T.C.
TRAKYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE İÇİN DÜŞÜNCE KUTUPLUK DERLEMİ OLUŞTURULMASI VE DÜŞÜNÇE ÇÖZÜMLEMESİ İÇİN BİR YÖNTEM GELİŞTİRİLMESİ

ERTAÇ PAPRAT
YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

TEZ DANIŞMANI: YRD. DOÇ. DR. AYDIN CARUS

EDİRNE - 2015

Trakya Üniversitesi Fen Bilimleri Enstitüsü Onayı

Prof. Dr. Mustafa Özcan
T.Ü. Fen Bilimleri Enstitüsü Müdürü

Bu tezin Yüksek Lisans tezi olarak gerekli şartları taşıdığını onaylarım.

Prof. Dr. Yılmaz Kılıçaslan
Anabilim Dalı Başkanı

Bu tez tarafımca okunmuş, kapsamı ve niteliği bakımından bir Yüskek Lisans tezi olarak kabul edilmiştir.

Yrd. Doç. Dr. Aydın Carus
Tez Danışmanı

Bu tez, tarafımızca okunmuş, kapsam ve niteliği açısından Bilgisayar Mühendisliği Anabilim Dalında bir Yüksek Lisans tezi olarak oy birliği ile kabul edilmiştir.

Jüri Üyeleri (Ünvan, Ad, Soyad): İmza

Prof. Dr. Yılmaz Kılıçaslan .....
Yrd. Doç. Dr. Aydın Carus .....
Yrd. Doç. Dr. Oğuzhan Erdem .....

Tarih: 7 Ocak 2015

T.Ü. FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI YÜKSEK LİSANS PROGRAMI
DOĞRULUK BEYANI

İlgili tezin akademik ve etik kurallara uygun olarak yazıldığını ve kullanılan tüm literatür bilgilerinin kaynak gösterilerek ilgili tezde yer aldığını beyan ederim.

9 Ocak 2015
Ertaç PAPRAT



Yüksek Lisans Tezi
Türkçe İçin Düşünce Kutupluk Derlemi Oluşturulması Ve Düşünce Çözümlemesi İçin Bir Yöntem Geliştirilmesi
Trakya Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

ÖZET

İnternet kullanımı, erişim olanaklarının çeşitlenmesi ve ucuzlaması ile günümüzde oldukça yaygınlaşmıştır. İnternet kullanımının yaygınlaşması beraberinde, kullanıcı tarafından üretilen içeriklerin de çok büyük boyutlara ulaşmasına neden olmuştur. Kullanıcıların oluşturduğu büyük boyutlardaki verilerden değerli bilgilerin çıkarımını yapabilmek oldukça önemli bir hal almıştır. Bu verilerin, oldukça düzensiz olması nedeniyle bilgi çıkarımı işleminden önce içeriklerine ya da çeşidine göre sınıflandırılması gerekmektedir.

Türkçe film yorumları içeren bir web sitesinden özdevimli yöntemler ile bir kutupluk derlemi oluşturularak, İngilizce düşünce analizi için önerilen yapay öğrenme ve doğal dil işleme yöntemleri Türkçe içeriklere uygulanmış, Türkçe dili özelinde sonuçlar tartışılmıştır. Türkçe belgelerde düşünceyi en belirgin şekilde belli eden sözcük öbekleri belirlenmiş, bunlara yapay öğrenme ve doğal dil işleme yöntemleri uygulanarak Türkçe belgeleri, taşıdığı düşünce kutupluğuna göre özdevimli sınıflandıran bir yöntem önerilmiştir.

Bu çalışmada, internet kullanıcıları tarafından üretilen metinsel verilerin taşıdığı yargının “olumlu” ya da “olumsuz” olması bakımından özdevimli sınıflandırılma üzerinde durulmuştur. Bilgi edinimi ile öznitelikler belirlenmiş ve film yorumunun düşünceye göre sınıflandırılmasında Naïve Bayes Çokterimli sınıflandırıcısı kullanılmıştır. Sinema filmi yorumları alanındaki derlemin unigram+bigram sözcük kümelerinden oluşturulan bir örneği %90 F-Ölçütü ile en yüksek başarımı sağlamıştır.

Yıl : 2015
Sayfa Sayısı : 95
Anahtar Kelimeler : Düşünce Çözümleme, Fikir Madenciliği, Yapay Öğrenme, Doğal Dil İşleme, Türkçe, Kutupluk Derlemi, Metin Sınıflandırma, Veri Madenciliği, Naïve Bayes

Master Thesis
Introducing Turkish Polarity Corpus and a Method for Sentiment Analysis in Turkish
Trakya University Graduate School of Natural and Applied Sciences
Department of Computer Engineering

ABSTRACT

Internet access facilities have become widespread due to the variety of internet utilization methods and low internet subscription costs. This widespread usage brings huge amount of user generated contents together. Extraction of valuable information from vast amount of contents that are produced by internet users had been an important task. Prior to the information extraction phase, this kind of data have to be classified by its content and type, as it is highly unstructured.

In this thesis, a Turkish polarity corpus has been created automatically from a Turkish movie review website. Machine learning and natural language processing methods have been applied to those movie reviews in Turkish by also considering and exploiting previous studies for English and then the results are discussed. The most effective word phrases which obviously designate sentiment of a document in Turkish have been extracted and a new sentiment analysis method is proposed for Turkish by applying natural language processing and machine learning methods to these phrases.

Furthermore, a Binary classification of textual data that is generated by internet users is studied. Information gain has been used to extract features of the movie reviews and then the well-known Naïve Bayes Multinomial have been utilized to classify sentiment of these reviews. Ultimately, F-Score of 90% has been achieved as the best score for unigram+bigram subsample of the corpus.

Year: 2015
Pages: 95
Keywords: Sentiment Analysis, Opinion Mining, Machine Learning, Natural Language Processing, Turkish, Polarity Corpus, Text Classification, Data Mining, Naïve Bayes

TEŞEKKÜR

Yüksek lisans öğrenimim boyunca tecrübe, bilgi ve birikimini her fırsatta paylaşan, değerli fikirleriyle yol gösteren danışman hocam Yrd. Doç. Dr. Aydın CARUS'a teşekkür ederim.

Tüm eğitim-öğrenim hayatım boyunca maddi, manevi desteklerini esirgemeyen, bugüne gelmemde en büyük pay sahipleri annem Nermin PAPRAT, babam Mümin Salih PAPRAT ve babaannem Merhume Hayriye PAPRAT'a teşekkür ederim.

İÇİNDEKİLER

ÖZET ..... **i**

ABSTRACT ..... **ii**

TEŞEKKÜR ..... **iii**

İÇİNDEKİLER ..... **iv**

ŞEKİLLER LİSTESİ ..... **vii**

TABLOLAR LİSTESİ ..... **viii**

SİMGELER DİZİNİ ..... **x**

Kısaltmalar ..... **x**

GİRİŞ **1**

GENEL ARKAPLAN BİLGİLERİ **4**

2.1 Veri Madenciliği **4**

2.2 Metin Madenciliği **5**

2.3 Derlem Dilbilim **5**

2.4 Düşünce Çözümleme **6**

2.4.1 Bilişsel Bilimler Açısından Düşünce 6

2.4.2 Bilgisayar Bilimleri Açısından Düşünce Çözümleme 7

2.4.3 Türkçe Odaklı Çalışmalar 12

BİLGİSAYARLI DÜŞÜNCE ÇÖZÜMLEMEDE KULLANILAN YÖNTEM VE ARAÇLAR **13**

3.1 İşaretli Derlem (Annotated Corpus) ve Önemi **13**

3.2 İşaretli Derlemin Amacının Belirlenmesi ve Derlem Oluşturma Aşamaları **14**

3.3 Derlemi Oluşturan Verinin Toplanması ve Ağ Örümcekleri (Web Spiders) **15**

3.4 Doğal Dil İşleme (Natural Language Processing) **17**

3.5 Yapay Öğrenme (Machine Learning) **19**

BİLGİSAYARLI DÜŞÜNCE ÇÖZÜMLEME İÇİN YAPILAN ÇALIŞMA **24**

4.1 Derlem Verisinin Elde Edilmesi **25**

4.1.1 Ağ Örümceği Gerçeklemesi 25

4.1.2 Elde Edilen Veri 26

4.2 Derlem Verisinin İşlenmesi **30**

4.2.1 Doğal Dil İşleme ... 30
4.2.2 Yazım Hatası Denetimi ... 30
4.2.3 Sözcük Kök ve Eklerinin Tespiti ... 31
4.2.4 Olumsuzluk Durumu ... 33
4.2.5 Sözcük Sınıflandırma ... 34
4.3 Yöney Uzayı Modeli (Vector Space Model) ... 36
4.3.1 N-Gram Modeli ... 36
4.3.2 Terim Frekansı – Ters Metin Frekansı (TF-IDF) ... 38
4.4 Yapay Öğrenme ... 41
4.4.1 Gözetimli Yapay Öğrenme ... 42
4.4.2 Naïve Bayes Çokterimli ... 45
4.4.3 K-Katlı Çapraz Doğrulama (N-Fold Cross Validation) ... 46
4.4.4 Başarımın Ölçülmesi ... 47
SONUÇLAR ... 50
5.1 Dayanak ... 50
5.2 N-Gram Sonuçları ... 51
5.3 Sözcük Sınıflarına Göre Sonuçlar ... 52
5.4 Sözcük Köklerine Göre Sonuçlar ... 55
TARTIŞMALAR ... 57
EK 1 – Türkçe En Güçlü 40 Unigram ve Kutupluğu ... 63
EK 2 – Türkçe En Güçlü 40 Bigram ve Kutupluğu ... 64
EK 3 – Türkçe En Güçlü 40 Trigram ve Kutupluğu ... 65
EK 4 – Türkçe En Güçlü 40 Fourgram ve Kutupluğu ... 66
EK 5 – Türkçe En Güçlü 40 1-4 Gram İçin Kutupluk ... 67
EK 6 – Türkçe En Güçlü 40 Fiil 1-2 Gram İçin Kutupluk ... 68
EK 7 – Türkçe En Güçlü 40 İsim 1-4 Gram İçin Kutupluk ... 69
EK 8 – Türkçe En Güçlü 40 İsim+Fiil 1-4 Gram İçin Kutupluk ... 70
EK 9 – Türkçe En Güçlü 40 İsim+Sıfat 1-4 Gram İçin Kutupluk ... 71
EK 10 – Türkçe En Güçlü 40 İsim+Sıfat+Fiil 1-4 Gram İçin Kutupluk ... 72
EK 11 – Türkçe En Güçlü 40 Sıfat 1-4 Gram İçin Kutupluk ... 73

EK 12 – Türkçe En Güçlü 40 Sıfat+Fiil 1-4 Gram İçin Kutupluk 74

EK 13 – Hatalı Kabul Edilen Belirtke Örnekleri 75

EK 14 – ARFF Dosyası İçerik Örneği 76

EK 15 – Derlem İçeriği Örnekleri 77

EK 16 – Beyazperde Sitesinde Bir Kullanıcıya Ait Yorumun HTML Kaynak Kodu Örneği 78

ÖZGEÇMİŞ 82

ŞEKİLLER LİSTESİ

Şekil 2.1 Genel Bir Düşünce Çözümleme Mimarisi Örneği 8

Şekil 3.1 Genel Bir Ağ Örümceği Mimarisi 16

Şekil 3.2 Weka Yazılımının Arayüzü 23

Şekil 4.1 Bir Kullanıcı Yorumunu Elde Etmek İçin Gerekli XPATH Kodu 26

Şekil 4.2 Beyazperde Sitesinden Bir Üye Yorumu 27

Şekil 4.3 Puan Üst Verisi İle Kaydedilen Yorum Örneği 28

Şekil 4.4 Zemberek2'nin Yazım Hatası Düzeltme Önerisi Örneği 30

Şekil 4.5 Zemberek2 İçin Bir Sözcük Çözümleme Örneği 31

Şekil 4.6 Dilbilimsel Açıdan Kirli Veri Örneği 32

Şekil 4.7 Türkçe Olumsuzluk Sağlayan Sözcükler 33

Şekil 4.8 Olumsuz Sözcük Çözümlemesi ve Derlemde Saklanış Şekli 34

Şekil 4.9 Sınıflandırılmış Sözcüklere Derlemden Örnekler 35

Şekil 4.10 Bir Cümleyi Oluşturan Sözcüklerin N-Gram Örneği 37

TABLOLAR LİSTESİ

Tablo 4.1 Kullanıcı Puanlarının 10'lu Ölçekte Dağılımı....28

Tablo 4.2 Gözetimli Öğrenme İçin Kullanılan Örneklerin Kutupluk Dağılımı....29

Tablo 4.3 Yeniden Örneklenen Veri Kümesinin Kutupluk Dağılımı....29

Tablo 4.4 Eğitim Derleminde Sözcük Kökünün Özdevimli Bulunabilme Durumuna Göre Özgün Belirtke Sayıları....32

Tablo 4.5 Olumsuzluk Taşıyan Köklerin İşaretli Bulunduğu Veri Kümelerine Ait İstatistikler....34

Tablo 4.6 Dilbilimsel Sözcük Sınıflarına Göre Oluşturulan Veri Kümelerine Ait Özgün Belirtke Adetleri....36

Tablo 4.7 Her Bir Belgesi Yöneye Çevrilmiş Veri Kümeleri ve Öznitelik Aday Sayıları....38

Tablo 4.8 TF-IDF Dönüşümü Öncesi Derlem....39

Tablo 4.9 Belirtke Mevcudiyetine Göre Belge Yöneyleri....40

Tablo 4.10 Belirtke Tekrarına Göre Belge Yöneyleri....40

Tablo 4.11 Üssel Belirtke Tekrarına Göre Belge Yöneyleri....40

Tablo 4.12 N-Gram Belirtkelerden Oluşan Veri Kümelerine Ait Öznitelik Sayıları....44

Tablo 4.13 K-Katlı Çapraz Doğrulama Sınama ve Eğitim Verileri Dağılımı....47

Tablo 4.14 İki Sınıflı Sınıflandırıcıya Ait Karışıklık Matrisi....47

Tablo 5.1 ASCII Olağanlaştırılmış Derlem İçin Unigram Öznitelik Adayı ve Öznitelik ile Belirtke Sayıları....50

Tablo 5.2 Unigram Öznitelikler ile Sınıflandırma için Karışıklık Matrisi....51

Tablo 5.3 Unigram Öznitelikler ile Sonuçlar....51

Tablo 5.4 N-Gram Öznitelikler ile Sonuçlar....52

Tablo 5.5 Dilbilimsel Sözcük Sınıflarına Göre Unigram+Bigram Öznitelik Sayıları....53

Tablo 5.6 Dilbilimsel Sözcük Sınıflarına Göre Unigram+Bigram+Trigram Öznitelik Sayıları....53

Tablo 5.7 Dilbilimsel Sözcük Sınıflarına Göre Unigram+Bigram Sonuçları 54

Tablo 5.8 Dilbilimsel Sözcük Sınıflarına Göre Unigram+Bigram+Trigram Sonuçları..54

Tablo 5.9 Sözcük Köklerine Göre N-Gram Öznitelikler ile Sonuçlar 55

Tablo 5.10 Sözcük Köklerinin Olumsuzluk Durumuna Göre N-Gram Öznitelikler ile Sonuçlar.....55

SİMGELER DİZİNİ

Kısaltmalar

ARFF ..... Attribute-Relation File Format

ASCII ..... American Standard Code for Information Interchange

CSV ..... Comma Seperated Value

GNU ..... GNU's Not Unix

F-Measure ..... F-Ölçütü

F1 Score ..... F1 Puanı

FN ..... False Negative

FP ..... False Positive

HTML ..... Hyper Text Markup Language

HTTP ..... Hyper-Text Transfer Protocol

MATTER ..... Model, Annotate, Train, Test, Evaluate, Revise

MaxEnt ..... Maximum Entropy

NLP ..... Natural Language Processing

NLTK ..... Natural Language Toolkit

OpenNLP .....Open Natural Language Processing

TF-IDF ..... Term Frequency–Inverse Document Frequency

URL ..... Uniform Resource Locator

urllib2 ..... Python Programlama Dili Kütüphanesi

XML ..... Extensible Markup Language

XPATH ..... XML Path Language

BÖLÜM 1

GİRİŞ

Her gün milyonlarca internet kullanıcısı terabaytlar mertebesinde içerik üretmektedir [1]. Oldukça düzensiz haldeki bu verileri düzenli hale getirip, kullanılabilir bilgi elde edebilmek ise her yıl yüzlerce bilimsel çalışmaya konu olan güncel bir araştırma alanıdır. Çok büyük miktarlardaki düzensiz veriyi kabul edilebilir başarımla yönetmek başlı başına zor bir problem iken bu verilerden anlamlı bilgi çıkarımı çabası problemi daha da zor bir hale getirmektedir.

Büyük veriden bilgi çıkarımı konusundaki ilk ve en önemli aşamalardan birisi düzensiz haldeki bu verinin düzenli hale getirilip sınıflandırılmasıdır. Bu işlem, verinin boyutları ve hacmi nispeten az olduğu durumlarda özdevimsiz yapılabilmekteyken verinin boyutu ve hacmi arttıkça ve yapısı karmaşıklaştıkça önce özdevimli (Automatically) işlemsel süreçler ile verileri sınıflandırma (Classification) gereksinimi doğmakta, işlemsel süreçlere ya da düzenli kurallara indirgenemeyen daha karmaşık durumlarda yapay öğrenme ve doğal dil işlemeyi de içeren istatistiksel yöntemler uygulanmaktadır.

İşlemsel süreçlere indirgenemeyen sınıflandırma problemlerinden birisi de elde edilen metinsel verinin, belgenin içerdiği düşünceye göre sınıflandırılmasıdır. Özellikle yaygınlaşan internet kullanımıyla birlikte sosyal medya kullanımının artması, alışılagelmiş ticaretin ikamesi olarak elektronik ticaretin internet kullanıcıları için cazip hale gelmesi, görüşlerin, fikirlerin internet üzerinden çok daha geniş kitlelere ulaşabiliyor olması fikri, insanların internet üzerinden düşüncelerini ifade etmeye ve karar verme süreçlerinde diğer insanların ne düşündüğünü araştırmaya yöneltmektedir.

İnsanların karar verme süreçlerini ve alışkanlıklarını değiştiren bu paradigma dönüşümü ile birlikte, insanların düşüncelerini, ürettikleri içeriklerden özdevimli elde edecek yapılar geliştirmek, şirketler ve araştırmacılar için son yılların önemli çalışma alanlarından biri olmuştur. Özellikle insanların geribildirimlerinin önemli olduğu durumlarda çabuk tepki gösterebilmek ya da geleceğe yönelik çıkarımlar yapabilmek için düşünce analizi, günümüzde oldukça popüler bir araştırma konusudur. Araştırma şirketlerinden, sosyal bilimlere, siyasetçilere, finans kurumlarına, devlet kurumlarına hatta güvenlik güçlerine, istihbarat teşkilatlarına kadar uzanan çok geniş bir yelpazedeki kurum ve kuruluşlar insanların internet üzerinden ifade ettikleri düşünceleri düşünce analizi yöntemleri ile tarayarak oldukça değerli geribildirimler ve geleceğe yönelik çıkarımlar elde etmeyi amaçlamaktadır ve bu konudaki ilgileri giderek artmaktadır.[1]

Bu çalışmada, detaylarından 2.4.2.1'de bahsedileceği üzere, bilgisayar bilimleri açısından düşünce çözümlenin bir alt konusu olan belge seviyesinde düşünce çözümleme için özdevimli yöntemler ile bir derlem oluşturulup bu derleme yapay öğrenme ve doğal dil işleme yöntemleri uygulanarak Türkçe için bir model önerilmesi amaçlanmıştır. Bu amaçla bir film sitesine ait kullanıcılar tarafından önceden puanlanmış kullanıcı yorumlarından oluşan metinler derlem oluşturulması amacıyla kullanılmıştır.

Bu konudaki çalışmaların büyük çoğunluğunun İngilizce için yapılması ve Türkçe özelinde yapılan çalışmaların İngilizce'ye nazaran çok az sayıda olması bu çalışmanın en büyük motivasyon kaynağını oluşturmaktadır. Bu çalışma ile birlikte Türkçe için düşünce çözümlemeye bir temel oluşturmak ve gelecekte yapılacak diğer çalışmalara bir başlangıç olması amaçlanmıştır.

Bundan sonra sırasıyla bölüm 3'te bilgisayarlı düşünce analizi bağlamında faydalanılan bilgisayar bilimleri konuları genel hatlarıyla anlatılmıştır. Bölüm 4.1'de doğal dil işleme ve metin madenciliği bağlamında derlemin oluşturulması için verinin

[1] http://www.google.com/trends/explore#q=sentiment%20analysis&cmpt=q

toplanmasından, temizlenmesinden ve düzenli hale getirilmesinde bu çalışma kapsamında yapılan işlemler anlatılmıştır. Bölüm 4.2'de düzenli hale getirilen bu verinin doğal dil açısından olağanlaştırılması, sınıflandırılması ve düzeltilmesi ile oluşturulan derlemlerin model oluşturma safhasına hazırlanışında izlediğimiz yöntemler anlatılmıştır. Bölüm 4.3'te ise dilbilimsel ve yapısal bakımdan olağanlaştırılmış derlemlerin, uygulanacak matematiksel modele uygun kurgulanması ile derlemin modelin gerçeklenmesine hazır edilişi hususunda izlediğimiz adımlar anlatılmıştır. Bölüm 4.4 ise oluşturduğumuz derlemleri kullanarak bilgi edinimi yoluyla öznitelik seçimi ve Naïve Bayes Çokterimli algoritması ile yapay öğrenme temelli düşünce analizi modelimiz ve bu modelin başarımının nasıl ölçümlendiği anlatılmıştır. Son olarak bölüm 5 ve 6'da sırasıyla uygulanan bu modele ait elde edilen sonuçlar ve değerlendirlemeleri, son olarak da gelecekte bu çalışmayı temel alarak yapılabilecek çalışmalar anlatılmıştır.

BÖLÜM 2

GENEL ARKAPLAN BİLGİLERİ

2.1 Veri Madenciliği

Her gün, her saniye insanlar tarafından üretilen terabayt mertebesindeki düzensiz veri tek başına hiç bir anlam ifade etmemektedir. Bu düzensiz ve anlamsız veri kümelerinin anlamlı olabilmesi için önceden kestirilemeyen ve bilinemeyen örüntülerin belirlenmesi gerekir. Bu örüntülerin belirlenmesinden sonra düzensiz veri yığınlarından bilgi edinmekten bahsedilebilir. Düzensiz veri yığınlarından bilgiye ulaşmayı hedefleyen veri biliminin bu alt çalışma alanına veri madenciliği adı verilir.

Veriden değerli bilgiye ulaşmak düzensiz ve büyük veri yığınları söz konusu olduğunda hemen ilk adımda gerçekleştirilemez. Bu yüzden büyük veri boyutları, veri çeşitliliği ve verinin üretim hızı problemi zorlaştırdığından elde edilmek istenen bilgi, verinin belli aşamaları geçmesinden sonra elde edilebilir.

Bu aşamalardan ilki elde edilmek istenen bilginin yani hedefin belirlenmesidir. Daha sonra edinilmek istenen bu bilginin bulunabileceği veri kaynakları belirlenmelidir. Bu kaynaklardan elde edilen veriler çok çeşitli yapılarda, içeriklerde olacağından araştırma yapılacak alana, hedefe yönelik başlangıç seçimleri yapmak gerekebilir. Yine verinin çeşidine göre istenmeyen, gereksiz, bilgi çıkarımında hiçbir etkisi olmayan hatta bilgi elde edilmesine mani olan gürültü adı verilen veri parçaları temizlenmelidir. Verinin tek seferde işlenebilecek örnekleme indirgenmesi, varsa eksik verinin tespit edilmesi, hangi tip verinin saklanacağı, ne şekilde saklanacağı bu aşamada kararlaştırılır.

Verinin hazırlanma süreci tamamlandıktan sonra bu veriden bilgi edinmeye yönelik bir model belirlenmelidir. Burada verinin çeşidine, elde edilmek istenen bilgiye göre tahmin edici ya da tanımlayıcı bir model seçilmelidir.

Tanımlayıcı modelde elde edilen verilerin ilişki ve kümelenme örüntüleri değerlendirilerek verinin ait olduğu sınıf bu örüntüler sayesinde tanımlanmaya çalışılır.

Bu çalışmada da kullanılan tahmin edici modelde ise elde edilen veriden çıkarılan örüntüler sayesinde önceden bilinen sonlu sayıdaki sınıfa ait bir sınıflandırma modeli elde edilir. Elde edilen bu model gelecekte benzer verilerin sınıflarının tahmin edilmesi için kullanılır. Bir metnin, belgenin ait olduğu sınıfı geçmiş metinlerdeki örüntülerden tecrübe ederek öğrenen ve bunu gelecekteki metinleri sınıflandırmada kullanan modeller tahmin edici modele örnek olarak verilebilir.

Son olarak değerlendirme aşamasında elde edilen bilginin doğruluğu, güvenirliği ve verimi değerlendirilerek modellemenin başarımı ölçülmeye çalışılır.

2.2 Metin Madenciliği

Diğer adıyla metin veri madenciliği veri madenciliğinin düzenli ya da düzensiz metin kümelerinden yüksek kaliteli bilgi edinimini hedefleyen bir alt dalıdır. Metinlerin içeriğinden bilgi edinimi, varlık ismi çıkarımı, öntanımlı örüntü çıkarımı (eposta, telefon numarası), eş atıf çıkarımı, ilişki, eylem, olgu çıkarımı ve son olarak bu çalışmanın da konusu olan düşünce analizi metin madenciliğinin başlıca konularıdır.

Veri madenciliği için tanımlanan veriden bilgiye ulaşma adımları metin madenciliği için de geçerli olduğundan metin madenciliğinde de ilk ve en önemli koşul başlangıç verisinin seçimidir. Metin madenciliğinde ya da metne özgü yapılan tüm çalışmalarda kullanılan başlangıç veri kümelerine derlem adı verilir.

2.3 Derlem Dilbilim

Doğa bilimlerinin aksine bilgisayarlı dilbilimin konusu olan doğal dil gibi yüksek entropiye sahip yapıları laboratuar ortamında soyutlayıp tekrar tekrar test etmek ya da

sistemi her açıdan açıklayan çok terimli matematiksel modellerini oluşturmak kolay bir problem değildir. Bu zor problemin çözümü için ilk ihtiyaç duyulan, entropisi bir miktar da olsa azaltılabilmiş, en azından ilk haline göre çok daha düzenli ve amaca uygun yapıya bürünmüş metin verisidir. Bu durumda doğa bilimleri çalışmalarında sıkça başvurulan laboratuar ortamına benzer şekilde, bilgisayarlı dilbilim, hipotezlerini test etmek amacıyla derlemleri kullanır.

Bilgisayarlı dilbilim açısından derlem, bir bakıma, ilgili konuya ait elemanlar ve bu elemanlara ait mantıksal aksiyomlardan oluşan bir ontolojidir.

1950'lerden itibaren derlemler, dilbilim için, bir dile, konuya özgü örnek metinlerden [2], cümle öğeleri işaretlenmiş metinlere [3] kadar çok çeşitli amaçlarda ve özelliklerde hazırlanmıştır. Ya da bu çalışmada da örnek alınan derlemlerde olduğu gibi taşıdığı düşünce veya yargı bakımından işaretlenmiş metinlerden oluşan derlemler [4, 5] de metin madenciliği ve derlem dilbilimin ilgi alanına dolayısıyla bu çalışmanın kapsamına girmektedir.

2.4 Düşünce Çözümleme

2.4.1 Bilişsel Bilimler Açısından Düşünce

Bilgisayar bilimleri yardımıyla düşünce çözümlemeyi tanımlamadan önce bilişsel bilimler bağlamında düşünceyi tartışmaya açmak faydalı olacaktır.

İnsan ve doğal olarak birey, akıl yürütme yetenekleriyle eyleme geçmesiyle daha sık anılsa da bir diğer özelliği de duygularının güdümünde hareket etmesidir. Akıl yürütme ya da duygularıyla hareket etmesi farketmeksizin insanın dış veya iç uyaranlar/olaylar karşısında düşünce dolayısıyla karşılık oluşturabilmesi için beş duyu organından gelen duyuları duyum, izlenim ve algıya dönüştürmesi gereklidir. İşte bu dönüşüm sürecine doğrudan veya dolaylı etki eden bileşenleri Scherer beşe ayrılmaktadır [6]:

- Duygu
- Ruh Hali
- Kişilerarası Tavır
- Tavır
- Kişisel Özellikler

Duygu, önemli bir olayın sonucunda ve eşzamanlı olarak kısa ve nettir. Örnekleri öfke, neşe, üzgün olma, korkma, utanma, gururlu olma olarak verilebilir. Ruh hali ise düşük yoğunluklu, sebepsiz ve uzun süreli olarak öznel hissiyattaki değişimdir. Ruh hali örnekleri de keyifli, ümitsiz, asabi, kayıtsız, kederli, kaygısız olma olarak verilebilir. Kişilerarası tavır; başka bir kişi ile oluşan özel bir etkileşim karşısında oluşan duygusal durumdur. Örneğin, arkadaşça, sıcak kanlı, mesafeli, soğuk, destekleyici, hor gören tavırlar kişilerarası tavıra örnek olarak verilebilir. Tavır; bir kişiye ya da objeye karşı oluşan uzun süreli, duygusal olarak taraflı inanışlardır. Örnekleri beğenmek, sevmek, nefret etmek, değer vermek, arzulamak olarak verilebilir. Kişisel özellikler de kalıcı kişilik ve kişisel davranış eğilimleridir. Sinirli, huzursuz, laubali, somurtkan, muhalif, kıskanç olma olarak örneklendirilebilir.

Bu çalışmada, düşüncesini yazılı olarak ifade eden bireyin oluşturduğu metinsel içerikten hareketle, konu bağlamında oluşturduğu tavrın bilgisayar yardımıyla özdevimli tespiti üzerinde durulmuştur.

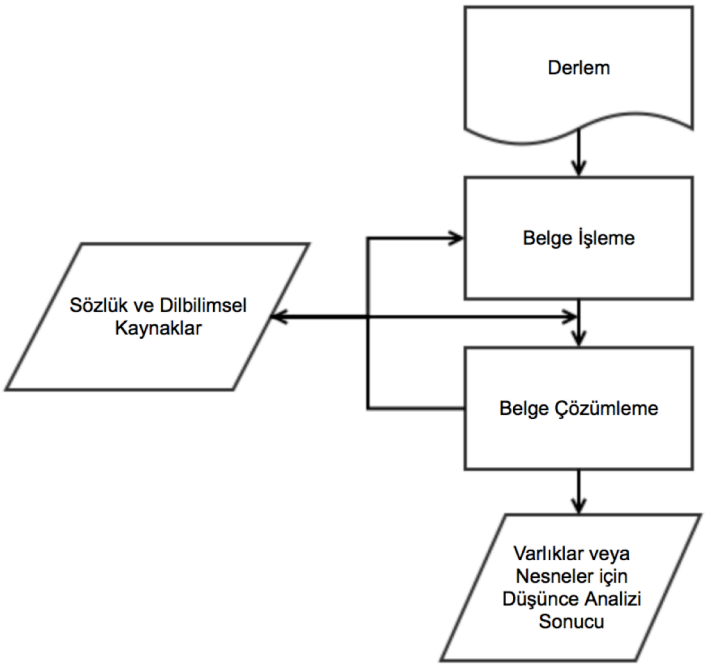
2.4.2 Bilgisayar Bilimleri Açısından Düşünce Çözümleme

Bilişsel bilimlerin bağlamı ile bilgisayar bilimleri açısından düşünce çözümleme, kişiler tarafından oluşturulmuş, daha önce tanımı verilen duygusal durumlardan biri olan tavırın tespit edilmesine yönelik bir metin madenciliği alt konusudur. Görüldüğü üzere veri madenciliğinin bir alt dalı olan metin madenciliği, hesaplamalı bilimler açısından disiplinlerarası bir çalışma konusu olduğu gibi düşünce çözümleme de hem hesaplamalı bilimlerin hem de bilişsel bilimlerin ilgi alanına giren disiplinler arası bir araştırma konusudur.

Bilgisayar bilimleri açısından düşünce çözümlemenin uygulamaları düşünsel tavrın tespitine yönelik dört konu ile ilgilenir:

- Tavrın kaynağı
- Tavrın hedefi
- Tavrın türü
    - Türü: beğeni, sevgi, nefret, değer, arzu
    - Tavrın kutupluğu: olumlu, olumsuz, tarafsız
- Tavrı içeren metin
    - Cümle ya da belge

Bilgisayar yardımıyla düşünce çözümlemede yukarıda bahsi geçen konuların tespitine yönelik bireylerin ürettiği farklı seviyelerdeki beş içeriği kullanacak yöntemler geliştirilmiştir [7]. Bunlar belge seviyesinde, cümle seviyesinde, nitelik temelli, kıyaslamalı, düşünce sözlüğü edinimli olarak sıralanabilir. Şekil 2.1'de genel hatlarıyla bir düşünce çözümleme sisteminin bileşenleri ve mimarisi verilmiştir.



Şekil 2.1 Genel Bir Düşünce Çözümleme Mimarisi Örneği

2.4.2.1 Belge Seviyesinde Düşünce Çözümleme

Gözetimli yapay öğrenme yöntemleri kullanarak bir belgenin yazarı tarafından belirtilen ve önceden bilindiği kabul edilen ana nesne hakkındaki düşüncesini tespit etmeye çalışır. Düşünce analizinin ilk ve en temel uygulama alanıdır. Basitçe her bir sınıf için eğitim verisi mevcut olan, sonlu sayıdaki sınıfa ait belgeleri sınıflandırmaya çalışır. Belgeleri olumlu ya da olumsuz düşünce barındırmasına göre ayırarak ya da ayrık tam sayı değerlerinden oluşturulmuş bir ölçekteki sınıflara göre sınıflandırmayı hedefler. Örneğin düşünce çözümlemedeki temel çalışmalardan biri olan B. Pang ve arkadaşlarının [4] çalışmasında da Amazon'un[2] kullandığı beş yıldızlı ürün puanlama sistemi, ürün hakkında yapılan yorumların taşıdığı düşünceyi belirlemek için kullanılmıştır. Daha sonra bahsedileceği üzere Türkçe için yaptığımız bu çalışmada da kutupluk derleminin oluşturulması için benzer bir yöntem uygulanmıştır.

2.4.2.2 Cümle Seviyesinde Düşünce Çözümleme

Bazı durumlarda tek bir belge düşünce bakımından birden fazla sınıfa ait özellik barındırabilir. Bu durumda belgeyi belge seviyesinde sınıflandırma yöntemleri ile sınıflandırmaya çalışmak yetersiz gelecektir. Bu gibi durumlarda belgenin alt kümesi sayılacak düşünce taşıyan elemanlarının düşünce özelliklerine odaklanmak faydalı olacaktır. Böylelikle belge düzeyinde düşünce kutupluğu aramanın sonuçsuz kaldığı durumlarda belgenin bir alt kümesi kabul edilebilecek cümle seviyesinde düşünce kutupluğu sorgulanır. Bu seviyede kullanılan yöntemlerin büyük çoğunluğu cümleleri olumlu ya da olumsuz olarak iki sınıfta değerlendiren gözetimli öğrenme kullanır [8].

Bu yaklaşımda her cümlenin bahsettiği varlığın önceden bilindiği ve yine her cümlenin tek sınıfa ait düşünce barındırdığı kabul edilir. Cümlenin düşünce kutupluğunun hesaplanabilmesi için cümlenin öznel olması gerekmektedir. Bu ön kabullerden anlaşılacağı üzere bu yöntem uygulanmadan önce gözetimli öğrenme için kullanılacak eğitim seti mutlaka insan kontrolünde işlenmeye muhtaçtır. Büyük eğitim setleri hazırlamada insan kontrolüne ihtiyacı azaltan bazı çalışmalar mevcuttur [9]. Yine

[2] http://www.amazon.com

Pang ve Lee'ye ait çalışmada [10] komşu cümlelerin aynı öznel sınıfa ait olduğu kabul edilerek özdevimli sınıflandırma denenmiştir. Bu yöntem için yapılan son çalışmalar [11] göstermiştir ki belge seviyesindeki sınıflandırmanın aksine farklı tiplerdeki cümlelere farklı yaklaşımlar gerekmektedir.

2.4.2.3 Nitelik Temelli Düşünce Çözümleme

Bundan önceki iki yaklaşım da tek bir varlığı işaret eden düşünceleri barındırsa da bir çok durumda insanlar bir ürün hakkında düşüncelerini ifade ederken ürüne ait her bir farklı nitelik için farklı düşünceler belirtirler.

"*Telefonun şarjı çok iyi gidiyor **fakat** çok ağır bir telefon."* (2.1)

*"Sürüş konforu mükemmel **ancak** yakıt sarfiyatı oldukça yüksek!"* (2.2)

Örnek 2.1 ve 2.2'den görüleceği üzere daha önceki iki yöntem ile bu cümleleri sınıflandırmayı tercih etmek kullanıcının tek bir cümlenin iki yarısında belirttiği çok değerli görüşlerin atlanmasına sebep olacaktır. Yukarıdaki iki örnekte değerli bilginin kaybolmaması için aynı varlığa ait farklı nitelikler hakkındaki görüşlerin tespit edilip sınıflandırılması gerekmektedir. Bu durumda uygulanan ilk yöntem ürün yorumları ihtiva eden ve frekansı deneysel olarak elde edilmiş bir eşik değeri geçen isim öbeklerinin tespit edilerek aynı ürünün farklı niteliklerine ait düşüncelerin tespit edilmesidir [12]. Bu yaklaşımda derlem yardımıyla eğitilen dizi sınıflandırıcı, Koşullu Rassal Alanlar (CRF) [13], niteliklerin bulunması için [14] kullanılan yöntemlerden biridir.

2.4.2.4 Kıyaslamalı Düşünce Çözümleme

İnsanların satın alma alışkanlıklarının doğası gereği bir ürün hakkında fikir beyan ederken yukarıdaki durumların dışında bir de başka ürünler ile kıyaslayarak ürün hakkında düşünce belirtmeleri de yaygın bir davranıştır.

"*Diğer telefon bence bundan çok daha iyi.*" (2.3)

Örnek 2.3'te görülen cümle klasik yaklaşımlar ile ele alındığında "çok" ve "iyi" sözcüklerini ihtiva etmesi bakımından belki de "olumlu" olarak sınıflandırılacaktır. Ancak gerçekte durum oldukça farklıdır. Ancak burada "daha" gibi küçük fakat çok etkili sözcükler öznitelik olarak belirlendiği takdirde bu tip cümlelerin ihtiva ettiği asıl düşünce doğru bir şekilde sınıflandırılabilmektedir [15].

2.4.2.5 Düşünce Sözlüğü Edinimi

Şimdiye kadar gördüğümüz tüm yaklaşımların olmazsa olmaz şartı düşünce bilgisini taşıyan sözcüklerin ya da sözcük öbeklerinin tespitine yöneliktir. Bu sözcüklerin tespiti için bir kaç yol mevcuttur. Bunlardan ilki bu sözcüklerin insan kontrolünde elde edilmesidir. Konu çeşitlilikleri, el ile işlenmesi gereken verinin miktarı göz önüne alındığında bunun çok da makul bir yöntem olmadığı aşikardır.

Bunun yerine küçük bir öznitelik kümesiyle başlayıp WordNet yardımıyla [16] ya da daha gelişmiş sözlük yaklaşımlarıyla [17] elde edilecek düşünce sözlükleri ile sınıflandırma yapmak mümkündür. Bunlarla birlikte doğal dil içinde mevcut "ve", "veya", "ne… ne de", "ya… ya da" gibi mantık operatörleri ile birbirine bağlı öznitelikler özdevimli tespit edilerek oluşturulan sözlük edinim yöntemleri de mevcuttur [18]. Bir konuda elde edilen düşünce yönünden belirleyici, sözcük ya da sözcük öbeklerinin diğer konulardaki etkilerini araştıran çalışmalar da yapılmıştır [19, 20].

Bunlarla birlikte doğal dil için mevcut başka bir mantıksal durum olan olumsuzlama için de Sanjiv ve Chen, Pang ve Lee'nin yaptığı çalışmalarda uygulanan yöntemin bir benzeri Türkçe için de bu çalışmada uygulanmıştır [4, 21].

İngilizce için yukarıda bahsedilen düşünce sözlüğü örnekleri [12, 22, 23, 24, 25] oldukça bol ve çeşitli özelliklerde mevcut iken Türkçe için umumun kullanımına açık böyle bir sözlük bulunmamaktadır. Bu çalışmanın amaçlarından bir diğeri de yukarıda örnekleri verilen düşünce sözlüğü uygulamalarının Türkçe için uygulanabilmesi açısından temel oluşturmaktır.

2.4.3 Türkçe Odaklı Çalışmalar

Türkçe için bilinen ilk çalışma Eroğul tarafından yapılmıştır [26]. Eroğul tezinde, destekçi vektör makineleri (SVM – Support Vector Machines) ve doğal dil işleme yöntemleri kullanarak film konusuna ait belgeleri %85 başarıyla sınıflandırmıştır.

Kaya ve arkadaşları Türkçe siyasi haber içeriklerinde Naïve Bayes, SVM, MaxEnt ve karakter tabanlı N-Gram yaklaşımlarını deneyerek %65-%77 arasında başarılı sınıflandırma sonuçları elde etmişlerdir [27].

Boynukalın ise tezinde [28] daha önceki çalışmaların aksine içeriğin kutupluğundan çok taşıdığı duyguya göre sınıflandırmayı denemiştir.

Çakmak ve arkadaşları ise yaptıkları iki çalışma [29, 30] ile Türkçe sözcük köklerinin ve cümlelerinin taşıdıkları özellikler ile ifade ettikleri duygu arasında çok güçlü bir korelasyon bulunduğunu ortaya koymuşlardır.

Kaya ve arkadaşları ise önceki çalışmalarına [29] ek olarak twitter'dan[3] elde ettikleri veriyle eğittikleri modellerini siyasi haber içeriklerini sınıflandırma sonuçlarını iyileştirmekte kullanmışlardır [31].

Son olarak Demirtaş ise tezinde İngilizce – Türkçe makine çevirisi yöntemlerini kullanarak kurguladığı bir düşünce çözümleme çözümünü denemiştir [32].

[3] http://www.twitter.com

BÖLÜM 3

BİLGİSAYARLI DÜŞÜNCE ÇÖZÜMLEMEDE KULLANILAN YÖNTEM VE ARAÇLAR

Bu bölümde sırasıyla, işaretli derlemlerden, işaretli derlemin oluşturulmasında dikkat edilecek hususlardan ve özdevimli olarak derlem oluşturmak için gerekli verileri toplamaya yarayan ağ örümceklerinden, oluşturulan derlemin kirli veriden arındırılması ve üst veri çıkarımı için gerekli bilgisayarlı doğal dil işlemeden, son olarak da belge seviyesinde sınıflandırma için gerekli yapay öğrenmeden bahsedilmiş olup bu konular hakkında genel bilgiler verilmiştir.

3.1 İşaretli Derlem (Annotated Corpus) ve Önemi

İnternetin her tür bilgi için oldukça önemli bir kaynak olduğu herkesin malumudur. Dileyen kişi bilgisayar programlamadan, yemek pişirmeye, bir müzik enstrümanı çalmaya kadar değişen bir yelpazede amacına uygun bilgiye ulaşabilir. Bu bilgilerin yanında internetin bir diğer bilgi katmanı da metin, fotoğraf, ses, video gibi muhtelif formlarda sunulan bilgilerin içerikle ve birbiriyle ilişkisidir. Bilgisayarlar, bu ilişkileri ve doğal dilin kendisini anlama kapasitesine sahip insana sunmakta oldukça başarılı iken, bunları kendi kendine anlama hususunda oldukça yetersizdir. İşte tam bu noktada çeşitli formatlardaki içeriklerin ilişkilerini, dilini, yapısını temsil eden, örneklerini içeren doğal dil veri kümelerine derlem (Corpus) adı verilir. Yapay öğrenme ve dolayısıyla bilgisayarın öğrenebilmesi için bu içeriklerin bilgisayarların anlayabileceği şekilde işaretlenmesi ve böylelikle işaretli derlemlerin oluşturulması gerekmektedir.

İnsanoğlu her gün yeni bir problemi bilgisayarlara çözdürmeyi başarıyor. Satrançtan, yüz tanımaya, en kısa yolu bulmaya kadar bir çok alandaki problemlerin bilgisayar yardımıyla çözümleri oldukça iyi sonuçlar elde etmemizi sağmaktadır. Fakat

tüm bunlara rağmen özellikle doğal dil alanında halen bilgisayarlar yardımıyla çözmeyi tam olarak başaramadığımız birçok zor problem mevcuttur. İstatistiksel yaklaşımlar bu tarz problemleri çözmeye yönelik en etkili yöntemlerden biri olarak kabul edilse de ileriki bölümlerde detaylarını vereceğimiz yapay öğrenme yöntemleri istatistiksel yöntemlerin aksine çok büyük hacimlerdeki verilere ihtiyaç bırakmadan daha kısıtlı veri kümeleri ile veriyle alakalı uygun işaretçileri kullanarak daha başarılı sonuçlar verebilmektedir. Doğal dil söz konusu olduğunda bu işaretçilerin (Annotation) çoğunlukla, doğal dil metnine ek bilgiler sağlayan üst veri (Metadata) olması beklenmektedir. Bununla birlikte bir bilgisayara etkin bir şekilde birşeyler öğretebilmek için atılması gereken en önemli adımlar, doğru verinin yeter miktarda sağlanmasıdır. Bu yüzden bilgisayarlı dil bilim ve dolayısıyla bilgisayar yardımı ile düşünce analizinin de en önemli aşamalarından bir tanesi, çözülmesi istenen probleme uygun olarak işaretli derlemin oluşturulmasıdır.

3.2 İşaretli Derlemin Amacının Belirlenmesi ve Derlem Oluşturma Aşamaları

Yapay öğrenmenin söz konusu olduğu çalışmalarda işaretçilerin ve işaretlemenin açık tanımını başlangıçta ortaya koymak oldukça önemli ve mutlaka yapılması gereken bir işlemdir. İşaretçi etiketlerin belirlenmesi sırasında, derlem yazım kuralları, işaretçiler ile nasıl çalışılacağı ve yapay öğrenme algoritmaları gibi detaylarla uğraşmak bir anda asıl ulaşılmak istenen hedeften sapmaya kolaylıkla sebep olabilir. Bu yüzden oluşturulacak derlemin hedefi çok net şekilde belirlenmeli ve ifade edilmelidir.

İngilizce kısaca MATTER [33, 34] olarak ifade edilen bir döngü ile yukarıda ifade ettiğimiz adımlar derlem oluşturulurken ve bu derlem ile çalışılırken hedeften sapmayı engelleyecek bir metodoloji ortaya koymaktadır. MATTER sırasıyla aşağıdaki adımlardan oluşmaktadır:

Model: Veri üzerinde yapılan ampirik gözlemler ile elde edilen kuramsal bilgiler ile ifade edilen yapısal açıklamalardan oluşur.

Annotate (İşaretleme): İşaretçi şeması gerekli öznitelikleri ya da hakkında bilgi edinilmek istenen etiketleri özel bir yapı ile ve modele uygun özellikler ile girdi datasında işaretlenmesidir.

Train (Eğitim): İşaretli derlemin yapay öğrenme algoritmaları kullanılarak hedef etiket konusunda eğitilmesidir.

Test (Sınama): Eğitim aşamasında algoritmanın henüz görmediği ancak etiketleri bilinen veri ile algoritmanın test edilmesidir.

Evaluate (Değerlendirme): Tekrarlanabilen hesaplamalar ile elde edilen sınama sonuçlarının değerlendirilmesidir.

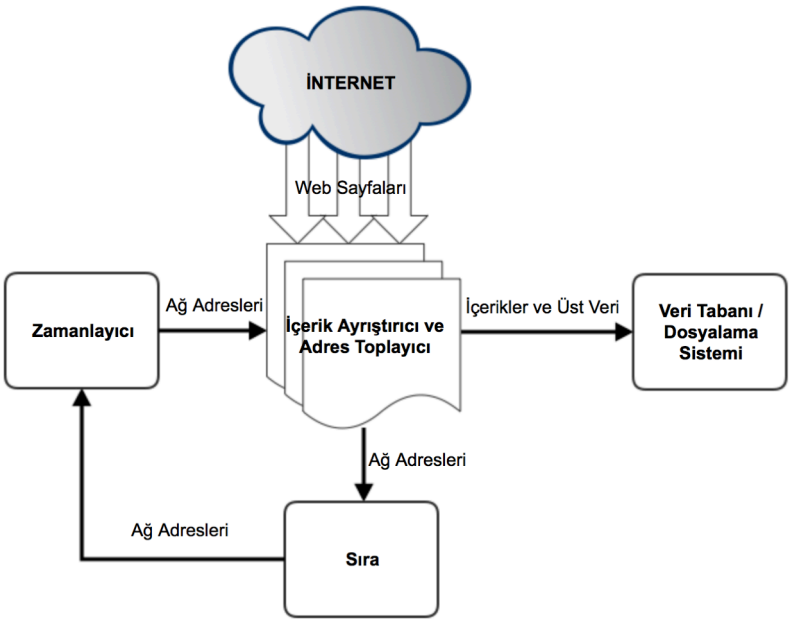
Revise (Düzeltme): Model ve işaretleme yöntemlerinin elde edilen sonuçları iyileştirmek için tekrar gözden geçirilmesi, gerekiyorsa düzeltilmesi aşamasıdır.

MATTER ile birlikte iyi bir derlemin mutlaka sahip olması gereken iki özellik daha vardır. Bunlar derlemin üzerine çalışılması planlanan problem hakkında karakteristik bilgileri taşıyor olması ve örneklem bakımından dengeli olmasıdır. Bunun sebebi doğal dilin sonsuz örneklere sahip olduğu gerçeğiyle birlikte, derlemin sonlu ve seçilmiş bir alt küme ile çözülmesi amaçlanan problemin çözümüne yardımcı olması gerekliliğidir. Her ne kadar bir kaç cümle ile derlemin karakteristik ve dengeli olması ifade edilebilse de bunu uygulamak pratikte oldukça zor ve zahmetli bir problemdir ve yıllardır derlem dilbilimcilerin üzerinde çalıştığı başlıca problemlerden biridir.

3.3 Derlemi Oluşturan Verinin Toplanması ve Ağ Örümcekleri (Web Spiders)

Metinsel işaretleme ile bir derlem oluşturulması hedefleniyorsa bu derlemi oluşturmak için gereken verinin internet aracılığıyla toplanması en yaygın olarak kullanılan yöntemlerden biridir. Bu sebeple verinin toplanması için gerekli URL'lerin elde edilmesi, sıralanması, ilgili URL'lerden elde edilen metinlerin HTML'den arındırılması, dosya sistemine ya da bir veritabanına kalıcılaştırılması için özelleştirilmiş bilgisayar yazılımlarına ihtiyaç vardır. Bu amaçla oluşturulmuş bilgisayar yazılımları günümüzde örümcek adıyla da bilinen ağ örümcekleridir.

Ağ örümcekleri, verilen bir ağ adresinden başlayarak var olan tüm ağ adreslerini özyinelemeli olarak bulup daha sonra ziyaret etmek üzere yığınlayarak biriktiren, bu yığındaki adresleri sırasıyla gezerek elde ettiği verileri veri tipine (HTML, Ses, Video, vb), içeriğine göre istenilen bir ortama kalıcılaştırmayı sağlayan bilgi edinimi (Information Retrieval) sistemlerine ait elemanlardır.



Şekil 3.1 Genel Bir Ağ Örümceği Mimarisi

Günümüzde internet kaynaklı büyük verinin el ile işlenemesi olanaksızdır. Bu durum, internet kaynaklı verileri özdevimli olarak elde edebilmek için geliştirilen bir ağ örümceğinin yetenekleri hususunda beklentileri oldukça yükseltmiştir. Bu bağlamda, bir ağ örümceğinden beklenen özellikler ikiye ayrılır. Sahip olması zorunlu özellikler ve mutlaka sahip olması gereken özellikler.

Zorunlu özellikler şunlardır:

- Dayanıklılık (Robustness): Bazı durumlarda web siteleri ağ örümceklerini sonsuz döngüye sokacak tuzaklara sahip olabilir. Bu ve benzeri durumlara dayanıklı olması bir ağ örümceğinden beklenen olmazsa olmaz özelliklerdendir.
- Kurallarra Riayet (Politeness): Websiteleri, trafik ve işlem yükünü dengeleyebilmek ve gerçek kullanıcılara daha performanslı hizmet sunabilmek için sistemlerini yorması, zorlaması muhtemel ağ örümceklerinin ziyaret ve işlem sıklığına bir sınırlama getirebilir. Bununla birlikte ağ örümceklerinin ziyaret edip kayıt altına almasını istemediği bölümleri, alanları belirtebilir. Ağ

örümceği web sitesi sahibinin örtük ya da açık olarak koyduğu bu kurallara mutlaka riayet etmelidir.

Olması gereken özellikleri ise şunlardır:

- Dağıtık Mimari: Birden çok makinede dağıtık bir şekilde ağ örümcekleri çalıştırılabilir olmalıdır.
- Ölçeklenebilirlik: Ek sunucu ve bant genişliği sağlanarak ağ örümceği kolaylıkla daha büyük verileri işlemek için ölçeklenebilir olmalıdır.
- Performans ve Etkinlik: Ağ örümceği, üzerinde çalıştığı sunucunun sistem kaynaklarını etkin biçimde kullanmalıdır.
- Kalite: Herhangi bir veri önceliklendirme anlayışına gerek duymadan bulduğu her veriyi kaydetmekten ziyade ihtiyaca yönelik ve istenilen kaliteli veriyi önceliklendirecek bir altyapısı olmalıdır.
- Güncellik: Veri topladığı kaynakların kopyalarını sürekli güncel tutmalıdır.
- Genişletilebilirlik: Ağ örümcekleri mutlaka çok çeşitli protokolleri, veri formatlarını ihtiyaca yönelik olarak kolayca destekleyebilecek özelliklere sahip olmalıdır.

3.4 Doğal Dil İşleme (Natural Language Processing)

Doğal dil işleme, insana özgü dillerin bilgisayar bilimleri açısından işlenmesi ve kullanılması hususunda araştırmalar yapılan bir bilgisayar bilimleri çalışma alanıdır. İnsan diline özgü, yazılı ya da sözlü parçacıkların anlam çözümlemesinden, sözdizimsel incelenmesine, cümle sınırı belirginleştirmesinden, cümle öğelerine ayrıştırılmasına kadar çok geniş bir uygulama ve araştırma alanı vardır.

Bilgisayar bilimleri, bilgisayarlı dilbilim, bilişsel bilimler ve yapay zeka gibi araştırma alanlarının doğrudan ilgi alanına giren bir çalışma alanı olarak kabul edilen doğal dil işlemenin en önemli amaçlarından biri makine insan etkileşimini doğal dil aracılığıyla gerçekleştirebilmektir. Doğal dil işleme bu amaca yönelik olarak günümüzde aşağıda bahsi geçen ana araştırma alanlarına ayrılmıştır.

- Soru Cevaplama Sistemi: Kişilerin başka bir arayüze ihtiyaç duymadan doğrudan doğal dil ile makinelere soru sorup cevap alabilmesini amaçlamaktadır.
- Özdevimli Metin Özeti Çıkarımı: Uzun ve büyük hacimde birikmiş metin kümelerinden kişilerin okuduğunda rahatlıkla anlayabileceği özet bilgiler derleyebilen bir yöntemdir.
- Makine Tercümesi: Doğal dil işlemenin en aktif çalışma konularından biridir. Kaynak doğal dilden hedef doğal dile başarılı bir şekilde özdevimli tercümeyi hedefler.
- Konuşma Tanıma: Doğal dil işlemenin en zorlu problemlerinden biri olarak bilinen bu alanın amacı insan sesini tanıyarak makine insan etkileşimini hedeflemektedir.
- Metin ve Belge Sınıflandırma: Doğal dil işlemenin en başarılı olduğu ve bu çalışmanın da özünü oluşturan alandır. Doğal dil metinlerini ya da bu metinleri içeren belgeleri önceden bilinen kategorilere özdevimli olarak ayırmayı hedeflemektedir.

Yukarıda bahsedilen çalışma alanlarında ilerleme kaydedebilmek için özellikle İngilizce özelinde oldukça kapsamlı ve etkili yazılım altyapıları ve kütüphaneleri geliştirilmiştir. Bunların başlıcaları OpenNLP ve NLTK'dir. Bu çalışmada da Türkçe özelinde doğal dil işleme için kullanılan yazılım kütüphanesi Zemberek olmuştur.

OpenNLP[4], The Apache Software Foundation[5] bünyesinde desteklenen ve geliştirilen açık kaynak kodlu, yapay öğrenme tabanlı bir doğal dil metin işleme kütüphanesidir. Java diliyle geliştirilmiştir. Temel doğal dil işleme işlemlerini başarıyla yapabilen ve çalışılan doğal dilden bağımsız altyapısı sayesinde istenilen amaca ve doğal dile uygun şekilde genişletilebilen bir yapısı vardır.

[4] https://opennlp.apache.org

[5] https://www.apache.org/

NLTK[6], Python[7] programla dili ile yazılmış daha çok İngilizce odaklı, içinde 50'den fazla derlem ve sözlük barındıran açık kaynaklı bir doğal dil işleme kütüphanesidir. Bu çalışmada, verileri elde etmek için yazılan ağ örümceği python programlama dili ile yazıldığından ilk aşamadaki belirtkelerine ayırma gibi basit, temel işlemleri yapmak amacıyla kullanılmıştır.

Zemberek2 [8] , Java [9] programlama dili ile yazılmış Türkçe'ye özgü bir biçimbilimsel doğal dil işleme kütüphanesidir. Türkçe sözcükler için kök bulma, eklerine ayırma, yazım hatası denetleme, heceleme gibi özellikler sunan oldukça kullanışlı ve Türkçe için şimdilik tek doğal dil işleme kütüphanesidir. 2011 yılından itibaren Zemberek-NLP[10] adı altında yeni bir sürümü geliştirilmeye başlandıysa da henüz olgunluk aşamasına ulaşmadığından eski sürüm olan Zemberek2 bu çalışmada Türkçe için uygulanan tüm doğal dil işleme işlemlerinde kullanılmıştır.

3.5 Yapay Öğrenme (Machine Learning)

Yapay öğrenme ya da diğer adıyla makine öğrenmesi bilgisayar bilimlerinin veriyle karşılaştıkça kendi kendine gelişen sistemlerin geliştirilmesine odaklanmış bir alt alanıdır. Nobel ödüllü ekonomist Herbert Simon'un "*Her hangi bir sistemin bir süreç sonunda performansını tecrübe ile arttırması öğrenmedir.*" sözünden ilhamla başlayacak olursak; bilgisayarların doğal dil öğrenmesi için tecrübe etmesi gereken girdi doğal dil metinleri, dolayısıyla işaretli derlemlerdir denilebilir. Ve pek tabii olarak doğal dil metni özelinde tecrübe ettiği ayrıntılar da daha önceden işaretçiler ile belirlenmiş ve derleme işlenmiş ya da bu işaretçiler yardımıyla edinilmiş, metnin amaca uygun öznitelikleri de tecrübeye ve öğrenme sürecine katkı veren en önemli unsuru oluşturur.

---

[6] http://www.nltk.org

[7] http://www.python.org

[8] http://code.google.com/p/zemberek/

[9] http://www.oracle.com/us/technologies/java/

[10] https://github.com/ahmetaa/zemberek-nlp/

Yapay öğrenme bilgisayarlı dilbilim açısından şu çalışmaları kapsar:

- Kelimelerin kategorilerini tespit etmek.
- Makale, e-posta, web sayfası konusunu tespit etmek.
- Bir kelime ya da kelime öbeğine anlambilimsel ya da varlıkbilimsel bir sınıf atamak.
- Doğal dil tespitini yapmak.
- Kelime telafuzu tanımlamak.
- El yazısı tanımak.
- Sözdizimsel yapıyı tanımak.
- Cümle tanımak.
- Bir olay ya da metnin zamanını tespit etmek.
- Metnin içinde geçen olayları tarihsel olarak ilişkilendirip sıralayaabilmek.
- Bir cümle içindeki anlambilimsel rolleri tespit edebilmek.
- İsimlendirilmiş varlık tanımlamak.
- Aynı kaynaktan atıf alan birden çok terimli doğal dil yapılarını çözümleyebilmek.
- Söylem yapısını tanımlayabilmek,
- vs.

Bu kadar çok çalışma alanı ilk başta karmaşık gelse de yapay öğrenme açısından bir kaç yaklaşımla bu problemleri çözmeye çalışmak mümkün olmaktadır. Bu yaklaşımlar sırasıyla şunlardır:

- Gözetimli Öğrenme (Supervised Learning): Önceden bilinen girdi ve girdi kategorisi ikililerinden tecrübe ederek öğrenme sürecinde kategorisi bilinmeyen bir girdinin kategorisini doğru olarak tespit edebilecek bir f fonksiyonunu elde etmeyi amaçlayan bir yöntemdir. Sınıflandırma yöntemi olarak da adlandırılır.
- Gözetimsiz Öğrenme (Unsupervised Learning): Önceden herhangi bir girdi kategorisi bilinmeden tüm girdilerin birbirleri ile ilişkilerinden, özelliklerinden hareketle benzer girdileri mümkün olduğu kadar kümeleme (Clustering) yöntemidir.

- Yarı Gözetimli Öğrenme (Semi Supervised Learning): Hem gözetimli hem de gözetimsiz öğrenme özelliklerinden aynı anda yararlanarak bir işlev ya da sınıflandırıcı oluşturmayı amaçlayan yöntemdir.

Bu çalışmada metin sınıflandırma amaçlandığından sadece gözetimli yapay öğrenme yöntemleri ele alınmış konunun odağının dağılmaması açısından metin sınıflandırmada sık kullanılan algoritmalar ele alınmıştır. Bu algoritmalar şu şekilde sıralanabilir:

- Karar Ağaçları: En yaygın ve temel sınıflandırma algoritmalarının başında gelir. Temel özelliği girdi veri kümesini kümeleme yöntemiyle özyinelemeli olarak kümelenen tüm alt küme elemanlarının aynı sınıf etiketine sahip olduğu duruma erişmeye çalışır.
- Naïve Bayes: Adını İngiliz matematikçi Thomas Bayes'ten alan ve yapay öğrenmeye olasılıkçı açıdan yaklaşan bir sınıflandırma algoritmasıdır. Bölüm 4.4.2'de ayrıntılı olarak incelemeden önce Naïve Bayes kısaca her bir özniteliğin birbirinden bağımsız olması gerekliliğini temel alır. Bu ilk bakışta yetersiz bir yaklaşım gibi görünse de pratikte oldukça başarılı sonuçlar vermektedir.
- Destek Vektör Makineleri: Sınıflandırmada kullanılan en etkili yöntemlerden biridir. Bir Düzlemde bulunan iki grup arasında bir çizgisel sınır çizerek iki grubu ayırmaya çalışır. Bu çizgi öyle bir yerden geçmelidir ki hem iki grubun üyelerini en etkin biçimde ayırabilmeli hem de bu çizginin uzaklığı iki grubun da elemanlarına olabilecek en uzak mesafede bulunmalıdır. Her ne kadar iki sınıf problemini çözen bir algoritma olsa da ek yaklaşımlarla çok sınıflı problemlerde de oldukça başarılı sonuçlar verebilmektedir.
- Maksimum Entropi (MaxEnt): Çokterimli lojistik regresyon olarak da bilinen ve lojistik regresyonun çok sınıflı sınıflandırma problemlerine uyarlanmış halidir. Naïve Bayes'e oldukça benzeyen bu yöntemin Naïve Bayes'ten farkı, özniteliklerin birbirinden bağımsız olduğu kabulünün bu yaklaşımda yapılmasına gerek olmamasıdır.

Yukarıda sayılan yapay öğrenme yaklaşımlarını ve bu yaklaşımlar için kullanılan muhtelif algoritmaları gerçeklemek hem zaman hem de gerçeklemenin etkin bir performansa sahip olması bakımından oldukça maliyetli süreçlerdir. Bu yüzden bu tarz çalışmalarda bu yaklaşımları sıfırdan gerçeklemek yerine önceden bir çok bilim adamı ve yazılım geliştiricinin katkılarıyla oluşturulan kütüphaneleri kullanmak oldukça akılcı bir yöntem olacaktır. Yapay öğrenme için ticari ve akademik amaçlarla kullanılan bir çok açık kaynak kütüphane bulunmaktadır. Bu kütüphanelerin en bilinenleri şunlardır:

- Apache Mahout[11]
- Scikit-Learn[12]
- Weka[13]

olarak sıralanabilir. Bu çalışmada hem içerdiği algoritmaların çeşitliliği bakımından hem de akademik olarak oldukça sık kullanılması bakımından Weka tercih edilmiştir.

Weka, Waikato Üniversitesinin geliştirdiği ve GNU Genel Kamu Lisansı ile dağıtımını yaptığı açık kaynaklı, java dilinde yazılmış bir yapay öğrenme kütüphanesidir. Weka, "Waikato Environment for Knowledge Analysis" kelimelerinin baş harflerinden oluşan bir kısaltmadır.

Tamamen genişletilebilir bir yazılım tasarımına ve görsel arayüze sahip olması Weka'yı birçok geliştirici ve akademisyen bakımından oldukça cazip hale getirmektedir. Yazılım girdi olarak CSV ve ARFF dosyalarını desteklemektedir. CSV bir çok sistem için ortak bir dosya formatı sayılabilirken ARFF Weka'ya özgü bir dosya formatıdır. Eklerde örnek bir ARFF dosyasının içeriği verilmiştir.

Weka bir veri madenciliği, yapay öğrenme uygulanmasından bekleneceği üzere temel olarak:

---

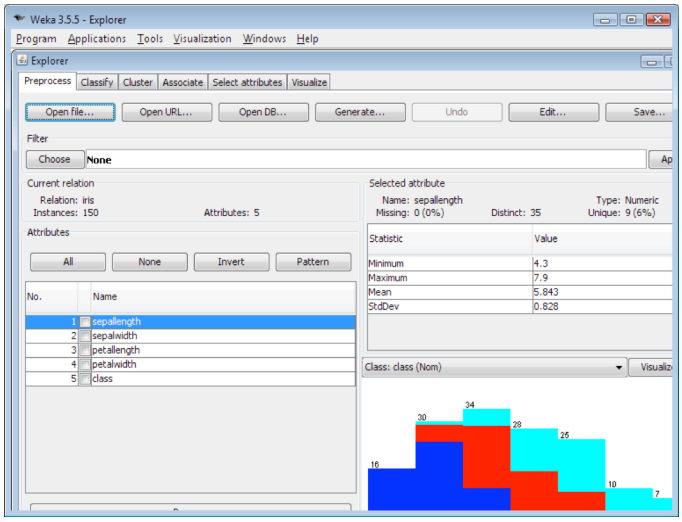
[11] http://mahout.apache.org

[12] http://scikit-learn.org/stable/

[13] http://www.cs.waikato.ac.nz/ml/weka/

- Veri Ön İşleme
- Veri Görselleme
- Sınıflandırma
- Kümeleme
- İlişkilendirme

işlemleri yapabilmektedir. Yazılımın dezavantajlı özellikleri olarak çoğu durumda çok çekirdekli çalışma prensibini ya da dağıtık çalışma prensibini benimsememesi gösterilebilir. Ancak bu dezavantajlarına rağmen bilinen birçok yapay öğrenme algoritmasının ve yönteminin gerçeklenmiş halde hazır sunması Weka'yı şimdilik bu konuda rakipsiz kılmaktadır. Şekil 3.2'de yazılımın görsel arayüzü görülmektedir.



Şekil 3.2 Weka Yazılımının Arayüzü

BÖLÜM 4

BİLGİSAYARLI DÜŞÜNCE ÇÖZÜMLEME İÇİN YAPILAN ÇALIŞMA

Bu çalışmada sırasıyla aşağıdaki yöntemler uygulanmıştır.

- Ağ örümceği ile bir sinema sitesinden kullanıcı yorumlarının toplanmış,
- Toplanan yorumların kutupluğu özdevimli olarak işaretlenmiş,
- İşaretlenen yorumlar ASCII -> Türkçe kontrolünden geçirilmiş ve olağanlaştırılmış,
- N-Gram(1-4 Gram) sözcük öbeklerinden,
- Olumsuzluk ekine sahip sözcüklerin işaretlenmesinden,
- İsim, sıfat, fiil,
- İsim + fiil,
- İsim + sıfat,
- İsim + sıfat + fiil,
- Sıfat + fiil

sözcük sınıflarından oluşan belirtkeleri içeren ayrı ayrı Türkçe kutupluk derlemleri oluşturulmuştur. Bu derlemlerin yapay öğrenme eğitim seti olarak kullanılabilmesi için derlemlerdeki her bir yorum örneği Terim Frekansı – Ters Metin Frekansı değerleri kullanılarak birer yöneye çevrilmiştir. Her bir yöneyde, sınıfın belirlenmesine katkı verebilecek öznitelikler bilgi edinimi yöntemiyle belirlenmiştir. Bu öznitelik

değerlerinden oluşan yöneylerin eğitim seti olarak kullanılmasıyla Naïve Bayes Çokterimli sınfılandırıcısını kullanan yapay öğrenme temelli bir düşünce çözümleme yöntemi Türkçe için önerilmiştir. Son olarak bu yöntemin başarımının ölçülmesinde kullanılan ölçütlerden bahsedilmiştir.

4.1 Derlem Verisinin Elde Edilmesi

Bölüm 2.3'ten hatırlanacağı üzere doğal dil işleme ile ilgili hipotezlerimizi gerçekleyebilmemiz ve sınayabilmemiz için derlemlere ihtiyaç duyulmaktadır. Bu bölümde ağ örümcekleri kullanarak istenilen konuya ait kaynakları tarayan ve istenilen konuda içerik ve bu içeriğe ait üst veriyi özdevimli yollara elde etmek amacıyla kullandığımız yöntemlerden bahsedilmiştir. Son olarak da elde ettiğimiz bu veriler ile ilgili bilgi verilmiştir.

4.1.1 Ağ Örümceği Gerçeklemesi

Bu çalışmada, bir film yorum sitesindeki filmlere ait ağ adreslerini ziyaret eden ve sadece kullanıcılara ait yorumları daha sonra kullanılmak üzere dosyalama sistemimize kalıcılaştıran bir ağ örümceği gerçekleyerek, derlemde kullanılmak üzere gerekli veriler elde edilmiştir.

Bu amaçla geliştirdiğimiz ağ örümceği Python[14] programlama diline ait bir kütüphane olan urllib2[15] kullanarak HTTP isteklerini gerçekleştirmiş ve ilk aşamada sitede bulunan filmlerin kullanıcı yorumlarına göre sıralandığı ağ adresini başlangıç noktası alarak sitede bulunan tüm filmlerin adreslerini Redis [16] veritabanına kalıcılaştırmıştır.

İkinci aşamada filmlere ait her ağ adresi sekiz eşzamanlı işlem arasında paylaştırılıp sırasıyla Redis veritabanından alınarak eşzamanlı işlenmiştir. Bu verilerin işlenme aşamasında saklanmak istenilen kullanıcı yorumuna ait metin ve bu yoruma

[14] http://www.python.org

[15] http://docs.python.org/2/library/urllib2.html

[16] http://redis.io

kullanıcı tarafından verilen puan üst verisi bir çeşit XML olan HTML içerikten bir XML sorgu dili olan XPATH ile ayrıştırılarak daha sonra kullanılmak üzere dosyalama sistemine kalıcılaştırılmıştır.

Bu tarz ağ örümceklerinde çoğu zaman düzenli ifadeler ile istenilen metnin HTML etiketlerinden ayrıştırılması tercih edilse de gerek kullanım kolaylığı gerekse okuma kolaylığı ve yazılımın bakımının kolaylığı açısından XPATH sorgusuyla bu metinlerin HTML etiketlerinden ayrıştırılması için oldukça iyi bir düzenli ifadeler alternatifidir.

EK – 16'dan görülebileceği gibi tek bir kullanıcıya ait bir yorumu barındıran HTML DIV etiketi bile düzenli ifadeler bakımından oldukça karmaşık bir problem oluşturabilir. Ancak aynı HTML sayfasından bir yorumun metin değerini XPATH ile elde etmek istediğimizde yazmamız gereken örüntü Şekil 4.1'deki gibidir.

```
//*[@id="col_main"]/div[12]/div/div/div[1]/div/p
```

Şekil 4.1 Bir Kullanıcı Yorumunu Elde Etmek İçin Gerekli XPATH Kodu

Örnekten görülebileceği gibi XPATH, HTML'in düzenli ve hiyerarşik yapısından faydalanarak; kolay okunabilir şekilde veriyi temizleme ve istenilen veriyi elde etme işlemini oldukça hızlandırmıştır.

4.1.2 Elde Edilen Veri

Gözetimli yapay öğrenme yöntemlerinin yapısı gereği eğitim derleminin önceden sınıflandırılmış olması gerekmektedir. Bu tip önceden işaretlenmiş eğitim derlemlerini oluşturmak için özdevimli yöntemler daha önceki çalışmalarda önerilmiştir [4].

Büyük ölçekli veri kümelerinin bulunabilirliği ve özdevinimli işaretlemeye yönelik puan ya da yıldız ölçekli üst veri barındırması açısından film yorumları düşünce çözümleme için gözetimli eğitim derlemi oluşturulması açısından oldukça popülerdir. Bu çalışmada da eğitim derleminin özdevimli hazırlanması amaçlandığından film yorumları konusu derlem hazırlanması için seçilmiştir. Film konusunda Türkçe için

yapılan önceki çalışmalarda [26] olduğu gibi Beyazperde[17] isimli sinema eleştiri web sitesi derlemin hazırlanması için kullanılmıştır.

Sitede her film için ayrı başlık ve bu başlık altında kullanıcıların oluşturduğu 1-5 yıldız ölçeğinde puanlama ile gönderilen yorumlar bulunmaktadır. Şekil 4.2'de kullanıcı yorumlarının bir örneği görülebilir.



Şekil 4.2 Beyazperde Sitesinden Bir Üye Yorumu

Sitede bulunan tüm filmler üyelerden alınan yorumlara göre sıralanmış şekilde bulunabilmektedir.[18] Bu adres, sitede bulunan kullanıcı yorumlarının toplanması için yazılan ağ örümceğinin başlangıç noktasını oluşturmaktadır. Aralık 2013 itibari ile bu adresten başlayarak 620 alt sayfada 12400 film ve bu filmlere ait kullanıcı yorumları sayfalarından[19] 218806 yorum toplanmıştır.

Kullanıcı yorumları toplanırken 1-5 yıldız ölçeği 1-10 puan ölçeğine çevrilmiş ve bu üst veri her yorumun sonuna eklenerek saklanmıştır. Şekil 4.3'te kullanıcı yorumunun 1-10 puan ölçeğinde puanlamasıyla birlikte saklanış biçimi gösterilmiştir.

[17] http://www.beyazperde.com

[18] http://www.beyazperde.com/filmler/tum-filmleri/yorumlar/

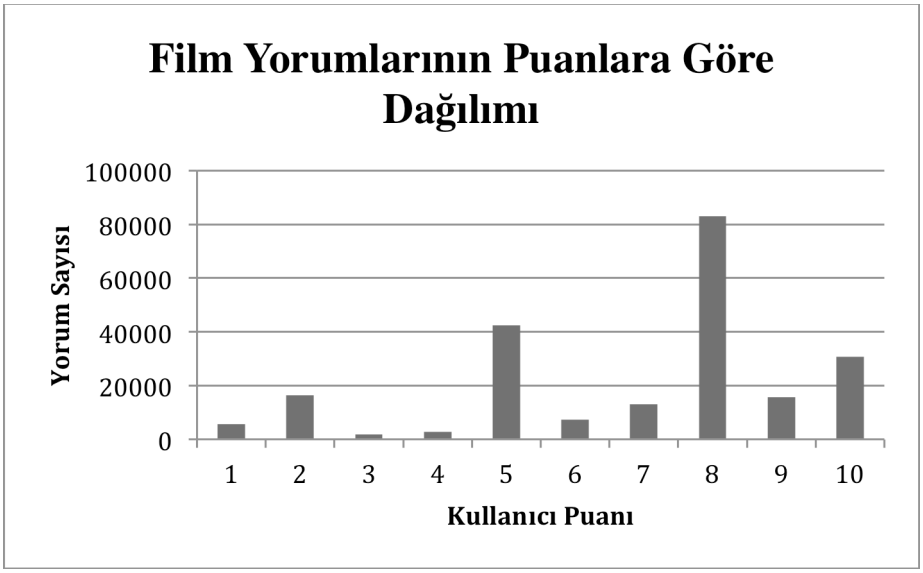
[19] http://www.beyazperde.com/filmler/film-8136/kullanici-elestirileri/

Joker i özletse de süper bir film üçlemeye yakışır bir başyapıt ve mükemmel bir son. İlk 1 saatinde güzelim seriye yazık olmuş dedim ama sonu muhteşemdi ***--userRating—8***

Şekil 4.3 Puan Üst Verisi İle Kaydedilen Yorum Örneği

Elde edilen tüm yorumların 1-10 puan sistemine çevrildikten sonraki dağılımları Tablo 4.1'deki gibi olmuştur.

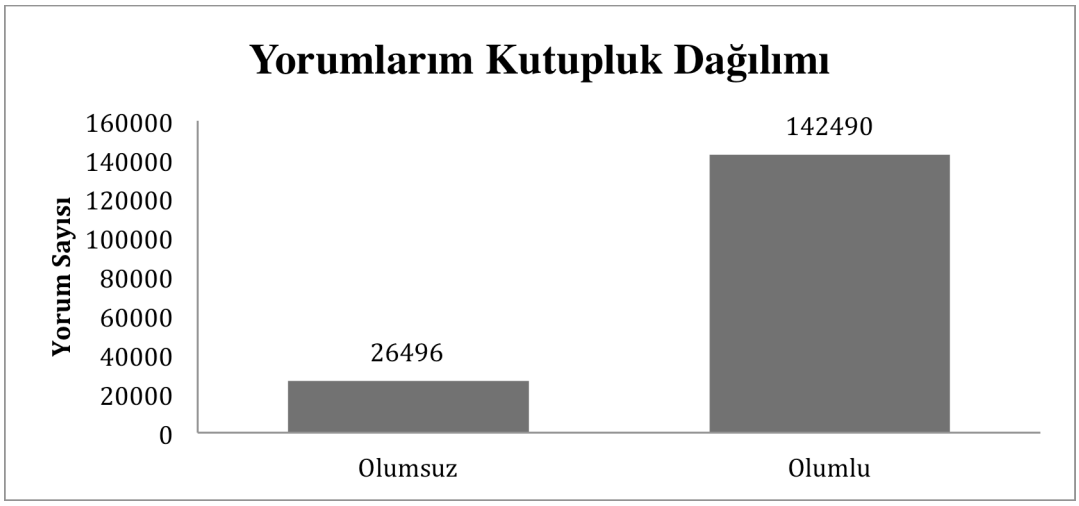
Tablo 4.1 Kullanıcı Puanlarının 10'lu Ölçekte Dağılımı



Tablo 4.1'den görülebileceği üzere 5 puan altındaki yorum sayısı oldukça azdır. İlginç bir şekilde kullanıcılar sitedeki puanlama sistemine göre 2.5, 4 ve 5 yıldız ile filmleri puanlama eğilimindedirler. Sitenin puanlama sistemine göre 4-5 yıldızlı, bizim dönüştürdüğümüz puanlamaya göre 8 ve 10 puanlı, yorumlarda taşıdığı düşünce kutupluğu açısından bir problem bulunmazken; bizim puanlama sistemimize göre 5-6 puana tekabül eden yorumlarda ise zıt sınıflara ait olması gerekirken aynı puanda bulunan oldukça fazla sayıda yorum bulunmaktadır. İki farklı sınıfa ait olması gereken bu yorumların eğitim derleminin sınıflandırma performansını oldukça fazla olumsuz yönde etkileyeceğinden bu yorumlar derlem oluştururken dikkate alınmamıştır. Böylece

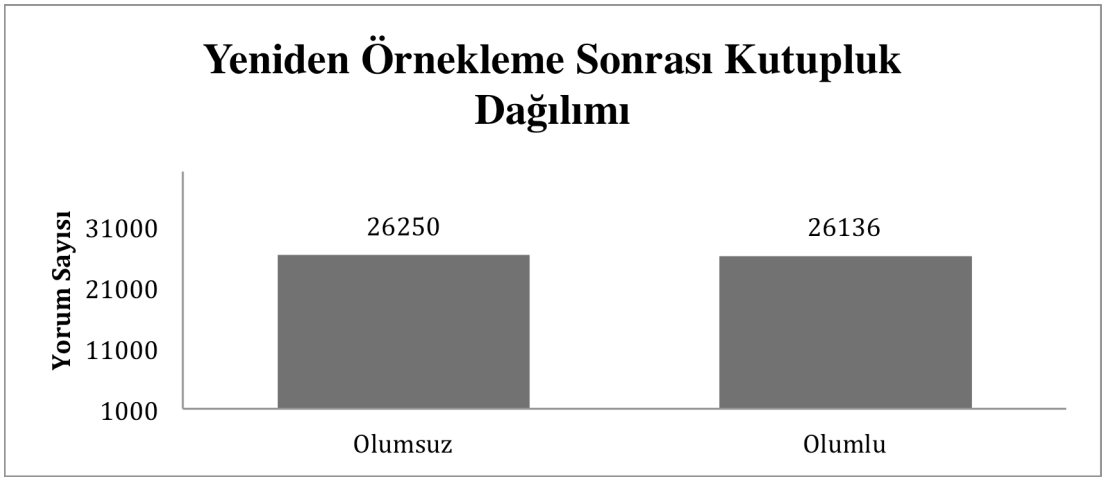
toplam 168986 kullanıcı yorumu kalmış, bunların 7 puan ve üstünde olan 142490 adedi "OLUMLU" olarak işaretlenmiş, 4 puan ve altındaki 26496 adedi "OLUMSUZ" olarak işaretlenmiştir.

Tablo 4.2 Gözetimli Öğrenme İçin Kullanılan Örneklerin Kutupluk Dağılımı



Tablo 4.2'den görülebileceği üzere, elde edilen yorumların kutupluğu oldukça dengesiz olduğundan bu durumu dengelemek için yorumlara rassal yeniden örnekleme uygulanmıştır. Bu şekilde dengeli miktarda olumlu ve olumsuz yorumun olduğu yeni bir veri kümesinin oluşturulması amaçlanmıştır. Yeniden örnekleme ile oluşturulan dengeli veri kümesinin kutupluk dağılımı Tablo 4.3'ten görülebilir.

Tablo 4.3 Yeniden Örneklenen Veri Kümesinin Kutupluk Dağılımı



4.2 Derlem Verisinin İşlenmesi

Bu bölümde önceki aşamada elde edilen ve belirli bir ölçüde düzenli hale getirilen verinin doğal dil işleme yöntemleri kullanılarak olağanlaştırılmasından ve son olarak istatistiksel dil modeli yaklaşımları kullanılmak üzere hazır hale getirilmesinden bahsedilmiştir.

4.2.1 Doğal Dil İşleme

Bu çalışmada doğal dil işlemeyi her bir kullanıcı yorumunda düşünceyi taşıyan belirtkelerin (Token) kökünün bulunması, bağlı bulunduğu dilbilimsel sınıfın tespiti, olumsuzluk eki taşıyıp taşımadığı gibi durumların tespiti ve kullanıcı yorumlarında var olabilecek basit düzensizliklerin, bozuklukların özdevimli olağanlaştırılması amacıyla kullanılmıştır.

4.2.2 Yazım Hatası Denetimi

Yazım hatası denetimi ve düzeltmesi bu tarz çalışmalarda verinin olağanlaştırılması için oldukça sık kullanılan bir yöntemdir. Zemberek2'nin işlemsel süreçleri gereği yazım hatası düzeltme işleminde kararı kullanıcıya bırakacak şekilde çalışması ve bu çalışmada tamamen özdevimli bir derlem ve model oluşturulması amaçlandığından yazım hatası denetimi elde edilen verilere uygulanmamıştır. Bu aşamada uyguladığımız işlemler tüm belirtkelerin küçük harfli hale getirilmisi ve Türkçe karakter kullanılmayan tüm belirtkelerin ASCII-Türkçe çevrim işlemine tabi tutulmasıdır.

Parmaklaini
[parmak laini, parmakla ini]

Şekil 4.4 Zemberek2'nin Yazım Hatası Düzeltme Önerisi Örneği

Zemberek2 kullanılarak belirlenen kökü bulunamayan belirtkeler hatalı kabul edilerek toplam 38097 benzersiz belirtke istatistiksel bilgi olması açısından kayıt altına alınmıştır.

4.2.3 Sözcük Kök ve Eklerinin Tespiti

Türkçe gibi bitişken dillerde aynı kökten onlarca hatta yüzlerce farklı sözcük türetilebilir. İstatistiksel metin sınıflandırıcılar aynı kökten türemiş bu sözcükleri farklı sözcükler olarak değerlendirme eğilimdedir. Bu tarz bir yaklaşım, kelime kökünün önemli olduğu durumlarda avantaj sağlayabilirken, olumsuzluk gibi Türkçe için fiilin ekinde taşınan bilginin gerekli olduğu durumlarda oldukça dezavantajlı bir durum yaratabilir.

Bu çalışmada, Türkçe için eklerin ve köklerin düşünce çözümleme açısından taşıdığı bilginin ölçümlenebilmesi açısından sözcüklerin köklerinden ve kökleri tespit edilemeyen belirtkelerin kaynaktan geldiği haliyle oluşturulan farklı derlemler hazırlanmıştır.

*anlam**landıramamışlarımızdan***
[Kok:anlam, Tip:ISIM |
Ekler: ISIM_KOK,
ISIM_DONUSUM_LE,
FIIL_EDILGENSESLI_N,
FIIL_ETTIRGEN_TIR,
FIIL_YETERSIZLIK_E,
FIIL_OLUMSUZLUK_ME,
FIIL_DONUSUM_MIS,
ISIM_COGUL_LER,
ISIM_SAHIPLIK_BIZ_IMIZ,
ISIM_CIKMA_DEN]

Şekil 4.5 Zemberek2 İçin Bir Sözcük Çözümleme Örneği

Türkçe'nin zengin biçimbilimsel yapısından dolayı, Zemberek2, çözümlemesini yapabildiği sözcükler için birden fazla kök önerisi döndürebilmektedir. Öneri olarak birden fazla kökün döndüğü durumlarda ise sıralama bu sözcüklerin Türkçe'deki tekrar sayısı istatistiklerine göre yapılmaktadır. Derlemin özdevimli oluşturulması amaçlandığından böyle durumlarda Zemberek'ten dönen ilk sıradaki kök dikkate alınmıştır.

Bununla birlikte Beyazperde kullanıcı yorumlarından elde edilen veri dilbilimsel açıdan oldukça kirli sayılabilecek durumdadır. Bu kirliliğe sebep olan durumlar yazım hataları, kullanıcıların sık kullandıkları internet jargonu, sözcük kısaltmaları ya da Şekil 4.6'daki gibi bir düşünceyi daha vurgulu ifade etmek amacıyla harf tekrarı ile uzatılan kelimeler olarak verilebilir.

*ciddiyim sinema **kolikseniz** bu filmi kaçırmayın dehşet ötesi **güzeeeeeeeeeeellllllll** ve ben iki kere gittim.*

Şekil 4.6 Dilbilimsel Açıdan Kirli Veri Örneği

Eğitim derlemi oluşturulurken köklerden oluşan derlem için kökü bulunabilen sözcüklerin kökü yazılmış, kökü bulunamayan sözcükler aynı şekilde bırakılmıştır. Bununla birlikte daha sonra bahsedileceği üzere sözcük sınıflarına göre kökler gruplanarak ayrı derlemler de oluşturulmuştur.

Tablo 4.4 Eğitim Derleminde Sözcük Kökünün Özdevimli Bulunabilme Durumuna Göre Özgün Belirtke Sayıları

Belirtke Durumu	**Özgün Adet**	**Oran (%)**
Kökü Bulunabilen	78448	67
Kökü Bulunamayan/Hatalı	38097	33
Toplam	116545	100

Tablo 4.4'den görülebileceği üzere her bir kullanıcı yorumundaki belirtkelerin köklerinden oluşturulması amaçlanan derlem, belirtkeler açısından %33 oranıyla kökü bulunamayan yapılardan oluşmaktadır. Bunlar genel itibari ile yabancı kelime, hatalı yazım, kısaltmalar, internet jargonuna özgü ifadeler ve anlamsız karakter dizilerinden oluşmaktadır. Türkçe için mevcut doğal dil işleme araçları ile dilbilimsel açıdan çözümlenemeyen ve özdevimli temizlenemeyen, kirliliğe sebep olan bu tip derlemden elde edilen belirtke örnekleri Ek-13'te verilmiştir.

4.2.4 Olumsuzluk Durumu

Türkçe'de olumsuzluk iki şekilde oluşturulur:

- Ekler ile
- Sözcükler ile

Ekler ile oluşturulan olumsuzluk *–mA* eki sağlanır. *–mAz*, *-mAdAn* gibi bu ekin başka eklerle kaynaşmasından oluşan durumlar da olumsuz olarak kabul edilebilir [35]. Sözcükler ile oluşturulan olumsuzluk ise mantıksal işleç vazifesi taşır.

değil
hiç
hayır
yok
ne … ne
aksi halde
yoksa
ama

Şekil 4.7 Türkçe Olumsuzluk Sağlayan Sözcükler

Sözcükler ile sağlanacak olumsuzluk durumlarının, daha sonra ayrı bir bölümde açıklanacak N-Gram modeli ile yakalanabileceği varsayımından hareketle, sadece Türkçe olumsuzluk taşıyan fiillerin belgenin olumsuz kutupluğuna etkisinin araştırılması amaçlanmıştır.

*ümit**lenmeseydim***

[Kok:ümit, Tip:ISIM |
Ekler:
ISIM_KOK,
ISIM_DONUSUM_LE,
FIIL_EDILGENSESLI_N,
FIIL_OLUMSUZLUK_ME,
FIIL_SART_SE,
IMEK_HIKAYE_DI,
ISIM_KISI_BEN_IM]

_ümit

Şekil 4.8 Olumsuz Sözcük Çözümlemesi ve Derlemde Saklanış Şekli

Bu sebeple daha önceki çalışmalara [4] benzer şekilde derlem içinde ekleriyle olumsuzluk taşıyan sözcükler Zemberek2 ile tespit edilmiş, bu kelimelerin başına "_" işareti getirilerek işaretlenmiş ve bu şekilde kök ve olumsuz köklerden oluşan bir derlem oluşturulmuştur.

Tablo 4.5 Olumsuzluk Taşıyan Köklerin İşaretli Bulunduğu Veri Kümelerine Ait İstatistikler

Kökler	**Özgün Adet**	**Oran %**
Olumsuzluk Taşıyan	692	9
Olumsuzluk Taşımayan	6877	91
Toplam	7569	100

4.2.5 Sözcük Sınıflandırma

Verilen bir metni oluşturan sözcüklerin dilbilimsel sınıflarının tespit edilmesi bilgisayarlı dilbilim açısından oldukça önemli bir işlemdir. Bu çalışmada dilbilimsel sözcük sınıflarının birbiri ile ve tek başlarına, içinde bulunduğu metnin düşünce kutupluğuna etkisinin araştırılması amaçlanmıştır.

Türkçe için mevcut bilgisayarlı dilbilim araçları derinlemesine dilbilimsel sözcük sınıflandırmasına imkan vermemektedir. Bu sebeple temel üç sözcük sınıfı olan **isim**, **sıfat** ve **fiil** kategorilerinde kelimelerin gruplandırılması amaçlanmıştır.

film
[Kok:film, Tip:**ISIM** |
Ekler:ISIM_KOK]

kötü
[Kok:kötü, Tip:**SIFAT** |
Ekler:ISIM_KOK]

beğenmedim
[Kok:beğen, Tip:**FIIL** |
Ekler:FIIL_KOK,
FIIL_OLUMSUZLUK_ME,
FIIL_GECMISZAMAN_DI,
FIIL_KISI_BEN]

Şekil 4.9 Sınıflandırılmış Sözcüklere Derlemden Örnekler

Sınıflandırılması yapılabilen her bir sözcüğe ait sınıf ya da sınıflar için sözcüklerin köklerinden oluşan derlem hazılanmıştır. Özellikle SIFAT+ISIM, ISIM, ISIM+FIIL gibi grupların oldukça güçlü düşünce kutupluluğu barındıracağı hipotezinden hareketle, sözcükler tek başlarına ait oldukları dilbilimsel sınıflarına göre ayrıldıkları gibi farklı farklı dilbilimsel sınıfların sözcüklerinden oluşan derlemler de düşünce kutupluğunun tespiti ve bu kutupluğun diğer dilbilimsel sözcük sınıfları ile karşılaştırılması açısından oluşturulmuştur.

Sözcüklerin ait oldukları dilbilimsel sınıfa göre gruplardırılması ile oluşturulan yedi veri kümesi ve bunlara ait özgün sözcük sayılarını gösteren istatistik Tablo 4.6'da verilmiştir.

Tablo 4.6 Dilbilimsel Sözcük Sınıflarına Göre Oluşturulan Veri Kümelerine Ait Özgün Belirtke Adetleri

Derlem	**Özgün Belirtke Adetleri**
Fiiller	36913
İsimler	36314
İsimler+Fiiller	73227
İsimler+Sıfatlar	38600
İsimler+Sıfatlar+Fiiller	75513
Sıfatlar	2286
Sıfatlar+Fiiller	39199

4.3 Yöney Uzayı Modeli (Vector Space Model)

Diğer adıyla terim yöneyi modeli, metin verilerinin yöney olarak ifade edilmesini sağlayan cebirsel bir modeldir.

Şu ana kadar sadece dilbilimsel ve bilgisayarlı dilbilim açısından ele aldığımız metin verisini gözetimsiz yapay öğrenme gibi matematiksel bir yöntem ile modelleyebilmek için önce bu metni cebirsel bir karşılığa çevirmemiz gerekmektedir. Bunun için izlediğimiz yol sırasıyla, sayısal olmayan metin verisini *N-Gram* modeli ile terimler dizisi haline getirilmesi ve daha sonra da bu dizilere *TF-IDF* dönüşümü uygulanarak belirtkelerden oluşan metin parçalarını sayısal bir karşılığa çevirilmesidir.

4.3.1 N-Gram Modeli

Doğal dilde dolayısıyla metinsel verilerde dilbilgisi olarak adlandırılan kurallar bütünü ile sözcükler belirli düzende sıralı olarak bulunur. Bu düzen anlambilimsel ya da sözdizimsel açıdan gerekli olsa da metin sınıflandırma problemini çözmek için sözcüklerin sırasının bir önemi yoktur. Bu yaklaşıma kelime torbası (bag of words) yaklaşımı da denir.

Bu durumda sınıflandırılmak istenen metni belirtkelerine ayırıp bu belirtkeleri ve belirtkelerin belge içindeki tekrar sayılarını metne ait öznitelik olarak kullanmak akla gelebilecek en naif yol olabilir. Ancak bu yaklaşım bazı belirtke öbeklerini hesaba katarken bir çoğunu gözardı ettiği için muazzam bir bilgi kaybına sebep olur.

Bu durumu engellemek için metin, *n* sayıda elemana sahip ardışık belirtke altkümelerine ayrılır. Burada *n* tamamen keyfi olup metne ait sıralı belirtke alt kümesinin eleman sayısını göstermektedir. Takdir edebileceginiz gibi burada en küçük *n* değeri 1, en büyük *n* değeri ise yine metnin toplam belirtke sayısına eşit olacaktır. Metne ait *n* elemanlı altküme, bir karakter dizisi olabileceği gibi bir kelime dizisi de olabilir. Bu seçim de *n* belirtke sayısı gibi keyfidir.

Bir metne ait bir elemanlı alt kümeleri unigram, iki elemanlı altkümeleri digram ya da bigram, üç elemanlı altkümeleri trigram olarak adlandırılır. Dört ve daha fazla sayıda elemana sahip altkümeler ise sırasıyla fourgram, fivegram gibi *n* altküme eleman sayısının ingilizce karşılıkları ile adlandırılır.

Çok kötü bir film.
Unigram: [‘çok’, ‘kötü’, ‘bir’, ‘film’]
Bigram: [‘çok kötü’, ‘kötü bir’, ‘bir film’]
Trigram: [‘çok kötü bir’, ‘kötü bir film’]
Fourgram: [‘çok kötü bir film’]

Şekil 4.10 Bir Cümleyi Oluşturan Sözcüklerin N-Gram Örneği

Bu çalışmada düşünce çözümlemesi yapılacak belgeye ait sözcüklerin N-Gram bakımından düşünce kutupluğuna etkileri incelenmiştir. Sırasıyla metinde bulunduğu şekliyle ya da ilgili dilbilimsel sözcük sınıfına ait unigram, bigram, trigram, fourgram, unigram+bigram, unigram+bigram+trigram, unigram+bigram+trigram+fourgram yapılar çıkarılmıştır.

Tablo 4.7 Her Bir Belgesi Yöneye Çevrilmiş Veri Kümeleri ve Öznitelik Aday Sayıları

N-Gram	**Öznitelik Adayları**
Unigram	35818
Bigram	97918
Trigram	89227
Fourgram	74984
Unigram+Bigram	133730
Unigram+Bigram+Trigram	222941
Unigram+Bigram+Trigram+Fourgram	297839

4.3.2 Terim Frekansı – Ters Metin Frekansı (TF-IDF)

Metin madenciliğinde TF-IDF (Term Frequency – Inverse Document Frequency) olarak da adlandırılan, belirtkeler dizisi olarak ifade edilebilen bir belgenin yöneye çevrilmesi için kullanılan bir yöntemdir.

Kısaca bir belirtkenin, önce incelenen belge içinde daha sonra tüm derlem çapında diğer belgelerde bulunup bulunmadığını, tercihe göre bulunuyorsa kaç kez tekrar (*frekans*) ettiğini araştırır.

d_j derlem içindeki bir belge ve w_i derlem içinde geçen bir belirtke olmak üzere derlem içindeki d_j belgesi denklem 4.1'deki gibi bir yöney ile temsil edilebilir.

$$d_j = \left(w_{1,j}, \quad w_{2,j}, \quad w_{3,j}, \quad w_{4,j}, \dots, w_{t,j}\right) \qquad (4.1)$$

Eğer w_i belirtkesi d_j belgesi içinde mevcut ise bu terim sıfırdan farklı bir değer alır. Mevcut değilse sıfır değeri alır. Burada yöneyin $w_{i,j}$ belirtkesinin değeri bir kaç farklı şekilde ele alınabilir. Bu değer, belirtkenin mevcudiyeti (*presence*) üzerinden 1 ve 0 değerini alabilirken, doğrudan belirtkenin tekrar sayısı olabileceği gibi,

$$tf(t,d) = \ln(1 + f_{t,d}) \quad (4.2)$$

$$tfidf(t,d,D) = tf_{t,d} * (f_{t,d} * \ln \frac{|D|}{|\{d \in D \mid t \in d\}|}) \quad (4.3)$$

şeklinde verilebilecek, bir belirtkenin belge içindeki tekrarına bağlı üssel bir ağırlık katsayısı olarak da belirtilebilir. Burada $f_{t,d}$, t belirtkesinin d belgesi içindeki tekrar sayısı, $|D|$ derlem içindeki toplam d belge sayısı ve $|\{d \in D \mid t \in d\}|$, t belirtkesinin tüm derlem içindeki tekrar ettigi belge sayısıdır.

Denklem 4.3'ten hemen anlaşılabileceği üzere bir belge içinde 100 kez tekrar eden bir belirtke ile derlem içinde toplamda 100 belge içinde birer kez geçen bir belirtke aynı öneme ve değere sahip olmayacaktır. Bu durumda derlem içinde farklı belgelerde geçen belirtkenin değeri ile tek belgede birden fazla kez tekrar eden bir belirtkenin değeri de farklı olacaktır.

Tablo 4.8 TF-IDF Dönüşümü Öncesi Derlem

Belge	İçerik	Sınıf
1	çok kötü bir film. kötü olmasının yanı sıra zaman kaybı bir film.	**olumsuz**
2	çok kötü bir film.	**olumsuz**
3	çok iyi bir film.	**olumlu**
4	bence iyi bir film	**olumlu**

İlk durumu Tablo 4.8'daki gibi olan bir derlem sırasıyla belirtke mevcudiyetine, belirtke tekrarına, üssel belirtke tekrarına göre TF-IDF dönüşümü ile ifade edilirse aşağıdaki gibi tablolarda veri kümesinin yeni durumları ifade edilir:

Tablo 4.9 Belirtke Mevcudiyetine Göre Belge Yöneyleri

belge	bir	film	kaybı	kötü	olmasının	sıra	yanı	zaman	çok	bence	iyi	**sınıf**
1	1	1	1	1	1	1	1	1	1	0	0	**olumsuz**
2	1	1	0	1	0	0	0	0	1	0	0	**olumsuz**
3	1	1	0	0	0	0	0	0	1	0	1	**olumlu**
4	1	1	0	0	0	0	0	0	0	1	1	**olumlu**

Tablo 4.10 Belirtke Tekrarına Göre Belge Yöneyleri

belge	bir	film	kaybı	kötü	olmasının	sıra	yanı	zaman	çok	bence	iyi	**sınıf**
1	2	2	1	2	1	1	1	1	1	0	0	**olumsuz**
2	1	1	0	1	0	0	0	0	1	0	0	**olumsuz**
3	1	1	0	0	0	0	0	0	1	0	1	**olumlu**
4	1	1	0	0	0	0	0	0	0	1	1	**olumlu**

Tablo 4.11 Üssel Belirtke Tekrarına Göre Belge Yöneyleri

belge	bir	film	kaybı	kötü	olmasının	sıra	yanı	zaman	çok	bence	iyi	**sınıf**
1	0	0	0.96..	0.76..	0.96..	0.96..	0.96..	0.96..	0.199..	0	0	**olumsuz**
2	0	0	0	0.48..	0	0	0	0	0.199..	0	0	**olumsuz**
3	0	0	0	0	0	0	0	0	0.199..	0	0.48..	**olumlu**
4	0	0	0	0	0	0	0	0	0	0.96..	0.48..	**olumlu**

Tablo 4.11'de görüleceği üzere üssel belirtke tekrarına göre oluşturulan veri kümesinde Tablo 4.9 ve Tablo 4.10'dan farklı olarak bazı belirtkeler tüm veri kümesi içinde mevcut olmasına rağmen 0 ile ifade edilmiştir. Denklem 4.3'ten kolayca hesaplanabileceği üzere bunun sebebi bu belirtkelerin veri kümesindeki her belgede tekrar etmesidir. Takdir edileceği üzere, veri kümesindeki her belgede tekrar eden belirtke, aynı zamanda her sınıf için ortak belirtkle anlamına geleceğinden bu tip belirtkelerin belgenin sınıfının tespitinde bir katkısı olmayacağı kabuluyle hesaplamalarda göz ardı edilmesi sağlanmaya çalışılmıştır. Bu aşama iki sebeple önemlidir, birincisi metin verilerinde yüzbinlerce benzersiz belirtke olabileceğinden bu tarz, sınıflandırmaya etkisi tartışmalı belirtkelerin yaratacağı zaman ve yer karmaşıklığını önceden engellemek ve olası aşırı uyum durumlarının önüne geçmek.

Türkçe için daha önce yapılan benzer çalışmaların [26, 27] aksine yukarıdaki sebeplerden bu çalışmada öznitelik çıkarımı için üssel belirtke tekrarı yöntemi kullanılarak, oluşturulan her derlem için $|\{d' \in D \mid t \in d'\}| \geq 3$ değerini gerçekleyen tüm belirtkeler öznitelik adayı sayılmıştır. Buradaki 3 değerenin seçimi tamamen ampirik olup aşırı uyumu engellemek açısından 1 ve 2 gibi düşük tekrara sahip belirtkelerin sınıflandırma performansına katkılarını gözardı etmek amaçlanmıştır.

4.4 Yapay Öğrenme

Tom Mitchell, yapay öğrenmeyi, "*Eğer bir bilgisayar programı, bir T görevini, P performansında yaparak, deneyiminde E artış sağlıyorsa, o bilgisayar programı, T görevinde, P performansıyla, E deneyimlerinden öğreniyordur denilir.*" şeklinde tanımlamıştır [36].

Daha genişleterek ifade etmek gerekirse, bir konuda yapay öğrenme kullanılarak sonuç alınabileceğini söylemek için öncelikle konu hakkında şu sorulara:

- Örüntü mevcut mu?
- Matematiksel olarak ifade edilemiyor mu?
- Veri mevcut mu?

"evet" cevabının verilmesi gereklidir. Bu üç sorunun cevabı evet olduğu takdirde yapay öğrenmenin gerçeklenebilmesinden bahsedilebilir. Bu şartlar sağlandığı takdirde:

$$Girdi: x \quad (4.4)$$

$$Çıktı: y \quad (4.5)$$

$$Bilinmeyen\ Hedef\ Fonksiyon: f: X \rightarrow Y \quad (4.6)$$

$$Veri: (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_3, y_3), \dots, (x_N, y_N) \quad (4.7)$$

$$Hipotez: g: X \rightarrow Y \quad (4.8)$$

Olmak üzere x_N, y_N şeklinde verilen, sırasıyla girdi ve çıktıdan oluşan çiftlerden keşfettiği örüntüleri, her yinelemede g hipotezini bilinmeyen f hedef fonksiyonuna

yaklaştırmada kullanılan $\mathcal{A}$ algoritması, öğrenme algoritması ve hipotezi hedef fonksiyona yaklaştıran her yineleme yapay öğrenme olarak tanımlanabilir.

Bölüm 4.3.2'deki derlem örneğinden hareketle;

$belge1$: ("çok kötü bir film. kötü olmasının yanı sıra zaman kaybı bir film.",olumsu (4.9)

$belge2$: ("çok kötü bir film.",olumsuz) (4.10)

$belge3$: ("çok iyi bir film.",olumlu) (4.11)

$belge4$: ("bence iyi bir film.",olumlu) (4.12)

şeklinde verilen derleme ait belgelerden öyle örüntüler keşfetmeliyiz ki $g(belge_i) = (olumsuz|olumlu)$ sonucunu en doğru şekilde verebilsin. Bu amaçla metin sınıflandırma için $\mathcal{A}$ algoritması olarak Naïve Bayes, Naïve Bayes Çokterimli, Destek Yöney Makineleri, Maksimum Entropi daha önceki çalışmalarda da görüleceği üzere en çok kullanılan yöntemler olup bu çalışmada sadece Naïve Bayes Çokterimli kullanılmıştır.

4.4.1 Gözetimli Yapay Öğrenme

("çok iyi bir film.", olumlu) örneğinde olduğu gibi denklem 4.5'te verilen y çıktısının eğitim için kullanılacak veride önceden bilindiği durumlarda $\mathcal{A}$ öğrenme algoritması gözetimli olarak öğreniyor ve girdi olarak x alan ve çıktı olarak y veren geneleştirilmiş bir f fonksiyonu vardır denilir. Doğru olduğu bilinen öncül önermelerden tümevarım sağlamaya çalışan mantıksal bir süreçtir.

Örneklerden açıkça anlaşılabiliceği üzere girdi verilerinde sınıfların önceden biliniyor olması ve ulaşılmak istenen hedef bakımından bu çalışmanın da gözetimli yapay öğrenme kapsamına girdiği kolayca görülebilir.

4.4.1.1 Öznitelik Çıkarımı (Feature Extraction)

Bir g hipotezinin f hedef fonksiyonuna en yakın sonuçları vermesi x girdi kümesininin elemanlarına ait özelliklerin çok iyi bilinmesi ile mümkün olabilir. (x_k, y_k)

girdi verilerindeki örüntülerden elde edilen ve hedef fonksiyona en yakın sonuçları veren x_k kümesinin elemanlarına ya da elemanlarının özelliklerine bu verinin öznitelikleri denir. Veriye özgü bu nitelikler tek tek seçilebileceği gibi özdevimli olarak da elde edilebilir. Diğer bir ifade ile, x_k'ya ait öyle özellikler bulmalıyız ki $g(x_k) = y_k$ sonucu bilinmeyen f fonksiyonuna denk sonuçlar verebilsin.

Bu çalışma kapsamında daha önce yöney uzayı konusunda bahsedilen N-Gram ve TF-IDF dönüşümleri ile elde ettiğimiz öznitelik adaylarından, belgenin sınıfı ile ilgili çıkarımları en verimli biçimde yapabilmemizi sağlayan özniteliklerin belirlenmesi hedeflenmiştir. TF-IDF dönüşümlerinden hatırlanacağı üzere belgeyi ifade ettiğimiz yöney, içinde barındırdığı belirtkelerin tekrarı bakımından sayısal değerler alıyordu. Tablo 4.12'den görüleceği üzere her bir belgeyi ifade eden yöneyin terim sayısı onbinlerce olabileceği için elimizde oldukça seyrek terimli belge yöneylerinden oluşan oldukça büyük boyutlu bir matris olacaktır. Katkısı nispeten az ya da olmayan bu seyrek terimleri eleyip sadece belgenin sınıfı açısından derlem genelinde en iyi bilgiyi sağlayan belirtkeleri tespit etmek hem işlemin hızı, hem de sınıflandırmanın performansı açısından oldukça önemlidir.

4.4.1.2 Öznitelik Seçimi: Bilgi Kazancı (Information Gain)

Oluşturduğumuz dil modeline göre uygun öznitelikleri seçmek bu çalışmanın kritik aşamalarından biridir. Metin verisi söz konusu olduğundan literatürde [37] en başarılı öznitelik seçim yönteminin bilgi kazancı olduğu görüldüğünden bu çalışmada da bu yöntem öznitelik seçiminde kullanılmıştır.

Bilgi teorisine göre bir rassal değişkenin sahip olduğu kesinsizlik ölçüsü entropi ile ifade edilir ve H ile gösterilir.

Kesinsizliğin bu ölçüsü ise:

$$\mathrm{H(X)} = -\sum_{x=i}^{n} p(x_i) \log_b p(x_i) \qquad (4.13)$$

Şeklinde hesaplanır. Bunu bizim örneğimizde ifade etmeye çalışırsak; daha önce TF-IDF ile bulduğumuz öznitelik adaylarının içinden öyle öznitelik adayları bulmalıyız ki bize $g(x_k) = y_k$ hipotezi için net bilgi sağlasın. Bu durumda net bilgi kazancını,

$$\text{NetKazanç}(\text{Sınıf}, \text{Öznitelik}) = \text{H}(Sınıf) - \text{H}(\text{Ö}znitelik|Sınıf) \quad (4.14)$$

Denklem 4.14 ile hesaplayabiliriz. Burada net kazanç eksi değerler de alabileceğinden yine ampirik olarak bize artı net kazanç sağlayan 0 eşik değeri üzerindeki öznitelik adaylarını öznitelik olarak seçeriz. Bu durumda özniteklik adayları şu şekilde belirlenmiş olur:

Tablo 4.12 N-Gram Belirtkelerden Oluşan Veri Kümelerine Ait Öznitelik Sayıları

N-GRAM	**Öznitelik Adayları**	**Öznitelik**
Unigram	**35818**	**14972**
Bigram	**97918**	**64321**
Trigram	89227	77024
Fourgram	74984	72528
Unigram+Bigram	**133730**	**79288**
Unigram+Bigram+Trigram	222941	156304
Unigram+Bigram+Trigram+Fourgram	297839	228772

İlk duruma göre öznitelik sayıları en çok azalan ilk üç N-Gram koyu renkte yazılmıştır. Daha az öznitelik, sınıflandırma yeteneği, daha genelleştirilmiş dolayısıyla f, fonksiyonuna daha yakın bir g fonksiyonu vaadetmesi bakımından önemlidir.

4.4.1.3 Sıralama

Belgelerin sınıfına dair verdikleri net bilgi bakımından özniteliklerin net kazançlarına göre sıralanmasından ibaret olan bu safha doğrudan sınıflandırmaya etki etmemektedir. Ancak bu sıralama Türkçe'ye ait taşıdığı düşünce kutupluğu bakımından en güçlü belirtkeleri belirlemesi açısından önemlidir. Bu tarz belirtkelere N-Gram ve dilbilimsel sözcük sınıfına göre en güçlü ilk kırk öznitelik örnekleri eklerde verilmiştir.

4.4.2 Naïve Bayes Çokterimli

Elde ettiğimiz metinsel verilere ait belirtkelerin, metnin ait olduğu sınıfa dair bilgi verip vermemesi, veriyorsa bu istatistiksel dağılımların keşfedilmesi çalışmadaki en önemli unsuru oluşturmaktadır. Bu durumların keşfedilmesi için ise istatistiksel bir kestirim yöntemi olan Naïve Bayes Çokterimli kullanılmıştır.

Metin sınıflandırma için ifade edecek olursak, X, belgenin ait olabileceği tüm sınıflar, x, bu sınıflardan biri ve θ, öncül bir bağımlılık durumu, olmak üzere bu öncül bağımlılık durumuna bağlı bir şekilde x'in gerçekleşmesini hesaplayan olabilirlik fonksiyonu:

$$\mathcal{L}(\theta|\mathrm{x}) = p_\theta(x) = P_\theta(\mathrm{X} = x) \qquad (4.15)$$

şeklinde verilir ve sıklıkla $P_\theta(\mathrm{X} = x|\theta)$ ve $P_\theta(\mathrm{X} = x;\theta)$ şeklinde de gösterilir.

McCallum ve Nigam'ın çalışmasından [38] hareketle Naïve Bayes Çokterimli sınıflandırıcımızı Denklem 4.15'te verilen koşullu olabilirlik üzerinden genelleştirmek ister isek, aynı V sözcük kümesine ait sıralı sözcüklerden oluşturulmuş bir d_i belgesi olduğunu ve bu belgenin uzunluğunun, ait olduğu c_j sınıfından bağımsız bir özellik olduğunu kabul edelim. Bayesçi yaklaşıma benzer şekilde, V sözcük kümesine ait bir w_t sözcüğünün bir d_i belgesi içinde bulunması durumunu bu sözcüğün d_i belgesi içindeki konumundan ve bağlamından bağımsız değerlendirdiğimiz varsayımını yapalım. Son olarak TF-IDF dönüşümünde örneğini verdiğimiz şekilde bir w_t sözcüğünün d_i belgesi içinde tekrarını N_{it} ile ifade edersek, bir d_i belgesinin verilen bir c_j belgesine ait olma olasılığı:

$$\mathrm{P}(d_i|c_j;\theta) = P(|d_i|)|d_i|!\prod_{t=1}^{|V|}\frac{P(w_t|c_j;\theta)^{N_{it}}}{N_{it}!} \qquad (4.16)$$

şeklinde ifade edilir. Denklem 4.16'dan görüleceği gibi $0 \leq \theta_{w_t|c_j} \leq 1$ ve $\sum_t \theta_{w_t|c_j} = 1$ olmak üzere, her sınıfa ait sözcük olasılıkları $\theta_{w_t|c_j} = P(w_t|c_j;\theta)$ şeklinde ifade

edilmiştir. $|\mathcal{D}|$ derlemdeki toplam belge sayısı olmak üzere bir sözcüğün bir sınıfa ait olma olasılığını veren bu denklem ise:

$$\hat{\theta}_{w_t|c_j} = P\left(w_t\middle|c_j;\hat{\theta}\right) = \frac{1 + \sum_{i=1}^{|\mathcal{D}|} N_{it} P(c_j|d_i)}{|V| + \sum_{s=1}^{|V|}\sum_{i=1}^{|\mathcal{D}|} N_{is} P(c_j|d_i)} \qquad (4.17)$$

şeklinde verilir. Denklem 4.16 ve 4.17 ile detayları verilen çokterimli Bayes sınıflandırıcımız bu durumda şu şekilde ifade edilmiş olur:

$$P\left(c_j\middle|d_i;\hat{\theta}\right) = \frac{P(c_j|\hat{\theta})\mathrm{P}\left(d_i\middle|c_j;\hat{\theta}\right)}{P(d_i|\hat{\theta})} \qquad (4.18)$$

4.4.3 K-Katlı Çapraz Doğrulama (N-Fold Cross Validation)

Bir sınıflandırıcının başarısının ölçülebilmesi için sınıfı önceden bilinen verileri sınıflandırması ve ürettiği sonuçların sınama verisindeki gerçek sınıfları ile karşılaştırılması gereklidir. Ancak burada dikkat edilmesi gereken husus sınıflandırıcıyı sınayacak verinin eğitim verisi içinde bulunmamasıdır. Sınama verisinin de eğitim için kullanıldığı durumda sınıflandırıcı gerçek başarımından daha üstün bir başarım elde edecektir. Bu, gerçek başarımın ölçülebilmesi açısından istenmeyen bir durumdur. Bu durumun önüne geçmek için derlem, k sayıda eşit parçaya bölünür ve her tekrarda, bir parça sınama için ayrılırken kalan k-1 parça sınıflandırıcının eğitim için kullanılır. Bu işlem, k sayıda tekrar edilerek her tekrardaki sonuçlar kayıt edilir. Bu tekniğe K-Katlı çapraz doğrulama adı verilir.

Tablo 4.13 K-Katlı Çapraz Doğrulama Sınama ve Eğitim Verileri Dağılımı

	1	2	3	4	5	6	7	8	9	10
Tekrar 1	**Sınama**	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
Tekrar 2	Eğitim	**Sınama**	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
Tekrar 3	Eğitim	Eğitim	**Sınama**	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
Tekrar 4	Eğitim	Eğitim	Eğitim	**Sınama**	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
Tekrar 5	Eğitim	Eğitim	Eğitim	Eğitim	**Sınama**	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim

Bu çalışmada sonuçları verilen tüm sınıflandırıcıların eğitiminde 10 katlı çapraz doğrulama tekniği kullanılarak oluşan sonuçların ağırlıklı ortalaması sınıflandırıcının başarımı olarak kabul edilmiştir.

4.4.4 Başarımın Ölçülmesi

Bu çalışmada da kullanılan iki sınıflı bir sınıflandırıcının verdiği sonuçlar dört şekilde gruplanabilir. Bunlar, 1. Sınıfa ait doğru sınıflananlar, TP (True Positive), 2. Sınıfa ait doğru sınıflananlar, TN (True Negative), 1. Sınıfa ait yanlış sınıflananlar, FP (False Positive), 2. Sınıfa ait yanlış sınıflananlar şeklinde olur, FN (False Negative) şeklinde verilebilir. Bu durumların bir tabloda gösterilmesi durumuna ise karışıklık matrisi adı verilir.

Tablo 4.14 İki Sınıflı Sınıflandırıcıya Ait Karışıklık Matrisi

		Sınıflandırıcı Sonucu	
		Pozitif	**Negatif**
Gerçek Sınıf	**Pozitif**	TP	FP
	Negatif	FN	TN

Karışıklık matrisinden gruplandırılan bu sonuçlar daha sonra başarımın ölçümlenmesinde aşağıda bahsedilecek hesaplamalarda kullanılır.

4.4.4.1 Doğruluk (Accuracy)

Sınıflandırıcı tarafından doğru sınıflandırılan belgelerin tüm belgelere oranıdır.

$$\text{Doğruluk} = \frac{\text{TP+TN}}{TP+TN+FN+FP} = \frac{doğru\ sınıflandırılmış\ belgeler}{derlemdeki\ tüm\ belgeler} \quad (4.19)$$

4.4.4.2 Duyarlılık (Precision)

Sınıflandırıcı tarafından doğru sınıflandırılmış pozitif sınıfa ait sonuçların, pozitif olarak sınıflandırılmış tüm belgelere oranıdır.

$$\text{Duyarlılık} = \frac{\text{TP}}{TP+FP} = \frac{doğru\ sınıflandırılmış\ pozitif\ belgeler}{pozitif\ olarak\ sınıflandırılmış\ tüm\ belgeler} \quad (4.20)$$

4.4.4.3 Geriçağırım (Recall)

Sınıflandırıcı tarafından doğru sınıflandırılmış pozitif belgelerin derlemdeki tüm pozitif belgelere oranıdır.

$$\text{Geriçağırım} = \frac{\text{TP}}{TP+FN} = \frac{doğru\ sınıflandırılmış\ pozitif\ belgeler}{pozitif tüm\ pozitif\ belgeler} \quad (4.21)$$

4.4.4.4 F-Ölçütü (F-Measure, F1 Score)

Geri çağırım ve duyarlılık sonuçlarının armonik ortalamasıdır.

$$\text{F} - \text{Ölçütü} = \frac{2*Duyarlılık*Geriçağırım}{duyarlılık+geriçağırım} \quad (4.22)$$

4.4.4.5 Kappa İstatistiği

Doğruluk ve F-Ölçütü ile birlikte sınıflandırıcının başarısını ölçtüğümüz bir diğer ölçüt ise Kappa istatistiğidir. Kısaca bir sınıflandırıcının tamamen şans eseri bir sonuçtan ne kadar daha başarılı bir sonuç verdiğini hesaplamaya çalışır [39]. Kappa istatistiği -1 ve 1 arası bir değer verir. -1 tamamen yanlış sonuçlar verdigini, 0 şans eseri bir sınıflandırma ile aynı başarımda olduğunu, 1 ise tamamen doğru sınıflandırdığını gösterir. $\Pr(a)$ sınıflandırıcının doğruluğu ve $\Pr(e)$ rastgelelikle elde edilen doğruluk oranı olmak üzere kappa istatistiği şu şekilde verilir.

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \qquad (4.23)$$

BÖLÜM 5

SONUÇLAR

Bu bölümde Bölüm 4.4'te bahsettiğimiz yöntemlerin uygulanması ile oluşan sonuçlar verilmektedir. Doğruluk, duyarlılık, geriçağırım, f-ölçütü ve kappa istatistiği sonuçları her deneme için bölüm 4.4.4'te açıklandığı şekliyle verilmiştir.

5.1 Dayanak

Sadece ASCII->Türkçe düzeltmesi yapılmış bunun dışında herhangi bir NLP müdahalesine uğramamış toplam 26250'si olumsuz 26136'sı olumlu olarak özdevimli işaretlenmiş ve oluşturulmuş toplam 52386 adet kullanıcı yorumundan oluşan derlem kullanılmıştır. Tüm derlem içinde Bölüm 4.3.2'de açıklaması verildiği şekliyle üssel tekrar ile hesaplanan eşik değeri ile belirlenen belirtkeleri öznitelik adayı olarak kabul ederek başlattığımız ilk denemede, diğer denemelerede de aynen uygulanacağı üzere, Naïve Bayes Çokterimli ile 10 katlı çapraz doğrulama uygulanmıştır. Sadece unigramlardan seçilen özniteliklerin ait olduğu derleme ait istatistikler aşağıdaki gibidir.

Tablo 5.1 ASCII Olağanlaştırılmış Derlem İçin Unigram Öznitelik Adayı ve Öznitelik ile Belirtke Sayıları

Derlem	Belirtke Sayısı	Öznitelik Adayları	Öznitelik
ASCII Olağanlaştırılmış	**116545**	**35818**	**14972**

Sınıflandırıcının bahsi geçen bu derlemin unigramlarına ait karışıklık matrisi (Confusion Matrix) şu şekilde oluşmuştur:

Tablo 5.2 Unigram Öznitelikler ile Sınıflandırma için Karışıklık Matrisi

		Sınıflandırıcı Sonucu	
		Olumsuz	**Olumlu**
Gerçek Sınıf	**Olumsuz**	22938	3312
	Olumlu	3502	22634

Tablo 5.3 Unigram Öznitelikler ile Sonuçlar

Sınıf	**TP Oran**	**FP Oran**	**Duyarlılık**	**Geriçağırım**	**F-Ölçütü**	**Doğruluk %**	**Kappa**
olumsuz	0.874	0.134	0.868	0.874	0.871		
olumlu	0.866	0.126	0.872	0.866	0.869		
Ağr. Ort.	0.870	0.130	0.870	0.870	**0.870**	**86.9927**	**0.7398**

Tablo 5.3'ten görülebileceği üzere 0.87 F-ölçütü ve 0.74 Kappa ile unigram, kabul edilebilir derecede başarılı bir sonuç vermiştir.

5.2 N-Gram Sonuçları

Bu sonuçlar Bölüm 4.3.1'den hatırlanacağı üzere, bir N-Gram belirtkenin belgenin kutupluluğunu belgenin içindeki konumundan bağımsız bir şekilde etkilemesi hipotezini sınaması açısından önemlidir. Sadece unigram ve sadece bigramdan oluşan derlemlerin bilgi kaybına yol açacağı hipoteziyle unigram+bigram şeklinde N-Gramların gruplanmasıyla olası bilgi kayıplarının önüne geçileceği düşünülmüştür.

Aşağıdaki tabloda ASCII->Türkçe dönüşümü yapılmış derlemden seçilen ve Tablo 4.12'te belirtilen sayıdaki N-Gram özniteliklere göre elde edilen sonuçlardan en yüksek başarımı elde eden ilk üç N-Gram için F-Ölçütü, Doğruluk ve Kappa değerleri koyu renk ile işaretlenmiştir.

Tablo 5.4 N-Gram Öznitelikler ile Sonuçlar

N-Gram	TP Oran	FP Oran	Duyarlılık	Geriçağırım	F-Ölçütü	Doğruluk %	Kappa
Unigram	0.870	0.130	0.870	0.870	0.870	86.9927	0.7398
Bigram	0.881	0.119	0.882	0.881	0.881	88.1419	0.7628
Trigram	0.800	0.201	0.813	0.800	0.798	79.9908	0.5996
Fourgram	0.643	0.358	0.748	0.643	0.601	64.3283	0.2856
1-2Gram	**0.902**	**0.098**	**0.903**	**0.902**	**0.902**	**90.1997**	**0.804**
1-3Gram	**0.899**	**0.101**	**0.902**	**0.899**	**0.899**	**89.8981**	**0.798**
1-4Gram	**0.884**	**0.116**	**0.890**	**0.884**	**0.883**	**88.3557**	**0.7672**

5.3 Sözcük Sınıflarına Göre Sonuçlar

Bu çalışmada görmek istediğimiz sonuçlardan birisi de dilbilimsel sözcük sınıflarının bir metnin kutupluğuna etkisi idi. Bu amaçla ASCII->Türkçe dönüşümü yaptığımız derlemden NLP araçlarımız tarafından dilbilimsel sözcük sınıfı tespit edilebilen belirtklelerin aynen bırakıldığı diğer belirtkelerin ise gözardı edildiği dilbilimsel sözcük sınıfına göre derlemler oluşturulmuştur. Sözcük sınıfına göre oluşturulan bu derlemler sırasıyla FIIL, ISIM, ISIM+FIIL, ISIM+SIFAT, ISIM+SIFAT+FIIL, SIFAT ve SIFAT+FIIL gruplarından oluşmaktadır.

Yukarıda verilen sözcük gruplarından oluşan derlemlere de aynen N-Gram derleminlerine uygulanan öznitelik seçim işlemleri ve kriterleri uygulanmıştır. Dayanak olarak aldığımız N-Gram sonuçlarında en yüksek başarımı 1-2 Gram ve 1-3 Gram grupları gösterdiği için sözcük sınıflarına göre oluşturulan derlemler de sadece bu özniteliklerin başarımları ölçülmüştür. Bunlara tek istisna Türkçe trigram formunda fiil grupları oluşmayacağı varsayımıyla fiil derlemidir. Fiiller sadece Unigram+Bigram öznitelikler açısından değerlendirilmiştir.

Oluşturulan dilbilimsel sözcük sınıfına göre derlemlerin unigram+bigram ve unigram+bigram+trigram öznitelikleri aşağıda verilmiştir.

Tablo 5.5 Dilbilimsel Sözcük Sınıflarına Göre Unigram+Bigram Öznitelik Sayıları

Sözcük Sınıfları 1-2 Gram	Öznitelik Adayları	Öznitelikler
Fiiller	32705	21138
İsimler	62763	35542
İsimler+Fiiller	99720	59538
İsimler+Sıfatlar	70800	40781
İsimler+Sıfatlar+Fiiller	107787	64717
Sıfatlar	4887	2809
Sıfatlar+Fiiller	42174	27120

Tablo 5.6 Dilbilimsel Sözcük Sınıflarına Göre Unigram+Bigram+Trigram Öznitelik Sayıları

Sözcük Sınıfları 1-3 Gram	Öznitelik Adayları	Öznitelikler
İsimler	100814	69577
İsimler+Fiiller	157585	112225
İsimler+Sıfatlar	117662	82394
İsimler+Sıfatlar+Fiiller	174955	125258
Sıfatlar	8471	5488
Sıfatlar+Fiiller	62723	46667

Tablo 5.7 Dilbilimsel Sözcük Sınıflarına Göre Unigram+Bigram Sonuçları

Sözcük Sınıfı 1-2 Gram	TP Oran	FP Oran	Duyarlılık	Geriçağırım	F-Ölçütü	Doğruluk %	Kappa
Fiil	0.739	0.262	0.742	0.739	0.738	73.8671	0.4772
İsim	0.808	0.192	0.809	0.808	0.808	80.7716	0.6155
İsim Fiil	**0.861**	**0.139**	**0.861**	**0.861**	**0.860**	**86.0535**	**0.7211**
İsim Sıfat	**0.861**	**0.139**	**0.862**	**0.861**	**0.861**	**86.1337**	**0.7227**
İsim Sıfat Fiil	**0.893**	**0.107**	**0.894**	**0.893**	**0.893**	**89.3235**	**0.7865**
Sıfat	0.682	0.319	0.703	0.682	0.673	68.1842	0.3632
Sıfat Fiil	0.829	0.171	0.829	0.829	0.829	82.9172	0.6583

Tablo 5.8 Dilbilimsel Sözcük Sınıflarına Göre Unigram+Bigram+Trigram Sonuçları

Sözcük Sınıfı 1-3 Gram	TP Oran	FP Oran	Duyarlılık	Geriçağırım	F-Ölçütü	Doğruluk %	Kappa
İsim	0.789	0.211	0.793	0.789	0.789	78.9199	0.5785
İsim Fiil	**0.846**	**0.154**	**0.849**	**0.846**	**0.846**	**84.6047**	**0.6922**
İsim Sıfat	**0.852**	**0.148**	**0.855**	**0.852**	**0.852**	**85.2041**	**0.7041**
İsim Sıfat Fiil	**0.885**	**0.115**	**0.887**	**0.885**	**0.885**	**88.5198**	**0.7704**
Sıfat	0.627	0.374	0.679	0.627	0.597	62.7057	0.2532
Sıfat Fiil	0.799	0.201	0.799	0.799	0.799	79.8897	0.5978

5.4 Sözcük Köklerine Göre Sonuçlar

Çalışmada araştırılması amaçlanan durumlardan bir tanesi de sondan eklemeli bir dil olan Türkçe'nin kelime köklerinin ve bu köklerin olumsuzluk eki aldığı durumların taşıdığı yargıya etkisi bakımından incelenmesidir. Bu amaçla hazırlanan veri kümelerinde kökler elde edilmek istense de Zemberek2'nin mevcut yetenekleri bakımından sözcüklerin çok büyük bir bölümünün köklerini bulmak mümkün olmamıştır. Ancak bu duruma rağmen kökleri bulunamayan belirtkelerin olduğu gibi dahil edilmesi ile aşağıda sonuçları verildiği gibi sonuçlar elde edilmiştir.

Tablo 5.9 Sözcük Köklerine Göre N-Gram Öznitelikler ile Sonuçlar

N-Gram	TP Oran	FP Oran	Duyarlılık	Geriçağırım	F-Ölçütü	Doğruluk %	Kappa
Unigram	0.832	0.168	0.832	0.832	0.832	83.152	0.663
Bigram	**0.871**	**0.129**	**0.871**	**0.871**	**0.871**	**87.134**	**0.7427**
Trigram	0.824	0.176	0.828	0.824	0.823	82.4075	0.6481
Fourgram	0.694	0.307	0.761	0.694	0.673	69.4384	0.3881
1-2Gram	**0.883**	**0.117**	**0.883**	**0.883**	**0.883**	**88.2659**	**0.7653**
1-3Gram	**0.890**	**0.110**	**0.892**	**0.890**	**0.889**	**88.9532**	**0.7791**

Tablo 5.10 Sözcük Köklerinin Olumsuzluk Durumuna Göre N-Gram Öznitelikler ile Sonuçlar

N-Gram	TP Oran	FP Oran	Duyarlılık	Geriçağırım	F-Ölçütü	Doğruluk %	Kappa
Unigram	0.840	0.160	0.840	0.840	0.840	83.9728	0.6795
Bigram	**0.878**	**0.122**	**0.878**	**0.878**	**0.878**	**87.8021**	**0.756**
Fourgram	0.696	0.305	0.765	0.696	0.674	69.5854	0.391
1-2Gram	**0.890**	**0.110**	**0.890**	**0.890**	**0.890**	**88.9684**	**0.7794**
1-3Gram	**0.895**	**0.105**	**0.897**	**0.895**	**0.895**	**89.4762**	**0.7896**

Bölüm 4.2.4'ten hatırlanacağı üzere olumsuzluk Türkçe'de iki durum ile sağlanabiliyordu. Bu bölümde yapılan çalışmada sözcükler ile oluşan olumsuzluk durumundan ziyade ekler ile oluşturulan olumsuzluk durumu gözlenmiştir. Türkçe'de ekler ile oluşturulan olumsuzluk fiillere gelen olumsuzluk ekleriyle mümkün olduğundan sadece olumsuzluk eki almış fiiller üzerinden bir sonuca ulaşmak sonuçlardan anlaşılacağı üzere İngilizce çalışmalardaki kadar etkili olamamıştır.

BÖLÜM 6

TARTIŞMALAR

Bu çalışmada Türkçe düşünce çözümleme için daha önce sinema filmlerinin kullanıcı yorumları alanında yapılan çalışmada [26] SVM (Support Vector Machines) sınıflandırıcısı ile elde edilen %85'lik F-Ölçütü değeri %90'lık F-Ölçütü ile geçilmiştir. Bu sonuç metin sınıflandırma problemlerinde Naïve Bayes Çokterimli algoritmasının SVM'e göre daha iyi performans göstermesi olarak yorumlanabileceği gibi aynı zamanda Eroğul'un çalışmasında bu performansın 5000 civarı öznitelik ile yakalandığını bu çalışmada ise yaklaşık 79000 öznitelik ile başarımın arttığı gözardı edilmemelidir. Bu çalışmada belirlenen 79000 öznitelik, sınıflandırıcının eğitim safhasında görmediği veriler üzerindeki başarımına oldukça olumlu katkı sağlamıştır. Bu bağlamda, önerilen yöntem ile hem aşırı uyumdan kaçınılmış hem de görülmemiş örnekler üzerinde tutarlı ve başarılı sonuçlar verecek bir yöntem önerilmiştir. On katlı çapraz doğrulama sayesinde toplam 52386 örnekten oluşan derlemden alınan on farklı grubun sırayla sınama verisi olarak kullanılması ile elde edilen sonuçların ağırlıklı ortalamasındaki yüksek değer bu başarımın en büyük göstergesi kabul edilebilir.

Yapılan çalışma Türkçe'nin kutupluğu hakkında unigram+bigram kelime gruplarının oldukça önemli bilgi sağladığını ortaya koyması açısından önemli olmuştur. Tablo 5.4'den görüleceği üzere tüm kelimeler üzerinden unigram+bigram grubunun %90 F-Ölçütü değerine ulaşması oldukça önemlidir. Bu N-Gramların ise isim+fiil, isim+sıfat ve isim+sıfat+fiil olduğu görülmektedir.

Eklerde görülebileceği üzere, kutupluk hakkında en güçlü kelimeleri incelediğimizde ise N-Gram sınıflandırma sonuçlarıyla bu kelimelerin örtüştüğü

görülmüştür. Ayrıca bu kelimeler gelecekte yapılabilecek çalışmalar hakkında da oldukça faydalı olmuştur.

Daha önce bahsedildiği gibi bu çalışmanın başlıca amaçlarından biri Türkçe düşünce çözümleme problemi konusunda gelecekteki çalışmalar için bir taban oluşturulması, karşılaşılabilecek muhtelif problemlerin araştırılması olmuştur. Gelecekteki çalışmalar adına ilk kullanılacak veri, eklerde verilen, Türkçe kutupluk kelimelerinin en güçlü olanlarının varyasyonları olmuştur. "izlenmesi gereken", "gereksiz film", "güzel bir", "muhteşem bir" kelimelerinden rahatlıkla anlaşılacağı üzere bu güçlü öznitelikler sadece sinema yorumları alanında değil diğer alanlarda da içinde bulunduğu metnin düşünce kutupluğunu etkileme gücüne sahiptir. Bu ve benzeri öznitelikler kullanılarak isim+fiil, isim+sıfat, isim+sıfat+fiil gruplarından diğer alanlar için özdevimli öznitelik sözlükleri oluşturulabilir. Ayrıca, elde edilen bu öznitelikler ile Türkçe için düşünce kelime ağı oluşturulabilir gözükmektedir.

Naïve Bayes Çokterimli bu çalışmada iki sınıf problemini çözmek amacıyla kullanılsa da çok sınıflı problemlerde de oldukça başarılı olabilen bir algoritmadır. Bu nedenle 1-10 arasında yıldız ile ölçeklenen kullanıcı yorumları sinema alanında değerlendirildiği gibi ürün yorumları alanında da benzeri çalışmalar yapılabilir. Aynı zamanda Scherer'in modeli [6] baz alınarak düşüncenin sahip olduğu duygunun çeşidi bakımından (örneğin öfke, mutluluk, sinirlilik vb.) çok sınıflı sınıflandırılmasına çalışılması bu çalışmanın tamamlayıcısı bir çalışma olacaktır.

Son olarak Pang ve arkadaşlarının [10] kullanıma sunduğu İngilizce kutupluluk derlemi benzeri bir derlem Türkçe için oluşturulup araştırmacıların kullanımına sunulduğu takdirde bu çalışmanın gelecekteki çalışmalara katkı vermesi bakımından en önemli çıktılardan birisi olacaktır.

KAYNAKÇA

1. SINTEF. ***Big Data, for better or worse: 90% of world's data generated over last two years***. ScienceDaily. 22 Mayıs 2013.
2. W.N. Francis and H. Kucera. ***A Standart Corpus of Present-Day Edited American English, for use Digital Computers***. Brown University, 1964, 1971, 1979.
3. M.P. Marcus, B. Santorini, M.A. Marcinkiewicz. ***Building a Large Annotated Corpus of English: The Penn Treebank***. Computational Linguistics, 1993.
4. B. Pang, L. Lee, and S. Vaithyanathan, ***Thumbs up? Sentiment Classification using machine learning techniques***. In Proceedings of EMNLP-02, 7th Conference on Empirical Methods in Natural Language Processing Philadelphia, PA, pages 79–86. 2002.
5. K. Dave, S. Lawrence, D.M. Pennock. ***Mining the peanut gallery: Opinion extraction and semantic classification of product reviews***. In proceedings of WWW, p. 519-528, 2003.
6. Scherer KR. ***Psychological models of emotion***. In The Neuropsychology of Emotion.ed. JC Borod, pp. 137–62. New York: Oxford Univ. Press. 2000.
7. Ronen Feldman, ***Techniques and applications for sentiment analysis*** Communications of the ACM, Vol. 56 No. 4, Pages 82-89. 2013.
8. H. Yu, and V. Hatzivassiloglou. ***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.
9. E. Riloff, and J. Wiebe. ***Learning extraction patterns for subjective eExpressions***. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2003.
10. B. Pang, and L. Lee. ***A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on minimum cuts***. In Proceedings of the Association for Computational Linguistics, 271–278. 2004.
11. R. Narayanan, B. Liu, and A. Choudhary, ***Sentiment analysis of conditional sentences***. In Proceedings of the 2009 Conference on Empirical Methods in

Natural Language Processing Association for Computational Linguistics, 180–189, Singapore. 2009.

12. M. Hu, and B. Liu. ***Mining and summarizing customer reviews***. In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 168–177. 2004.
13. J. Lafferty, A. McCallum, and F. Pereira. ***Conditional random fields: Probabilistic models for segmenting and labeling sequence data***. In Proc. 18th International Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA, 282–289, 2001.
14. N. Jakob, and I. Gurevych. ***Extracting opinion targets in a single-and cross-domain setting with conditional random fields***. In Proceedings of Conference on Empirical Methods in Natural Language Processing, 2010.
15. N. Jindal, and B. Liu. ***Identifying comparative sentences in text documents***. In Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval, 2006.
16. J. Kamps, M. Marx, R.J. Mokken, and M. de Rijke. ***Using WordNet to measure semantic orientation of adjectives***. LREC, 2004.
17. Peng, W. and Park, D.H. ***Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization***. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011.
18. V. Hatzivassiloglou, and K. McKeown. ***Predicting the semantic orientation of adjectives***. In Proceedings of the Joint ACL/EACL Conference, 174–181, 1997.
19. W. Du, S. Tan, X. Cheng, and X. Yun. ***Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon***. In Proceedings of ACM International Conference on Web Search and Data Mining, 2010.
20. S. Feng, R. Bose, and Y. Choi. ***Learning general connotation of words using graph-based algorithms***. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics Edinburgh, Scotland, UK, 1092–1103, 2011.

21. Sanjiv Das, and Mike Chen. ***Yahoo! for Amazon: Extracting market sentiment from stock message boards***. In Proceedings of the Asia Pacific Finance Association Annual Conference APFA, 2001.
22. Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. ***The General Inquirer: A Computer Approach to Content Analysis***. MIT Press. 1966.
23. J.W. Pennebaker, R.J. Booth & M.E. Francis. ***Linguistic Inquiry and Word Count: LIWC*** Austin, TX, 2007.
24. Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. ***Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis***. Proc. of HLT-EMNLP, 2005.
25. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. ***SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining***. LREC, 2010.
26. U. Eroğul. ***Sentiment analysis in Turkish***. Yüksek Lisans Tezi. Ortadoğu Teknik Üniversitesi, 2009
27. M. Kaya, G. Fidan, and I. Toroslu. ***Sentiment analysis of Turkish political news***. Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on, 1:174–180, 2012.
28. Z. Boynukalin. ***Emotion analysis of Turkish texts by using machine learning methods***. Middle East Technical University, 2012.
29. O. Cakmak, A. Kazemzadeh, D. Can, S. Yildirim, and S. Narayanan, ***Root-word analysis of Turkish emotional language***. Corpora for Research on Emotion Sentiment & Social Signals, 2012.
30. O. Cakmak, A. Kazemzadeh, S. Yildirim, and S. Narayanan. ***Using interval type-2 fuzzy logic to analyze Turkish emotion words***. Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, pages 1–4, 2012.
31. M. Kaya, G. Fidan, and I. Toroslu. ***Transfer learning using twitter data for improving sentiment classification of Turkish political news***. to appear in ISCIS 2013, 2013.
32. E. Demirtaş. ***Cross-Lingual Sentiment Analysis with Machine Translation***. Yüksek Lisans Tezi. Eindoven University of Technology, 2013.

33. James Pustejovsky. ***Unifying Linguistic Annotations: A TimeML Case Study***. In Proceedings of the Text, Speech, and Dialogue Conference. 2006.
34. James Pustejovsky and Amber Stubbs. ***Natural Language Annotation for Machine Learning***. O'Reilly Publishers. 2012.
35. N. İlhan. ***Türkçe'de olumsuzluk***. Karaman Dil Kültür ve Sanat Dergisi, S. 1, Ankara, s. 271-279. 2005.
36. T. Mitchell. ***Machine Learning***, McGraw Hill. ISBN 0-07-042807-7, p.2. 1997.
37. Y. Yang and J. O. Pedersen. ***A Comparative Study on Feature Selection in Text Categorization***. ICML Proceedings of the Fourteenth International Conference on Machine Learning P. 412-420 1997
38. A. Mccallum, K. Nigam: ***A Comparison of Event Models for Naive Bayes Text Classification***. In: AAAI-98 Workshop on 'Learning for Text Categorization', 1998.
39. J. Cohen. ***A coefficient of agreement for nominal scales***, Educational and Psychological Measurement Vol.20, No.1, pp.37-46. 1960.

EK 1 – Türkçe En Güçlü 40 Unigram ve Kutupluğu[20]

Unigram	Olumsuz	Olumlu
Kötü	0,005923305	0,001435402
Berbat	0,003181958	1,47E+12
Harika	3,64E+12	0,003990755
Güzel	0,002619884	0,006471254
Yazık	0,003013491	2,94E+11
Kaybı	0,002485136	1,13E+12
Mükemmel	3,75E+12	0,003386003
Saçma	0,003122457	5,66E+11
Hiç	0,00486452	0,002010982
Sıkıcı	0,002990841	5,63E+11
Mutlaka	2,45E+12	0,002596169
Beğenmedim	0,002237239	1,98E+12
Ederim	3,59E+11	0,002684665
Süper	4,98E+11	0,0029903
İyi	0,003320012	0,006461626
Yok	0,004606423	0,002134347
Gereksiz	0,002184398	3,36E+12
Vasat	0,002129979	3,17E+11
Beğendim	1,93E+12	0,002047177
Muhteşem	3,39E+11	0,002344234
Başarılı	9,11E+11	0,003323975
Gereken	5,07E+11	0,002529381
İzlenmesi	1,57E+12	0,001799474
Müthiş	2,51E+11	0,002060364
Hayal	0,002452014	7,24E+11
Kırıklığı	0,001609503	2,10E+12
Değmez	0,001159175	2,77E+11
Para	0,001823854	3,39E+10
Eğlenceli	3,63E+12	0,002002956
İzleyin	9,83E+11	0,002948481
İzlenmeli	4,84E+10	0,001239428
Etkileyici	3,87E+11	0,002001092
Bile	0,004023548	0,002222516
İzlemeyin	0,0013643	1,54E+12
Rezalet	0,001066536	2,86E+11
Basit	0,002036659	5,81E+11
Zaman	0,003100558	0,001837012
Ne	0,004923191	0,003139052
Kötüydü	0,001171818	8,94E+10

[20] Unigram öznitelikler ve sınıflandırmaya dair net bilgi kazancı etkileri.

EK 2 – Türkçe En Güçlü 40 Bigram ve Kutupluğu[21]

Digram	Olumsuz	Olumlu
En İyi	3,66E+12	0,004386397
Çok Güzel	3,83E+11	0,004226196
Kötü Bir	0,002035125	3,15E+11
Tavsiye Ederim	1,51E+11	0,003048318
En Kötü	0,001688522	1,45E+10
Çok Kötü	0,001899093	2,95E+11
Güzel Bir	6,02E+10	0,004298605
Zaman Kaybı	0,001408558	7,84E+10
Harika Bir	1,11E+12	0,002483113
Çok İyi	5,76E+11	0,003497829
Mükemmel Bir	1,07E+10	0,002002716
Hayal Kırıklığı	0,001260891	2,52E+11
Berbat Bir	9,10E+11	3,73E+10
Çok Beğendim	4,31E+11	0,00162343
İzlenmesi Gereken	5,51E+10	0,001648118
Çok Sıkıcı	9,04E+11	5,55E+10
Gereken Bir	1,91E+12	0,002059354
Mutlaka İzleyin	3,00E+11	0,001420999
Hiç Beğenmedim	8,39E+11	4,43E+11
Başarılı Bir	1,42E+12	0,001884751
Tavsiye Etmiyorum	7,73E+11	2,32E+11
Saçma Sapan	0,001017163	1,72E+12
Kadar Kötü	0,001074478	2,20E+12
Bu Kadar	0,002899296	0,002355577
Vakit Kaybı	8,21E+10	7,82E+08
Çok Başarılı	1,55E+12	0,001782237
Süper Bir	5,08E+11	0,001346545
Sıkıcı Bir	7,91E+11	1,16E+12
Gereksiz Bir	6,83E+10	5,30E+11
Bir Hayal	7,79E+11	1,28E+12
Vasat Bir	7,80E+10	1,40E+12
Hiç Bir	0,001257634	5,76E+11
Herkese Tavsiye	1,48E+11	9,79E+11
Tavsiye Etmem	6,32E+11	5,06E+11
Muhteşem Bir	7,87E+10	0,001261259
Saçma Bir	6,24E+11	4,77E+10
Müthiş Bir	6,03E+10	0,001194327
Çok Güzeldi	9,99E+10	0,001291918
Hoş Bir	5,87E+11	0,001114959
En Güzel	2,22E+12	0,001643207

[21] Bigram öznitelikler ve sınıflandırmaya dair net bilgi kazancı etkileri.

EK 3 – Türkçe En Güçlü 40 Trigram ve Kutupluğu[22]

Trigram	Olumsuz	Olumlu
Kötü Bir Film	0,00142663	3,39E+12
Çok Güzel Bir	1,03E+11	0,003851744
Güzel Bir Film	2,96E+12	0,004979118
İzlediğim En Kötü	9,30E+11	1,92E+10
Çok Kötü Bir	9,47E+10	8,22E+09
İzlediğim En Iyi	2,37E+09	0,002649546
Harika Bir Film	4,62E+11	0,002753882
Mükemmel Bir Film	1,89E+11	0,00215856
Gereken Bir Film	1,54E+12	0,003127543
Bu Kadar Kötü	7,90E+10	1,46E+11
Berbat Bir Film	6,44E+11	2,72E+10
Bir Hayal Kırıklığı	6,46E+11	9,65E+10
İzlenmesi Gereken Bir	3,98E+10	0,001999195
Güzel Bir Filmdi	1,77E+11	0,001770829
Tam Bir Hayal	5,76E+11	4,73E+10
Gereksiz Bir Film	5,01E+11	1,53E+10
Sıkıcı Bir Film	5,30E+11	8,10E+09
Vasat Bir Film	5,89E+10	1,49E+11
Süper Bir Film	2,89E+11	0,001606583
Saçma Sapan Bir	4,44E+11	3,64E+10
Herkese Tavsiye Ederim	1,42E+11	0,001330377
Başarılı Bir Film	4,51E+11	0,001592971
İzlediğim En Güzel	2,51E+11	0,001428639
En İyi Filmlerden	7,61E+09	0,001187239
Saçma Bir Film	3,75E+11	9,00E+08
En İyi Filmi	7,02E+10	0,001680245
En Kötü Filmlerden	3,65E+11	9,03E+09
En Kötü Film	3,89E+12	3,02E+11
İzlemenizi Tavsiye Ederim	2,78E+11	0,001300355
Çok Başarılı Bir	3,23E+11	0,001333658
Tek Kelimeyle Mükemmel	5,89E+08	9,19E+11
Çok Iyi Bir	1,81E+12	0,002132365
En Kötü Filmdi	2,96E+12	1,79E+10
Muhteşem Bir Film	2,84E+11	0,001178283
Gerçekten Çok Güzel	1,49E+11	0,001072673
Eğlenceli Bir Film	3,71E+09	0,00124207
Hayal Kırıklığına Uğradım	4,65E+11	1,89E+12
Kesinlikle Tavsiye Etmiyorum	2,84E+12	1,79E+10
Hiç Bir Şey	4,94E+10	2,32E+12
En Iyi Filmlerinden	7,92E+08	9,57E+10

[22] Trigram öznitelikler ve sınıflandırmaya dair net bilgi kazancı etkileri.

EK 4 – Türkçe En Güçlü 40 Fourgram ve Kutupluğu[23]

Fourgram	Olumsuz	Olumlu
çok güzel bir film	3,11E+10	0,004639285
çok kötü bir film	7,36E+10	1,03E+12
tam bir hayal kırıklığı	5,25E+11	3,04E+10
hayatımda izlediğim en kötü	4,75E+11	3,53E+10
izlenmesi gereken bir film	3,37E+11	0,003261353
bu kadar kötü bir	3,25E+12	3,32E+10
izlediğim en iyi filmlerden	6,44E+08	0,001569648
izlediğim en kötü film	2,87E+12	1,87E+11
çok güzel bir filmdi	3,42E+09	0,001558283
izlediğim en kötü filmlerden	2,76E+12	1,89E+11
izlediğim en kötü filmdi	2,44E+11	3,53E+10
bir film 10 10	6,44E+08	0,001299162
hayatımda izlediğim en iyi	6,44E+08	0,001272393
güzel bir film olmuş	6,44E+08	0,001281671
saçma sapan bir film	2,49E+12	1,92E+10
gerçekten çok güzel bir	6,44E+08	0,001122889
en iyi filmlerden biri	3,57E+10	0,001175288
kesinlikle izlenmesi gereken bir	6,44E+08	0,001081398
en kötü filmlerden biri	1,90E+12	3,53E+10
kadar kötü bir film	2,72E+12	1,40E+12
ve gereksiz bir film	1,82E+11	3,53E+10
çok kötü bir filmdi	1,82E+11	3,53E+10
bence çok güzel bir	6,44E+08	9,83E+11
çok güzel bi film	3,62E+10	0,001065066
bir film tavsiye ederim	6,55E+09	0,001104273
izlemesi gereken bir film	6,60E+09	0,001048735
en iyi filmlerinden biri	3,67E+09	9,91E+11
çok başarılı bir film	1,23E+11	0,001140854
hiç bir şey yok	2,31E+12	1,13E+12
izlenmeye değer bir film	6,64E+09	0,001006682
kaybından başka birşey değil	1,56E+12	3,53E+10
ancak bu kadar kötü	1,53E+12	3,53E+10
çok sıkıcı bir film	1,53E+12	3,53E+10
vasatın altında bir film	1,48E+12	3,53E+10
izlediğim en güzel filmlerden	1,25E+11	0,001044062
hayatımda izlediğim en berbat	1,61E+11	2,04E+11
çok eğlenceli bir film	6,44E+08	7,81E+11
hayatımda izlediğim en saçma	1,40E+12	3,53E+10
izlediğim en iyi türk	6,44E+08	7,51E+11
tek kelimeyle mükemmel bir	6,44E+08	7,51E+11

[23] Trigram öznitelikler ve sınıflandırmaya dair net bilgi kazancı etkileri.

EK 5 – Türkçe En Güçlü 40 1-4 Gram İçin Kutupluk[24]

Fourgram	Olumsuz	Olumlu
kötü	0,001301678	5,86E+11
harika	8,03E+10	0,001613998
berbat	6,93E+11	6,01E+11
güzel	5,78E+11	0,002626362
yazık	6,63E+11	1,20E+11
kaybı	5,39E+11	4,65E+10
en iyi	1,01E+12	0,001450922
mükemmel	8,33E+10	0,001348184
çok güzel	1,05E+12	0,001392605
saçma	6,86E+11	2,31E+12
kötü bir	5,59E+11	1,04E+12
hiç	0,001072494	8,19E+11
tavsiye ederim	4,16E+08	0,001008316
sıkıcı	6,52E+11	2,30E+11
en kötü	4,64E+12	4,79E+10
çok kötü	5,18E+11	9,80E+10
mutlaka	5,41E+10	0,001056781
güzel bir	1,66E+12	0,001417672
beğenmedim	4,90E+11	7,92E+10
ederim	7,92E+10	0,001092804
zaman kaybı	3,83E+12	2,60E+10
süper	1,10E+12	0,001197901
iyi	7,31E+10	0,002631137
yok	0,001015574	8,69E+11
harika bir	3,02E+11	8,17E+11
gereksiz	4,78E+12	1,36E+12
beğendim	4,25E+10	8,33E+11
vasat	4,62E+12	1,30E+11
kötü bir film	3,77E+12	5,92E+11
muhteşem	7,51E+10	9,45E+11
başarılı	2,01E+12	0,001352068
gereken	1,12E+11	0,001029595
izlenmesi	3,47E+11	7,31E+10
müthiş	5,54E+10	8,35E+11
çok güzel bir	2,72E+11	6,73E+11
güzel bir film	7,83E+10	8,67E+11
çok iyi	1,58E+11	0,001155406
hayal	5,38E+10	2,95E+11
izlediğim en kötü	2,45E+12	3,35E+09
kırıklığı	3,52E+11	8,56E+10

[24] 1-4 Gram aralığındaki tüm öznitelikler birlikte değerlendirildiği durumda sınıflandırmaya dair net bilgi kazancı etkileri.

EK 6 – Türkçe En Güçlü 40 Fiil 1-2 Gram İçin Kutupluk[25]

Fourgram	Olumsuz	Olumlu
beğenmedim	0,004312248	5,51E+11
beğendim	3,72E+11	0,005882203
gereken	0,001007188	0,007461004
izlenmesi	3,06E+12	0,005257407
değmez	0,002299584	8,19E+10
izlenmesi gereken	1,37E+12	0,003954909
etkileyici	7,53E+11	0,005758047
izleyin	0,001934361	0,008311492
izlemeyin	0,002604537	4,59E+11
izlenmeli	9,38E+10	0,003353593
gitmeyin	0,002283747	2,91E+12
etmiyorum	0,001958267	1,52E+11
olmamış	0,002776617	7,21E+11
sıkılmadan	1,84E+12	0,002896913
harcamayın	0,001294771	1,91E+10
sıkıldım	0,001887418	4,01E+11
etmem	0,001553353	2,31E+12
kaçırmayın	1,70E+12	0,002582967
izlemeye değmez	9,78E+11	1,12E+11
olmayan	0,00286084	0,001734167
izlemesi	1,65E+12	0,002154884
verdiğim	0,001406492	3,39E+11
uğradım	0,001328319	3,18E+10
bayıldım	7,18E+09	0,001677126
boşa harcamayın	7,81E+11	1,16E+11
sıkıcıydı	0,001004919	1,25E+11
çalışmışlar	0,001016244	1,25E+12
bekliyorum	3,24E+11	0,002361061
gittim	0,002403764	0,001506767
izlemeli	1,19E+12	0,001696891
anlamadım	0,00187522	9,60E+11
izlemesi gereken	4,69E+11	0,001314652
acıdım	8,27E+11	8,48E+10
izlenilmesi	3,04E+11	0,001224231
unutulmaz	1,91E+11	0,001161666
yapmaya	0,001395345	5,55E+11
kurtaramamış	7,63E+11	6,76E+10
izlerken sıkıldım	6,82E+10	3,10E+10
çıkarmış	2,86E+11	0,002036062
izlemenizi	4,26E+12	0,002364008

[25] Zemberek2 tarafından kökü fiil olarak işaretlenen sözcüklerden oluşturulan veri kümesindeki 1-2 Gram aralığındaki tüm öznitelikler ve sınıflandırmaya dair net bilgi kazancı etkileri.

EK 7 – Türkçe En Güçlü 40 İsim 1-4 Gram İçin Kutupluk[26]

Fourgram	Olumsuz	Olumlu
yazık	0,001483302	2,92E+12
kaybı	0,00116078	1,16E+10
hiç	0,002419874	0,002032273
tavsiye ederim	9,42E+10	0,002463317
mutlaka	1,22E+12	0,002507705
ederim	1,79E+12	0,002668475
yok	0,002318915	0,002122332
süper	2,54E+11	0,002862768
gereksiz	0,001053117	3,44E+12
vasat	0,001017863	3,19E+11
müthiş	1,27E+12	0,002049039
hayal	0,001203909	7,38E+11
para	9,05E+11	3,45E+11
kırıklığı	7,82E+11	2,13E+12
hayal kırıklığı	7,63E+11	2,08E+12
eğlenceli	1,83E+11	0,001891382
ne	0,002467666	0,003159097
sapan	6,38E+11	1,46E+12
özellikle	4,69E+12	0,002744569
rezalet	4,95E+11	2,96E+11
gerçekten	0,001036082	0,004160741
vakit kaybı	4,83E+12	6,55E+10
diye	0,001625228	0,001744003
paranıza	4,26E+12	2,77E+11
sadece	0,001591627	0,001724657
gereksiz film	3,94E+11	2,28E+10
hoş	1,88E+11	0,001643684
paraya	4,87E+11	1,11E+11
tavsiye	7,14E+11	0,003071228
süper film	4,00E+10	9,76E+11
gayet	3,14E+09	0,002028318
değer	2,16E+12	0,0016399
vasat film	4,29E+12	8,03E+10
paranıza yazık	3,36E+11	1,22E+11
hiçbir	8,89E+11	6,75E+11
değil	0,001968978	0,002685008
en	0,001823384	0,005576819
fiyasko	3,52E+11	3,47E+10
sürükleyici	1,20E+12	0,00118909
anlamsız	4,44E+12	1,41E+12

[26] Zemberek2 tarafından kökü isim olarak işaretlenen sözcüklerden oluşturulan veri kümesindeki 1-4 Gram aralığındaki tüm öznitelikler ve sınıflandırmaya dair net bilgi kazancı etkileri.

EK 8 – Türkçe En Güçlü 40 İsim+Fiil 1-4 Gram İçin Kutupluk[27]

Fourgram	Olumsuz	Olumlu
yazık	9,58E+11	1,92E+12
kaybı	7,62E+11	7,60E+10
hiç	0,00156191	0,001323165
tavsiye ederim	6,10E+10	0,001615044
sıkıcı	9,39E+11	3,73E+11
mutlaka	7,94E+10	0,001686301
beğenmedim	7,06E+11	1,29E+12
ederim	1,16E+12	0,001750612
yok	0,001483914	0,001390552
süper	1,62E+11	0,001901363
beğendim	6,24E+10	0,001332727
gereksiz	6,86E+11	2,22E+12
vasat	6,63E+11	2,10E+11
gereken	1,62E+12	0,001659092
izlenmesi	5,08E+10	0,001175479
müthiş	8,11E+10	0,00134139
hayal	7,80E+11	4,78E+11
para	5,85E+11	2,23E+12
kırıklığı	5,06E+11	1,39E+11
değmez	3,64E+11	1,84E+11
hayal kırıklığı	4,94E+11	1,36E+11
izlenmesi gereken	2,22E+11	8,76E+11
etkileyici	1,23E+12	0,00130123
bile	0,001289107	0,001463816
izleyin	3,16E+11	0,001906711
izlenmeli	1,50E+11	7,84E+11
eğlenceli	1,17E+12	0,00126119
gereken film	6,75E+10	0,001045138
mutlaka izleyin	1,21E+11	7,53E+11
hiç beğenmedim	3,35E+12	2,37E+11
izlemeyin	4,30E+11	1,02E+12
ne	0,001581438	0,002064956
tavsiye etmiyorum	3,08E+11	1,24E+11
rezalet	3,26E+11	1,91E+11
sapan	4,10E+12	9,47E+10
boş	5,81E+11	2,98E+12
gitmeyin	3,73E+11	6,47E+10
etmiyorum	3,24E+12	3,38E+11
özellikle	3,02E+11	0,001786724
boşa	4,01E+10	1,01E+12

[27] Zemberek2 tarafından kökü isim+fiil olarak işaretlenen sözcüklerden oluşturulan veri kümesindeki 1-4 Gram aralığındaki tüm öznitelikler ve sınıflandırmaya dair net bilgi kazancı etkileri.

EK 9 – Türkçe En Güçlü 40 İsim+Sıfat 1-4 Gram İçin Kutupluk[28]

Fourgram	Olumsuz	Olumlu
kötü	0,002419236	0,001099061
berbat	0,001273028	1,12E+12
harika	1,51E+11	0,002956495
yazık	0,001231177	2,21E+11
güzel	0,001088302	0,004861387
en iyi	1,90E+11	0,002710039
kaybı	9,76E+11	8,68E+10
çok güzel	2,06E+12	0,00260297
mükemmel	1,58E+11	0,002460391
saçma	0,00127459	4,34E+12
hiç	0,002002028	0,001530999
tavsiye ederim	7,83E+10	0,001889555
en kötü	8,68E+11	8,99E+10
çok kötü	9,62E+11	1,87E+12
kötü film	9,57E+11	1,97E+12
mutlaka	1,02E+12	0,001929932
ederim	1,48E+12	0,002044419
yok	0,001906666	0,001612923
süper	2,08E+11	0,002186995
iyi	0,001372851	0,004914357
güzel film	2,35E+12	0,002138326
gereksiz	8,86E+10	2,57E+10
vasat	8,45E+11	2,44E+12
başarılı	3,78E+11	0,002513481
muhteşem	1,41E+11	0,001723731
çok iyi	2,98E+12	0,002174834
müthiş	1,05E+12	0,001548676
hayal	1,00E+12	5,55E+11
kırıklığı	6,53E+11	1,61E+12
harika film	2,65E+10	0,001049921
para	7,47E+11	2,60E+12
hayal kırıklığı	6,38E+11	1,57E+12
berbat film	4,35E+11	1,41E+11
çok güzel film	2,52E+10	9,80E+11
eğlenceli	1,49E+12	0,001467376
basit	8,35E+11	4,42E+11
ne	0,002030953	0,002382632
mükemmel film	1,85E+11	8,69E+11
sapan	5,30E+11	1,10E+12
kötüydü	4,74E+12	6,68E+10

[28] Zemberek2 tarafından kökü isim+sıfat olarak işaretlenen sözcüklerden oluşturulan veri kümesindeki 1-4 Gram aralığındaki tüm öznitelikler ve sınıflandırmaya dair net bilgi kazancı etkileri.

EK 10 – Türkçe En Güçlü 40 İsim+Sıfat+Fiil 1-4 Gram İçin Kutupluk[29]

Fourgram	Olumsuz	Olumlu
kötü	0,001657193	7,73E+11
harika	1,02E+12	0,002106247
berbat	8,78E+11	7,95E+10
güzel	7,38E+10	0,003447243
yazık	8,46E+11	1,56E+11
en iyi	1,29E+11	0,001910115
kaybı	6,80E+11	6,16E+10
mükemmel	1,07E+12	0,001752814
çok güzel	1,36E+12	0,00182632
saçma	8,73E+11	3,04E+11
hiç	0,001369464	0,001079153
tavsiye ederim	5,33E+11	0,001327968
en kötü	5,96E+11	6,30E+10
sıkıcı	8,29E+11	3,03E+12
mutlaka	6,93E+10	0,001388072
beğenmedim	6,23E+11	1,05E+12
çok kötü	6,56E+11	1,31E+12
kötü film	6,34E+11	1,24E+12
ederim	1,01E+12	0,00143861
süper	1,42E+12	0,001558183
iyi	9,35E+11	0,003463261
yok	0,001298848	0,001141804
güzel film	1,48E+12	0,001470649
gereksiz	6,09E+11	1,80E+12
beğendim	5,45E+11	0,001094733
vasat	5,86E+10	1,72E+12
muhteşem	9,64E+10	0,001231969
başarılı	2,57E+11	0,001778139
gereken	1,42E+11	0,001354497
izlenmesi	4,45E+10	9,59E+11
müthiş	7,10E+10	0,001095853
çok iyi	2,02E+11	0,001523049
hayal	6,86E+11	3,89E+11
izlediğim en kötü	3,15E+12	4,42E+09
harika film	1,73E+10	7,55E+08
kırıklığı	4,47E+12	1,13E+12
para	5,12E+11	1,82E+12
değmez	3,22E+12	1,49E+11
hayal kırıklığı	4,37E+12	1,10E+12
çok beğendim	1,52E+11	7,09E+11

[29] Zemberek2 tarafından kökü isim+sıfat+fiil olarak işaretlenen sözcüklerden oluşturulan veri kümesindeki 1-4 Gram aralığındaki tüm öznitelikler ve sınıflandırmaya dair net bilgi kazancı etkileri.

EK 11 – Türkçe En Güçlü 40 Sıfat 1-4 Gram İçin Kutupluk[30]

Fourgram	Olumsuz	Olumlu
kötü	0,022988672	0,006971793
berbat	0,009220944	6,97E+11
harika	0,001617858	0,013397584
çok kötü	0,010224459	0,002151852
çok güzel	0,004009226	0,016519122
saçma	0,010379869	0,002822262
mükemmel	0,001716853	0,011024094
güzel	0,011762288	0,026574971
başarılı	0,003824673	0,014731744
çok iyi	0,005739341	0,016980713
iyi	0,015772928	0,030999731
çok başarılı	0,001395673	0,008274502
muhteşem	0,001464838	0,008159485
çok güzel çok	0,001152694	0,006904141
kötüydü	0,003962984	3,44E+12
kötü çok	0,006938907	0,002323226
güzel çok	0,003556692	0,010922574
basit	0,007541751	0,002860047
kötü kötü	0,004538044	5,42E+11
harika çok	4,54E+12	0,004382888
çok kötüydü	0,002740057	1,55E+12
çok	0,029981878	0,043932678
başarısız	0,003457985	6,12E+11
tamamen	0,006041913	0,002353374
güzeldi	0,002285079	0,0076785
sıradan	0,005951193	0,002520804
başarılı çok	0,001140709	0,005222289
çok saçma	0,003873334	0,001055195
çok güzeldi	0,001081346	0,004993907
iyi çok	0,004697232	0,01099186
çok harika	3,05E+12	0,003029936
çok basit	0,003550996	0,001032694
güzel çok güzel	4,35E+12	0,003334725
berbat çok	0,00211023	2,21E+12
çok çok güzel	7,96E+10	0,003923251
saçmalık	0,002370795	3,24E+11
çok kötü çok	0,002923891	6,89E+11
çok başarılı çok	3,42E+12	0,002872159
mükemmel çok	4,25E+12	0,003094879
iğrenç	0,00298811	6,16E+11

[30] Zemberek2 tarafından kökü sıfat olarak işaretlenen sözcüklerden oluşturulan veri kümesindeki 1-4 Gram aralığındaki tüm öznitelikler ve sınıflandırmaya dair net bilgi kazancı etkileri.

EK 12 – Türkçe En Güçlü 40 Sıfat+Fiil 1-4 Gram İçin Kutupluk[31]

Fourgram	Olumsuz	Olumlu
kötü	0,004603898	0,002336011
berbat	0,002324639	2,44E+12
harika	2,92E+12	0,005777743
güzel	0,002137306	0,00987539
mükemmel	3,07E+12	0,004806123
çok güzel	5,19E+10	0,005453329
saçma	0,002388712	9,23E+11
çok kötü	0,001886224	4,32E+11
sıkıcı	0,002251856	8,89E+11
beğenmedim	0,001663558	3,06E+11
iyi	0,002703739	0,010203709
beğendim	1,54E+12	0,003179525
gereken	4,00E+11	0,003930447
başarılı	7,36E+11	0,005191869
izlenmesi	1,12E+12	0,002748209
muhteşem	2,85E+11	0,003423423
izlediğim kötü	8,79E+10	2,57E+11
çok beğendim	4,98E+10	0,002097375
değmez	8,76E+11	4,59E+11
çok iyi	7,94E+11	0,00478937
izlenmesi gereken	5,36E+11	0,002052372
çok sıkıcı	9,02E+11	9,34E+10
bile	0,003214014	0,003557757
etkileyici	3,12E+10	0,003052097
izlediğim iyi	3,30E+10	0,001780061
basit	0,001607517	9,26E+11
kötüydü	9,02E+10	1,20E+12
izleyin	7,77E+11	0,004369831
çok başarılı	1,76E+12	0,002444294
izlemeyin	0,001017149	2,48E+11
izlenmeli	3,75E+10	0,001713936
gitmeyin	9,19E+11	1,61E+11
boş	0,001416532	7,24E+11
etmiyorum	7,62E+09	8,08E+10
boşa	9,62E+11	2,46E+11
olmamış	0,001102063	3,91E+09
güzel olmuş	4,26E+10	0,00162367
iğrenç	8,27E+10	1,61E+12
sıradan	0,001340415	8,70E+11
harcamayın	4,97E+11	1,06E+11

[31] Zemberek2 tarafından kökü sıfat+fiil olarak işaretlenen sözcüklerden oluşturulan veri kümesindeki 1-4 Gram aralığındaki tüm öznitelikler ve sınıflandırmaya dair net bilgi kazancı etkileri.

EK 13 – Hatalı Kabul Edilen Belirtke Örnekleri

Hatalı Kabul Edilen Belirtke Örnekleri
ounun
huhaha
braz
hralde
yazılarnları
breavheart
somewhere
hunter1979
anlıyosunuzki
frransayı
hatırlıorum
andreuccio
sonuuu
brat
komedei
heelaal
beyenmemde
>>>
değiştiricek
stevenin
gilmi
ayaqta
1900lerin
parmaklaini
böylebir
finney
kayınvalide
kamuran
diebilseydim
hunter1989
2000lerde
everbody
kurtarıyormu
fusst
efemm
eddie
soudtrackları
yarısındada
danzelin
çocuğada

EK 14 – ARFF Dosyası İçerik Örneği

@relation 'Sentiment_beyazperde_islenmis_poz-neg-weka.filters.supervised.instance.Resample-B1.0-S4-Z31.0'

@attribute text string
@attribute @@class@@ {neg,poz}

@data
' yasındayım ilk kez yıl önce izlemiş olsam hala denize girdiğim zaman korkuyla etrafıma bakmama neden oluyor',poz
' arkadaşlar filmi dublajlı izlemeyin filmi tam olarak anlamanız ve orada verilmek istenen duyguları hissetmeniz için altyazıyı tercih edin bırakın ses orjinal kaşlın filmi kesinlikle tavsiye ediyorum ',poz
' çocuk filmi',neg
' ders çıkarılması gereken bir film ama vasatı aşamamış ',poz
' acaba john cena kemerini alabilecekmi',poz
' beklediğimin çok altında en kötü spider man filmi incisinde sinemada izlediğimde koltuğuma çivilenmiştim keza ikincisindede öyle ama üçüncüsü öyle olmadı çok fazla karakter sığdırmaya çalışmışlar',neg
' ilk iki filmde olduğu gibi hareket komedi macera aşk daha ne olsun',poz
' kim bu puanı verdi buna ',neg
' alperin yaşadığı duygusal boşluklara gerilime boşvermişliğe pornografi olarak bakmazsanız filmin konusunu basit olarak değerlendirmeden önce konunun işleniş biçimine dikkat ederseniz her filmin sizin hayatınızı anlatması gerektiğini düşünmeden sadece düş ürünü olduğundan hareketle ancak böyle yaşayan insanların olduğunu da bilerek değerlendirmeler yaparsanız film de kullanılan müziklere ve mekanlara dikkatlice bakıp özenli seçimlerin farkına varırsanız film sizin için oldukça anlamlı olabilir film çıkışı üzerinde düşündükçe daha çok anlaşılan ve böyle bir konunun bu kadar güzel işlenmiş olması dolayısıyla hayranlık uyandıran bir film oldu benim için emeğin önünde saygıyla eğiliyorum',poz

EK 15 – Derlem İçeriği Örnekleri

47-film-10080.txt

1995 ten bu yana bu tür filmler arttı zamanlar başarılı sayılabilecekse de suan emin değilim konusu baya sıradan

61-film-10126.txt

hayatımın çok kötü geçen saat 25 dakikası konusu olmayan amacı olmayan mesaj vermeyen hadi mesajı geçtim hiç bişey vermeyen bir film dansları pek sevmedğim için midir nedir filmdeki dansın da bi olayı yok bence çok daha güzel dansları gördüm filmin sonundaki müzik biraz güzeldi kadar hayatımın iki buçuk saatini ziyan etmek ve bişeyi yaparken çok sıkılmak istiyorum derseniz tam size göre

401-film-105557.txt

yine son zamanlarda yapılan dünyanın sonu veya felaket filmlerinden bi tanesi aksiyon beklntisi olanlar filmi izlemeken zevk almayabilir çünkü aksiyon yok denecek kadar az öylesine izlenecek bi film yani aman amn bi film değil çok sıkılıyo insan kaşıyo biraz ben beğenmedim

EK 16 – Beyazperde Sitesinde Bir Kullanıcıya Ait Yorumun HTML Kaynak Kodu Örneği

```
<div class="box_06 margin_20b j_entity_container"
itemprop="review" itemscope=""
itemtype="http://schema.org/Review">
<meta itemprop="interactionCount" content="UserInteraction:0">
<div class="review_box review_users">
<div class="img_side_content">
<a href="/uye-T20141208122026220711599/" class="xXx ">
<img
src="http://graph.facebook.com/100001843362283/picture?width=60&
amp;height=65" alt="Ayşe B.">
</a>
<div class="content">
<a href="/uye-T20141208122026220711599/" class="xXx tt_18
j_entities" data-
entities="{&quot;entityType&quot;:&quot;User&quot;,&quot;avatar&
quot;:&quot;http://graph.facebook.com/100001843362283/picture&qu
ot;,&quot;entityId&quot;:&quot;T20141208122026220711599&quot;,&q
uot;label&quot;:&quot;Ayşe B.&quot;}" itemprop="author">
Ayşe B.
</a>
<p class="lighten">
0 takipçi
|
<a
href="/community/mitglieder/T20141208122026220711599/reviews/"
class="xXx ">
Onun 1 yorumunu gör
</a>|
<span class="jsu" style="display: inline;">
```

<span class="btn-rounded btn-social e_over btn-small jmy j_w" data-w="relationship" data-relationship="{"label":"Aktivitelerini takip et"}" data-action="hide" id="relationship-11-0-5828308">

Aktivitelerini takip et

<!--/simple_tooltip-->

</p>

   5 - Başyapıt<meta itemprop="bestRating" content="5">
<p itemprop="description" class="margin_10b j_entities" data-entities="{"userId":"T20141208122026220711599","userNickName":"Ayşe B.","ratingDecimal":5.0,"entityId":1397352,"entityType":"Review"}">
hayatım da izlediğim en güzel sinema filmiydi. Murat Boz'un oyunculuğuna bayıldım. Büşra Pekin zaten deha bence herkes izlemeli Hadi İnşallah bol gişeli olsun
</p>

```
<span class="lighten fs11" itemprop="datePublished"
content="2014-12-08">
Eklenme Tarihi 08 Ara 2014, saat 12.24
</span>
<div class="sep_02_10v"></div><!-- sep -->
<span class="jsu" style="display: inline;">
<span class="btns-useful">
<span class="btn-rounded btn-useful-y e_over btn-medium j_w"
data-w="socialAction" data-
socialaction="{&quot;action&quot;:&quot;Helpful&quot;,&quot;disp
lay&quot;:&quot;Count&quot;,&quot;warningMessage&quot;:&quot;Ken
di eleştirini \&quot;faydalı oldu\&quot; şeklinde
işaretleyemezsin.&quot;}" id="socialAction-12-0-5828310">
<i class="icon-useful-y"> </i>
<span class="j_label">0</span>
<span class="jui simple_tooltip  st_arrow_b" data-ui="tooltip"
data-tooltip="{&quot;label&quot;:&quot;Bu eleştiri faydalı
oldu.&quot;}" style="margin-left: -65px; left: 50%;">
<span style="left: 50%;"></span>
<b>Bu eleştiri faydalı oldu.</b>
</span><!--/simple_tooltip-->
</span>
<span class="btn-sep-v"></span>
<span class="btn-rounded btn-useful-n e_over btn-medium j_w"
data-w="socialAction" data-
socialaction="{&quot;action&quot;:&quot;Unhelpful&quot;,&quot;di
splay&quot;:&quot;Count&quot;,&quot;warningMessage&quot;:&quot;K
endi eleştirini \&quot;faydalı olmadı\&quot; şeklinde
işaretleyemezsin.&quot;}" id="socialAction-13-0-5828313">
<i class="icon-useful-n fr"> </i>
<span class="j_label">0</span>
```

```
<span class="jui simple_tooltip  st_arrow_b" data-ui="tooltip"
data-tooltip="{&quot;label&quot;:&quot;Bu eleştiri hiç faydalı
değil.&quot;}" style="margin-left: -75px; left: 50%;">
<span style="left: 50%;"></span>
<b>Bu eleştiri hiç faydalı değil.</b>
</span><!--/simple_tooltip-->
</span>
</span>
<a href="#" class="xXx fr lighten fs11 j_m" data-
right="loginRequired" data-
comp="{&quot;type&quot;:&quot;dialog&quot;,&quot;content&quot;:&
quot;report_abuse&quot;,&quot;settings&quot;:{&quot;query&quot;:
{&quot;entityid&quot;:230875,&quot;entitytype&quot;:&quot;Movie&
quot;,&quot;opinion&quot;:1397352}}}" id="101397352">Kurallara
uymuyor ise tıklayın</a>
</span>
<div class="breaker"></div>
</div>
</div>
<div class="breaker"></div>
</div>
<!-- /img_side_content -->
</div>
<!-- /review_box -->
</div>
```

ÖZGEÇMİŞ

Adı Soyadı: Ertaç PAPRAT
Adres: Şükrüpaşa Mahallesi Ören Kent Sitesi D Blok No:8 Merkez EDİRNE
Telefon: 532 777 24 89
e-posta: epaprat@gmail.com

1982 Yılında Edirne'de doğdu. İlk okulu Edirne Şükrüpaşa İlk Okulu'nda, orta okulu Edirne Yüksel Yeşil İlk Öğretim Okulu'nda okudu. Ege Üniversitesi Fen Fakültesi Fizik Bölümü'nde ve Anadolu Üniversitesi İktisat Fakültesi İktisat Bölümü'nde lisans öğrenimini tamamlamıştır. 2010 yılından bu yana Trakya Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Ana Bilim Dalı'nda tezli yüksek lisans öğrenimine devam etmektedir.

2013 yılında Innotrend Bilişim Sanayi ve Danışmanlık A.Ş.'de danışman olarak çalışmış olup halen yapay öğrenme, veri madenciliği, metin madenciliği konularında çeşitli firmalara danışmanlık yapmaktadır.