T.C
MALTEPE ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

TÜRKÇE METİNLERDE DUYGU ANALİZİ

ENDER AHMET YURT

Yüksek Lisans Tezi

Tez Danışmanı

Yrd. Doç. Dr. Volkan TUNALI

İSTANBUL - 2015

T.C.
MALTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

TÜRKÇE METİNLERDE DUYGU ANALİZİ

YÜKSEK LİSANS TEZİ

ENDER AHMET YURT

Tez Danışmanı
Yrd. Doç. Dr. Volkan TUNALI

İSTANBUL – 2015

ÖZET

Yüksek Lisans Tezi, Türkçe Metinlerde Duygu Analizi, Maltepe Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı.

Bu tez çalışmasında Türkçe metinlerde önceden tasarlanmış Doğal Dil İşleme algoritmalarını kullanarak duygu analizlerinin başarıları test edilmiştir. Türkçe metinler ağ ortamından çekilip, ön işlemelerden geçtikten sonra Veri Madenciliği ve Makine Öğrenmesi konularında yardımcı olan araçlar ile analizler yapılmış ve çıkan sonuçlar tartışılmıştır.

Bu tez dört bölümden oluşmaktadır. Birinci bölümde tez ile ilgili genel kavramların açıklamalarına yer verilmiştir. İkinci bölümde literatür taraması, Duygu Analizinin çalışma alanları ve tezde kullanılan araçlardan bahsedilmiştir. Üçüncü bölümde ise tezde kullanılan verinin çevrimiçi ortam çekilmesi, ön işlenmesi ve Duygu Analizinde yapılan deneyler ve bu deneylerin sonuçları açıklanmıştır. Son bölümde deneylerin sonuçları karşılaştırılmış, değerlendirmeler yapılmış ve gelecek çalışmalar için öneriler sunulmuştur.

Bu tez 2015 yılında yapılmıştır ve 38 sayfadan oluşmaktadır.

Anahtar kelimeler: Veri Madenciliği, Metin Madenciliği, Türkçe, Duygu Analizi, Düşünce Analizi, Makine Öğrenmesi, Doğal Dil İşleme

ABSTRACT

Master Thesis, Sentiment Analysis in Turkish Documents, Maltepe University, Institute of Natural Sciences, Department of Computer Engineering.

In this master thesis, The Sentiment Analysis success was tested for Turkish documents by NLP (Natural Language Processing) Algorithms that were designed before. Turkish documents were fetched from web pages and after the processing, NLP and other Machine Learning software tools were used.

This thesis consists of 4 sections. The first section includes the general information about the thesis. In the second section, the previous studies about the general concepts are presented and the tool which are used in the thesis. The data which that used in the thesis and preprocessing process on it and the experiments about Sentiment Analysis are explained and and their results at the third section. The last section includes the result of the all experiments, comparison of them and the suggestions for the future academic works.

This thesis has been completed in 2015 and consists of 38 pages.

Keywords: Data Mining, Text Mining, Turkish, Sentiment Analysis, Opinion Mining, Machine Learning, Natural Language Processing.

TEŞEKKÜR

Tez konusunu seçmemde beni yönlendiren, tez süreci boyunca destek ve yardımlarını esirgemeyen, değerli bilgilerinden istifade ettiğim danışman hocam Yrd. Doç. Dr. Volkan Tunalı'ya, tez süresince yardımlarını aldığım insanlara ve ilerlediğim bu yolda her zaman yanımda olup, beni hiç yalnız bırakmayan aileme sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

ŞEKİLLER

DENKLEMLER

ÇİZELGELER

KISALTMALAR

Kısaltma	İngilizcesi	Türkçesi
NLP	Natural Language Process	Doğal Dil İşleme
IMDB	Internet Movie Database	Internet Film Veri tabanı
PDF	Portable Document Format	Taşınabilir Belge Biçimi
HTML	Hyper Text Makeup Language	Hiper Metin İşaretleme Dili
XML	Extensible Markup Language	Genişletilebilir İşaretleme Dili
TF	Term Frequency	Terim Frekansı
TFIDF	Term Frequency Inverse Document Frequency	Terim Frekansı Ters Metin Frekansı
TFIDF-NORM	Term Frequency Inverse Document Frequency Normalization	Terim Frekansı Ters Metin Frekansı Normalizasyonu
PMI	Pointwise Mutual Information	Karşılıklı Bilgi
POS	Part of Speech	Cümlenin Öğeleri
CSV	Comma-Separated Values	Virgül ile Ayrılmış Değerler
ARFF	Attribute Relation File Format	Özellik İlişkili Dosya Formatı
CSS	Cascading Style Sheets	Basamaklı Stil Sayfaları

1. GİRİŞ

Her geçen gün atılan tweet'ler, yazılan bloglar, haberler, sosyal medya paylaşımları, filmlere ve ürünlere yapılan yorumlar, çevrimiçi ortamda ciddi miktarda verinin birikmesine neden olmaktadır. Bu verilerin alınıp en doğru şekilde işlenerek analiz edilmesi ve yorumlanması gerekmektedir. Amerika'da yapılan bir çalışmaya göre orada yaşayan insanların %73'ü çevrimiçi ortamda okudukları ürün yorumlarını ciddiye almaktadır [1]. Özellikle büyük şirketler bu bilgilerin yorumlanıp, analiz edilmesine oldukça özen göstermektedirler. Şirketler, ürünleri ve kendileri hakkındaki bu yorumları internetten hızlı bir şekilde elde edip, analiz etmek için Duygu Analizini kullanmaya başlamışlardır. Duygu Analizi üzerinde daha çok pazarlama, insan ilişkileri ve reklam firmaları çalışmaktadırlar. Hatta Duygu Analizi, son zamanlarda politikada bile kendine yer bulmaktadır. Siyasiler, seçimlerden önce çevrimiçi ortamda kendileri hakkındaki yorumları önemsemekte ve seçim taktilerini bu yorumlara göre belirlemektedirler.

Son zamanların en popüler araştırma alanlarından bir olan Duygu Analizi hakkında yayınlanmış 7000'den fazla makale bulunmaktadır. Birçok yeni girişim şirketi ve köklü firma, yeni çözümler üretmek adına ciddi yatırımlar yapmakta ve yeni departmanlar kurarak Duygu Analizi konusunda çalışmalarına ağırlık vermektedirler [8].

Duygu Analizi (Sentiment Analysis / Opinion Mining) konusunda yapılan birçok çalışmada bir ürünün ya da servisin yorumlarını analiz etmek, açıklayabilmek için basit ifadeler kullanır. Buna rağmen diller arası farklılıklar ve bir kelimenin bir cümle içindeki kullanımına göre farklı anlamlara gelmesi, yazı dilindeki kültürel farklar ve kelime kullanımlarına göre içeriğin farklılaşması Duygu Analizi çalışmaları için birer problem haline gelebilmektedir [9]. Çoğu zaman bir cümle içindeki kelimelerden anlam bütünlüğü çıkarmak bir dokümana bakıp yorumlamaktan daha zor hale gelebilmektedir. Duygu ve düşüncelerini metin

dokümanlarına yansıtan insanların yazdıklarını analiz etmek ve doğru kararı vermek bilgisayarlar için her zaman kolay olmamaktadır.

Veri kirliliği günümüzün en büyük sorunudur. Doğru veriyi bulmak, doğru şekilde işlemek ve analiz etmek çok kolay değildir. Çevrimiçi ortama göre İngilizce içerikli veriye ulaşmak ve onu işlemek diğer dillere nispeten daha kolay olmaktadır. İngilizce içerikli olarak yapılan Duygu Analizi çalışmaları, Türkçe içerikli yapılan çalışmalara göre sayıca daha fazladır. Bu durumun en temel nedenlerinden biri Türkçe metinler üzerinde dil bilgisi çalışmalarının azlığıdır. Üstünde akademik çalışmalar yapılacak Türkçe içerikli dokümanların yapısal olarak bozukluğu da çalışmaların azlığına neden olmaktadır. Bu tez çalışmasında İngilizce metinler üzerinde uygulanmış teknikler, Türkçe metinler üzerinde uygulanmıştır.

Duygu Analizi birçok çalışmanın bir araya gelmesi ile oluşmuştur. Asıl olarak Metin Madenciliği yöntemlerinden faydalanılsa da Duygu Analizinde Veri Madenciliği, Makine Öğrenmesi ve Doğal Dil İşleme yöntemleri oldukça sık kullanılmaktadır. Verinin bir kaynaktan çekilmesi, işlenmesi ise Metin Madenciliği çalışma alanlarının konusuna girmektedir. Ayrıca dokümanlardaki kelimelerin analiz edilmesi, Doğal Dil İşleme ve bu sürecin otomatik hale getirilerek bilgisayarlar tarafından yapılabilmesi Makine Öğrenmesinin çalışma alanına girmektedir.

Bu tez çalışmasında, Metin Madenciliği ve Makine Öğrenmesi Teknikleri kullanılarak Türkçe metinler üzerinde Duygu Analizi çalışmaları yapılmıştır. Veri Madenciliği'nin alt dallarından biri olan Metin Madenciliği, önceden bilinmeyen ancak kullanılabilir ve üzerinde çalışıldığı zaman anlamlı olacak bilgileri çok büyük veriler içinden çıkarma yöntemidir [2]. Bu tez çalışmasında çevrimiçi ortamdan çekilen büyük ve kirli veriden, Metin Madenciliği yardımıyla anlamlı veriler elde edilmeye çalışılmış ve makine öğrenmesi algoritmaları kullanılarak da Duygu Analizi çalışmaları yapılmıştır.

Makine Öğrenmesi alanında üç yaygın yöntem; Naïve Bayes, Maksimum Entropi (Maximum Entropy Classification) ve Destek Vektör Makinesi (Support

Vector Machine) yöntemleridir. Bu tez çalışmasında Naïve Bayes yöntemi kullanılarak Türkçe metinler üzerinde deneyler gerçekleştirilmiştir.

2. LİTERATÜR TARAMASI

2.1. Veri Madenciliği

Veri Madenciliği, birçok disiplinin bir araya gelmesi ile oluşmuş bir araştırma sahasıdır. Bu disiplinler arasında; Yapay Zeka, Makine Öğrenmesi, İstatistik ve Veri Tabanı Sistemlerine Erişim Yöntemleri yer almaktadır [3]. Son zamanlarda çevrimiçi ortamdaki verinin öneminin artmasından dolayı Veri Madenciliği'ne olan ilgi de aynı ölçüde artmış ve artmaya devam etmektedir.

Karmaşık veriden anlamlı bir bilgi çıkarmak için izlenmesi gereken adımlar şunlardır;

1. Veri Seçimi: Üzerinde çalışılacak verinin veri tabanından ya da herhangi bir kaynaktan alınmasıdır.
2. Veri Entegrasyonu: Eğer veri birden çok kaynaktan alınıyorsa, bu verilerin birbirleri ile olan birleştirme işlemidir.
3. Veri Temizleme: Veri içindeki tutarsızlıkların ve gürültünün giderilmesidir.
4. Veri Dönüştürme: Verinin özetleme veya derleme işlemlerine tabi tutularak, kullanıma uygun hale getirilmesidir.
5. Veri Madenciliği: Veri örüntülerini ortaya çıkarmak için akıllı yöntemlerin uygulandığı önemli bir süreçtir.
6. Örüntü Değerlendirmesi: Bilgiyi temsil eden ilginç örüntülerin özel ölçümlere dayanarak belirlenmesi işlemidir.
7. Bilgi Sunumu: Ortaya çıkarılan bilginin görselleştirme ve bilgi sunum yöntemleri kullanılarak kullanıcıya gösterilmesi adımıdır [10].

1. ve 4. adımlar arası "Veri Ön İşleme Süreci" olarak adlandırılmaktadır. Bu süreç, üzerinde madencilik yapılacak verinin temizlenmesi ve işleme hazır hale getirilmesi için gereken adımları içermektedir. "Veri Madenciliği" adımı ise kullanıcı ile etkileşimli bir şekilde gerçekleştirilebilir. Bu adımdan sonraki örüntülerin

değerlendirilmesi aşamasında, ilgi alanı dışındaki örüntüler belirli ölçümlerle ayıklanır. Önemli olanlar ise bir sonraki aşamada kullanıcıya değişik yollarla gösterilebilir. Veri Madenciliği, çok büyük veri kümelerinde standart yöntemlerle görülemeyecek bilgi ve örüntüleri ortaya çıkardığı için önemlidir. Veri Madenciliği genellikle küçük veri kümeleri ile ilgilenmez [11].

2.2. Metin Madenciliği

Metin Madenciliği, farklı yazılı kaynakların bir araya getirdiği verinin, otomatik olarak alınması ve o veri içinde yeni bir bilgi keşfetme işidir [4]. Asıl amacı belli bir metin üzerinde belli bir yapısı olan veriyi bulup, o verinin ilgili metin içinden çıkarılmasıdır. Metin Madenciliği Teknikleri dört temel kategoriye ayrılır: Sınıflandırma (Classification), Birliktelik Analizi (Association Analysis), Bilgi Çıkarım (Information Extraction) ve Kümeleme (Clustering). Sınıflandırma işlemi, nesnelerin daha önceden bilinen sınıflara ya da kategorilere dahil edilmesidir. Birliktelik Analizi ise sıklıkla birlikte yer alan ya da gelişen sözcük veya kavramların belirlenmesini amaçlar. Böylece doküman içeriğinin ya da doküman kümelerinin anlaşılmasını sağlar. Bilgi Çıkarım Teknikleri yardımlarıyla dokümanların içerisindeki yararlı veri ya da ifadeler bulunmaya çalışılır. Kümeleme Analizi, doküman kümelerinin temelini oluşturan yapıların keşfedilmesi amacıyla uygulanmaktadır [13].

Metin Madenciliği çalışmaları, metin kaynaklı literatürdeki diğer bir çalışma alanı olan Doğal Dil İşleme (Natural Language Processing, NLP) çalışmaları ile çoğu zaman beraber yürütülmektedir. Doğal Dil İşleme çalışmaları daha çok Yapay Zeka altındaki dil bilimine dayalı çalışmaları kapsamaktadır. Metin Madenciliği çalışmaları ise daha çok istatistiksel olarak metin üzerinden sonuçlara ulaşmayı hedefler. Metin Madenciliği çalışmaları sırasında çoğu zaman Doğal Dil İşleme Teknikleri kullanılarak, özellik çıkarımı yapılmaktadır [5].

2.3. Makine Öğrenmesi

Bilgisayar bilimlerindeki çoğu araştırma ve geliştirme çalışmalarında Makine Öğrenmesi algoritmaları kullanılır. Ayrıca Makine Öğrenmesi, İstatistik, Olasılık Kuramı gibi diğer alanlarla iç içe geçmiştir. En popüler uygulama alanları olarak Doğal Dil İşleme (Natural Language Processing), Bilgisayarlı Görme (Computer Vision) ve Arama Motorları gösterilebilir. Makine Öğrenmesi'nde kullanılan teknikler, Karar Ağacı Öğrenmesi (Decision Tree Learning), Birliktelik Kuralı Öğrenmesi (Association Rule Learning), Yapay Sinir Ağları (Artificial Neural Networks), Destek Vektör Makineleri (Support Vector Machines), Bayes Ağları (Bayesian Networks), Kümeleme (Clustering) ve Genetik Algoritmalar (Genetic Algorithms) olarak sıralanabilir [6].

2.4. Duygu Analizi

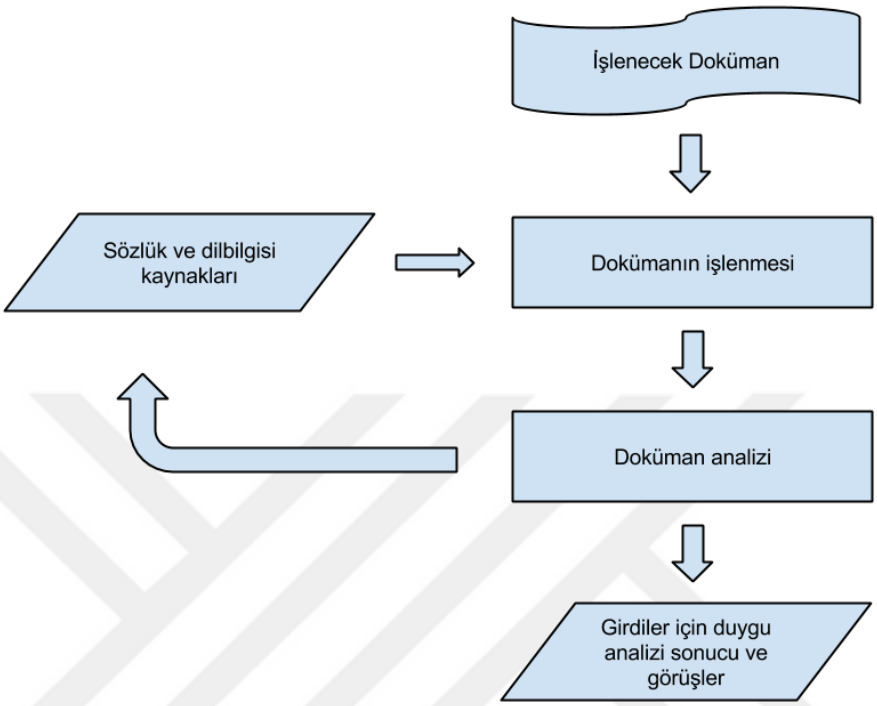
Duygu Analizi'nin amacı bir konu hakkında yapılan yorumların pozitif mi, negatif mi, tarafsız mı olduğunu belirlemektir. Bu yorumlar yazılı ya da sözlü olabilir. Dokümanların konusu hakkında ise herhangi bir sınırlama yoktur. Buradaki asıl amaç, makinelere bu yorumların pozitif mi negatif mi olduğunu, bir insan gibi tahmin ettirebilmek ve bunu otomatik olarak yapabilmesini sağlamaktır.

İnternetin gelişmesi ile Duygu Analizi konusunda çalışmalar son zamanlarda hız kazanmıştır. Özellikle Web 2.0 ile dinamik içerikli sayfaların üretilmesi ve veri tabanına olan önemin artması, veriyi daha anlamlı ve kullanışlı hale getirmiştir. 2002 yıllında Bo Pang ve Lillian Lee tarafından "Thumbs up? Sentiment Classification using Machine Learning Techniques" başlıklı çalışması Duygu Analizi konusunun yapı taşlarından biri olarak görülmektedir [19]. Bu çalışmada Internet Film Veri Tabanı (IMDB Internet Movie Database)'dan veriler çekilmiş ve Naïve Bayes, Destek Vektör Makinesi (Support Vector Machine), Maksimum Entropi, Maksimum Entropi (Maximum Entropy Classification) Makine Öğrenmesi Algoritmaları ile Duygu Analizi çalışmaları yapılmıştır. İngilizce dışında başka dillerde farklı çalışmalar da gerçekleştirilmeye çalışılmıştır [15-17]. Türkçe için Duygu Analizi

konusunda şimdiye kadar yapılmış akademik çalışma sayısı çok azdır. Türkçe metinler üzerinde yapılan bir tez çalışmasında İngilizce çalışmalarda denenen yöntemler, iki yeni Türkçe veri seti üzerinde denenmiş ve %85 başarı sağlanmıştır [18].

Genel yorumlar ve görüşler için Duygu Analizinin yapılması zor ve anlamsızdır. Kişisel görüşlerin belirtildiği dokümanları yorumlamak ve analiz etmek Duygu Analizi'nin çalışma alanına girer. Örneğin, bir kullanıcı bir cep telefonu hakkında "*Ciddi anlamda güzel bir telefon. Hem hafif, hem çok dayanıklı. Hep böyle bir telefona sahip olmak istemiştim. Fiyatı da performansına göre çok uygun. Bir kaç gün kullandıktan sonra batarya süresinin çok gitmediğini fark ettim. Böylesi bir telefonun batarya süresinin neden az olduğunu anlamıyorum. Ekran çözünürlüğünün diğer telefonlardan yüksek olması iyi ancak batarya süresinin kısalığı kullanımı zaman zaman zorlaştırıyor. Eğer yeni bir telefon almak istiyorsanız bu telefonu deneyebilirsiniz.*" yorumunu yapmış olabilir. Burada Duygu Analizi konusunda birçok çıkarım yapılabilir. Cep telefonu hakkındaki iyi ve kötü yorumlar bu metin içinde mevcuttur. Ancak cümle olarak bakıldığında, yorumların hem iyi hem de kötü olduğu görülmektedir. Ayrıca bahsedilen telefon diğer telefonlarla karşılaştırılmıştır. Telefonun bazı özellikleri kendi içindeki diğer özellikleri ile de karşılaştırılmıştır. Duygu Analizi çalışmalarını yapmadan önce doküman üzerinde ne gibi bir yaklaşım sergileneceğini belirlenirse daha iyi sonuç alınacağı kesindir.

Duygu Analizi çalışmalarına herhangi bir dosya türünde (pdf, html, xml, word vb.) doküman dizisi alınarak başlanır. Bu doküman girdisi, ön işleme yöntemlerinden olan hecelere ayırma, kelime grubu etiketleme, bilgi çıkarma ve kelimeler arası ilişki kurma kullanılarak sadeleştirilir. Dokümanı daha anlaşılır kılmak için dokümanın kendi diline ait sözlükler ve dil ile alakalı diğer kaynaklar kullanılabilir. Sadeleşen ve sözlükler üzerinden analizleri yapılan doküman üzerinde artık hangi Duygu Analizi Yaklaşımı uygulanacağına karar verilir. Bu yöntemler araştırmanın şekline göre farklılık gösterebilir. Uygulanan yöntemden sonra çıkan veri, son kullanıcının anlayacağı şekilde hazırlanır ve sunulur. Duygu Analizi çalışmalarının akış diyagramı Şekil 2.1'deki gibidir.



Şekil 2.1 Duygu Analizi Sistem Mimarisi [12]

Duygu Analizi'nin alt çalışma alanlarını, Doküman Seviyesinde (Document-Level), Cümle Seviyesinde (Sentence-Level), Özellik Temelli (Aspect-Based) ve Karşılaştırmalı (Comparative) olarak sıralayabiliriz.

2.4.1. Doküman Seviyesinde Duygu Analizi (Document-Level)

Duygu Analizinin en yayın olduğu çalışma alanıdır. Analiz yapılırken bir dokümanın bütünü ele alınır ve ona göre çalışmalar yapılır. Doküman Temelli Duygu Analizi, Denetimli Öğrenme (Supervised Learning) ve Denetimsiz Öğrenme (Unsupervised Learning) yaklaşımları olarak ikiye ayrılır.

Denetimli Öğrenme alanında sonlu bir doküman kümesinden elde edilen eğitim verisi ile çalışmalar yapılır. Bu durumda bir dokümanı pozitif, negatif ya da tarafsız olarak yorumlamak kolaydır. Sınıflandırmalar yapılırken Destek Vektör Makinesi

(Support Vector Machines) ve Naïve Bayes gibi Makine Öğrenmesi algoritmalarından yararlanılır. “Thumbs Up? Sentiment Classification Using Machine Learning Techniques” başlıklı araştırmada [19] Amazon’daki ürün yorumlarına bakılarak çıkarılan bag of words verisi ile gayet iyi bir kesinlik oranı (accuracy) sonucu elde edilmiştir. Dokümanlar üzerinde Terim Frekansı Ters Metin Frekansı (Term Frequency Inverse Document Frequency), Cümlenin Öğeleri (Part of Speech) ve Duygu Sözlüğü (Sentiment Lexicons) gibi yaklaşımlar daha ileri çalışma teknikleri olarak gösterilebilir.

Denetimsiz Öğrenme (Unsupervised Learning) ise Denetimli Öğrenme’den farklı olarak, verileri önceden etiketlemek yerine veri içerisinde bulunan yapıların öğrenilmesidir. En yaygın yöntemlerden biri olan PMI (Pointwise Mutual Information), verilen bir kelime grubunun doküman içindeki kelimeler ile olan farklılıklarının hesaplanmasıdır [20].

2.4.2. Cümle Seviyesinde Duygu Analizi (Sentence-Level)

Doküman Seviyesinde Duygu Analizi yaklaşımının benzeri bir yaklaşıma sahip olan Cümle Temelli Duygu Analizinde çalışmalar daha çok pozitiflik ve negatiflik üzerinden yapılmaktadır. Yöntem olarak ise Denetimli Öğrenme Yöntemi yaygın olarak kullanılmaktadır [21]. Daha basit bir yaklaşım ise bir dokümandaki cümleleri tek tek ele almaktır. Doküman içindeki cümleler ve onları takip eden cümlelerin birbirleri ile içerik olarak alakalı olmasından dolayı bu cümleleri birlikte ele almak daha az veri ile daha hızlı çalışmayı mümkün kılmaktadır [22, 23]. Fakat günümüzdeki çalışmalar gösteriyor ki her cümlenin yapısı aynı değildir. Özellikle iğneleyici (sarcasm) yaklaşımları olan cümleler üzerinde Duygu Analizinin yapılması oldukça zordur. Bu konuda Tsur tarafından yapılmış bir çalışma mevcuttur [24].

2.4.3. Özellik Temelli Duygu Analizi (Aspect-Based)

Çevrimiçi ortamda insanlar ürün yorumlarını sadece pozitif ve negatif yorumlar halinde yapmak yerine yorum yaptıkları ürünün özellikleri hakkında yapabilirler. Özellikle tartışma forumlarında ürünlerin sekmelerinde bu tür yorumlarla karşılaşmak mümkündür. Eğer bu yorumlara pozitif mi negatif mi olarak bakılırsa, bu durum yanıltıcı olabilir. Bir ürünün özelliklerinin üzerine yorumların yapıldığı dokümanlarda Özellik Temelli Yaklaşım ile Duygu Analizi yapmak en doğrusudur. En yaygın çalışmalardan biri, dokümanlardan ürün ile alakalı isim öbeklerini (noun phrase) çıkararak bir yaklaşım elde etmektir [25].

2.4.4. Karşılaştırmalı Duygu Analizi (Comparative):

Kimi zaman kullanıcılar bir ürün hakkında doğrudan yorum yapmaktansa, başka ürünler ile yorum yapacakları ürünü karşılaştırırlar. Karşılaştırmalı Duygu Analizi çalışma alanında hedef, cümleler içinde karşılaştırma kelimelerine odaklanarak nasıl karşılaştırmalar yapıldığını bulabilmektir. İngilizce birçok karşılaştırma kelimesini çıkarmış ve karşılaştırma üzerine çalışma yapmış olan Jindal ve Liu'ya [26] göre 'more', 'less' gibi ifadeleri belirtmek için kelimeler '-er' eki ile biter. Diğer taraftan 'most', 'least' gibi ifadeler için ise kelimeler '-est' ile biter. Bütün bunların dışında özel durumlar için de kelime kümeleri çıkartılmıştır.

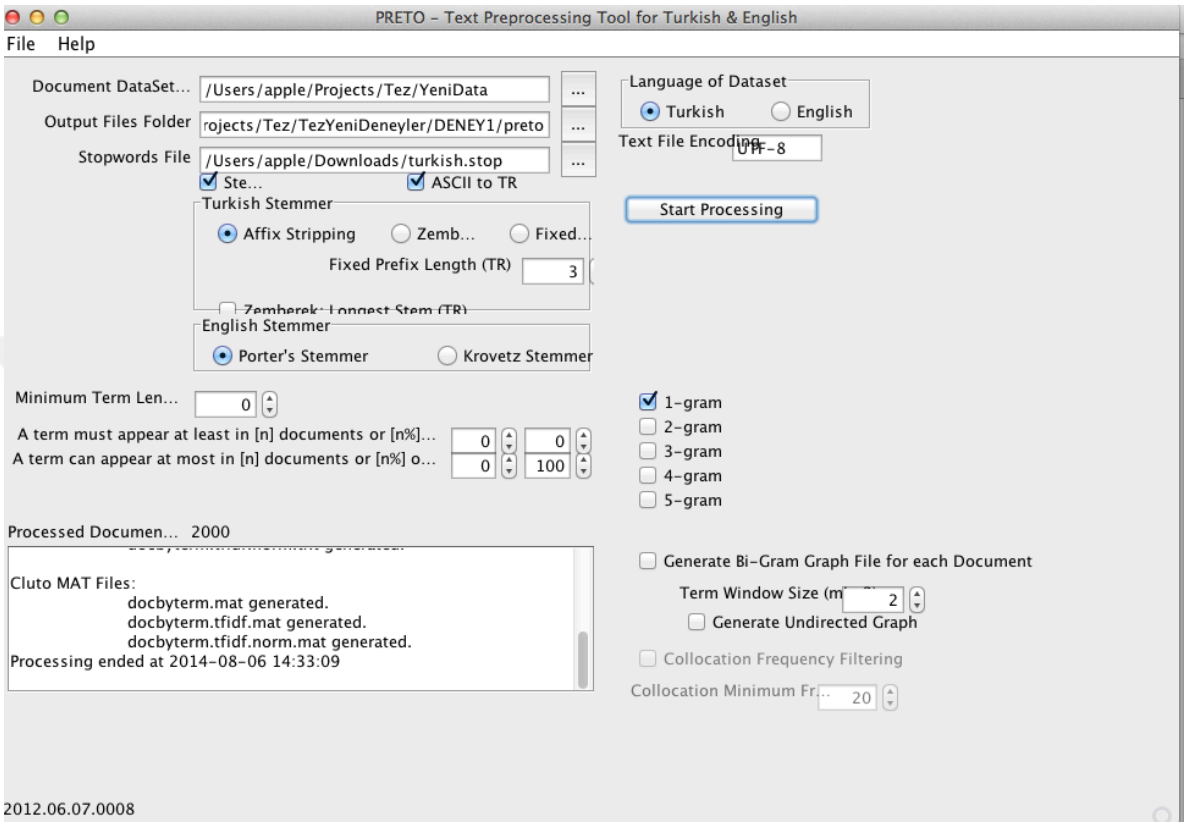
2.5. Kullanılan Araçlar

Bu bölümde tez çalışmasında kullanılan araçlardan bahsedilmiştir.

2.5.1. Preto

Preto, Türkçe metinlerde Metin Madenciliği'nin ön işleme operasyonları için geliştirilmiş açık kaynak kodlu ve platformdan bağımsız çalışabilen bir araçtır. İçerisinde Kök Bulma (Stemming), Durak Sözcük Filtreleme (Stopword Filtering), İstatistiksel Terim Filtreleme (Statical Term Filtering) ve N-gram oluşturma gibi birçok metin işleme teknikleri bulunur [27]. Preto, farklı terim ağırlıklandırma

yöntemlerine göre çıktılar üretebilmekte ve bunları bu isimlere sahip dosyaları çıktı olarak vermektedir.



Şekil 2.2 Preto Ekran Görüntüsü

2.5.2. WEKA

Makine Öğrenmesinin en popüler araçlarından biri olan WEKA (Waikato Environment for Knowledge Analysis) hemen hemen birçok makine öğrenmesi algoritmasını bünyesine barındırır. Java Programlama Dili ile geliştirilmiştir ve açık kaynak kodludur.

WEKA, tamamen modüler bir tasarıma sahip olup, içerdiği özelliklerle veri kümeleri üzerinde görselleştirme, veri analizi, iş zekası uygulamaları gibi işlemler yapabilmektedir. WEKA yazılımının kendisine özgü olarak bir .arff dosya biçimi vardır. Ayrıca WEKA yazılımının içerisinde CSV dosyalarını da ARFF dosya formatına çevirmeye yarayan özellikler mevcuttur.

Temel olarak aşağıdaki üç Veri Madenciliği işlemi WEKA ile yapılabilir:

- Sınıflandırma (Classification)
- Kümeleme (Clustering)
- Birliktelik Kuralı Analizi (Association Rule Analysis)

Ayrıca yukarıdaki işlemlere ilave olarak, Veri Ön İşleme (Data Pre-Processing), Görselleştirme (Visualization) yardımıyla veri kümeleri üzerinde ön ve son işlemler yapılabilir. Son olarak WEKA kütüphanesinde veri kümelerini içeren dosyalar üzerinde çalışan çok sayıda hazır fonksiyon bulunmaktadır.

```
@relation havatahmini

@attribute nem numeric
@attribute sıcaklık numeric
@attribute basınç numeric
@attribute tahmin numeric

@data 53,25,1013,1 41,22,1011,-1 54,18,1012,-1 67,23,1000,1
```

Şekil 2.3 Arff Veri Dosyası Örneği [28]

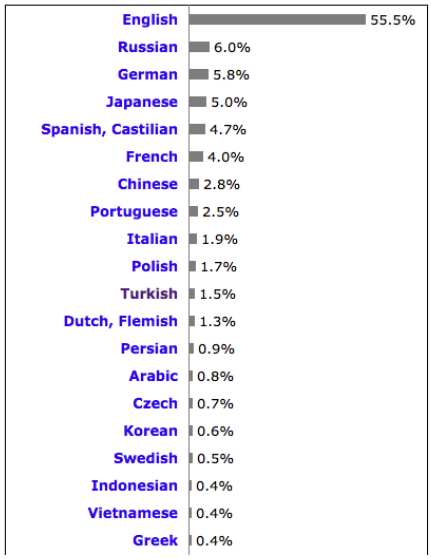
ARFF (Attribute Relationship File Format) dosya yapısı, WEKA'ya özel olarak geliştirilmiştir ve dosya metin yapısında tutulmaktadır. Dosyanın ilk satırında, dosyadaki ilişki tipi (relation) tutulmakta olup, ikinci satırdan itibaren de veri kümesindeki özellikler (attributes) ve türleri yazılmaktadır. Özelliklerin hemen ardından veri kümesi yer alır ve veri kümesindeki her satır bir örneği (instance) ifade etmektedir. Veri kümesindeki her örneğin her özelliği arasında virgül ayıracı kullanılmaktadır.

Şekil 2.3'teki örnek kodda, hava tahmini için kullanılan nem, sıcaklık ve basınç değerleri bir dosya içerisinde dört örnek içerecek şekilde gösterilmiştir. Bu değerler, tip olarak sayısal değerler olduğundan "numeric" olarak ifade edilmiştir. Bu değerler bir sonraki sayfadaki gibi tiplerde de olabilir:

- Küme Değerleri: Tahmin değeridir ve bir tanım kümesi alır. Örneğin, tahmin şeklinde tanımlanan bir değer, tanım kümedeki {güneşli, yağmurlu, sisli} değerlerinden birisini alabilir.
- Real: [Reel Sayılar] kümesinden bir değer verileceğinde kullanılır. Örneğin, sıcaklık değeri 22,8 şeklinde ondalıklı değer olarak ifade edilmek istenirse, tip olarak nümerik yerine reel kullanabiliriz.
- String: Veri kümesinin bu özelliğinin serbest yazı şeklinde olabileceğini ifade eder. Özellikle Metin Madenciliği çalışmaları için sıkça kullanılan bir tiptir.
- Date: Veri kümesinin bu özelliğinin tarih olduğunu ifade eder. Örneğin, veri kümesindeki kişilerin doğum tarihi veya örneklerin toplanma tarihi gibi özelliklerin tutulmasında kullanılabilir [28].

3. TÜRKÇE DOKÜMANLAR ÜZERİNDE DUYGU ANALİZİ

Çevrimiçi ortamda Türkçe içerik sayısının hızla artması [29] ve buna bağlı olarak şirketlerin kendi ürünleri hakkında yapılan yorumların ne gibi bir düşünce içerdiğine (pozitif/negatif) önem vermesi bu tezin motivasyonu olmuştur. İngilizce içerik yüzdesi göz ardı edildiğinde, Türkçenin çevrimiçi ortamdaki yerinin Şekil 3.1'de de görüldüğü gibi azımsanamayacak ölçüde iyi olduğu görülmektedir. Türkçe adına yapılan çalışmaların ışığında geliştirilen yeni yöntemler ile Duygu Analizi hakkında yapılan çalışmaların sayısı arttırılmaya çalışılmıştır.



Şekil 3.1 Web Sitelerinin Dilleri [29]

3.1. Çalışmada Kullanılan Veri

Bu tez çalışması için gerekli olan pozitif ve negatif içerikli verinin, film içeriklerine yapılan yorumların bulunduğu sitelerden elde edilebileceği düşünülmüştür. Türkiye'nin en büyük web sitelerinden biri olan mynet.com'un [30] sahip olduğu film içerik ve yorum sitesi olan beyazperde.com'dan kullanıcıların filmler hakkında yaptıkları yorumlar alınmıştır.

Şekil 3.2 Beyazperde.com Ekran Görüntüsü

Kullanıcılar beyazperde.com üzerinde ilgilendikleri filmler üzerine yorumlarını Şekil 3.2'deki görüldüğü gibi iletebilmektedirler. İsterlerse yorumladıkları filmleri oylayabilmektedirler. Bu yorumların direk olarak pozitif, negatif ya da tarafsız olarak yapılıyor olması bu tezin konusu için uygun bir veri kaynağı olduğunu göstermektedir.

3.2. Verinin Çekilmesi

Bu tez çalışmasında kullanılacak veri, siteye yapılan sorgular ile HTML (Hyper Text Markup Language) sayfaları halinde çekilerek ayıklanmış ve yorumların olduğu yerlerden metinler elde edilmiştir. Bu yöntemi uygulamak için Ruby Programlama Dili [31] kullanılmıştır. İstenilen verinin alınacağı alan belirlendikten sonra yorum ID'lerine göre web sitelerine 'get' metodu ile sorgu yapılmıştır. Beyazperde.com'da her filme bir ID atadığı için ilk olarak filmlerin kendilerine ait sayfalara gidilmiştir. Sonrasında dönen sayfadan yorumların olduğu ilgili link, sayfa tarama yöntemiyle bulunup, bilgisayar tarafından tetiklendikten sonra ilgili filmin kullanıcı

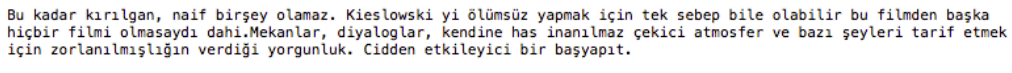
yorumlarının olduğu sayfaya gidilmiştir. Yorumlar ilgili tablodan alınmış ve daha önceden belirlenmiş yöntem ile ayıklanmıştır. Yorumların belli bir CSS (Cascading Style Sheet) class'ları içinde olduğunun anlaşılmasıyla o class içindeki metinler çekilmiştir. Alınan yorumlar herhangi bir ön işlem veya başka bir teknik kullanılmadan farklı dosyalar halinde yazılmıştır. Her yorumun kendine ait bir dosyası olması ve bütün dosya adlarının farklı olması gerektiğinden her dosya kendine ait bir ID ile isimlendirilmiştir.

```ruby
require 'rubygems'
require 'nokogiri'
require 'open-uri'

link = "http://www.beyazperde.com/filmler/"

START_MOVIE_ID  = 100000
FINISH_MOVIE_ID = 150000

comment_id = 3980
movie_count = 0

(START_MOVIE_ID..FINISH_MOVIE_ID).each do |movie_id|
  url = "#{link}film-#{movie_id}/kullanici-elestirileri/"
  puts url
  begin
    doc = Nokogiri::HTML(open(url))
    movie_count+=1
    doc.css(".box_06.margin_20b.j_entity_container").each do |item|
      file_path = "comments_part11/comment_#{comment_id}.txt"
      File.open(file_path, 'w') do |f|
        comment = item.css(".box_07.j_entity_container p").text()
        f.write(comment)
        comment_id+=1
      end
    end
  rescue
    puts "URL BROKEN!!"
  end
end

statistics = File.open('statistics_file_part11.txt', 'w') {|file| file.write("Movie Count = #{movie_count}, Total Comment = #{comment_id}")}

# 200000-280000
```

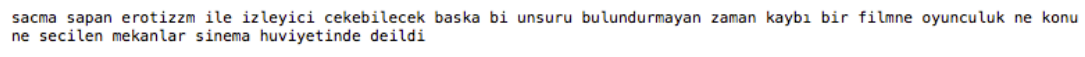
Şekil 3.3 Kullanıcı Yorumlarını Çeken Ruby Scripti

Yaklaşık 6000 adet yorum dosyası oluşturulmuştur. Bu dosyalar tek tek incelenmiş, hangi yorumun pozitif, hangisinin negatif olduğuna karar verilmiş ve yorumlar ayrıştırılmıştır. Bu yöntemde yorumlar tamamen insanlar tarafından okunmuş ve ayıklanmıştır. Burada amaç bir insanın bir yoruma verdiği kararı, bir makinenin ne kadar yüksek doğrulukta verebileceğini test etmek olduğu için verinin önce insanlar tarafından ayıklanması doğru bulunmuştur. Bütün yorumların pozitif ve negatif olmadığını düşündüğümüzde, pozitif ya da negatif kutbuna karar verilemeyen yorumlar tarafsız olarak değerlendirilmiştir.

Bu tezde başlangıç için 1000 pozitif ve 1000 de negatif olmak üzere toplam da 2000 adet yorum dosyası hazırlanmıştır.



Şekil 3.4 Beyazperde.com'dan Alınmış Pozitif İçerikli Bir Yorum



Şekil 3.5 Beyazperde.com'dan Alınmış Negatif İçerikli Bir Yorum

3.3. Verinin Ön İşlemesi (Pre-processing)

Çevrimiçi ortamdan çekilen verinin, çok dağınık ve karmaşık bir yapıda olduğu için verinin ön işlemeye sokulması ve temizlenmesi gerekliliği doğmuştur. Ön İşleme sadece verinin temizlenmesi değil, çalışmaya uygun hale getirilmesidir. Çalışmaya uygun şekilde bir verinin oluşturulması için öncelikle çalışmada nasıl bir veri kullanmak gerektiğine karar verilmiştir. Beyazperde.com'dan alınan pozitif ve negatif olarak ikiye ayrılan yorumlar bir araya getirilerek, bir dosya halinde deney durumlarına göre ön işleme tabi tutulmuştur.

Tekrarlanan kelimeler ve Türkçe'deki etkisiz kelimeler (stop word) çıkarımı için bir dosya (Ek-1) kullanılmıştır. Bu dosya Preto Programı'nın "stop word" kısmına verilmiştir.

Stemming için ise Prefix ve Suffix için kullanılan ve birçok Avrupa dili için uygulanan Affix Stemming Yöntemi [32] seçilmiştir. Stemming, aslında kök bulma tekniği olarak kullanılır. Bir kelime birçok kelimeye dönüşebilir ancak bu kelime temelinde aynıdır. Bu yöntem, Metin Madenciliğinde ön işleme sırasında en çok kullanılan yöntemlerdendir.

Preto'da dokümanın hangi N-gram seviyesinde ön işlemeye tabi tutulacağı da belirlenmiştir. Deneylerde farklı N-gramlarda alınan sonuçlar tartışılmıştır. N-gram, sıralı kelimelerin kaç tanesinin birlikte alınacağına karar vermeye yardımcı olur. "Ali eve gel çünkü dışarısı çok soğuk." Bu cümle için 1-gram tekniği uygularsak, "Ali",

“eve”, “gel”, “çünkü”, “dışarısı”, “çok”, “soğuk” şeklinde olacaktır ve bütün kelimeler kendi içlerinde anlamlarına göre yorumlanacaktır. 2-gram tekniğinde ise “Ali eve”, “eve gel”, “gel çünkü”, “çünkü dışarısı,” “dışarısı çok”, “çok soğuk” şeklinde gruplanacaktır. N-gram sayısını istediğimiz kadar arttırabiliriz ancak en yaygın kullanım 3-gram’a kadardır.

Birinci deney için ön işlemede Preto’ya verilen Türkçe veri setinin, belirlenen ön işleme kriterlerine göre işlenmesi sağlanmıştır. Sonrasında bir Doküman Terim Matrisinin çıkarılması beklenmiştir. Doküman Terim Matrisi Şekil 3.6’daki gibi matematiksel olarak bir dizi doküman içindeki terimlerin bir matriste gösterilmesidir. Genelde satırlar dokümanları, sütunlar ise terimleri ifade eder [33].

- D1= “I like databases”
- D2= “I hate databases”

	I	**like**	**hate**	**databases**
D1	1	1	0	1
D2	1	0	1	1

Şekil 3.6 Doküman Terim Matris Örneği [33]

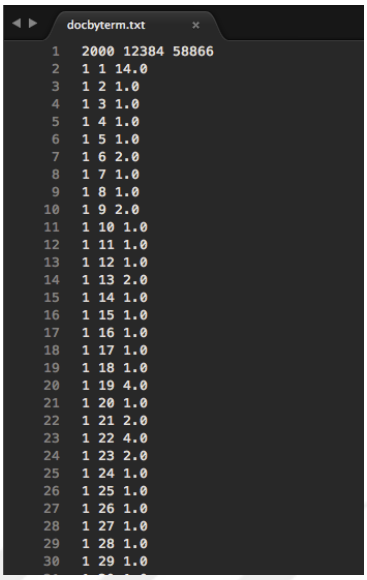
Preto’nun çıktı olarak verdiği dosyalar;

- **Docbyterm.mat:** Doküman Terim Matrisinin .mat olarak uzantılı bulunduğu dosyadır.
- **Docbyterm.tfidf.mat:** Doküman Terim Matrisi’nin Terim Frekansı - Ters Metin Frekansı (Term Frequency - Inverse Document Frequency)’nın .mat olarak uzantılı bulunduğu dosyadır. Terim Frekansı - Ters Metin Frekansı, bir terimin bir doküman içinde kaç kere geçtiğinin ve bütün dokümanlar içinde ne kadar sıklıkta geçtiğinin hesaplanmış halidir.
- **Docbyterm.tfidf.norm.mat:** Doküman Terim Matrisinin Terim Frekansı - Ters Metin Frekansının normalize edilmiş halinin .mat uzantılı olarak bulunduğu dosyadır.
- **Docbyterm.txt:** Şekil 3.7’deki gibi Doküman Terim Matrisinin bulunduğu dosyadır.

- **Docbyterm.tfidf.txt:** Doküman Terim Matrisinin Terim Frekansı - Ters Metin Frekansının normalize edilmiş halinin .txt uzantılı olarak bulunduğu dosyadır.
- **Docbyterm.tfidf.norm.txt:** Doküman Terim Matrisinin Terim Frekansı - Ters Metin Frekansının normalize edilmiş halinin .txt uzantılı olarak bulunduğu dosyadır.
- **Documents_fullname.txt:** Üzerinde çalışan dokümanların o makine üzerindeki açık adreslerini listeleyen dosyadır.
- **Documents.txt:** Üzerinde çalışılan dokümanların isimlerinin listeli olarak tutulduğu dosyadır.
- **Term_detailed.txt:** Hangi kelimeden toplamda kaç adet ve kaç tane doküman üzerinde geçtiğini tutan dosyadır.
- **Terms.txt:** İşlenen dokümanda geçen tüm kelimelerin listelendiği dosyadır.

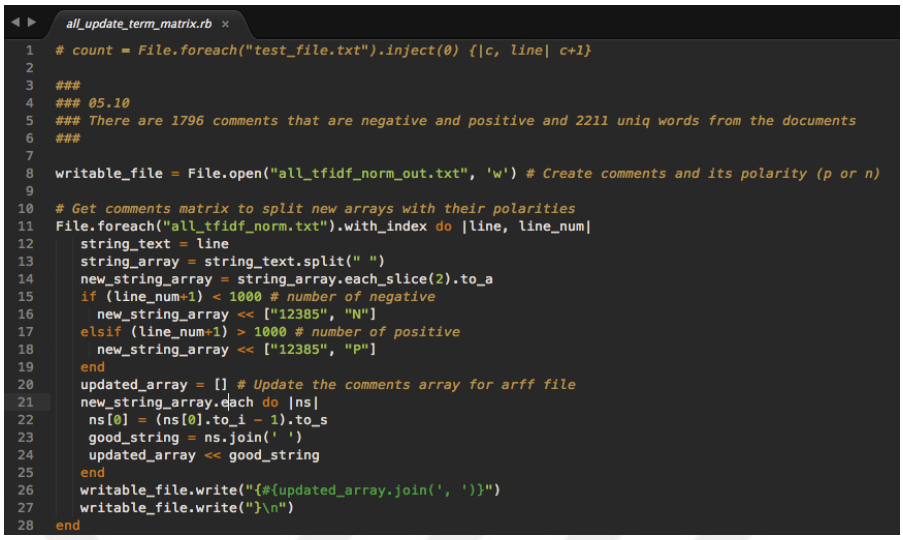
3.4. Deneyler

Dokümanları ön işleme sürecinden geçirdikten sonra WEKA ile analiz yapılmıştır. Daha önce de belirtildiği gibi WEKA ile analiz yapabilmek için dosyaların formatının .arff olması gerekmektedir. Eldeki .txt Doküman Terim Matris dosyalarını .arff uzantılı Şekil 3.8 ‘deki gibi bir dosya haline çevirmek için Ruby Programlama Dili ile çeşitli script’ler Şekil 3.9’daki gibi yazılarak, WEKA için bir format hazırlanmıştır.



Şekil 3.7 Preto'dan Çıkan docbyterm.txt Dosyası

Şekil 3.8 Ruby Script'i ile Düzenlenen docbyterm.txt Dosyası



Şekil 3.9 WEKA İçin Uygun Formata Çeviren Ruby Kodu

Bütün bu işlemler üç doküman terim matris türü için de tek tek yapılmıştır. Buradaki amaç WEKA ile kullanılan Naïve Bayes algoritmasının hep sabit tutulup, ön işleme adımlarının değiştirilerek hangi terim ağırlıklandırma yönteminin daha iyi sonuç vereceğini anlayabilmektir.

Bu tezdeki deneylerde, temelde Bayes Teorem'ini (Denklem 3.2) kullanan Naïve Bayes'in Sınıflandırıcı Yöntemi kullanılmıştır. Bu teknik adını İngiliz matematikçi Thomas Bayes'den (yak. 1701 - 7 Nisan 1761) alır. Naïve Bayes, sınıflandırıcı örüntü tanıma problemine kısıtlayıcı bir önerme getiren olasılıkçı bir yaklaşımdır. Bu önerme, örüntü tanımada kullanılacak her bir tanımlayıcı nitelik ya da parametrenin istatistik açıdan bağımsız olması gerekliliğidir. Her ne kadar bu önerme; Naïve Bayes Sınıflandırıcısının kullanım alanını sınırlasa da, genelde istatistik bağımsızlık koşulu esnetilerek kullanıldığında daha karmaşık Yapay Sinir Ağları gibi metotlarla karşılaştırabilir sonuçlar vermektedir [37]. Her özellik için sınıflar içinde bulunma olasılıkları ve sınıfların veri üzerinde görülme olasılıklarını hesaplayarak karar veren bir modeldir. "Koşullu Bağımsızlık Kabulü" ile bir özelliğin bir sınıfta belirli bir olasılıkla geçmesi, bir başka özelliğin aynı sınıfta geçiş olasılığından etkilenmez ve o olasılığı etkilemez [38].

T öğrenme kümesinde bulunan her örnek n boyutlu uzayda tanımlı olsun, X = (x_1, x_2, …, x_n). Veri kümesinde m adet sınıf bulunuyor olsun, C_1, C_2, …, C_m. Sınıflamada son olasılığı büyütme aranır (The Maximal $P(C_i|X)$) Bayes Teoremi'nden türetilebilir. P(X) olasılığı bütün sınıflar için sabit olduğuna göre, sadece olasılığı için en büyük değer aranır. Eğer bu basitleştirilmiş ifadede bütün özellikler bağımsız ise $P(X|C_i)$ Denklem 3.1'deki gibi yazılabilir [38].

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

Denklem 3.1 Bayes Sınıflandırıcı Formülü

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Denklem 3.2 Bayes Formülü

- P(A|B) ; B olayı gerçekleştiği durumda A olayının meydana gelme olasılığıdır (Bknz. Koşullu Olasılık)
- P(B|A) ; A olayı gerçekleştiği durumda B olayının meydana gelme olasılığıdır
- P(A) ve P(B) ; A ve B olaylarının önsel olasılıklarıdır.

[39] makalesinde Naïve Bayes algoritmasının nasıl çalıştığını gösteren örnek incelenirse, havanın durumuna, sıcaklığa, neme ve rüzgarın şiddetine göre tenis oynanabilme durumun kararı verilmektedir. Veri kümesinin Çizelge 3.1'deki şeklinde olduğu düşünüldüğünde

Günler	Hava	Sıcaklık	Nem	Rüzgar	Durum
Gün 1	Güneşli	Sıcak	Yüksek	Az	Hayır
Gün 2	Güneşli	Sıcak	Yüksek	Güçlü	Hayır
Gün 3	Bulutlu	Sıcak	Yüksek	Az	Evet
Gün 4	Yağmurlu	Ilık	Yüksek	Az	Evet
Gün 5	Yağmurlu	Soğuk	Normal	Az	Evet
Gün 6	Yağmurlu	Soğuk	Normal	Güçlü	Hayır
Gün 7	Bulutlu	Soğuk	Normal	Güçlü	Evet
Gün 8	Güneşli	Ilık	Yüksek	Az	Hayır
Gün 9	Güneşli	Soğuk	Normal	Az	Evet
Gün 10	Yağmurlu	Ilık	Normal	Az	Evet
Gün 11	Güneşli	Ilık	Normal	Güçlü	Evet
Gün 12	Bulutlu	Ilık	Yüksek	Güçlü	Evet
Gün 13	Bulutlu	Sıcak	Normal	Az	Evet
Gün 14	Yağmurlu	Ilık	Yüksek	Güçlü	Evet

Çizelge 3.1 Tenis Oynama Durumu Örnek Veri Kümesi

eldeki bu veriler ile 15. Gün havanın güneşli, sıcaklığın soğuk, nemin yüksek ve rüzgarın da güçlü olduğunu bilirsek, tenis oynanabilme durumun hesaplanması yapılabilir.

Hava güneşli iken tenis oynanabilir çıkma olasılığı : 2/9
Sıcaklık soğuk iken tenis oynanabilir çıkma olasılığı : 3/9
Nem yüksek iken tenis oynanabilir çıkma olasılığı : 3/9
Rüzgar güçlü iken tenis oynanabilir çıkma olasılığı : 3/9

Dolayısıyla sonucun tenis oynanabilir çıkma olasılığı,
P(y) = 9/14 x 2/9 x 3/9 x 3/9 x 3/9 = **0,0053**

Hava güneşli iken tenis oynanamaz çıkma olasılığı : 3/5
Sıcaklık soğuk iken tenis oynanamaz çıkma olasılığı : 1/5
Nem yüksek iken tenis oynanamaz çıkma olasılığı : 4/5
Rüzgar güçlü iken tenis oynanamaz çıkma olasılığı : 3/5

Dolayısıyla sonucun tenis oynanamaz çıkma olasılığı
P(n) = 5/14 x 3/5 x 1/5 x 4/5 x 3/5 = **0,0205**

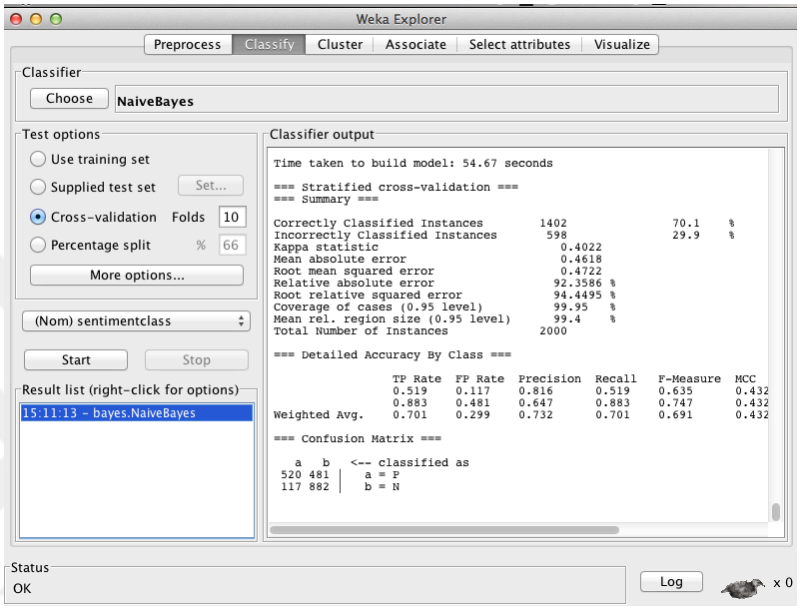
Tenis oynanamaz çıkma olasılığı, oynanabilir çıkma olasılığından büyük olduğundan **(P(n)>P(y))** algoritmanın öğrendiği verilerden tahmini, tenis maçının oynanmayacağı yönündedir.

WEKA'da testler yapılırken Naïve Bayes Algoritması için bir Çapraz Geçerleme (Cross-Validation) değeri verilmiştir. K-Kat Çapraz Geçerleme (K-Fold Cross-Validation), X veri kümesi rastgele olmak üzere K tane eşit büyüklükte $X_{i\ (i = 1, \ldots, K)}$ parçalarına bölünür. Her katta, K parçadan biri geçerleme için ayrılırken, kalan (K-1) parça birleştirilerek öğrenme kümesi oluşturulur. Böylece K tane çift elde edilmiş olur.

$$V_1 = X_1 \qquad T_1 = X_2 \cup X_3 \cup \ldots \cup X_K$$
$$V_2 = X_2 \qquad T_2 = X_1 \cup X_3 \cup \ldots \cup X_K$$
$$\vdots$$
$$V_K = X_K \qquad T_K = X_1 \cup X_2 \cup \ldots \cup X_K$$

K değeri sıklıkla 10 ya da 30 alınmaktadır. Bu değer arttıkça öğrenme kümelerinin yüzdesi artar ve daha etkin kestiriciler oluşur. Ancak geçerleme kümesi küçülecektir. Ayrıca, K değeri artarken, toplam öğrenme ve geçerleme karmaşıklığı da artar. Veri kümesi büyüklüğü olan N değeri artarken, K değeri küçülebilmektedir.

N değeri küçük ise öğrenmede yeterli veri olabilmesi adına K değerinin daha büyük olması gerekmektedir. En uç durumda K = N alınarak, her katta geçerleme için yalnızca bir örnek dışarıda bırakılır ve N-1 örnekle öğrenme yapılır [35].



Şekil 3.10 Örnek WEKA Çıktısı

WEKA Sınıflandırma Yöntemi ile Şekil 3.10'daki gibi çıktılar üretilmektedir. Bu tezdeki deneylerde de bu ve benzeri çıktılar üretilmiştir. Ayrıca deneylerde F-Skor değeri dikkate alınmıştır. Bu değerin 1'e olan yakınlığı, deneyin o kadar başarılı olduğunu ifade etmektedir. F-Skor değerini bulmak için Kesinlik (Precision) ve Hassasiyet (Recall) değerlerinin bilinmesi gerekmektedir.

- **Kesinlik (Precision):** "Getirilen bilginin ne kadarı istenilen bilgiyle ilgilidir" sorusunun cevap değeridir.

$$Kesinlik(Precision) = \frac{\{ilgili\ getirim\} \cap \{b\ddot{u}t\ddot{u}n\ veri\ \text{çıkarımı}\}}{\{b\ddot{u}t\ddot{u}n\ veri\ \text{çıkarımı}\}}$$

Denklem 3.3 Kesinlik Formülü

- **Hassasiyet (Recall):** "Getirilmesi gereken bilginin ne kadarı getirilmiştir" sorusunun cevap değeridir.

$$Hassasiyet(Recall) = \frac{\{ilgili\ getirim\} \cap \{b\ddot{u}t\ddot{u}n\ veri\ çıkarımı\}}{\{ilgili\ veri\ çıkarımı\}}$$

Denklem 3.4 Hassasiyet Formülü

- **F-Skoru (F-Skor):** Kesinlik (Precision) ve Hassasiyet (Recall) değerlerinin harmonik ortalamasıdır.

$$F_{1^{Skoru}} = 2\frac{Kesinlik\ .\ Hassasiyet}{Kesinlik + Hassasiyet} = 2\frac{pr}{p+r}$$

Denklem 3.5 F-Skor Formülü

3.4.1. Deney-1

Bu deneyde Ön İşleme adımında Affix Stemming ve N-gram değeri 1 seçilmiştir. Stop word dosyası olarak "turkish_stop_word" dosyası Preto'ya verilmiştir. 2000 adet pozitif ve negatif yorumun bulunduğu TF, TFIDF ve TFIDF-Norm Doküman Terim Matrisi dosyası sıra ile WEKA'ya verilmiştir. Sınıflandırma için Naïve Bayes algoritması kullanılarak 10-Kat Çapraz Geçerleme seçilmiştir. Amaç N-gram değeri 1 olan dokümanların, değişik Doküman Terim Matrislerinde sonuç olarak ne kadar iyi F-Skorlar vereceğini görmektir.

Yöntem	Kesinlik	Hassasiyet	F- Skoru
TF	0,749	0,742	0,740
TFIDF	0,745	0,740	0,739
TFIDF-NORM	0,789	0,779	0,777

Çizelge 3.2 1-gram Terimlerle Deney Sonuçları

F-Skor sonuçları dikkate alındığında, TFIDF-NORM Terim Ağırlıklandırma Yöntemi ile Doküman Terim Matrisi gerçeğe yakın bir sonuç vermektedir. Bunun anlamı %77,7 oranında dokümanın pozitif veya negatif bilgisinin doğru tahmin edilmiş olduğudur.

3.4.2. Deney-2

Bu deneyde ön işleme adımında Affix Stemming ve N-gram değeri olarak 2 seçilmiştir. Stop word dosyası olarak "turkish_stop word" dosyası Preto'ya verilmiştir. 2000 adet pozitif ve negatif yorumun bulunduğu TF, TFIDF ve TFIDF-Norm Doküman Terim Matrisi dosyası sıra ile WEKA'ya verilmiştir. Sınıflandırma için Naïve Bayes algoritması kullanılarak 10-Kat Çapraz Geçerleme seçilmiştir. Amaç N-gram değeri 2 olan dokümanların değişik Doküman Terim Matrislerinde ne kadar iyi F-Skor vereceğini görmektir.

Yöntem	Kesinlik	Hassasiyet	F- Skoru
TF	0.742	0.738	0.720
TFIDF	0.748	0.743	0.742
TFIDF-NORM	0.685	0.632	0.604

Çizelge 3.3 2-gram Terimlerle Deney Sonuçları

İkinci deneyin gösterdiği bilgi, 2-gram için TFIDF-NORM Terim Ağırlıklandırma Yöntemi, Doküman Terim Matrislerinde iyi sonuçlar vermediğidir. Birinci deney, diğer deneylerden daha başarılı sonuçlar vermiştir. İkinci deney için en başarılı Terim Ağırlıklandırma Yöntemi, TFIDF olduğu görülmektedir.

3.4.3. Deney-3

Bu deneyde ön işleme adımında Ek Çıkaran Kök Bulucu (Affix Stripping Stemmer) ve N-gram değeri olarak 3 seçilmiştir. Stop word dosyası için "turkish_stop_word" dosyası Preto'ya verilmiştir. 2000 adet pozitif ve negatif yorumun bulunduğu TF, TFIDF ve TFIDF-Norm Doküman Terim Matrisi dosyası sıra ile WEKA'ya verilmiştir. Sınıflandırma için Naïve Bayes algoritması kullanılarak 10-Kat Çapraz Geçerleme seçilmiştir. Amaç N-gram değeri 3 olan dokümanların değişik Doküman Terim Matrislerinde ne kadar iyi F-Skor vereceğini görmektir.

Yöntem	Kesinlik	Hassasiyet	F- Skoru
TF	0.732	0.701	0.691
TFIDF	0.732	0.701	0.691
TFIDF-NORM	0.676	0.587	0.527

Çizelge 3.4 3-gram Terimlerle Deney Sonuçları

3.4.4. Deney-4

Bu deneyde ön işleme adımında Ek Çıkaran Kök Bulucu (Affix Stripping Stemmer) ve N-gram değeri olarak 1 ve 2 seçilmiştir. Stop word dosyası için "turkish_stop_word" dosyası Preto'ya verilmiştir. 2000 adet pozitif ve negatif yorumun bulunduğu TF, TFIDF ve TFIDF-Norm Doküman Terim Matrisi dosyası sıra ile WEKA'ya verilmiştir. Sınıflandırma için Naïve Bayes algoritması kullanılarak 10-Kat Çapraz Geçerleme seçilmiştir. Amaç N-gram değeri 1 ve 2 olan dokümanların değişik Doküman Terim Matrislerinde ne kadar iyi F-Skor vereceğini görmektir.

Yöntem	Kesinlik	Hassasiyet	F- Skoru
TF	0.757	0.750	0.748
TFIDF	0.750	0.746	0.744
TFIDF-NORM	0.782	0.774	0.772

Çizelge 3.5 1 ve 2-gram Terimlerle Deney Sonuçları

4. SONUÇLAR, DEĞERLENDİRMELER ve ÖNERİLER

Bu bölümde tez boyunca yapılan çalışmaların ve deneylerin sonuçları karşılaştırılmış, ileride uygulanabilecek çalışmalar ve neler yapılabileceğine dair önerilere yer verilmiştir.

4.1. Sonuçlar ve Değerlendirmeler

Çevrimiçi ortamlardan elde edilen film yorumları, Ön İşleme Yöntemleri ile daha anlaşılabilir bir veri haline getirilmiştir. Daha sonra makine öğrenmesi yöntemlerini uygulamak için WEKA adlı yazılıma uygun .arff uzantılı dosyalar oluşturulmuştur.

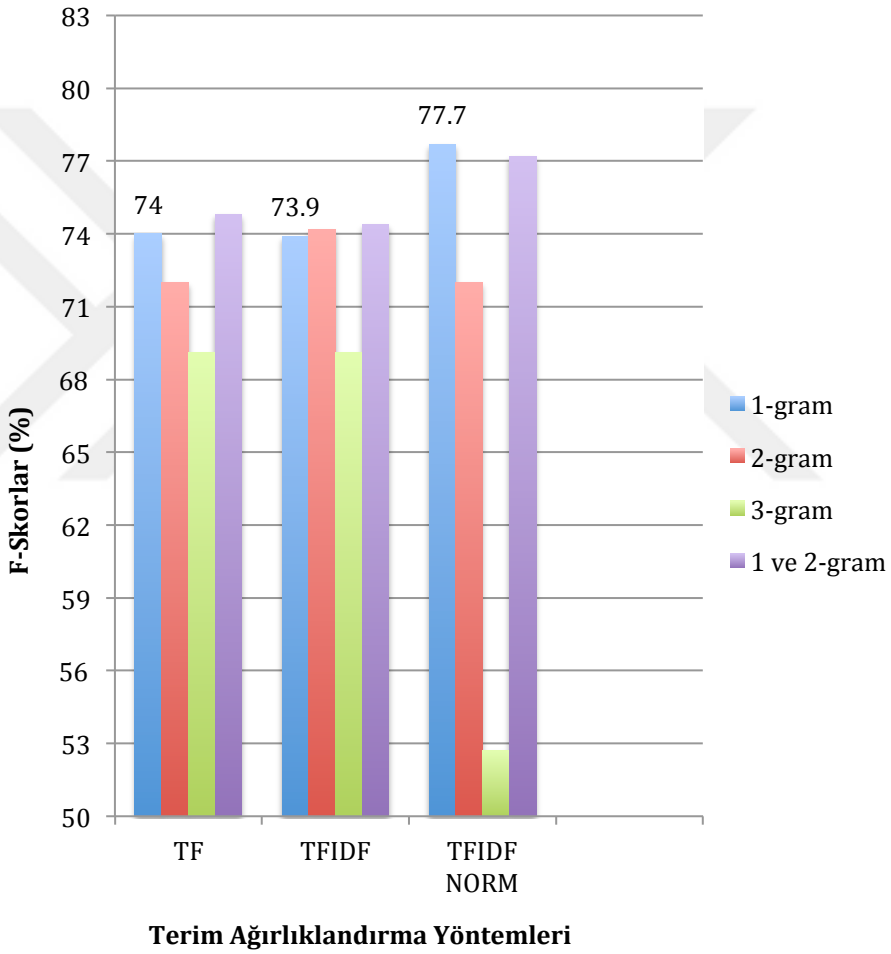
Deneyler, Ön İşleme Yöntemlerinden N-gram sayısı değiştirilerek gerçekleştirilmiştir. İşlenmiş dosyalar sırası ile Naïve Bayes Algoritması kullanılarak, 10-Kat Çapraz Geçerleme ile makine öğrenmesi uygulanmıştır. TF, TFIDF ve TFIDF NORM Terim Ağırlıklandırma Yöntemlerinin çıktılarının F-Skorları karşılaştırıldı. N-gram sayısındaki değişimlerin ve Terim Ağırlıklandırma Yöntemlerindeki değişikliklerinin F-Skor üzerinde gösterdiği farklılığın ve gerçek sonuca ne kadar yakın olduğuna bakıldı.

Yöntem	1-gram	2-gram	3-gram	1 ve 2-gram
TF	0.740	0.720	**0.691**	0.748
TFIDF	0.739	**0.742**	**0.691**	0.744
TFIDF-NORM	**0.777**	0.604	0.527	**0.772**

Çizelge 4.1 Farklı N-gramlarda F-Skorları

Yukarıdaki tabloya göre gerçeğe en yakın sonucun, 1-gram ile TFIDF-NORM Terim Ağırlıklandırma Yöntemi ile olduğu görülmektedir. Kelimeler üzerinde 1-gram ön işleme yöntemi kullanılarak, terim doküman ağırlıklandırma yöntemlerinden TDIDF dosyasının normalize edilmiş halinin Duygu Analizi için

gerçeğe en yakın sonucu verdiği görülmektedir. Tekil N-gram'lar arasında en kötü sonucu veren 3-gram TFIDF-NORM deneyidir. 1 ve 2 gramların bir arada alınıp, makine öğrenmesi uygulandığı deneyde ise ortaya çıkan sonuç, TF ve TFIDF Terim Ağırlıklandırma Yöntemlerinde diğer N-gram'lara göre en iyisi olarak görülmektedir. TFIDF-NORM Terim Ağırlıklandırma Yönteminde 1-gram'ın F-Skor değerini geçememiştir.



Şekil 4.2 Farklı N-gram'lardaki F-Skorları

4.2. Öneriler

Çevrimiçi ortamdan çekilen verilerin ön işleme adımlarını değiştirerek daha temiz bir veri ile çalışmak daha iyi sonuçlar almaya yardımcı olabilir. Bunların yanında diğer makine öğrenmesi algoritmaları kullanılarak yapılan çalışmalar, Naïve Bayes'e göre daha iyi sonuçlar verebilmektedir. Ayrıca daha çok veri üzerinde çalışmak, daha iyi sonuçlar almak adına etkili olacaktır. Türkçe'nin dil bilgisi yapısını incelemek ve pozitif – negatif kelimelerin listesini çıkarmak ve bu listeye göre uygun algoritmalar geliştirmek Duygu Analizi adına verimliliği artıracaktır.

5. KAYNAKLAR

1. ComScore/the Kelsey group, Online consumer-generated reviews have significant impact on offline purchase behavior. Press Release, November 2007.
2. Volkan Tunalı, "Türkçe Metinlerin Kümelenmesinde Farklı Kök Bulma Yöntemlerinin Etkisinin Araştırılması", ELECO '2012 Elektrik - Elektronik ve Bilgisayar Mühendisliği Sempozyumu, 29 Kasım Bursa
3. Soumen Chakrabarti, Martin Ester, Usama Fayyad, J "Data Mining Curriculum", ACM SIGKDD, Nisan 2006
4. http://people.ischool.berkeley.edu/~hearst/text-mining.html (Erişim Tarihi: 23.08.2014)
5. http://mis.sadievrenseker.com/2014/06/metin-madenciligi-text-mining/ (Erişim Tarihi: 23.08.2014)
6. http://www.wikiwand.com/en/Machine_learning#/Approaches (Erişim Tarihi: 23.08.2014)
7. http://wordnet.princeton.edu/ (Erişim Tarihi: 23.08.2014)
8. Commnunications of the ACM “Techniques and Applications for Sentiment Analysis”, April 2013, Vol 56, No:4, Sayfa 82
9. http://www.nytimes.com/2009/08/24/technology/internet/24emotion (Erişim Tarihi: 23.08.2014)
10. Han J. Kamber M., "Data Mining Concepts and Techniques, Second Edition", Morgan Kaufmann, ISBN 13: 978-1-55860-901-3, San Francisco, 2006.
11. Mustafa Koray Aytekin, Yüksek Lisans Tezi “Vekil sunucu verisi üzerinde ile kullanıcı sorguları kümelemesi”, Maltepe Üniversitesi, 2012
12. Commnunications of the ACM “Techniques and Applications for Sentiment Analysis”, April 2013, Vol 56, No:4, Sayfa 84
13. http://www.vtunali.com/tr/index.php/2009/10/metin-madenciligi-text-mining-nedir/ (Erişim Tarihi: 25.08.2014)

14. Jaime Carbonell. Subjective Understanding: Computer Models of Belief Systems. PhD thesis, Yale, 1979.
15. Mihalcea, C. Banea and J. Wiebe. 2007. Learning multilingual subjective language via crosslingual projections. In Proceedings of ACL-2007.
16. Banea, R. Mihalcea, J. Wiebe and S. Hassan. 2008. Multilingual subjectivity analysis using machine translation. In Proceedings of EMNLP-2008.
17. Wan, X. 2009. Co-training for cross-lingual sentiment classification. In Proceedings of the ACL, 235–243
18. Umut Eroğul, Sentiment Analysis in Turkish, METU Master's Thesis, 2009.
19. Pang, b., Lee, L. and Vaithyanathan, S. "thumbs up? sentiment Classification using machine learning techniques." in Proceedings of EMNLP-02, 7th Conference on Empirical Methods in Natural Language Processing (Philadelphia, Pa, 2002). association for Computational Linguistics, morristown, nj, 79–86.
20. Turney, P. "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews." in Proceedings of the Association for Computational Linguistics (2002), 417–424.
21. Yu, H. ve Hatzivassiloglou, V. "Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences." in Proceedings of the Conference on Empirical Methods in Natural Language Processing (2003).
22. Pang, B. and Lee, L. "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts." in Proceedings of the Association for Computational Linguistics (2004), 271–278.
23. Riloff, R. and Wiebe, J. "Learning extraction patterns for subjective expressions." in Proceedings of the Conference on Empirical Methods in Natural Language Processing (2003).
24. Tsur, O., Davidov, D. ve Rappoport, A. "A great catchy name: semi-supervised recognition of sarcastic sentences in online product reviews." in Fourth International AAAI Conference on Weblogs and Social Media (2010).

25. Netzer, O., Feldman, R., Fresko, M. ve Goldenberg, Y. “Mine your own business: market structure surveillance through text mining.” Marketing Science, 2012.
26. Pang, B. ve Lee, L. “A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts.” in Proceedings of the Association for Computational Linguistics (2004), 271–278.
27. Volkan Tunalı, Turgay Tugay Bilgin, "PRETO: A High-performance Text Mining Tool for Preprocessing Turkish Texts", International Conference on Computer Systems and Technologies (CompSysTech), Ruse, Bulgaria, June 22-23, 2012, 134-140.
28. http://www.wikiwand.com/tr/WEKA
(Erişim Tarihi: 26.02.2015)
29. http://w3techs.com/technologies/details/cl-tr-/all/all
(Erişim tarihi: 28.02.2015)
30. http://www.alexa.com/siteinfo/mynet.com (Erişim tarihi: 28.02.2015)
31. https://www.ruby-lang.org/tr/ (Erişim Tarihi: 28.02.2015)
32. Jongejan, B.; and Dalianis, H.; Automatic Training of Lemmatization Rules that Handle Morphological Changes in pre-, in- and Suffixes Alike, in the Proceeding of the ACL-2009, Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, August 2–7, 2009, pp. 145-153
33. http://www.wikiwand.com/en/Document-term_matrix
(Erişim tarihi: 28.02.2015)
34. http://cis.poly.edu/~mleung/FRE7851/f07/NaïveBayesianClassifier.pdf
(Erişim tarihi: 02.03.2015)
35. Ethem Alpaydın, Yapay Öğrenme (2007), s: 416-417
36. https://tr.wikipedia.org/wiki/Naive_Bayes_s%C4%B1n%C4%B1fland%C4%B1r%C4%B1c%C4%B1 (Erişim tarihi: 02.03.2015)
37. Diri B. "Doküman Sınıflandırma Sunumu”, Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği Bölümü, 2014

38. Albayrak S. “Sınıflama ve Kümeleme Yöntemleri Sunumu”, Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği Bölümü, 2014
39. http://www.tameroz.com/content/files/NaiveBayes.pdf (Erişim Tarihi: 06.09.2015)

6. EKLER

acaba	altmış	altı	ama	ancak	arasında
artık	aynı	bana	bazı	başka	belki
ben	benden	beni	benim	beş	bile
bin	bir	biri	birkaç	birkez	birlikte
birşey	birşeyi	biz	bizden	bizi	bizim
bu	bugün	buna	bunda	bundan	bunu
bunun	böyle	bütün	büyük	da	daha
dahi	de	dedi	defa	devam	değil
diye	diğer	doksan	dokuz	dört	dün
eden	elli	en	en gibi	eski	etti
eğer	gelen	geçen	gibi	göre	gün
hem	hep	hepsi	her	hiç	iki
ile	ilgili	ilk	ise	iyi	için
içinde	iş	kabul	kadar	karşı	katrilyon
kendi	kez	ki	kim	kimden	kime
kimi	konusunda	kırk	mi	milyar	milyon
mu	mü	mı	nasıl	ne	neden
nedeniyle	nerde	nerede	nereye	niye	niçin
o	olan	olarak	oldu	olduğu	olduğunu
on	ona	ondan	onlar	onlardan	onları
onların	onu otuz	ortaya	pek	sadece	sanki
sekiz	seksen	sen	senden	seni	senin
siyasi	siz	sizden	sizi	sizin	son
sonra	söyledi	tam	tarafından	tek	trilyon
tüm	var	ve	veya	ya	yani
yaptığı	yapılan	yedi	yeni	yer	yetmiş
yine	yirmi	yok	yüz	yüzde	zaman
çok	çünkü	önce	önemli	özel	üzerine
üç	şey	şeyden	şeyi	şeyler	şimdi
şu	şuna	şunda	şundan	şunu	şöyle

Ek-1 Türkçe'deki Durak Sözcükler

7. ÖZGEÇMİŞ

Ender Ahmet Yurt, 1986 yılı Ordu doğumludur. Ünye Mehmet Refik Güven Anadolu Öğretmen Lisesi'nden mezun olduktan sonra 2005 yılında Fevziye Mektepleri Vakfı Işık Üniversitesi Bilgisayar Mühendisliği'ni kazandı ve 2011 yılında fakülte derecesinde mezun oldu. 2011-2013 yılları arasında çeşitli firmalarda web uygulama geliştiriciliği yaptı. 2013 senesinden itibaren Phonoclick şirketinde Yazılım Mühendisi olarak çalışmaktadır. 2012 yılında başladığı Maltepe Üniversitesi Bilgisayar Mühendisliği Yüksek Lisans programını 2015 senesinde tamamladı.