz ve Nötr) sınıflamada ise en iyi %69 F_1 değerlerini elde etmişlerdir.

Almanya'da Narr ve ekibi [21] tarafından yapılan çalışmada, dört dilde toplam 10000 adet Twitter yorumunu Mekanik Türk Platformundaki kullanıcılarla etiketlemişlerdir. Daha sonrasında bu etiketlenen yorumların hepsini bir arada kullanarak duygu sembollerini ve NB (*Naive Bayes*) sınıflayıcısını kullanıp, bazı sonuçlar elde etmişlerdir. Elde edilen sonuçların iş yükü açısından çok bir gayret gerektirmediğini ve başarı oranlarının en iyi İngilizce'de yapılan ikili sınıflama için %81,3 olarak tespit ettiklerini paylaşmışlardır.

Pak ve ekibi [22] 300.000 adet karışık olarak dağılmış Olumlu, Olumsuz veya duygu içermeyen Twitter paylaşımlarını veri seti olarak kullanarak duygu içeren sembollerini, *N-gram* yöntemlerini uygulayarak ve yaptıkları ağırlıklandırmada sadece en sık geçen kelime frekanslarını referans alıp SVM sınıflayıcısı ile %61 civarı başarı elde etmişlerdir.

Pang ve ekibi [23] başparmak yukarı ve başparmak aşağı şeklinde duyguları olumlu veya olumsuz şeklinde etiketlendirerek kullanmış oldukları film yorumlarını N-gram, NB ve SVM yöntemlerini kullanarak test edip ikili (Olumlu, Olumsuz) sınıflamada Naive Bayes için en iyi %86,4, SVM sınıflayıcı içinse %86.15 doğruluk değerlerine ulaşmışlardır.

Saif ve ekibi [24] 60.000 adet Twitter yorumunu eğitim seti olarak olumlu ve olumsuz olacak şekilde ve her kategorideki yorum sayısı eşit olacak şekilde etiketlemişlerdir. Daha sonrasında duygu sembollerini, *N-gram* ve NB yöntemlerini kullanarak 1,54 milyon twitter yorumunu sınıflamışlardır. Bu çalışmada en yüksek başarıyı ikili sınıflamada %86,3 doğrulukla tespit etmişlerdir.

Genel olarak literatür çalışmalarına bakıldığında SVM ve benzer sınıflama yöntemleri kullanılarak %60 ile %86 F_1 değerleri aralığında sonuçlar elde edilmiştir. Ayrıca karşılaşılan çalışmalarda veri olarak film değerlendirmeleri, seyahat önerisi verileri, ürün değerlendirmeleri, sosyal medya verileri ve SMS verileri kullanılmıştır. Cümleler köklerine doğru bir şekilde ayrıldıktan sonra klasik *bag-of-words* (Kelime Öbekleri) [24] yöntemiyle öznitelik seçme işlemi yapılıp, bazılarında ise kelime grupları ağırlıklandırılarak değerlendirilmiştir. Öznitelik seçme işlemi yaparken *N-gram* yaklaşımı ve genel olarak POS (*Part-of-speech tagging*) etiketleri göz önünde bulundurulmuştur.

Türkçe DA konusunda ise oldukça az çalışmaya rastlanmıştır. Bu konuda ilk çalışan Eroğul [26] Türkçe olarak yapılmış film değerlendirmelerini toplamış, SVM ve bazı NLP (Doğal Dil İşleme) teknikleri ile SA yapmaya çalışmıştır. Olumlu ve olumsuz olarak yapılan iki kategorideki sınıflamada %85 başarı elde etmiştir.

Boynukalın [27] tez çalışmasında Türkçe SA konusunda yeni bir veri seti oluşturarak, Türkçe'nin morfolojik yapısından faydalanmış ve yeni değişkenler ekleyerek, sonuçları iyileştirmeye çalışmıştır.

Türkçede kelime kökleri, duygu ifadeleri ve cümle arasındaki ilişkiyi gözler önüne koyan Çakmak ve ekibi [28] 197 adet duygu içeren kelimeyi değerliği, aktivitesi ve baskınlık yönleri olmak üzere üç boyutlu olarak 0-100 aralığında kullanıcılara önceden puanlandırmışlardır. Daha sonra *Fuzzy Logic* (Bulanık Mantık) kullanarak kelimelerin cümle içerisinde belirtmiş olduğu duyguyu sayısal değerler olarak analiz etmişlerdir. Sonuçlarının cümledeki baskınlığı haricinde diğer iki boyutta iyi değerler verdiğini paylaşmışlardır.

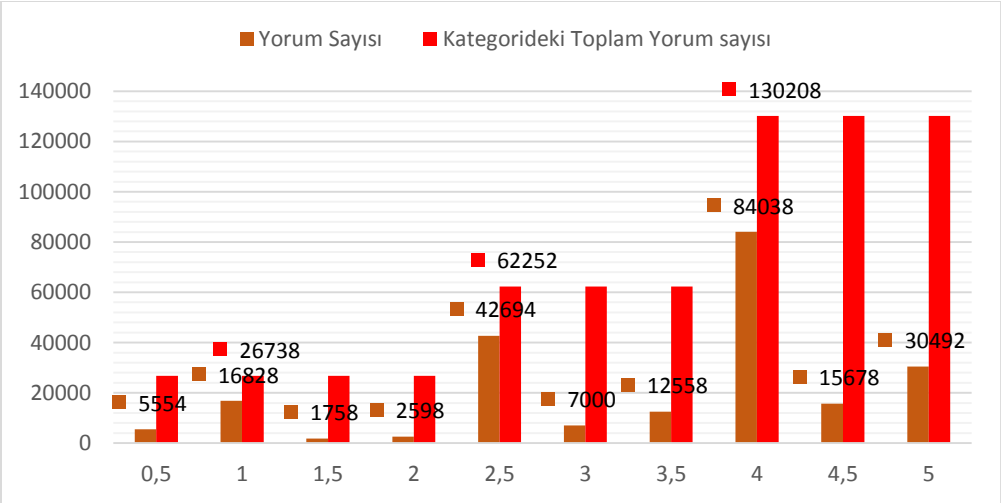
Özsert ve ekibi [29] da İngilizce dili için yapılmış *Seed Words* (Kelime Ekme) yöntemi ile duygu içeren kelime tanımlama yöntemi ile yeni kelimeler ekleyerek çalışmıştır.

Kaya ve ekibi [30] ise politika haberleri verisi ile NLP ve makine öğrenmesi kullanarak çalışmışlar ve nispeten başarısız sonuçlar elde etmişlerdir.

3. Veri Hazırlama ve Deney Süreci

3.1. Tez Çalışmasında Kullanılan Veri

Duygu Analizi konusunda yapılmış olan ilk çalışmaların büyük bir kısmında film yorumları kullanılmıştır. Çünkü tez çalışmasının eğitim ve test kümelerini oluşturarak yöntemin ne oranda başarılı olduğunu tespit edebilmek için, önceden hazır olacak şekilde etiketlenmiş film yorumlarına ihtiyaç duyulmuştur. Bunun sağlanabilmesi için, büyük bir Türkçe yorum arşivi araştırma ihtiyacı görülmüştür. Var olan verilerin doğru sınıflanabilmesi için gerekli olan veri seti "BeyazPerde.com" [31] adresinde bulunmuştur. Web sitesi üzerinde konu başlığı içerisinde filmleri izlemiş olduğu varsayılan kişiler tarafından toplam 5.662 filme ait 219.198 adet yorum bulunmaktadır. Kullanıcılar, kendi yazmış oldukları yorumları 0,5 ile 5,0 aralığında puanlar vererek etiketlemişlerdir.



Şekil 3.1 Elde Edilen Bütün Yorumların Kategorilere Göre Yayılım Tablosu

Şekil 3.1'de gösterilen tabloda '0,5' ile '5' aralığında değişen yorum puanlarının her birine ait olan yorum sayısı gösterilmektedir. Kahverengi renk ile belirtilen kısımda eğitim ve test verisi olarak rastgele seçilmiş yorumların hangi kategoriden kaç adet alındığı gösterilmektedir. Rastgele alınan bu veriler tamamen veri tabanı yazılımı tarafından herhangi bir koşula bakılmaksızın seçilmiştir. Kırmızı renk ile belirtilen kısımda ise üç kategoriye ayrı ayrı ait olmak üzere eğitim ve test veri seti olarak kullanılabilecek toplam yorum sayısı gösterilmektedir.

Veri seti oluşturulması çalışmasında '0,5' – '2,0' aralığı olumsuz yorumlar, '2,5' ile '3,5' aralığı duygu içermeyen (Nötr), '4,0' ve '5,0' aralığında puan alanlar ise olumlu yorumlar olarak kabul edilmiştir. Bu yorumların en kısası '1' kelimeden, en uzunu ise '587' kelimeden oluşmaktadır. Olumlu, olumsuz ve nötr duygulara sahip yorumların sayıları değişkenlik göstermektedir. Sitedeki en düşük yorum sayısına sahip olan olumsuz yorumların sayısı olan '26.700' adet yorum temel alınarak, diğer türdeki yorumlardan da aynı sayıda yorum rastgele puanlamalara bakılmaksızın seçilip bir yorum havuzu oluşturulmuştur.

Sonuç olarak hazırlanan veri setinde üç kategorinin her birinde eşit sayıda olmak üzere toplam '80.100' adet yorum kullanılmıştır. Bu veri setinin her bir kategoriden eşit sayıda olmak üzere yarısı test için diğer yarısı da eğitim için ayrılmıştır. Kullandığımız filtreleme yazılımı tarafından eğitim seti için kullanılan '40.050' adet yorumdan, sadece '44' adedinde hiç bir kelime kökü tespit edilememiştir. Yorum başına düşen ortalama kelime sayısı ise '26' kelimedir.

3.2. Örümcek Yazılım ve Filtreleme Yazılımı Yapısı

3.2.1. Örümcek Yazılım (Webcrawler)

İnternet üzerinde var olan web sitelerinin geneli metin tabanlı olan HTML (*Hyper Text Markup Language*) dilini kullanmaktadırlar. Html'nin metin tabanlı olmasından dolayı, buradaki verilerin '*Webcrawler*' olarak bilinen ve Türkçesi örümcek yazılım olan, web sitelerin içerisindeki istenilen verilere ulaşmayı sağlayan uygulamalar ve programlama dili kütüphaneleri üretilmiştir. Bu tez

çalışmasında da örümcek yazılım kütüphanesi kullanılarak, hazır kullanıcılar tarafından puanlanmış yorumlar site üzerinden alınıp, önceden oluşturulan veri tabanı yapısı içerisine yorumun aslı, puanı ve filtre edilmiş hali olarak kaydedildi. Bu tezde örümcek yazılım olarak C# ile yazılmış "HTMLAgilitypack" [32] kütüphanesi kullanıldı. Bu sayede yorumların hepsinin veri tabanına aktarılması otomatik bir şekilde gerçekleştirildi.

3.2.2. Zemberek

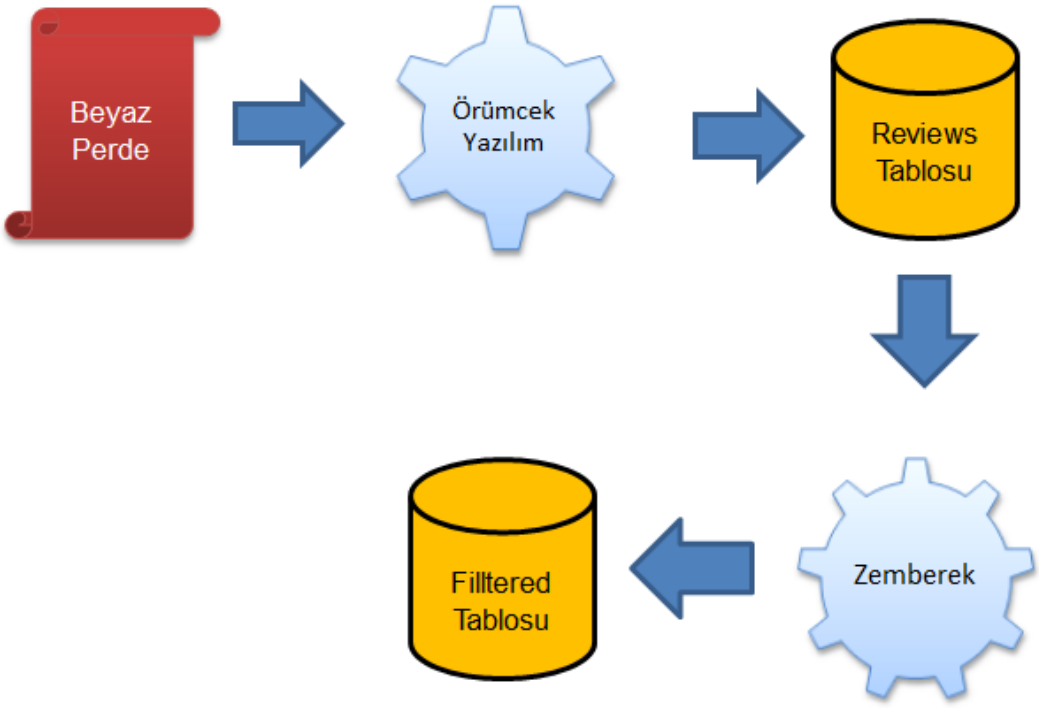
Zemberek "TÜBİTAK BILGEM" yardımlarıyla gerçekleştirilmiş olan 2005 yılında "LKD 4. Linux ve Özgür Yazılım" şenliğinde yılın en iyi özgür yazılımı ödülünü kazanmış olan ilk ve tek açık kaynak kodlu Türkçe doğal dil işleme kütüphanesidir. Zemberek'in ismi "OpenOffice" adlı açık kaynak kodlu uygulamanın içerisindeki Türkçe dil hataları kontrol eden kütüphaneden bilinmektedir. Aynı zamanda Türkçe Linux işletim sistemi olarak bilinen ve yine TÜBİTAK tarafından geliştirilen "Pardus"'un yazım denetleme işini de gerçekleştiren kütüphanedir.

Zemberek doğal dil işleme (NLP) alanında sıkça kullanılan ve dillerdeki kelimeleri içerisinde barındıran '*lemmatizer*' olarak da isimlendirilen uygulamalardan bir tanesidir. Lemmatizer uygulamaları çoğu dil için mevcuttur. Genel olarak içlerinde kullanılacağı dile ait kelimeleri ve bu kelimelerin köklerini barındırır. Lemmatizer'ın gelişmişliğine göre içlerinde ayrıca köklerin kelime türünü barındırabilmekte ve kullanıcı tarafından yanlış yazılan kelimelerdeki harf hatalarını düzeltebilmektedir.. Bu tür uygulamaların amacı analiz yapılacak olan metin verilerinin en doğru bir biçimde değerlendirilmesini sağlayabilmektir.

Lemmatizer uygulamalarına alternatif olarak '*N-gram*' adı verilen yöntem gösterilebilir. Bu yöntem isminden de tahmin edilebileceği üzere kökün sahip olacağı ilk 'n' adet harfin kök olarak kabul edileceğini göstermektedir. Bu yöntem lemmatizer yöntemine göre hızlı çalışmakta fakat uzun köke sahip kelimeler için yapılacak filtreleme işlemlerinde başarısı çok düşük olacaktır. Daha ayrıntılı bir süzgeç görevi yapması için lemmatizer uygulamaları tercih sebebi olmaktadır. Bu tezde filtre işlemi, Türkçe için geliştirilmiş olan Zemberek [33] isimli açık kaynak

kodlu kütüphane kullanılmıştır. Açık kaynak kodlu olduğu için çeşitli yazılım kütüphanelerinde bu uygulamaya erişmek mümkündür.

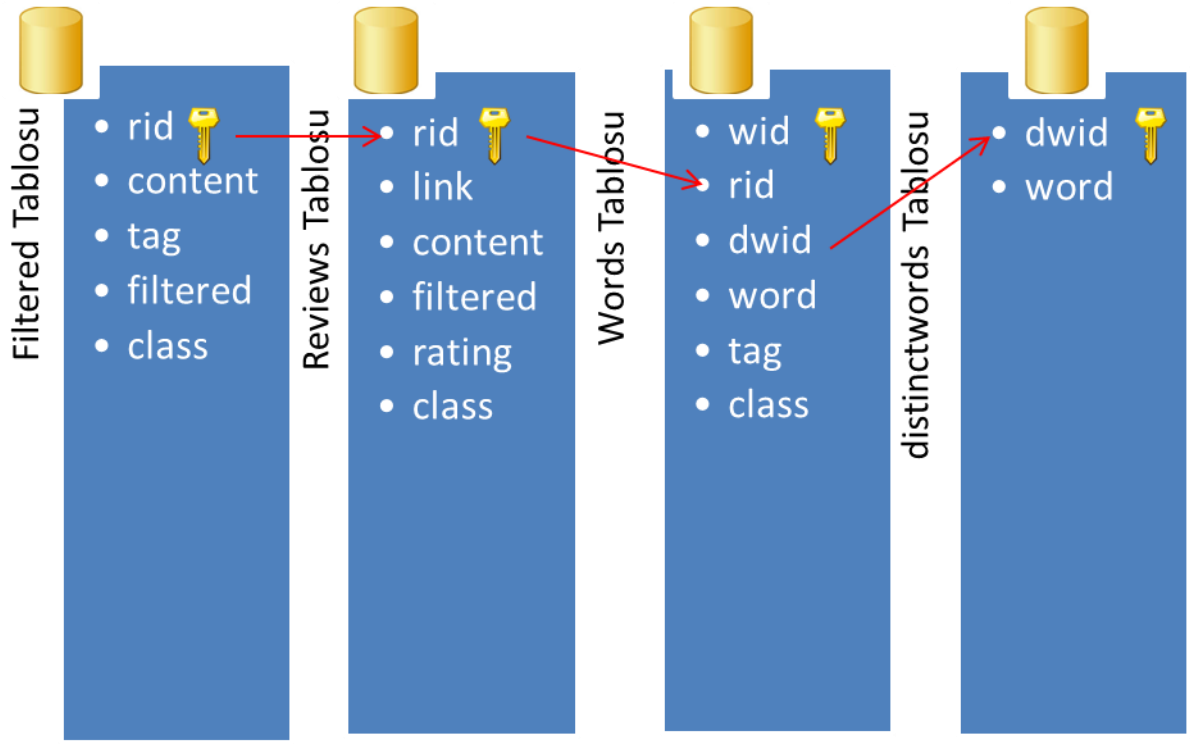
Tez çalışmasının başında veri seti üzerinde uygulanan *N-gram*'ın kelime kökü ayrıştırma esnasındaki başarısının düşük olduğu gözlenince, Zemberek uygulamasının bu tezde kullanılması kararlaştırıldı. Nitekim Zemberek'in kelimeleri köklerine ayrıştırma başarısı diğer yönteme göre daha başarılıydı. Kelimedeki sadece bir harf hatasını tolere edebilip ve onu düzeltebilmesi, bunun yanı sıra sınıflamanın kolay yapılabilmesi için gerekli olan kelime kökünü sunabilmesi, kökü bulamadığındaysa o kelimeye kendi sözlüğünde en benzer olan kökü bulup karşımıza getirmesi Zemberek'in ne kadar gelişmiş bir lemmatizer olduğunu kanıtlamaktadır. Burada anlatılan Zemberek kütüphanesine ait tüm bu özellikler bu tez içerisinde aktif olarak kullanılmıştır. Zemberekte kök bulma işleminden geçirilmiş olan cümlelere ait olan kelimeler "Filtered" adlı tabloda veri tabanına kaydedildi. İçerisinde kelime olmayan anlamsız ve sadece noktalama işaretlerinin bulunduğu yorumlar bu sayede filtre edilerek elendiler.



Şekil 3.2 Örümcek Yazılım, Zemberek ve Veri Tabanı İlişkisi

Özet olarak filtreleme işlemi için oluşturulan tasarımda hedef web sitesi olan ‘Beyazperde.com’ adresindeki film yorumları örümcek yazılım ile veri tabanına aktarılır. Daha sonra yapılan işlemlerde filmlere ait yorumlar eğitim ve test kümesi olarak kullanılabilmesi için her kategoriden eşit yorumlar seçilerek ‘Reviews’ tablosunda eğitim veya test sınıfı belirtilir şekilde depolanır. Daha sonradan Zemberek kütüphanesi kullanılarak seçilen yorumlar köklerine ayrıştırılır ve Şekil 3.2’de gösterildiği üzere ‘Filtered’ tablosunda sadece cümlenin kelime köklerini barındıran, yani herhangi bir çekim eki almamış hali depolanır.

3.2.3. Veri tabanı Tasarımı



Şekil 3.3 Tezde Kullanılan Veri Şeması

Bütün tablo ve view yapıları Ms-SQL Server 2012 Enterprise üzerinde oluşturulmuştur. Bu kapsamda oluşturulan veri tabloları ve her birinin amacı takip eden kısımlar kısaca açıklanmıştır. Tabloların birbirleriyle olan ilişkileri ve bağlantı nokraları Şekil 3.3'te gösterilmiştir. Kısaca veri tabanı dört adet birbiriyle ilişkili tablolardan oluşmaktadır. 'Filtered' tablosu 'Reviews' tablosundaki yorumlardaki kelimelerin filtreleme yazılımından geçirildikten sonraki aktarıldığı tablodur. Bu filtreleme işleminden sonra 'Words' tablosunda her bir yoruma ait cümlelerdeki kelimeler tutulmaktadır. Kelimeleri vektör uzay matrisinde temsil edebilmek için birbirinden farklı her bir kelimeye farklı bir numara atanmıştır. Bu numaralar da 'distinctwords' tablosunda kayıt altına alınmıştır.

Reviews Tablosu: Örümcek yazılım ile alınan film yorumlarının ait olduğu bağlantı adresini, içeriğinin, puanının, hangi kategoriye ait olduğunun ve Zemberek işleminden geçip köklerden oluşan cümle halini barındıran kısımların tutulduğu tablodur.

rid: Kelimenin ait olduğu yorum numarasıdır. Aynı zamanda eşsiz anahtar numarasıdır.
link: Yorumun bulunduğu bağlantı adresi depolanır.
content: Yorumun herhangi bir filtreleme işleminden geçmemiş yani olduğu gibi saklandığı bölümdür.
filtered: Yorumun Zemberek'te filtrelendikten sonraki köklerden oluşan cümle halidir.
rating: Yorumcunun filmi yorumladıktan sonra vermiş olduğu 0.5 ile 5 aralığındaki sayısal puan değeridir.
class: Test ve eğitim için ayrılmış olan yorumları gösteren sütun.

Words Tablosu: Zemberek tarafından filtrelenmiş olan 80.100 adet film yorumunun içerisindeki kelimelerin bulunduğu tablodur. Bu tabloda aşağıdaki alanlar bulunmaktadır.

wid: Her bir kelime için atanmış olan, biricik anahtar (*Primary Key*) numarasıdır, aynı numaralı iki kelime bu tabloda bulunamaz.
rid: Kelimenin ait olduğu yorum numarasıdır.

dwid: Her bir kelimeye verilmiş olan eşsiz anahtar numarasıdır. distinctword tablosu ile ilişkilidir. Farklı yorumlarda yer alabilir.

word: Kelimenin kendisinin tutulduğu sütun.

tag: Yorumun ait olduğu kategorinin (Olumlu, Olumsuz ve Nötr) tutulduğu sütun.

class: Test ve eğitim için ayrılmış olan yorumları gösteren sütun.

Filtered Tablosu: Depolanan yorumların orijinallerinin ve filtre yazılımı tarafından köklerine ayrılarak içerisinde tutulmuş olduğu tablodur. Bu tabloda aşağıdaki alanlar bulunmaktadır.

rid: Her bir yorum için atanmış olan eşsiz anahtar numarasıdır. Tabloda var olan yorumların biricik anahtarının "reviews" tablosundaki "rid" numaralarıyla aynı değere sahiptir.

content: Web sitesinden alınmış olan yorumların asıl hallerinin bulunduğu sütundur.

tag: Yorumun ait olduğu kategorinin (Olumlu, Olumsuz ve Nötr) tutulduğu sütundur.

filtered: Filtreleme uygulaması tarafından, her bir yorumda bulunan cümlelere ait olan kelimelerin sadece kök olarak tutulduğu sütundur.

class: Test ve eğitim için ayrılmış olan yorumları gösteren sütundur.

distinctwords Tablosu: Yorumlarda kullanılmış olan bütün kelime köklerinin bulunduğu tablodur. Bu tabloda aşağıdaki alanlar bulunmaktadır.

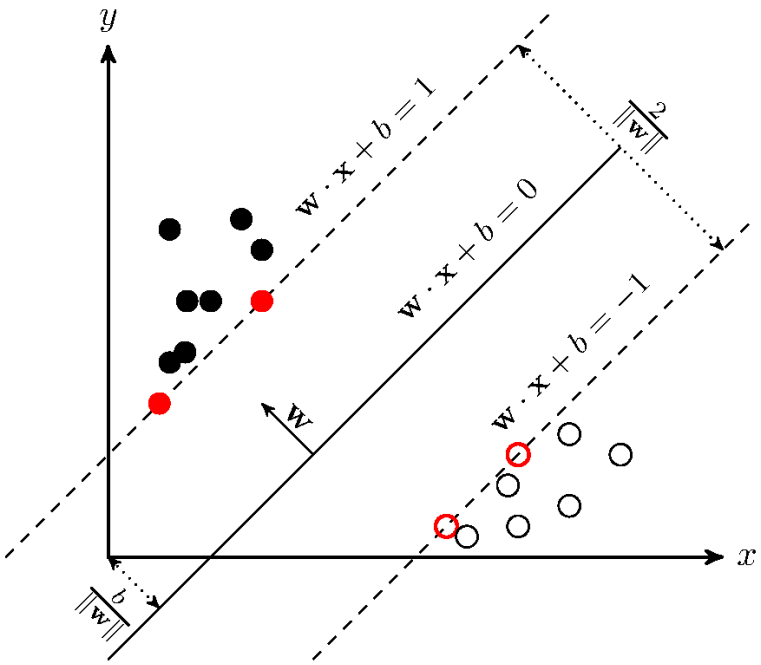
dwid: Kelime için atanmış eşsiz anahtar numarasının bulunduğu sütundur. Yorumlarda geçmiş her bir kelimenin sadece bir adet dwid numarası vardır.

word: Yorumlarda kullanılmış olan kelimenin filtrelenmiş olarak bulunduğu sütundur.

4. Uygulanan Yöntemler ve Sonuçlar

4.1. Makine Öğrenimi Yöntemleri ve SVM

Makine öğrenimi yöntemleri insanlar üzerinde var olan iş yükünü neredeyse tamamen kaldırmayı hedefleyen sistemleri oluşturmak için kullanılmaktadırlar. Bu yöntemler belirli algoritmalar ve hesaplamaları bilgisayarların işlem gücünü kullanarak, verilmiş olan problemlerin çözümüne ulaşmaya çalışırlar. Makine öğrenimi yöntemleri Gözetimli (Supervised) ve Gözetimsiz (Unsupervised) yöntemler olarak ayrılmaktadır. Bu tezde var olan çalışmanın başarısını tespit edebilmek amacıyla, gözetimli yöntemlerden biri olan ve karmaşık verileri sınıflama problemleri üzerindeki başarısı bilinen, doğrusal olarak ayrılabilen veya ayrılamayan veriler üzerinde çalışabilen ve kolay uygulanabilirliği yönüyle veri madenciliği alanında sık sık kullanılan, Türkçesi destek vektör makinesi olan SVM (*Support Vector Machine*) [16] yöntemi kullanılmıştır.



Şekil 4.1 Hiper Düzlem Oluşturma İşlemi

SVM, hiper düzlemleri birbirinden düzgün olarak ayıran bir sınıflayıcı olarak tanımlanmaktadır. Başka bir deyişle, SVM eğitim kümesi olarak verilmiş etiketli veriler üzerinde yapmış olduğu model hesaplama denklemlerine göre doğrusal hiper düzlemler oluşturur (Şekil 4.1 [34]). Doğrusal hiper düzlem oluşmadığı takdirde ise verileri orijinal hiper düzlemin üzerinde oluşturmuş olduğu yeni uzaya aktarır. Böylelikle yeni gelen örnek verileri uygun olan kategorilere yerleştirir. SVM İngilizce için neredeyse iki ayrı kategoride %80 ve üç ayrı kategorideyse %60 doğruluk oranıyla çalışabilmektedir [17]. Deneylerde kullanılan SVM içerisinde birçok kernel ve yöntemler mevcuttur. Bu tezde önceden yapılmış olan aynı alanda var olan çalışmalar gibi; Weka [35] adlı araç üzerinde bulunan LibSVM [36] uygulaması üzerinde, "C-SVC" kernel'i ve "Lineer ($u.v'$)" "5 Fold" seçeneği kullanılmıştır. Burada belirtilen "Fold" sayısı eğitim kümesi için oluşturulacak olan modeli, test kümesi içindeki verilerin kaç parça halinde bu model üstünde test edileceğini belirlemek için kullanılır. Yani "5 Fold" için örnek verilirse, 20.000 adet test verisi rastgele olacak şekilde 4.000 adetlik beş parçaya bölünüp, model üstünde test edilecek ve akabinde oluşan beş adet test sonucunun aritmetik ortalaması temel alınarak sonuç tayin edilecektir.

Tez çalışmasının ilerideki test çalışmalarında, yine Weka içerisinde var olup ve istatiksel olarak hesaplama yaparak çalışan ML yöntemlerinden biri olan NB yönteminin sınıflama başarısı da önerilen yöntem üzerinde kısmen değerlendirilmiştir.

4.2. Öznitelik Seçme Metrikleri

Öznitelik seçme (*Feature Selection*) metrikleri, metin madenciliğinde kullanılan ve kategoriler halinde yapılan sınıflamalardaki en başlıca kullanılan aktivitelerdir. Öznitelik seçme metrikleri, her-bir kategorideki en önemli kelimeleri büyük bir hassaslık oranıyla ayrıştırabilmekte ve metin madenciliği alanında çok yaygın kullanılmaktadır. Böylelikle metin içerisinde geçen en ayrıştırıcı kelimeler, belirli hesaplamalar sonucunda tespit edilmektedir. *Information Gain* (IG), *Chi-Square*,

Odd Ratio gibi yöntemler öznitelik seçme metrikleri arasında en yaygın olarak kullanılan yöntemlerdir.

$$IG = \frac{a}{N} * \log \frac{a * N}{(a+c) * (a+b)} + \frac{b}{N} * \log \frac{b * N}{(b+d) * (a+b)} + \frac{c}{N} * \log \frac{c * N}{(a+c) * (c+d)} + \frac{d}{N} * \log \frac{d * N}{(b+d) * (c+d)}$$

Eşitlik 4.1 Information Gain Metriği Denklemi

$$Chi\ Square = N * \frac{(a * d - b * c)^2}{(a+c) + (b+d) + (a+b) + (c+d)}$$

Eşitlik 4.2 Chi-Square Metriği Denklemi

Bu tezde, yaygın olarak kullanılan metriklerden IG ve Chi-Square, düşük hesaplama maliyetleri ve kolay uygulanabilir olmalarından dolayı kullanılmıştır. Tezde kullanılan metriklerin denklemleri Eşitlik 4.1 ve Eşitlik 4.2 de yer almaktadır.

Eşitlik 4.1 ve Eşitlik 4.2'de gösterildiği üzere; N toplam yorum sayısı, a olumlu yorumlar kategorisi içerisinde o terimi barındıran doküman sayısı, b olumlu yorumlar kategorisi içerisinde o terimi barındırmayan doküman sayısı, c olumsuz yorumlar kategorisi içerisinde o terimi barındıran doküman sayısı, d olumsuz yorumlar kategorisi içerisinde o terimi barındırmayan doküman sayısını göstermektedir.

Üçlü sınıflama için hesaplanacak yeni IG ve Chi-Square sonuçları için, her bir kategori için ayrı ayrı a, b, c ve d değerleri hesaplandı. Yani her bir kategoride var olan a, b, c ve d değişkenleri olumlu, olumsuz ve nötr kategorisi ayrı ayrı olmak üzere hesaplandı. Sonra bu kelimelerin her kategoride oluşan puanları toplanarak, en ayrıştırıcı kelimeler bu yolla hesaplanmış oldu. Üçlü kategorilerde yapılan sınıflamalar için kullanılan denklemler Eşitlik 4.3 ve Eşitlik 4.4 de yer almaktadır.

$$Chi\ Square(Total) = Chi\ Square(Negative) + Chi\ Square(Positive) + Chi\ Square(Neutral)$$

Eşitlik 4.3 Üçlü sınıflama için kullanılan Chi-Square Metriği Denklemi

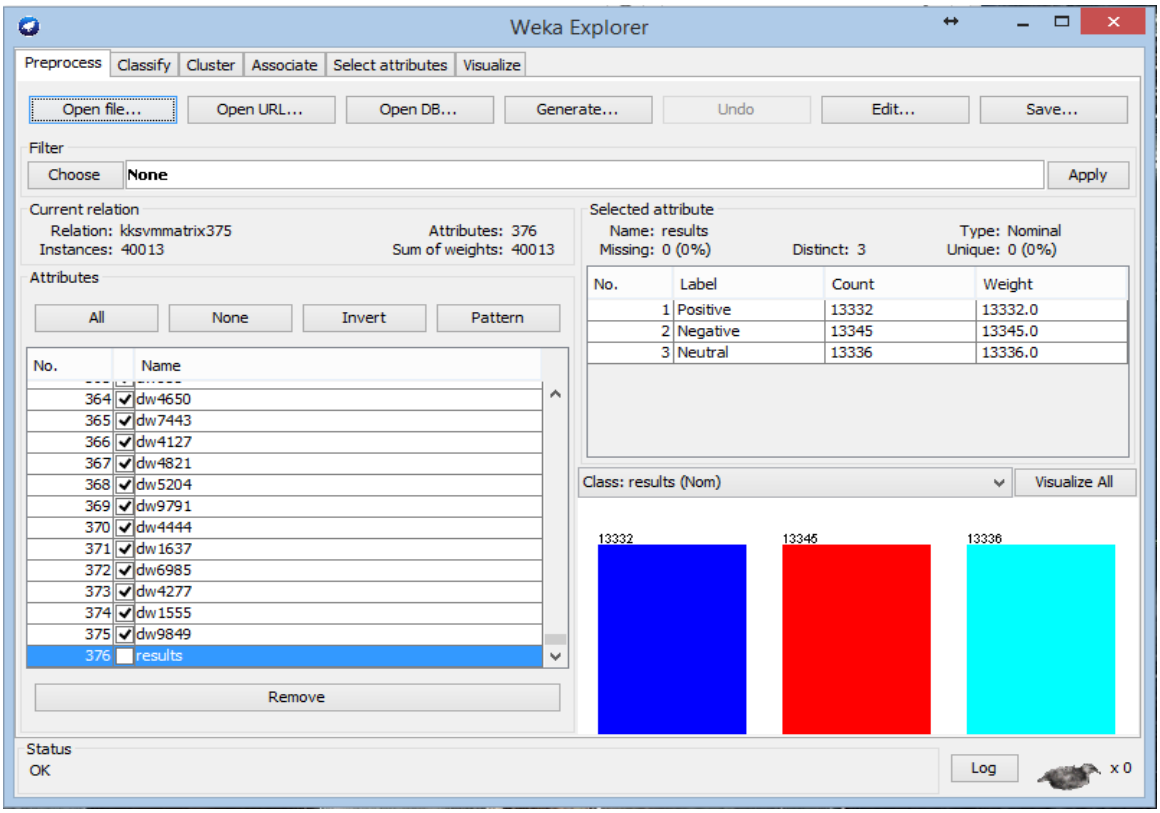
$$\mathrm{IG(Total)} = \mathrm{IG(Negative)} + \mathrm{IG(Positive)} + \mathrm{IG(Neutral)}$$

Eşitlik 4.4 Üçlü sınıflama için kullanılan Information Gain Metriği Denklemleri

4.3. Weka

Veri madenciliği dalında yapılan çalışmaların ve yöntemlerin test ve analizlerinde Waikato Üniversitesince oluşturulan "Weka Tools" ismi sıkça geçmektedir. Weka ML yöntemlerini en basite indirgeyerek test ve analiz yapmak isteyen kullanıcılarca en yaygın kullanılan araçlardan biridir. Bunun nedeni ise mevcutta var olan veri madenciliğinin ön plana çıkmış yöntemlerinin hepsinin test ve analiz araçlarını kendi içerisinde barındırmasıdır. Ayrıca JAVA tabanlı bir uygulama olması nedeniyle kullanım alanı çok geniştir. Weka içerisinde bulunan sayıca yüz adet civarı fonksiyon aracılığıyla istenilen ML yöntemini uygulamak ve bu yöntemleri birbirileriyle kıyaslamak çok vakit almamaktadır. Bu yazılım VSM haline dönüştürülmüş eğitim veya test verilerini çok kolay bir şekilde istenilen yöntem ile analizini gerçekleştirmektedir.

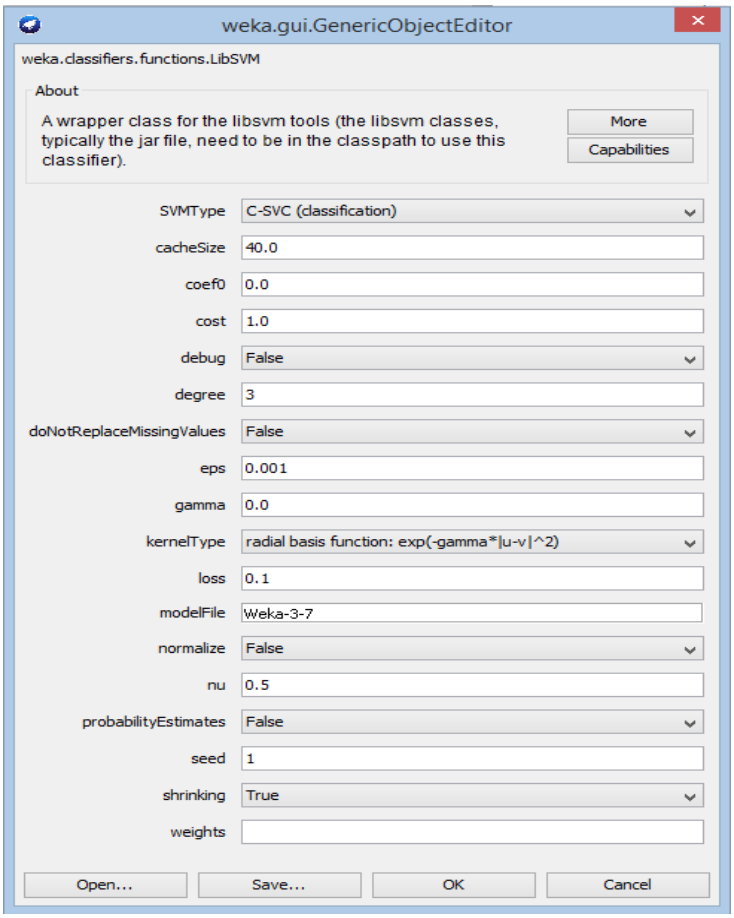
Weka'nın içerisindeki yöntemlerin hepsi akademik olarak alanlarının uzmanları tarafından hazırlanmıştır. Böylelikle sonuçların yanlış hesaplanması gibi testi bireysel hataların oluşmasının önüne geçilmiştir. Ayrıca test edilecek liste halindeki verilerin "MS Excel" dosyasından kendi uyumlu olduğu format olan ".arff" uzantısına dönüşümünü kendi gerçekleştirmektedir. Bu şekilde analize uygun hale gelen verilerin girişi uygulama üzerinden gerçekleştirilir. Burada değişkenler (*Attributes*) ve sınıf (*Class*) seçilir. Bu işlem gerçekleştirildikten sonra analiz yapılacak sınıflara ait veri sayılarını grafiksel olarak kullanıcıya aktarır (Şekil 4.2).



Şekil 4.2 Weka Veri Giriş Arayüzü

Hemen hemen her yapay zekâ modelinin kendine özgün hassaslık ayarlamaları vardır. Weka bu ayarlamaların hepsini kullanıcıya basit bir ara yüz aracılığıyla ulaşmalarını sağlamaktadır. Şekil 4.3'de "LIBSVM" yönteminde kullanılan SVM türlerinden biri olan 'C-SVC' matematiksel fonksiyonuna ait ince ayarlar gösterilmiştir. Weka'da bunun gibi dört farklı matematiksel fonksiyon daha mevcuttur. Bu farklı matematiksel fonksiyonlardan çıkacak analiz sonuçlarının başarıları değişkenlik gösterebilmektedir. En uygun olan fonksiyonun kararını analizi yapacak olan kişiler vermektedir.

Daha detaylı olarak Weka sınıflayıcısının araçlarının gerçekleştirebildiği işlemler ve onlara ait olan araçlar Çizelge 2'de liste şeklinde yer almaktadır.



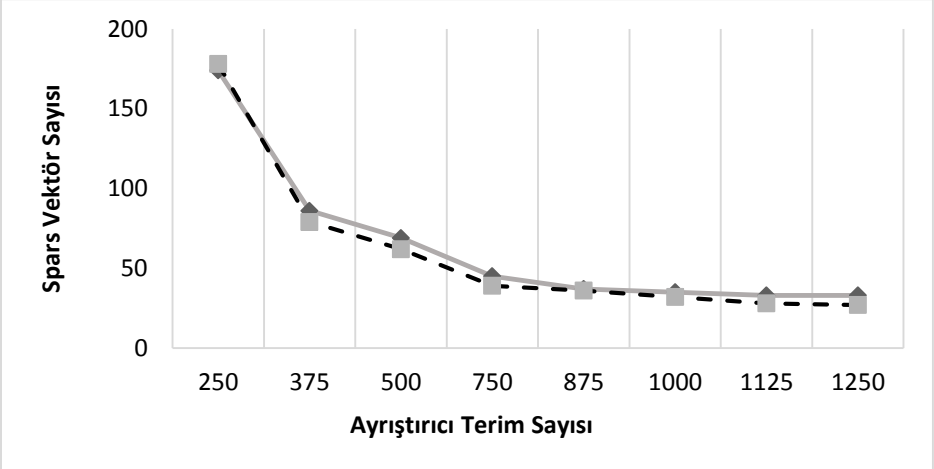
Şekil 4.3 Weka Yöntem Hassaslık Ayarları Arayüzü

4.4. SVM ve Öznitelik Seçme Metriklerinin Uygulanması

Oluşturulan eğitim veri seti içerisindeki köklerine ayrılmış olarak depolanan yorumların hepsi için Chi-Square ve IG denklemleri uygulandı. Böylelikle yorumların içerisindeki kelimelerin ne kadar olumlu, olumsuz ve nötr puanlarına sahip oldukları gözlenebildi. Öznitelik seçme işlemi uygulanan her bir kelimenin kendi kategori (Olumlu, Olumsuz, Nötr) puanlarından en yüksek olduğu kategori puanı, o kelimenin kendi kategorisini belirlemiş oldu. Daha sonrasında her bir yöntem için oluşan, olumlu, olumsuz ve nötr kategorilerine alınan kelimeler puanlama sırasına göre ayrıştırıcılıkları tespit edildi. Daha sonrasında en değerliden değersize sıralanacak şekilde bir liste halinde veri tabanına aktarıldı. Puanları en yüksek olan ilk 250, 375, 500, 750, 875, 1.000, 1.125 ve 1.250 değerli

kelime için, dwid (Ayrık terim ID) sırasına göre eğitim ve test setlerinde yer alan yorumlar, vektörler şeklinde veri tabanında view yapısı içerisinde oluşturuldu. Yani bir yorumda ayrıştırıcı olan o kelime var ise ‘1’ yok ise ‘0’ olacak şekilde her bir yorum için ayrı ayrı olmak üzere, seçilen en etkili kelime sayıları kadar uzunluğunda olacak şekilde yorum matrisleri oluşturuldu. Bu matrislere VSM yani Vektör Uzay Modeli adı verilir.

Bu aşamada bazı yorumların içeriğinin hepsi ‘0’ olduğundan, ayrıştırıcı terimlerle ifade edilemediler. Bu tür vektörlere “Seyrek” (Sparse) vektörler denmektedir ve sonuçların başarısına herhangi bir katkı sağlamamaktadırlar. Seyrek vektörlerin sayısının, ayrıştırıcı terim sayısına göre değişimi Şekil 4.4’de gösterilmiştir. Kesikli çizgi le ifade edilen değişim test kümesindeki seyrek vektör sayısını, düz çizgi le ifade edilen değişim eğitim kümesindeki seyrek vektör sayısını temsil etmektedir. Görüldüğü üzere ayrık terimlerin sayısı arttıkça, seyrek vektör oluşumu iki kümede de belirgin bir şekilde azalmıştır.



Şekil 4.4 Seyrek Vektör Sayısının Ayrıştırıcı Terimlerin Sayısına Göre Değişimi

4.5. Deney ve Sonuçlar

Bu aşamada oluşturulan ve içerisinde sadece olumlu ve olumsuz kategorilerindeki kelimeleri ve yorumları temsil eden yorum vektörlerinin ne kadar sağlıklı

ayrıştırılabildiğini ve elimizdeki veri tabanının içerisindeki veri kirliliğini ne kadar engelleyebildiğimizi ölçebilmek için sadece olumlu ve olumsuz kategorilerde var olan eğitim veri seti için ayrılmış olan yorumları, test için ayrılmış olan yorum vektörleri ile beraber SVM sınıflayıcı üzerinde test edildi.. Her bir ayrıştırıcı terim sayısı için ayrı ayrı sonuçlar alınmıştır. Bu sonuçlar Tablo 4.1'de gösterilmişlerdir.

	Chi Square				**Information Gain**			
# İlk Terim Sayısı Sıralaması	# Seyrek Vektör Sayısı	Olumlu F_1	Olumsuz F_1	Ortalama F_1	# Seyrek Vektör Sayısı	Olumlu F_1	Olumsuz F_1	Ortalama F_1
250	159	0.833	0.844	0.838	174	0.833	0.844	0.838
375	85	**0.834**	**0.845**	**0.839**	86	**0.834**	**0.845**	**0.839**
500	66	0.833	0.843	0.838	69	0.833	0.843	0.838
750	40	0.832	0.841	0.837	45	0.832	0.841	0.836
875	36	0.831	0.840	0.836	37	0.831	0.840	0.835
1000	34	0.829	0.838	0.834	35	0.829	0.838	0.833
1125	33	0.827	0.835	0.831	**33**	0.827	0.836	0.831
1250	**32**	0.825	0.834	0.830	33	0.825	0.834	0.830

Tablo 4.1 Seyrek vektör sayısı ve Olumlu, Olumsuz Kategorilerdeki SVM Sonuçları

Tablo 4.1'deki elde edilen tabloda, en iyi başarının ilk 375 ayrık terimin kullanılması ile alındığı tespit edilmiştir. 750 ve ayrık terim sayısının daha da artması ile beraber seyrek vektör sayısının düştüğü fakat bunun çok az bir başarı düşüşüne neden olduğu gözlemlenmiştir. Bunun nedeni olarak da ayrık terim sayısının artmasının, bir noktadan sonra sonuçların başarısını bilgi kirliliğinden dolayı düşürmesi olarak yorumlanması gösterilebilir.

Bu aşamadan sonra deneyler tablodaki sonuçlara istinaden iki aşamada devam edilmiştir. İlk aşamada Tablo 4.1'deki en yüksek F_1 skor değerine sahip olan ilk 375 ayrıştırıcı kelime sayısı ve ikinci aşama olarak, seyrek vektör sayısı bakımından az ve başarı bakımındansa diğerlerinden daha iyi olan 1.000 ayrıştırıcı kelime sayısı üzerine odaklanıldı.

Sonraki adımda ise üçlü sınıflama için öznitelik seçme metrikleri üç kategori için uygulanmıştır. Bunun için Eşitlik 4.1 ve Eşitlik 4.2'de yer alan a,b,c,d değerleri

a_negative, a_positive, a_neutral, b_negative, b_positive, b_neutral, c_negative, c_positive, c_neutral olarak yeniden hesaplanmıştır (Çizelge 1).

Böylelikle üç kategori için en ayrıştırıcı kelimeler puanlama önceliğine göre yukarıdan aşağıya sıralanmıştır. Önceki iki kategoride yapılmış olan sınıflamadan alınan referans noktalarına göre *Chi-Square* ve *IG* için oluşan en ayrıştırıcı ilk 375 ve ilk 1000 terim için SVM sınıflama işlemi tekrar gerçekleştirilmiştir.

Tablo 4.2'de görülen üç kategori için yapılmış deneyde Chi-Square 375 terim için elde edilmiş olan sonuçlara göre, temel başarı ölçütümüz olan F değeri olumsuz yorumların sınıflandırılması esnasında en yüksek değer %69'a ulaşmış.

Tüm Eğitim Seti	**TP Oranı**	**FP Oranı**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Alanı**	**PRC Alanı**	**Sınıfı**
	0.745	0.212	0.637	0.745	0.687	0.515	0.766	0.560	Olumlu
	0.675	0.138	0.711	0.675	0.692	0.545	0.769	0.588	Olumsuz
	0.479	0.201	0.543	0.479	0.509	0.288	0.639	0.434	Nötr
Ağırlıklı Ort.	0.633	0.184	0.184	0.630	0.633	0.630	0.449	0.725	
Tüm Test Seti	**TP Oranı**	**FP Oranı**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Alanı**	**PRC Alanı**	**Sınıfı**
	0.610	0.138	0.689	0.610	0.647	0.488	0.736	0.550	Olumlu
	0.767	0.195	0.664	0.767	0.712	0.555	0.786	0.587	Olumsuz
	0.524	0.217	0.547	0.524	0.535	0.310	0.653	0.445	Nötr
Ağırlıklı Ort.	0.634	0.183	0.633	0.634	0.631	0.451	0.725	0.527	

Tablo 4.2 Chi-Square Yöntemi İçin En Ayrıştırıcı 375 Adet Terime Ait Ayrıntılı Test Sonuçları

Tüm Eğitim Seti	**TP Oranı**	**FP Oranı**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Alanı**	**PRC Alanı**	**Sınıfı**
	0.744	0.211	0.638	0.744	0.687	0.515	0.766	0.560	Olumlu
	0.677	0.137	0.712	0.677	0.694	0.547	0.770	0.589	Olumsuz
	0.479	0.201	0.544	0.479	0.510	0.289	0.639	0.434	Nötr
Ağırlıklı Ort.	0.634	0.183	0.631	0.634	0.630	0.450	0.725	0.528	
Tüm Test Seti	**TP Oranı**	**TP Oranı**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Alanı**	**PRC Alanı**	**Sınıfı**
	0.609	0.137	0.690	0.607	0.647	0.489	0.736	0.551	Olumlu
	0.766	0.194	0.640	0.766	0.711	0.554	0.786	0.587	Olumsuz
	0.525	0.218	0.546	0.525	0.535	0.310	0.653	0.445	Nötr
Ağırlıklı Ort.	0.634	0.183	0.633	0.634	0.631	0.451	0.725	0.527	

Tablo 4.3 Information Gain Yöntemi İçin En Ayrıştırıcı 375 Adet Terime Ait Ayrıntılı Test Sonuçları

Tüm Eğitim Seti	TP Oranı	FP Oranı	Precision	Recall	F-Measure	MCC	ROC Alanı	PRC Alanı	Sınıfı
	0.751	0.227	0.623	0.751	0.681	0.503	0.762	0.551	Olumlu
	0.659	0.134	0.711	0.659	0.684	0.535	0.762	0.582	Olumsuz
	0.470	0.198	0.542	0.470	0.503	0.282	0.636	0.431	Nötr
Ağırlıklı Ort.	0.627	0.187	0.625	0.627	0.623	0.440	0.720	0.521	
Tüm Test Seti	**TP Oranı**	**FP Oranı**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Alanı**	**PRC Alanı**	**Sınıfı**
	0.619	0.149	0.675	0.619	0.646	0.481	0.735	0.545	Olumlu
	0.750	0.189	0.665	0.750	0.705	0.546	0.781	0.582	Olumsuz
	0.526	0.215	0.550	0.526	0.538	0.314	0.655	0.447	Nötr
Ağırlıklı Ort.	0.631	0.184	0.630	0.631	0.629	0.447	0.724	0.525	

Tablo 4.4 Chi-Square Yöntemi İçin En Ayrıştırıcı 1000 Adet Terime Ait Ayrıntılı Test Sonuçları

İlk 375 terim için I.G. yönteminin en yüksek başarı değeri olan %71 F-değerinin tekrardan olumsuz kategorisine ait olduğu Tablo 4.3'te görülmektedir. 1.000 adet terim değerlendirmeye katılarak yapılmış olan deneylerde çıkan sonuçları gösteren Tablo 4.4 ve Tablo 4.5'te yine olumsuz kategoriler %70 başarı değerleriyle en iyi sınıflanan yorumlar olmuşlardır. Sonuçlardan anlaşılacağı üzere bu yöntemde üç kategori için yapılacak analizlerde en başarılı olarak olumsuz kategorisindeki veriler sınıflanmıştır.

Tüm Eğitim Seti	TP Oranı	FP Oranı	Precision	Recall	F-Measure	MCC	ROC Alanı	PRC Alanı	Sınıfı
	0.750	0.227	0.622	0.750	0.680	0.503	0.761	0.550	Olumlu
	0.659	0.135	0.710	0.659	0.684	0.535	0.762	0.582	Olumsuz
	0.470	0.199	0.541	0.470	0.503	0.282	0.635	0.431	Nötr
Ağırlıklı Ort.	0.626	0.187	0.625	0.626	0.622	0.440	0.720	0.521	
Tüm Test Seti	**TP Oranı**	**FP Oranı**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Alanı**	**PRC Alanı**	**Sınıfı**
	0.619	0.150	0.674	0.619	0.645	0.480	0.734	0.544	Olumlu
	0.749	0.188	0.665	0.749	0.705	0.546	0.780	0.582	Olumsuz
	0.526	0.215	0.550	0.526	0.537	0.314	0.655	0.447	Nötr
Ağırlıklı Ort.	0.631	0.184	0.630	0.631	0.629	0.446	0.723	0.524	

Tablo 4.5 Information Gain Yöntemi İçin En Ayrıştırıcı 1000 Adet Terime Ait Ayrıntılı Test Sonuçları

SVM	Precision			Recall			F-Measure			
#Terim (C.S.)	Olumlu	Olumsuz	Nötr	Olumlu	Olumsuz	Nötr	Olumlu	Olumsuz	Nötr	Ort.
375	0.702	0.661	0.544	0.605	0.793	0.511	0.65	0.721	0.527	**0.633**
1000	**0.717**	0.652	0.536	0.578	**0.814**	0.506	0.64	**0.724**	0.521	0.628

Tablo 4.6 Üç Kategoride Chi-Square Yöntemi İçin Elde Edilmiş SVM Sınıflayıcı Sonuçları

SVM	Precision			Recall			F-Measure			
#Terim (C.S.)	Olumlu	Olumsuz	Nötr	Olumlu	Olumsuz	Nötr	Olumlu	Olumsuz	Nötr	Ort.
375	0.704	0.66	0.544	0.603	0.794	0.512	0.65	0.721	0.527	**0.633**
1000	**0.717**	0.651	0.536	0.578	**0.815**	0.504	0.64	**0.724**	0.52	0.628

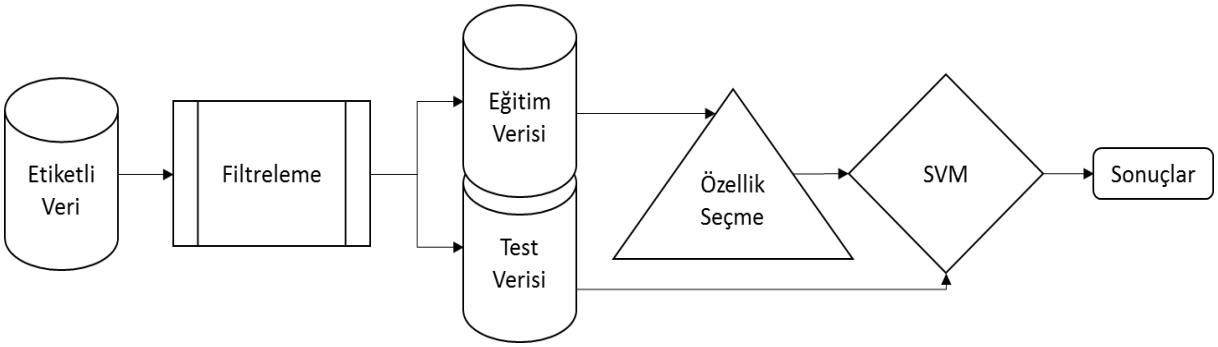
Tablo 4.7 Üç Kategoride IG Yöntemi İçin Elde Edilmiş SVM Sınıflayıcı Sonuçları

Tablo 4.6 ve Tablo 4.7'de gösterilen tablolarda üç kategoride yapılmış olan sınıflamaların terim sayısına göre F değeri değişimi karşılaştırılmıştır. 375 terim ile yapılan sınıflamanın 1.000 adet terimle yapılan sınıflamaya göre çok küçük bir farkla daha başarılı olduğu gözlenmiştir. Ayrıca 1.000 adet terimi değerlendirmeye alarak yapılan bir sınıflamanın daha uzun bir süre alacağı da göz ardı edilmemelidir. Bu nedenle 375 adet terimi sınıflamak daha avantajlı görünmektedir.

Seyrek Matris Sayısı				
	Information Gain		Chi-Square	
#Terim	Test	Eğitim	Test	Eğitim
375	87	105	86	104
1000	27	33	27	32

Tablo 4.8 Üç Kategoride Uygulanan IG ve Chi-Square Yöntemi Sonrasında Oluşan Seyrek Vektör Sayısı

Tablo 4.8'de yöntemlere göre seyrek vektör sayılarının oluşma miktarları karşılaştırılarak gösterilmiştir. Bu tabloda görünen seyrek vektör sayılarının veri seti sayısının yüksek olması nedeniyle oranlarının birbirinden çok farklı olmadığı görülmekte fakat 375 terim kullanılarak oluşturulan vektör uzay modelinde daha fazla seyrek vektör oluştuğu görülmektedir.



Şekil 4.5 Tezde Uygulanan Sistem Tasarımı Şematiği

Bu tezde yapılmış olan tüm deney ve sonuç aşamaları Şekil 4.5'te gösterilmektedir. Özet olarak etiketli veriler örümcek yazılımlar aracılığıyla hedef olarak seçilen web sitesi üzerinden toplanır ve veri tabanına kaydedilir. Sonrasında bu verilerin içerisindeki cümlelere ait kelimeler filtreleme işleminden geçirilerek üzerinde analiz yapılabilecek eğitim ve test verileri oluşturulur. Daha sonrasındaysa makine öğrenimi yöntemine bu verileri direk olarak vermek yerine öznitelik seçme yöntemleri uygulanır ve en uygun öznitelik sayısının tespiti SVM kullanılarak gerçekleştirilir. En son işlem olarak cümleler VSM içerisinde temsil edilebilecek hale getirilerek SVM makine öğrenimi yöntemi uygulanır.

5. Tartışma ve Değerlendirme

Bu tezde, literatür konu başlığı altında yer alan ve makine öğrenimi yöntemlerinin başarısını arttırmaya yönelik uygulanmış yöntemlerin ışığında elde edilen deneyimler mevcut film yorumları veri setine uygulanmış olup, daha önceden

uygulanmış olan yöntemlerin çoğuna kıyasla daha iyi bir başarı oranı ile sınıflanabilmiştir.

Veri setimizde seyrek vektör olarak sınıflanan yorumlar incelendiğinde daha çok noktalama işaretleri ve Türkçe içerisinde yer almayan sokak dilinde kullanılmakta olan cümleler tespit edilmiştir. Çok az sayıda olan bu yorumlar, kullanmış olduğumuz Zemberek kütüphanesi tarafından kök hallerinde tespit edilememiştir. Benzeri çalışmalarda başlıca İngilizce ve diğer diller için elde edilmiş sonuçlar ile bu tezde elde edilen sonuçlar karşılaştırıldığında hemen hemen aynı başarı yakalanmış olup, uygulanabilirlik açısından diğer yöntemlere kıyasla daha kolay uygulanabilirliği ile ön plana çıkmaktadır.

Ayrıca yapılan deney değerlendirmesinde Türkçe dilinde yazılmış olan cümleler içerisindeki duyguların tespiti sırasında kullanılması gereken en uygun kelime sayısının da tespiti yapılmaya çalışılmıştır. Burada en başarılı sonuçların alındığı gözlenen 375 terim sayısı, Türkçe dilindeki duygu analizlerinde yapılacak diğer çalışmalara referans olacaktır.

Diğer yapılan bir değerlendirmede elde edilen sonuçların başarısının arttırılabileceği düşünülmektedir. Bunun gerçekleştirilmesi için eğitim setinde kullanılan verilerin karışık olarak kategori aralığından seçilmesi yerine, en uç noktalardan yani sadece 0,5, 3,0 ve 5,0 gibi yorumu tam olarak niteleyen puanlardan seçilerek oluşturulması ile sağlanabilir. Yapılan çalışmada doğal en baştan doğal yani müdahale olmaksızın yapılacak bir duygu analizi kararlaştırıldığı için bu yöntem uygulanmamıştır.

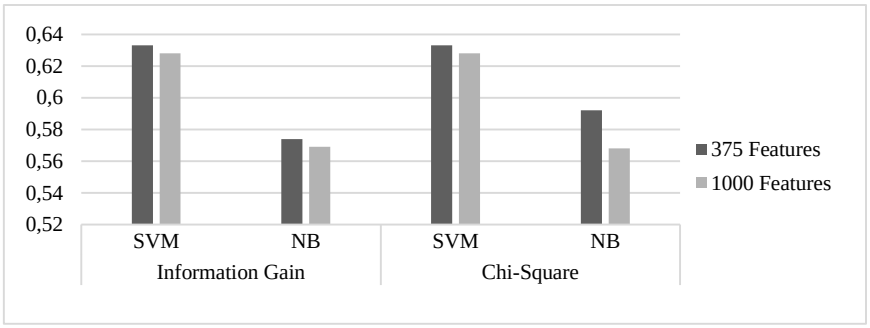
6. Gelecek Çalışmalar

Yapılan tez çalışmasındaki kısıtlı süre sıkıntısı nedeniyle, önerilen yöntemin daha farklı veri setlerinde uygulanması mümkün olmamıştır. Önerilen yöntemin dil ve veri tipi bakımından bağımsız olduğundan, sadece film yorumları değil diğer farklı türlerdeki yorum verilerinin duygu analizlerinde de bu tezde önerilen yöntemin uygulanarak başarı alınması teorik olarak mümkündür. Gelecek çalışmalarda, var olan başka veri setleri ile ve bu tezde önerilen yönteme ek olarak “Duygu

Sembolleri” (Emotion Tags) ve *Bag-of-words* gibi yöntemlerin de sonuçları iyileştirmesine yönelik çalışmaların yapılması ön görülmektedir. Ayrıca eğitim veri seti oluşturulması için kategorilerdeki en belirleyici olan puan aralıklarının seçilip tekrar sistem dizaynının test edilmesi faydalı görülmüş ve yapılacak gelecek çalışmalar içerisinde yer almaktadır.

7. Tez Kapsamında yapılan Ek Testler

Tez kapsamında uygulanan makine öğrenimi yöntemlerinden olan SVM sınıflayıcısının başarısına ek olarak, istatistiki olarak sınıflama yapan Naive Bayes (NB) sınıflayıcısının, öznitelik seçme metriklerinin uygulanması sonrası oluşan sınıflama başarı değerleri de gözlenmiştir. Gözlemlenen sonuçlarda beklendiği üzere SVM sınıflayıcısının başarısı, NB sınıflayıcısına oranla üçlü sınıflamalarda açık ara başarılı olduğu gözlemlenmiştir. Karşılaştırma sonuçları Şekil 7.1’de yer almaktadır.



Şekil 7.1 SVM ve Naive Bayes Test Sonuçları

İkinci bir test olarak; önceden öznitelik seçme metrikleriyle hesaplanmış olan ağırlıklandırma puanları, sınıflayıcıya verilen matrislerdeki ‘1’ ve ‘0’ yerlerine kullanıldı. Bu şekil sonuçlarda bir iyileşme olup olmadığı araştırıldı. Chi-Square yönteminden elde edilen en ayrıştırıcı 375 kelime için yapılan testin sonuçlarına göre SVM üzerinde alınan %63,6 doğru sınıflama başarısıyla, önceki sonuca ek

%0,3 gibi bir iyileşme gözlemlendi. Bunun da genel başarı oranına kıyasla, işlem maliyeti hesabı açısından önemli bir başarı artışı olarak değerlendirilmedi. Bu deney için var olan ayrıntılı sonuçlar Tablo 7.1'de gösterilmektedir.

Tüm Eğitim Seti	TP Oranı	FP Oranı	Precision	Recall	F-Measure	MCC	ROC Alanı	PRC Alanı	Sınıfı
	0.636	0.194	0.621	0.636	0.628	0.439	0.796	0.656	Olumlu
	0.678	0.189	0.642	0.678	0.66	0.483	0.832	0.695	Olumsuz
	0.468	0.226	0.509	0.468	0.488	0.248	0.678	0.511	Nötr
Ağırlıklı Ort.	0.594	0.203	0.591	0.594	0.592	0.39	0.769	0.621	
Tüm Test Seti	**TP Oranı**	**FP Oranı**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Alanı**	**PRC Alanı**	**Sınıfı**
	0.497	0.139	0.641	0.497	0.56	0.385	0.763	0.653	Olumlu
	0.735	0.241	0.604	0.735	0.663	0.474	0.825	0.676	Olumsuz
	0.498	0.254	0.495	0.498	0.497	0.244	0.675	0.488	Nötr
Ağırlıklı Ort.	0.577	0.212	0.58	0.577	0.573	0.368	0.754	0.606	

Tablo 7.1 Chi-Square Metriği için en ayrıştırıcı 375 adet terimin ağırlıklandırılmış haldeki ayrıntılı test sonuçları

KAYNAKLAR

[1] Donovan, John J. Database system approach the management decision support. ACM Transactions on Database Systems (TODS), 1.4: 344-369, **1976**.

[2] Sultan, M. U. And Fredrik S.. Consumers' Attitude towards Online Shopping Factors influencing Gotland consumers to shop online, VT2011 Master Thesis in Business Administration, **2011**.

[3] A Tutorial on Clustering Algorithms - http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/, **2011**.

[4] Han, Jiawei, and Micheline Kamber. Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan kaufmann, **2006**.

[5] Vural, AG, et al., A Framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish, Computer and Information Sciences III, Springer London, pp437-445, **2013**.

[6] ConnotationWordNet: Learning Connotation over the Word+Sense Network – www.aclweb.org/anthology/P/P14/P14-1145.xhtml, **2014**.

[7] Esuli, A., and Sebastiani, F.. Sentiwordnet: A publicly available lexical resource for opinion mining. ,Proceedings of LREC. Vol. 6., **2006**.

[8] Whissell, C.. The dictionary of affect in language. Emotion: Theory, research, and experience 4, pp113-131, **1989**.

[9] Ghorbel, H., and David J.. Sentiment analysis of French movie reviews. Advances in Distributed Agent-Based Retrieval Tools, Springer Berlin Heidelberg, pp97-108, **2011**.

[10] Valdivia, M., et al.. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. , Expert Systems with Applications, **2012**.

[11] Denecke, K.. Using SentiWordNet for multilingual sentiment analysis.,Data Engineering Workshop. ICDEW 2008. IEEE 24th International Conference on. IEEE, **2008**.

[12] Pang, B., et. al.. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.,Proceedings of the

42nd annual meeting on Association for Computational Linguistics,Association for Computational Linguistics, **2004**.

[13] Lewis, DD.. Naive (Bayes) at forty: The independence assumption in information retrieval. Machine learning: ECML-98, Springer Berlin Heidelberg, pp4-15, **1998**.

[14] Berger, Adam L., Vincent J. Della Pietra, and Stephen A. Della Pietra. "A maximum entropy approach to natural language processing." Computational linguistics 22.1, **1996**.

[15] Cover, T., and Hart, P.. Nearest neighbor pattern classification. Information Theory, IEEE Transactions on, 13, pp21-27, **1967**.

[16] Cortes, C and Vapnik, V.. Support vector machine. Machine learning 20.3, pp273-297, **1995**.

[17] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics, **2011**.

[18] Becker, Lee, et al. "Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion." Atlanta, Georgia, USA, **2013**.

[19] Habernal, I. et. al.. Sentiment Analysis in Czech Social Media Using Supervised Machine Learning, WASSA: 65, **2013.**

[20] Suen, C. Y.. N-gram statistics for natural language understanding and text processing. Pattern Analysis and Machine Intelligence, IEEE Transactions on, (2), 164-172,**1979**.

[21] Narr, S., Hülfenhaus, M., & Albayrak, S.. Language-independent Twitter sentiment analysis. Knowledge Discovery and Machine Learning (KDML), LWA, **2012**.

[22] Pak, A., and Paroubek, P.. Twitter as a Derlem for Sentiment Analysis and Opinion Mining. ,LREC, **2010**.

[23] Pang, B., et. al.. Thumbs up: sentiment classification using machine learning techniques.,Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, **2002**.

[24] Saif, H. et. al.. Alleviating data sparsity for twitter sentiment analysis.,The 2nd Workshop on Making Sense of Microposts, **2012**.

[25] Harris, Zellig S. "Distributional structure." Word, **1954**.

[26] Erogul, U.,Sentiment Analysis In Turkish.,Master's thesis,Middle East Technical University, **2009**.

[27] Boynukalin, Z. Emotion Analysis of Turkish texts by using machine learning methods.,Master's thesis,Middle East Technical University, **2012**.

[28] Cakmak, O. et al.. Using interval type-2 fuzzy logic to analyze Turkish emotion words.,Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC),Asia-Pacific. IEEE, **2012**.

[29] Ozsert, CM, and Arzucan O.. Word polarity detection using a multilingual approach. Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, pp75-82, **2013**.

[30] Kaya, M., et. al.. Sentiment Analysis of Turkish Political News.,Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01,IEEE Computer Society, **2012**.

[31] BeyazPerde - http://www.beyazperde.com,15.01.2014.

[32] HTML Agility Pack DOM parser - http://htmlagilitypack.codeplex.com/, **2012**.

[33] Akin, AA and Akin, MD. Zemberek, an open source NLP framework for Turkish Languages Online Available at: https://code.google.com/p/zemberek/, **2007**.

[34] SVM trained with samples from two classes - http://blog.pengyifan.com/tikz-example-svm-trained-with-samples-from-two-classes/, **2013**.

[35] Holmes, G., et al.. Weka: A machine learning workbench. In Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on (pp. 357-361). IEEE, **1994**.

[36] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology (TIST) 2.3, **2011**.

ÇİZELGELER

Çizelge 1. IG ve Chi-Square İçin En Ayrıştırıcı 375 Terim Listesi ve Puanları

Kelime	IG_Pos	IG_Neg	IG_Neu	Toplam Skor	Değer Sırası	Kelime	CS_Pos	CS_Neg	CS_Neu	Toplam Skor
kötü	0.006	0.009	0.001	0.016	1	kötü	669	1219	82	1970
mükemmel	0.007	0.006	0.000	0.013	2	mükemmel	929	549	50	1528
berbat	0.003	0.008	0.002	0.013	3	berbat	298	964	190	1452
harika	0.005	0.006	0.000	0.011	4	harika	694	555	8	1257
sıkıcı	0.004	0.005	0.000	0.010	5	sıkıcı	410	707	40	1158
saçma	0.003	0.005	0.001	0.009	6	saçma	288	655	75	1018
kayıp	0.002	0.005	0.001	0.008	7	kayıp	194	627	123	945
hiç	0.002	0.005	0.001	0.007	8	hiç	243	605	81	929
değil	0.005	0.001	0.001	0.007	9	ama	545	45	277	867
ama	0.005	0.000	0.002	0.007	10	değil	574	145	142	862
kırık	0.003	0.003	0.000	0.007	11	para	170	528	99	796
para	0.002	0.004	0.001	0.007	12	yazık	202	498	66	766
yazık	0.002	0.004	0.001	0.006	13	kırık	296	443	15	753
vasat	0.004	0.002	0.000	0.006	14	git	311	409	7	727
git	0.003	0.003	0.000	0.006	15	en	483	119	123	725
en	0.004	0.001	0.001	0.006	16	başyapıt	412	236	25	673
muhteşem	0.003	0.003	0.000	0.006	17	muhteşem	384	277	9	670
başyapıt	0.003	0.002	0.000	0.006	18	vasat	388	247	16	652
mutlaka	0.003	0.003	0.000	0.005	19	mutlaka	363	258	9	629
hayal	0.003	0.003	0.000	0.005	20	hayal	257	353	8	618
müthiş	0.002	0.002	0.000	0.005	21	müthiş	311	230	6	547
süper	0.002	0.002	0.000	0.004	22	süper	268	247	0	515
basit	0.002	0.002	0.000	0.004	23	harca	132	326	43	501
harca	0.001	0.002	0.000	0.004	24	basit	209	277	5	491
iyi	0.000	0.003	0.001	0.004	25	iyi	43	299	115	458
yok	0.001	0.002	0.000	0.004	26	yok	135	277	25	437
sapan	0.001	0.002	0.000	0.003	27	güzel	14	245	141	400
güzel	0.000	0.002	0.001	0.003	28	sapan	85	258	47	390
sıradan	0.002	0.001	0.000	0.003	29	yine	53	76	256	385
puan	0.002	0.001	0.000	0.003	30	biraz	131	14	232	377
yine	0.000	0.001	0.002	0.003	31	puan	202	165	2	369
biraz	0.001	0.000	0.002	0.003	32	vakit	147	206	5	359
beklenti	0.002	0.000	0.000	0.003	33	ol	236	84	39	359
vakit	0.001	0.002	0.000	0.003	34	beğen	180	176	0	356
saçmalık	0.001	0.002	0.000	0.003	35	sadece	155	198	3	356
beğen	0.002	0.001	0.000	0.003	36	saçmalık	91	229	31	351

Kelime	IG_Pos	IG_Neg	IG_Neu	Toplam Skor	Değer Sırası	Kelime	CS_Pos	CS_Neg	CS_Neu	Toplam Skor
ol	0.002	0.001	0.000	0.003	37	sıradan	209	119	13	340
sadece	0.001	0.002	0.000	0.003	38	boş	127	202	9	337
boş	0.001	0.002	0.000	0.003	39	beklenti	216	50	58	324
iğrenç	0.001	0.002	0.000	0.003	40	kelime	215	47	61	322
abart	0.002	0.000	0.001	0.003	41	ben	184	126	5	315
ben	0.002	0.001	0.000	0.003	42	iğrenç	73	208	34	315
kelime	0.002	0.000	0.001	0.003	43	tarihi	205	71	35	310
rezalet	0.001	0.001	0.000	0.003	44	bile	44	196	54	295
tarihi	0.002	0.001	0.000	0.002	45	abart	178	9	107	294
konu	0.002	0.001	0.000	0.002	46	rezalet	63	192	35	290
maalesef	0.001	0.001	0.000	0.002	47	konu	181	91	15	287
bile	0.000	0.002	0.000	0.002	48	tamamen	82	183	20	285
tamamen	0.001	0.001	0.000	0.002	49	boşa	59	184	34	278
boşa	0.001	0.001	0.000	0.002	50	anla	78	176	20	274
fiyasko	0.001	0.001	0.000	0.002	51	etkile	89	166	12	267
etkile	0.001	0.001	0.000	0.002	52	maalesef	126	141	0	267
anla	0.001	0.001	0.000	0.002	53	kurtar	85	167	14	266
kurtar	0.001	0.001	0.000	0.002	54	fiyasko	73	163	18	254
yılmaz	0.001	0.001	0.000	0.002	55	diye	77	160	15	252
bekle	0.001	0.001	0.000	0.002	56	bekle	160	73	17	250
fena	0.001	0.000	0.001	0.002	57	birey	106	140	2	249
birey	0.001	0.001	0.000	0.002	58	izle	104	138	2	244
diye	0.001	0.001	0.000	0.002	59	om	156	68	18	242
izle	0.001	0.001	0.000	0.002	60	yılmaz	75	152	13	241
pek	0.001	0.000	0.001	0.002	61	fena	116	0	115	230
om	0.001	0.001	0.000	0.002	62	pek	142	10	78	230
arşiv	0.001	0.001	0.000	0.002	63	arşiv	137	75	9	221
uğra	0.001	0.001	0.000	0.002	64	hoş	36	36	146	219
eğlence	0.000	0.001	0.001	0.002	65	değ	41	146	32	219
değ	0.000	0.001	0.000	0.002	66	eğlence	0	101	115	216
açık	0.001	0.000	0.000	0.002	67	uğra	83	128	5	216
fakat	0.001	0.000	0.001	0.002	68	bu	14	136	63	213
hoş	0.000	0.000	0.001	0.002	69	fakat	98	0	112	211
aldan	0.001	0.001	0.000	0.002	70	fazla	99	0	108	207
bu	0.000	0.001	0.001	0.002	71	hayat	131	12	63	206
fazla	0.001	0.000	0.001	0.002	72	aldan	49	133	20	203
hayat	0.001	0.000	0.001	0.002	73	yap	32	135	35	202
klişe	0.001	0.000	0.000	0.002	74	açık	130	57	15	202
çalış	0.001	0.001	0.000	0.002	75	amerikan	80	115	3	198
amerikan	0.001	0.001	0.000	0.002	76	çalış	86	111	2	198
kaçır	0.001	0.001	0.000	0.002	77	gerçek	112	78	3	194
yap	0.000	0.001	0.000	0.002	78	her	111	78	3	192
gerçek	0.001	0.001	0.000	0.002	79	hiçbir	35	127	29	190

Kelime	IG_Pos	IG_Neg	IG_Neu	Toplam Skor	Değer Sırası	Kelime	CS_Pos	CS_Neg	CS_Neu	Toplam Skor
her	0.001	0.001	0.000	0.002	80	kaçır	91	99	0	190
dış	0.001	0.000	0.000	0.002	81	dış	117	43	18	178
hiçbir	0.000	0.001	0.000	0.002	82	hala	117	33	26	175
zayıf	0.001	0.000	0.000	0.001	83	klişe	115	42	18	175
yapıt	0.001	0.001	0.000	0.001	84	keyif	3	70	101	173
mantık	0.001	0.000	0.000	0.001	85	nasıl	27	115	31	173
keyif	0.000	0.001	0.001	0.001	86	arkadaş	36	114	22	172
gibi	0.001	0.000	0.000	0.001	87	gibi	109	47	13	170
hala	0.001	0.000	0.000	0.001	88	yapıt	86	83	0	169
uzak	0.001	0.001	0.000	0.001	89	başka	18	111	40	168
nasıl	0.000	0.001	0.000	0.001	90	uzak	77	86	0	164
arkadaş	0.000	0.001	0.000	0.001	91	resmen	36	108	19	163
küfür	0.000	0.001	0.000	0.001	92	boşuna	39	107	17	162
boşuna	0.000	0.001	0.000	0.001	93	aksiyon	89	1	72	161
başka	0.000	0.001	0.000	0.001	94	mantık	98	57	5	161
aksiyon	0.001	0.000	0.001	0.001	95	küfür	42	103	14	159
hüsran	0.000	0.001	0.000	0.001	96	zayıf	102	26	26	154
abartı	0.001	0.000	0.000	0.001	97	kadar	56	92	4	153
resmen	0.000	0.001	0.000	0.001	98	başarı	4	56	91	152
sıfır	0.000	0.001	0.000	0.001	99	hüsran	40	98	13	151
favori	0.001	0.001	0.000	0.001	100	favori	90	52	5	146
kadar	0.000	0.001	0.000	0.001	101	abartı	97	23	26	146
başarı	0.000	0.000	0.001	0.001	102	sıfır	38	95	13	146
yarı	0.001	0.001	0.000	0.001	103	uyu	32	95	16	143
kopuk	0.001	0.000	0.000	0.001	104	sırf	50	88	5	143
sırf	0.000	0.001	0.000	0.001	105	yarı	69	74	0	142
şaheser	0.001	0.000	0.000	0.001	106	için	94	21	26	141
uyu	0.000	0.001	0.000	0.001	107	şaheser	89	43	8	140
için	0.001	0.000	0.000	0.001	108	kot	37	88	11	136
kot	0.000	0.001	0.000	0.001	109	korku	47	83	5	135
ancak	0.001	0.000	0.000	0.001	110	herkes	82	48	5	134
korku	0.000	0.001	0.000	0.001	111	ancak	83	6	44	133
fragman	0.001	0.001	0.000	0.001	112	kopuk	67	65	0	132
herkes	0.001	0.000	0.000	0.001	113	tüy	87	25	19	130
alaka	0.001	0.000	0.000	0.001	114	fragman	53	75	2	130
belli	0.001	0.000	0.000	0.001	115	belli	68	59	0	127
rezil	0.000	0.001	0.000	0.001	116	ver	67	57	0	125
tüy	0.001	0.000	0.000	0.001	117	rezil	31	81	12	124
yani	0.001	0.000	0.000	0.001	118	alaka	63	61	0	123
ver	0.001	0.000	0.000	0.001	119	gayet	18	23	82	123
morto	0.001	0.000	0.000	0.001	120	yani	73	46	3	122
yüksek	0.001	0.000	0.000	0.001	121	defa	81	25	16	122
tamam	0.001	0.000	0.000	0.001	122	morto	69	48	2	119

Kelime	IG_Pos	IG_Neg	IG_Neu	Toplam Skor	Değer Sırası	Kelime	CS_Pos	CS_Neg	CS_Neu	Toplam Skor
gayet	0.000	0.000	0.001	0.001	123	söz	78	20	19	117
defa	0.001	0.000	0.000	0.001	124	olağanüstü	69	43	3	115
olağanüstü	0.001	0.000	0.000	0.001	125	kesin	76	26	13	115
sıkıntı	0.000	0.001	0.000	0.001	126	sıkıntı	38	71	5	114
bomboş	0.000	0.001	0.000	0.001	127	tamam	70	38	5	113
söz	0.001	0.000	0.000	0.001	128	yüksek	75	21	16	112
kesin	0.001	0.000	0.000	0.001	129	genel	30	9	72	111
daha	0.001	0.000	0.000	0.001	130	daha	69	5	36	110
ney	0.000	0.000	0.000	0.001	131	ney	44	63	2	109
genel	0.000	0.000	0.001	0.001	132	geçmiş	68	35	5	108
var	0.001	0.000	0.000	0.001	133	var	62	43	2	107
ucuz	0.000	0.001	0.000	0.001	134	ucuz	25	70	12	106
geçmiş	0.001	0.000	0.000	0.001	135	bomboş	24	70	12	106
sahne	0.001	0.000	0.000	0.001	136	kusur	65	34	5	103
eder	0.000	0.001	0.000	0.001	137	sahne	68	22	13	103
kusur	0.000	0.000	0.000	0.001	138	sakın	20	68	15	103
falan	0.001	0.000	0.000	0.001	139	eder	15	67	19	100
müzik	0.000	0.001	0.000	0.001	140	çok	64	8	27	99
sakın	0.000	0.001	0.000	0.001	141	müzik	39	58	2	99
mahsus	0.000	0.000	0.000	0.001	142	sık	27	8	64	99
saçmala	0.000	0.000	0.000	0.001	143	falan	57	39	2	97
çok	0.001	0.000	0.000	0.001	144	ne	9	63	25	97
hata	0.000	0.000	0.000	0.001	145	ağla	65	17	15	97
sık	0.000	0.000	0.000	0.001	146	zaman	30	61	5	96
özel	0.000	0.000	0.000	0.001	147	hata	53	41	1	95
numara	0.000	0.000	0.000	0.001	148	özel	2	56	37	95
ne	0.000	0.001	0.000	0.001	149	amaç	32	58	4	94
performans	0.000	0.001	0.000	0.001	150	dünya	62	17	14	94
şen	0.000	0.001	0.000	0.001	151	neden	51	41	1	92
zaman	0.000	0.000	0.000	0.001	152	numara	47	46	0	92
efsane	0.000	0.000	0.000	0.001	153	efsane	48	44	0	92
ağla	0.000	0.000	0.000	0.001	154	mahsus	18	61	13	92
neden	0.000	0.000	0.000	0.001	155	saçmala	37	53	1	92
amaç	0.000	0.000	0.000	0.001	156	performans	16	61	15	91
unut	0.000	0.000	0.000	0.001	157	unut	51	40	1	91
dünya	0.000	0.000	0.000	0.001	158	sinema	33	3	55	91
sinema	0.000	0.000	0.000	0.001	159	geçir	20	10	59	90
komedi	0.000	0.000	0.000	0.001	160	kez	55	30	4	89
kez	0.000	0.000	0.000	0.001	161	biri	57	7	24	88
geçir	0.000	0.000	0.000	0.001	162	komedi	54	4	29	88
reklam	0.000	0.000	0.000	0.001	163	artık	24	56	7	86
biri	0.000	0.000	0.000	0.001	164	salon	31	52	3	86
idare	0.000	0.000	0.000	0.001	165	dayan	19	57	10	86

Kelime	IG_Pos	IG_Neg	IG_Neu	Toplam Skor	Değer Sırası	Kelime	CS_Pos	CS_Neg	CS_Neu	Toplam Skor
salon	0.000	0.000	0.000	0.001	166	şen	4	54	28	86
kan	0.000	0.000	0.000	0.001	167	kan	30	52	3	85
artık	0.000	0.000	0.000	0.001	168	ya	14	56	14	84
harikulade	0.000	0.000	0.000	0.001	169	reklam	35	48	1	83
dayan	0.000	0.000	0.000	0.001	170	ilginç	37	0	45	83
ilginç	0.000	0.000	0.000	0.001	171	sürükleyici	1	35	47	83
sürükleyici	0.000	0.000	0.000	0.001	172	hayret	14	55	13	83
mide	0.000	0.000	0.000	0.001	173	saat	38	44	0	82
nezt	0.000	0.000	0.000	0.001	174	nezt	37	45	0	82
tatmin	0.000	0.000	0.000	0.001	175	harikulade	52	22	6	81
saat	0.000	0.000	0.000	0.001	176	idare	34	1	45	80
sahnele	0.000	0.000	0.000	0.001	177	mide	26	49	4	79
ya	0.000	0.000	0.000	0.001	178	tatmin	42	0	36	79
atatürkçü	0.000	0.000	0.000	0.001	179	sahnele	52	9	18	79
hayret	0.000	0.000	0.000	0.001	180	şahane	48	27	3	78
şahane	0.000	0.000	0.000	0.001	181	anlam	21	50	6	78
desen	0.000	0.000	0.000	0.001	182	göre	39	0	38	77
göre	0.000	0.000	0.000	0.001	183	bit	34	43	0	77
diken	0.000	0.000	0.000	0.001	184	desen	19	50	8	77
zorla	0.000	0.000	0.000	0.001	185	nere	28	46	2	76
yapmacık	0.000	0.000	0.000	0.001	186	nadir	49	21	6	75
bit	0.000	0.000	0.000	0.001	187	işkence	19	49	7	75
tol	0.000	0.000	0.000	0.001	188	zorla	37	38	0	75
ban	0.000	0.000	0.000	0.001	189	şey	22	48	5	75
maron	0.000	0.000	0.000	0.001	190	ban	49	7	19	75
anlam	0.000	0.000	0.000	0.001	191	maron	48	21	6	74
komik	0.000	0.000	0.000	0.001	192	joker	9	15	49	74
tarz	0.000	0.000	0.000	0.001	193	tarz	44	2	29	74
işkence	0.000	0.000	0.000	0.001	194	atatürkçü	11	49	14	74
nadir	0.000	0.000	0.000	0.001	195	branda	49	15	10	73
nere	0.000	0.000	0.000	0.001	196	komik	44	27	2	73
branda	0.000	0.000	0.000	0.001	197	diken	47	22	5	73
ortalama	0.000	0.000	0.000	0.001	198	eksik	15	10	49	73
şey	0.000	0.000	0.000	0.001	199	seyirlik	8	16	48	73
joker	0.000	0.000	0.000	0.001	200	yapmacık	28	42	1	72
geri	0.000	0.000	0.000	0.001	201	gül	22	45	4	71
görsel	0.000	0.000	0.000	0.001	202	geri	38	33	0	71
seyirlik	0.000	0.000	0.000	0.001	203	yıl	28	42	1	71
yıl	0.000	0.000	0.000	0.001	204	özgür	45	21	5	70
eksik	0.000	0.000	0.000	0.001	205	niye	12	47	11	70
hayır	0.000	0.000	0.000	0.001	206	görsel	46	9	14	70
hariç	0.000	0.000	0.000	0.001	207	lütfen	10	46	13	70
orta	0.000	0.000	0.000	0.001	208	oku	25	42	2	69

Kelime	IG_Pos	IG_Neg	IG_Neu	Toplam Skor	Değer Sırası	Kelime	CS_Pos	CS_Neg	CS_Neu	Toplam Skor
yavan	0.000	0.000	0.000	0.001	209	efekt	40	28	1	69
gül	0.000	0.000	0.000	0.001	210	kimse	7	45	17	68
efekt	0.000	0.000	0.000	0.001	211	ortalama	37	0	30	68
özgür	0.000	0.000	0.000	0.001	212	orta	44	6	18	68
atatürk	0.000	0.000	0.000	0.001	213	tempo	14	8	45	68
sinir	0.000	0.000	0.000	0.001	214	kenar	15	8	44	67
aşırı	0.000	0.000	0.000	0.001	215	derece	33	34	0	67
niye	0.000	0.000	0.000	0.001	216	hadi	22	42	3	67
oku	0.000	0.000	0.000	0.001	217	sinir	27	39	1	67
lütfen	0.000	0.000	0.000	0.001	218	hayır	40	25	2	66
yaratık	0.000	0.000	0.000	0.001	219	bazı	24	2	40	66
derece	0.000	0.000	0.000	0.001	220	hariç	42	19	5	66
uzat	0.000	0.000	0.000	0.001	221	tam	10	44	12	66
hadi	0.000	0.000	0.000	0.001	222	bap	38	26	1	65
kimse	0.000	0.000	0.000	0.001	223	yaratık	26	38	1	65
tempo	0.000	0.000	0.000	0.001	224	tol	43	12	10	64
bap	0.000	0.000	0.000	0.001	225	fark	16	6	41	64
kenar	0.000	0.000	0.000	0.001	226	millet	10	42	11	63
bazı	0.000	0.000	0.000	0.001	227	et	39	22	2	63
sonuç	0.000	0.000	0.000	0.001	228	uç	18	41	5	63
tam	0.000	0.000	0.000	0.001	229	jeyn	38	23	2	63
jeyn	0.000	0.000	0.000	0.001	230	atatürk	13	42	8	63
sıkıl	0.000	0.000	0.000	0.001	231	sonuç	41	17	5	63
erdoan	0.000	0.000	0.000	0.001	232	sıkıl	42	12	9	63
espri	0.000	0.000	0.000	0.001	233	bak	35	26	1	62
et	0.000	0.000	0.000	0.001	234	aşırı	39	20	3	62
bari	0.000	0.000	0.000	0.001	235	olay	39	19	4	62
olay	0.000	0.000	0.000	0.001	236	bari	21	38	3	62
bak	0.000	0.000	0.000	0.001	237	uzat	41	7	13	62
fark	0.000	0.000	0.000	0.001	238	espri	35	26	1	61
hart	0.000	0.000	0.000	0.001	239	yavan	33	29	0	61
millet	0.000	0.000	0.000	0.001	240	oyun	4	39	18	61
uç	0.000	0.000	0.000	0.001	241	düşün	40	7	14	61
üçle	0.000	0.000	0.000	0.001	242	ibaret	16	39	5	60
oyun	0.000	0.000	0.000	0.001	243	üçle	34	26	1	60
düşün	0.000	0.000	0.000	0.001	244	belgesel	16	39	5	60
dizi	0.000	0.000	0.000	0.000	245	savaş	40	8	12	60
gişe	0.000	0.000	0.000	0.000	246	rol	0	32	27	60
rol	0.000	0.000	0.000	0.000	247	tekrar	37	20	2	59
ibaret	0.000	0.000	0.000	0.000	248	dizi	29	30	0	59
belgesel	0.000	0.000	0.000	0.000	249	alla	5	38	15	58
savaş	0.000	0.000	0.000	0.000	250	testere	20	36	2	58
merak	0.000	0.000	0.000	0.000	251	rağmen	8	12	38	58

Kelime	IG_Pos	IG_Neg	IG_Neu	Toplam Skor	Değer Sırası	Kelime	CS_Pos	CS_Neg	CS_Neu	Toplam Skor
tekrar	0.000	0.000	0.000	0.000	252	erdoan	18	37	3	58
testere	0.000	0.000	0.000	0.000	253	teşekkür	38	11	8	58
ayrı	0.000	0.000	0.000	0.000	254	patla	17	37	4	57
çek	0.000	0.000	0.000	0.000	255	dk	23	33	1	57
dk	0.000	0.000	0.000	0.000	256	çek	32	24	1	57
alla	0.000	0.000	0.000	0.000	257	merak	38	12	7	57
karşıla	0.000	0.000	0.000	0.000	258	gişe	34	21	2	57
patla	0.000	0.000	0.000	0.000	259	tür	13	6	37	56
rağmen	0.000	0.000	0.000	0.000	260	işe	15	36	5	56
sahan	0.000	0.000	0.000	0.000	261	ayrı	18	34	2	55
yakış	0.000	0.000	0.000	0.000	262	yakış	31	24	1	55
ilerle	0.000	0.000	0.000	0.000	263	ilerle	32	1	23	55
teşekkür	0.000	0.000	0.000	0.000	264	felaket	17	34	3	55
olumlu	0.000	0.000	0.000	0.000	265	hart	36	13	6	55
artı	0.000	0.000	0.000	0.000	266	propagand a	9	36	9	55
age	0.000	0.000	0.000	0.000	267	yoksa	25	28	0	53
işe	0.000	0.000	0.000	0.000	268	TRUE	35	8	9	53
vampir	0.000	0.000	0.000	0.000	269	usta	24	29	0	53
tür	0.000	0.000	0.000	0.000	270	böyle	0	27	25	53
felaket	0.000	0.000	0.000	0.000	271	bedel	33	18	2	53
ağır	0.000	0.000	0.000	0.000	272	karşıla	35	11	7	52
TRUE	0.000	0.000	0.000	0.000	273	ağır	34	7	11	52
usta	0.000	0.000	0.000	0.000	274	buka	10	34	7	52
facia	0.000	0.000	0.000	0.000	275	facia	15	33	4	52
yoksa	0.000	0.000	0.000	0.000	276	yürek	34	11	7	52
propaganda	0.000	0.000	0.000	0.000	277	vampir	24	28	0	51
amatör	0.000	0.000	0.000	0.000	278	sahan	9	34	8	51
çizgi	0.000	0.000	0.000	0.000	279	tartışma	33	13	5	51
bedel	0.000	0.000	0.000	0.000	280	sanki	34	10	7	51
böyle	0.000	0.000	0.000	0.000	281	biz	14	33	4	51
düzgün	0.000	0.000	0.000	0.000	282	senaryo	33	12	5	50
sanki	0.000	0.000	0.000	0.000	283	salak	15	32	3	50
salak	0.000	0.000	0.000	0.000	284	organ	33	12	5	50
dest	0.000	0.000	0.000	0.000	285	port	33	13	5	50
senaryo	0.000	0.000	0.000	0.000	286	batman	9	7	33	50
buka	0.000	0.000	0.000	0.000	287	artı	33	8	9	50
fatih	0.000	0.000	0.000	0.000	288	çizgi	33	4	13	50
yürek	0.000	0.000	0.000	0.000	289	olumlu	32	15	3	50
seyirci	0.000	0.000	0.000	0.000	290	seyirci	25	25	0	50
tartışma	0.000	0.000	0.000	0.000	291	kurt	10	33	6	50
biz	0.000	0.000	0.000	0.000	292	düzgün	26	22	0	49
port	0.000	0.000	0.000	0.000	293	recep	9	32	7	49
organ	0.000	0.000	0.000	0.000	294	age	32	12	5	48

Kelime	IG_Pos	IG_Neg	IG_Neu	Toplam Skor	Değer Sırası	Kelime	CS_Pos	CS_Neg	CS_Neu	Toplam Skor
ilgi	0.000	0.000	0.000	0.000	295	çıktı	14	31	4	48
erotik	0.000	0.000	0.000	0.000	296	ilgi	32	8	8	48
recep	0.000	0.000	0.000	0.000	297	can	13	31	4	48
hitap	0.000	0.000	0.000	0.000	298	çocuk	23	25	0	48
renk	0.000	0.000	0.000	0.000	299	amatör	22	25	0	48
batman	0.000	0.000	0.000	0.000	300	çöp	8	31	8	47
kurt	0.000	0.000	0.000	0.000	301	karakter	16	2	29	47
çöp	0.000	0.000	0.000	0.000	302	yad	24	23	0	47
enfes	0.000	0.000	0.000	0.000	303	erotik	12	30	4	47
çocuk	0.000	0.000	0.000	0.000	304	renk	18	27	1	47
korkut	0.000	0.000	0.000	0.000	305	hitap	27	19	1	46
yad	0.000	0.000	0.000	0.000	306	ender	28	16	2	46
çıktı	0.000	0.000	0.000	0.000	307	korkut	18	27	1	46
can	0.000	0.000	0.000	0.000	308	liste	30	10	5	46
dindar	0.000	0.000	0.000	0.000	309	tavsiye	30	11	5	46
ırmak	0.000	0.000	0.000	0.000	310	fatih	15	29	2	46
ender	0.000	0.000	0.000	0.000	311	oda	15	28	2	46
çağa	0.000	0.000	0.000	0.000	312	dindar	9	30	6	44
karakter	0.000	0.000	0.000	0.000	313	değer	1	16	27	44
absürt	0.000	0.000	0.000	0.000	314	dest	25	18	1	44
sergile	0.000	0.000	0.000	0.000	315	gider	14	28	3	44
tavsiye	0.000	0.000	0.000	0.000	316	uğrat	16	27	1	44
oda	0.000	0.000	0.000	0.000	317	enfes	22	21	0	44
liste	0.000	0.000	0.000	0.000	318	dalga	15	27	2	43
ticari	0.000	0.000	0.000	0.000	319	bık	27	2	14	43
dalga	0.000	0.000	0.000	0.000	320	dolu	21	22	0	43
uğrat	0.000	0.000	0.000	0.000	321	sergile	7	29	7	43
değer	0.000	0.000	0.000	0.000	322	yorum	8	29	7	43
gider	0.000	0.000	0.000	0.000	323	ırmak	23	19	0	43
dönem	0.000	0.000	0.000	0.000	324	ticari	17	25	1	43
dolu	0.000	0.000	0.000	0.000	325	çağa	24	18	0	42
illa	0.000	0.000	0.000	0.000	326	dönem	1	25	16	42
bık	0.000	0.000	0.000	0.000	327	gidi	16	25	1	42
hatır	0.000	0.000	0.000	0.000	328	mutaf	10	28	5	42
esaret	0.000	0.000	0.000	0.000	329	esaret	25	16	1	42
bol	0.000	0.000	0.000	0.000	330	duygu	19	23	0	42
mutaf	0.000	0.000	0.000	0.000	331	gör	17	24	1	42
korkunç	0.000	0.000	0.000	0.000	332	absürt	20	22	0	42
neat	0.000	0.000	0.000	0.000	333	bol	24	17	1	41
yorum	0.000	0.000	0.000	0.000	334	ev	11	27	3	41
gidi	0.000	0.000	0.000	0.000	335	tüm	27	6	8	41
duygu	0.000	0.000	0.000	0.000	336	yaz	6	27	8	41
şişir	0.000	0.000	0.000	0.000	337	korkunç	17	23	0	41

Kelime	IG_Pos	IG_Neg	IG_Neu	Toplam Skor	Değer Sırası	Kelime	CS_Pos	CS_Neg	CS_Neu	Toplam Skor
bunal	0.000	0.000	0.000	0.000	338	say	22	0	19	41
kombi	0.000	0.000	0.000	0.000	339	kült	27	8	5	41
say	0.000	0.000	0.000	0.000	340	üzül	4	27	10	41
sağlam	0.000	0.000	0.000	0.000	341	da	21	0	19	41
sihirbaz	0.000	0.000	0.000	0.000	342	illa	19	21	0	40
sıcak	0.000	0.000	0.000	0.000	343	sihirbaz	25	13	2	40
gör	0.000	0.000	0.000	0.000	344	argo	9	26	5	40
şahım	0.000	0.000	0.000	0.000	345	sağlam	1	24	15	40
dram	0.000	0.000	0.000	0.000	346	yerin	19	21	0	40
da	0.000	0.000	0.000	0.000	347	dakika	23	17	0	40
ev	0.000	0.000	0.000	0.000	348	dram	13	25	2	40
örümcek	0.000	0.000	0.000	0.000	349	film	19	21	0	39
tüm	0.000	0.000	0.000	0.000	350	şiddet	26	5	8	39
yaz	0.000	0.000	0.000	0.000	351	son	26	4	10	39
dakika	0.000	0.000	0.000	0.000	352	şaş	7	26	6	39
argo	0.000	0.000	0.000	0.000	353	bunal	13	24	2	39
büyüle	0.000	0.000	0.000	0.000	354	iste	18	21	0	39
üzül	0.000	0.000	0.000	0.000	355	sıcak	0	20	19	39
heba	0.000	0.000	0.000	0.000	356	final	1	23	14	39
kült	0.000	0.000	0.000	0.000	357	öldür	5	26	8	38
yerin	0.000	0.000	0.000	0.000	358	büyüle	20	18	0	38
son	0.000	0.000	0.000	0.000	359	değişik	8	5	25	38
film	0.000	0.000	0.000	0.000	360	ırak	4	25	9	38
final	0.000	0.000	0.000	0.000	361	şahım	15	1	22	38
ırak	0.000	0.000	0.000	0.000	362	vahşet	6	25	6	38
sığ	0.000	0.000	0.000	0.000	363	aynı	25	5	7	38
iste	0.000	0.000	0.000	0.000	364	ısparta	11	24	3	38
prestij	0.000	0.000	0.000	0.000	365	neat	14	23	1	38
ilkokul	0.000	0.000	0.000	0.000	366	halbuki	12	23	2	38
boz	0.000	0.000	0.000	0.000	367	ilkokul	8	25	5	38
şaş	0.000	0.000	0.000	0.000	368	kadın	19	19	0	38
şiddet	0.000	0.000	0.000	0.000	369	şişir	19	18	0	37
aynı	0.000	0.000	0.000	0.000	370	hızlı	16	0	21	37
skeç	0.000	0.000	0.000	0.000	371	bula	21	16	0	37
bay	0.000	0.000	0.000	0.000	372	mi	9	25	4	37
eşkıya	0.000	0.000	0.000	0.000	373	hatır	25	7	6	37
halbuki	0.000	0.000	0.000	0.000	374	boz	19	18	0	37
tabii	0.000	0.000	0.000	0.000	375	tabii	2.36	23.64	11.07	37.06

Çizelge 2. Weka İçerisinde Bulunan Uygulamalar ve İşlevleri

CLOPE	Clustering
DMNBtext	Text classification
DTNB	Classification
DilcaDistance	Distance
DistributionBasedBalance	Preprocessing
EMImputation	Preprocessing
EvolutionarySearch	Attribute selection
GPAttributeGeneration	Classification, Preprocessing
IWSS	Attribute selection
J48graft	Classification
JDBCDriversDummyPackage	Misc
LVQ	Clustering
LibLINEAR	Classification
LibSVM	Classification
MODLEM	Classification, Ensemble learning
NNge	Classification
PCP	Visualization
PSOSearch	Attribute selection
RBFNetwork	Classification/regression
RPlugin	R integration
RerankingSearch	Attribute selection
SMOTE	Preprocessing
SPegasos	Classification
SSF	Attribute Selection
SVMAttributeEval	Attribute selection
SelfOrganizingMap	Clustering
SparseGenerativeModel	Text classification
TPP	Visualization
WekaExcel	Converter
WekaODF	Converter
XMeans	Clustering
alternatingDecisionTrees	Classification
anonymizationPackage	Preprocessing
associationRulesVisualizer	Visualization
attributeSelectionSearchMethods	Attribute selection
averagedOneDependenceEstimators	Classification
bayesianLogisticRegression	Text classification
bestFirstTree	Classification
cascadeKMeans	Clustering
cassandraConverters	Converters
chiSquaredAttributeEval	Attribute selection
citationKNN	Multi-instance learning
classAssociationRules	Associations
classificationViaClustering	Classification
classificationViaRegression	Classification
classifierBasedAttributeSelection	Attribute selection
classifierErrors	Visualization

clojureClassifier	Classification
complementNaiveBayes	Classification
conjunctiveRule	Classification
consistencySubsetEval	Attribute selection
costSensitiveAttributeSelection	Attribute selection
dagging	Ensemble learning
decorate	Ensemble learning
denormalize	Preprocessing
ensembleLibrary	Ensemble learning
ensemblesOfNestedDichotomies	Ensemble learning
extraTrees	Classification
fastCorrBasedFS	Attribute selection
filteredAttributeSelection	Attribute selection
functionalTrees	Classification
fuzzyLaticeReasoning	Classification
fuzzyUnorderedRuleInduction	Classification
gaussianProcesses	Regression
generalizedSequentialPatterns	Associations
grading	Ensemble learning
gridSearch	Classification
hiddenNaiveBayes	Classification
hiveJDBC	Misc
hotSpot	Associations
hyperPipes	Classification
isolationForest	Outlier
isotonicRegression	Regression
jfreechartOffscreenRenderer	KnowledgeFlow
jsonFieldExtractor	Knowledge Flow
kernelLogisticRegression	Classification
kfGroovy	KnowledgeFlow
kfKettle	KnowledgeFlow
kfPMMLClassifierScoring	KnowledgeFlow
latentSemanticAnalysis	Preprocessing
lazyAssociativeClassifier	Classification
lazyBayesianRules	Classification
leastMedSquared	Regression
levenshteinEditDistance	Distance measure
linearForwardSelection	Attribute selection
localOutlierFactor	Outlier
massiveOnlineAnalysis	Data streams
metaCost	Classification
multiBoostAB	Ensemble learning
multiInstanceFilters	Preprocessing
multiInstanceLearning	Multi-instance learning
multiLayerPerceptrons	Classification/Regression
multilayerPerceptronCS	Classification
naiveBayesTree	Classification
normalize	Preprocessing
oneClassClassifier	Classification
optics_dbScan	Clustering
ordinalClassClassifier	Classification
ordinalLearningMethod	Classification

ordinalStochasticDominance	Classification
paceRegression	Regression
partialLeastSquares	Preprocessing
predictiveApriori	Associations
prefuseGraph	Visualization
prefuseGraphViewer	KnowledgeFlow
prefuseTree	Visualization
probabilisticSignificanceAE	Attribute Selection
raceSearch	Attribute Selection
racedIncrementalLogitBoost	Ensemble learning
realAdaBoost	Ensemble learning
regressionByDiscretization	Regression
ridor	Classification
rotationForest	Ensemble learning
sasLoader	Converter
scatterPlot3D	Visualization
scriptingClassifiers	Classification
sequentialInformationalBottleneckClusterer	Clustering
simpleCART	Classification
simpleEducationalLearningSchemes	Classification
stackingC	Ensemble learning
supervisedAttributeScaling	Preprocessing
tabuAndScatterSearch	Attribute selection
tertius	Associations
thresholdSelector	Classification
timeseriesForecasting	Time series
userClassifier	Classification/regression
votingFeatureIntervals	Classification
wavelet	Preprocessing
wekaServer	Server
winnow	Classification

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı : Fırat AKBA

Doğum Yeri : Diyarbakır

Medeni Hali : Bekâr

E-posta : firatakba@gmail.com

Adresi : Höyük cd. Bağlıca Mah. No:33/19 Etimesgut/ANKARA

Eğitim

Lise : Özel Dicle Lisesi

Lisans : Çankaya Üniversitesi

Yüksek Lisans : Hacettepe Üniversitesi

Yabancı Dil ve Düzeyi

İngilizce: Okuma, yazma – İyi

İş Deneyimi

Deneyim Alanları

Yazılım, Duygu Analizi, Makine Öğrenimi Yöntemleri, Bilgi Erişimi Sistemleri, Veri Madenciliği, Bilgi ve İletişim Teknolojileri

Tezden Üretilmiş Projeler ve Bütçesi

Tezden Üretilmiş Yayınlar

- Akba, F; Uçan, A.; Sezer, E. A. & Sever, H.: Assessment of feature selection metrics for sentiment analyses: Turkish movie reviews, 8th European Conference on Data Mining, 180-184, **2014**.

Tezden Üretilmiş Tebliği ve/veya Poster Sunumu ile Katıldığı Toplantılar

- The European Conference on Data Mining (ECDM'14)