T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

N-SEVİYELİ GİZLİ DİRİCHLET AYIRIMI DESTEĞİ İLE TÜR VE DUYGU SINIFLANDIRMA

ZEKERİYA ANIL GÜVEN

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
PROF. DR. BANU DİRİ

İSTANBUL, 2018

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

N-SEVİYELİ GİZLİ DİRİCHLET AYIRIMI DESTEĞİ İLE TÜR VE DUYGU SINIFLANDIRMA

Zekeriya Anıl GÜVEN tarafından hazırlanan tez çalışması 12.06.2018 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Prof. Dr. Banu DİRİ

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Prof. Dr. Banu DİRİ

Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Göksel BİRİCİK

Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Tuğba YILDIZ

İstanbul Bilgi Üniversitesi _____

ÖNSÖZ

Gelişmekte olan teknolojiyle akıllı telefonlar, tabletler ve dizüstü bilgisayarlar ve sınırsız internet hayatımıza hızla girmiştir. Bunların sonucu olarak erişilecek bilgi miktarı da yüksek hızda artmakta ve uzaktan bilgi akışı kısa süre içinde gerçekleşmektedir.

Teknolojiyle beraber, gelişen sosyal medya araçlarıyla insanlara sınırsız ifade özgürlüğü de sunulmuştur. Geçmiş yıllarda bir konu hakkındaki düşünceleri paylaşabilecek platformlar bulunmazken; artık her konuda istediğimiz an duygularımızı paylaşabildiğimiz platformlara sahibiz. Eskiden hızlıca ulaşılamayan, merak edilen bilgilere sosyal medya paylaşımları sayesinde anında ulaşabilmekteyiz. Ayrıca, platformlar insanların "sosyal medya fenomeni" olması için, korkmadan duygu ve düşüncelerini yazabildikleri ortamlardır. Sosyal medya, insanların gerektiğinde bir yerlerde toplanmasına, gülmesine, sinirlenmesine, üzülmesine; yani kısacası tüm duyguların ve olayların yaşanmasına katkı sağlamaktadır. Bu tez çalışması ile sosyal medyanın vazgeçilemez olduğu bir ortamda paylaşımlardaki duyguları ve konularını daha iyi tespit edebilmek için bir yöntem geliştirmek istedim. Yapmış olduğum çalışmayı, inceleyip daha ileri götürecek çalışmaların yapılacağını ümit etmekteyim.

Tezimi yapmamda büyük emeği ve desteği olan, tezimde bana danışmanlık yapan hocam Prof. Dr. Banu Diri'ye, e-postalar ile sürekli benimle bilgi alışverişinde bulunan Tolgahan Çakaloğlu'na, tez aşamasında bana yardımcı olan tüm arkadaşlarıma, kullandığım tivitleri paylaşan Twitter kullanıcılarına ve haberleri paylaşan sitelere teşekkürlerimi sunarım. Ayrıca, her zaman yanımda olan, inancımı asla kaybettirmeyen aileme her şey için teşekkür ederim.

Haziran, 2018

Zekeriya Anıl GÜVEN

İÇİNDEKİLER

Sayfa

KISALTMA LİSTESİ vi

ŞEKİL LİSTESİ vii

ÇİZELGE LİSTESİ viii

ÖZET ix

ABSTRACT xi

BÖLÜM 1 1

GİRİŞ 1

1.1 Literatür Özeti 2
1.2 Tezin Amacı 10
1.3 Hipotez 11

BÖLÜM 2 13

MATERYALLER 13

2.1 Veri Setleri 13
2.1.1 Türkçe Haber Veri Seti 13
2.1.2 İngilizce Haber Veri Seti 14
2.1.3 Türkçe Tivit Veri Seti 14
2.2 Kök Bulma Araçları 14
2.2.1 Zemberek 14
2.2.2 Snowball Stemmer 15
2.2.3 Kelimenin İlk Beş Harfinin Kök Kabul Edilmesi 16
2.2.4 Porter Stemmer 16
2.3 Programlama Dilleri 17
2.3.1 Java 17
2.3.2 Python 17
2.4 Weka 18
2.5 Metotlar ve Algoritmalar 18
2.5.1 Gizli Dirichlet Ayırımı Algoritması 19

2.5.2 Naive Bayes 21
2.5.3 Multinomial Naive Bayes 22
2.5.4 Destek Vektör Makineleri 22
2.5.5 Rastgele Orman Algoritması 24
2.5.6 Multilayer Perceptron 25
2.5.7 k-Katlı Çaprazlı Doğrulama 26

BÖLÜM 3 28

SİSTEM TASARIMI VE UYGULANMASI 28

3.1 Veri Seti Oluşturma 29
3.2 Veri Ön İşleme 30
3.2.1 Türkçe Veri Setinde Ön İşlem 30
3.2.2 İngilizce Veri Setinde Ön İşlem 31
3.3 Sistemin GDA ile Modellenmesi 31
3.4 Uygun Konu Sayısının Bulunması 32
3.4.1 Perplexity Değeri Hesaplanması 32
3.4.2 Coherence Değeri Hesaplanması 33
3.5 Uygun Modelin Seçilmesi 35
3.6 N-seviyeli GDA Yönteminin Geliştirilmesi 38
3.7 Weka'da Sınıflandırma Yapabilmek için Dosya Oluşturma 40
3.8 Sınıflandırma İşlemi 41

BÖLÜM 4 42

DENEYSEL ÇALIŞMALAR 42

4.1 Türkçe Haber Veri Seti için Sonuçlar 42
4.2 İngilizce Haber Veri Seti için Sonuçlar 44
4.3 Türkçe Tivit Veri Seti için Sonuçlar 45

BÖLÜM 5 49

SONUÇ VE ÖNERİLER 49

KAYNAKLAR 51

EK-A 55

KONULARIN GRAFİKLERİ VE ETİKETLERİ 55

A-1 Türkçe Haber Veri Seti 55
A-2 İngilizce Haber Veri Seti 62
A-3 Türkçe Tivit Veri Seti 67

ÖZGEÇMİŞ 73

KISALTMA LİSTESİ

BOW	Bag of Words
DVM	Destek Vektör Makineleri
GDA	Gizli Dirichlet Ayırımı
K-KÇD	K-Katlı Çapraz Doğrulama
K-NN	K-En Yakın Komşu
MP	Multilayer Perceptron
NB	Naive Bayes
MNB	Multinomial Naive Bayes
RF	Random Forest
RO	Rastgele Orman
SVM	Support Vector Machines
UCI-P	Veri Seti Porter Stemmer
VS-5	Veri Seti-5 Kök
VS-S	Veri Seti-Snowball
VS-Z	Veri Seti-Zemberek

ŞEKİL LİSTESİ

Sayfa

Şekil 2. 1 Türkçe tivit veri setinde zemberek ile kök bulma örneği 15
Şekil 2. 2 Türkçe tivit veri setinde snowball ile kök bulma örneği 16
Şekil 2. 3 Türkçe tivit veri setinde ilk beş harf ile kök bulma örneği 16
Şekil 2. 4 İngilizce haberler veri setinde porter stemmer ile kök bulma örneği 17
Şekil 2. 5 Gizli dirichlet ayırımı süreci 19
Şekil 2. 6 Destek vektör makineleri 23
Şekil 2. 7 Multilayer perceptron yapısı 25
Şekil 2. 8 MP'nin bir düğümü; yapay nöron 26
Şekil 2. 9 K-katlı çapraz doğrulama çalışma yapısı 27
Şekil 3. 1 Sistemin tasarımı 28
Şekil 3. 2 Örnek bir aşama olarak ön işleme 30
Şekil 3. 3 Örnek perplexity grafiği 33
Şekil 3. 4 Tutarlık değeri örnek sıralama 35
Şekil 3. 5 Türkçe tivit veri seti için örnek konu gösterimi 35
Şekil 3. 6 pyLDAvis ile çizilien Türkçe tivit veri setinde örnek grafik 37
Şekil 3. 7 N-seviyeli yöntemin işleyişi 38

ÇİZELGE LİSTESİ

Sayfa

Çizelge 3. 1 Türkçe tivit VS-Z verileri üzerinde örnek modelin etiketlenmesi 36
Çizelge 3. 2 Türkçe tivit veri setinde kök bulma araçlarına göre N-GDA için sözlükteki kelime sayısı (5 sınıf için) 39
Çizelge 3. 3 Türkçe tivit VS-Z veri setinde "kork" kelimesinin aşamalardaki ağırlıkları 39
Çizelge 4. 1 Sınıf sayılarında tutarlık değerine göre konu sayıları (VS-Z için) 42
Çizelge 4. 2 Kök bulma araçlarına göre sınıflarda GDA'nın başarısı 43
Çizelge 4. 3 Geliştirilen 2-GDA yöntemi sonucunda sistemin başarısı 43
Çizelge 4. 4 2-GDA ile oluşturulan dosyanın sınıflandırma algoritmalarında başarısı 44
Çizelge 4. 5 Sınıf ve bulunan konu sayıları göre GDA'nın başarısı 44
Çizelge 4. 6 GDA ve 2-GDA başarısının karşılaştırılması 45
Çizelge 4. 7 2-GDA ile oluşturulan dosyanın sınıflandırmadaki başarısı 45
Çizelge 4. 8 Sınıf sayısına göre elde edilen konu sayıları (VS-Z için) 46
Çizelge 4. 9 Kullanılan kök bulma araçlarına göre GDA'nın başarısı 46
Çizelge 4. 10 Geliştirilen 2-GDA yönteminin başarısı 47
Çizelge 4. 11 Oluşturulan modellerdeki başarı oranları 47
Çizelge 4. 12 3-GDA'dan elde edilen dosyanın sınıflandırmadaki başarısı 48

ÖZET

N-SEVİYELİ GİZLİ DİRİCHLET AYIRIMI DESTEĞİ İLE TÜR VE DUYGU SINIFLANDIRMA

Zekeriya Anıl GÜVEN

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Prof. Dr. Banu DİRİ

Haber başlıklarının türü ve sosyal medyada yapılan paylaşımların duygu durumlarına göre sınıflandırılması gelişen teknoloji ile beraber medya sektöründe kullanım açısından büyük önem taşımaktadır. Bir haberin hangi tür olduğunu anlamanın yanı sıra, kişinin çok ziyaret ettiği haber türü bulunabilmekte ve o kişiye özgü ilgi çekebilecek reklamlar gösterilebilmektedir. Ayrıca, haber ajansları için haberlerin otomatik olarak sınıflandırılması önemlidir. Sosyal medya artık iletişim için kullanılmanın ötesinde birçok alanda etkili hale gelmiştir. Kullanıcılar Facebook, Twitter, Blog gibi sosyal medya araçlarında bir olayla alakalı duygusunu, düşüncesini ve deneyimlerini paylaşabilmektedir. Ayrıca, bu araçlar haber paylaşmak ve organizasyon düzenlemek için de kullanılmaktadır. Sosyal medyada yapılan paylaşımlar ile kişi hakkında bilgi de edinilebilmektedir. Yapılan paylaşımlardaki duygulardan yola çıkarak kişinin ruh hali tahmin edilmektedir. Böylece kişiye özgü sayfalar önerilebilmektedir.

Çalışmada haberlerin türlerini ve Twitter'dan paylaşılan tivitlerin hangi duyguya sahip olduğunu tespit etmek amaçlanmıştır. Yöntem olarak konu modelleme algoritması Gizli Dirichlet Ayırımı (GDA), N seviyeli bir yapıda geliştirilerek kullanılmıştır. Haberler için oluşturulan veri seti Milliyet, Mynet gibi sitelerden yararlanılarak oluşturulmuştur. Tivitlerin duygu tespitinde kullanılan veri seti de Türkçe tivitlerden oluşturulmuştur. Haber veri seti en fazla 7 sınıflı iken; tivit veri seti kızgın, korku, mutlu, üzgün ve şaşkın duygu türü olmak üzere 5 sınıftan oluşmuştur. Sistemi modellerken kelimelerin kökleri alınmıştır. Köklerin çıkarılması işleminde Zemberek, Snowball ve kelimenin ilk 5

karakterini kök alan yöntemler kullanılmıştır. Haber ve Tivit veri setleri için önce klasik GDA yöntemi ile haberler için konu, tivitler için duygu ataması yapılmış ve sonrasında gerçek etiket değerleri ile karşılaştırarak bir başarı hesaplanmıştır. Klasik GDA yöntemini referans alarak aşamalı olarak değiştirilen GDA yöntemi ile tekrardan konu ve duygu belirleme işlemi yapılarak başarının arttığı gözlemlenmiştir. N-seviyeli GDA yöntemi kullanılarak her haber ve tivit için çıkarılmış olan özellikler kullanılarak Naive Bayes, Multinomial Naive Bayes, Destek Vektör Makineleri, Rastgele Orman ve Çok Katmanlı Algılayıcı gibi makine öğrenmesi yöntemleri kullanılarak sınıflandırıcılar ile sistemin başarısı ölçülmüştür.

Anahtar Kelimeler: Konu modelleme, Duygu Analizi, Haber Analizi, Gizli Dirichlet Ayırımı, Sosyal Medya, Doğal Dil İşleme

ABSTRACT

GENRE AND EMOTION CLASSIFICATION BY SUPPORT OF N-STAGE LATENT DIRICHLET ALLOCATION

Zekeriya Anıl GÜVEN

Department of Computer Engineering

MSc. Thesis

Adviser: Prof. Dr. Banu DİRİ

Classification of news headlines and social media based emotions are of great importance in terms of their use in the media sector, along with developing technology. In addition to understanding what kind of news is, you can find the type of news that the person has visited and can show ads that can attract that person's interest. It is also important to automatically classify news for news agencies. Social media has become more effective in many areas beyond communication. Users can share their feelings, thoughts and experiences with an event in social media tools like Facebook, Twitter, Blog. These tools are also used to share news and organize events. Information about the person can also be obtained through sharing in the social media. The mood of the person is estimated by going out of the shared feelings. Thus, personalized pages can be offered.

The purpose of the study was to determine the types of news and the feelings of the tweets shared on Twitter. As a method, the subject modeling algorithm has been developed and used in a N-stage structure of the Latent Dirichlet Allocation (LDA). The data set created for the news was created by using sites like Milliyet, Mynet. The dataset used in the detection of the feelings of tweets was also created from Turkish tweets. While the news dataset has a maximum of 7 classifications; the dataset consists of five classes: angry, fear, happy, sad and confused. When modeling the system, the words are rooted. In the process of removing the roots, Zemberek, Snowball and the first 5 characters of words are used. For the news and tweets data sets, the classical LDA

method was used first for the news, the emotion assignment for the tweets, and then a success by comparing with the actual label values. It has been observed that success has been achieved by using the GDA method, which was gradually changed with reference to the classical GDA, by performing the subject and feeling determination process from the other. Using the N-stage GDA method, the performance of the system with classifiers was measured using machine learning methods such as Naive Bayes, Multinomial Naive Bayes, Support Vector Machines, Random Forest and Multilayer Perceptron, using features extracted for each news and tweets.

Keywords: Topic modeling, Emotion analysis, News analysis, Latent dirichlet allocation, Social media, Natural language processing

BÖLÜM 1

GİRİŞ

Günümüzde internet kullanımının artmasıyla yaşamımıza yeni alanlar ve meslekler girmeye başlamıştır. Yaşamımıza giren en önemli alanlardan biri sosyal medya olmuştur. Sosyal medya, herkesin birbirine doğrudan ulaşmasını sağlamaktadır. Ayrıca, sosyal medyayı kullanan herkes fikirlerini istediği gibi paylaşabilmektedir. Sosyal medya ile beraber kullanıcılar arasında bilgi alışverişi kısa sürede sağlanmakta ve merak edilen bilgiye de anında ulaşılabilmektedir. Bunlar sonucunda sosyal medya alanı hızla büyüme göstermiştir. Büyüme ile beraber yenilikler hayatımıza girse de bazı problemler de ortaya çıkmıştır. Sorunlara örnek olarak kullanıcıların zamanlarının çoğunu internet başında harcaması, sosyal medya ile insanların kolayca manipüle edilmesi ve sosyal yaşamımızın zayıflaması verilebilir. Ancak sorunların en önemlisi verilerin saklanması ve korunması olmuştur. Büyük boyutlu verilerin saklanamama sorunu yeni teknolojilerin ve çözümlerin oluşmasına olanak sağlamıştır. Böylece Büyük Veri, Bulut Mimarisi ve Siber Güvenlik kavramları yaşamımıza girmiştir. Verinin büyük olması nedeniyle dağıtık şekilde saklanmasını Bulut Mimari sağlamaktadır. Bulut Mimari ile sunucular arasında yatay ve dikey olarak hızlı şekilde veri geçişi gerçekleşmektedir. Verilerin korunması için ise Siber Güvenlik alanı önlemler almıştır. Bu alanlar sayesinde artık veriler saklanabilmekte ve korunabilmektedir.

Sosyal medyayı her alandaki kurumların, kişilerin etkili bir şekilde kullanmasından dolayı bir ürün hakkında kullanıcıların yorumları da önemli hale gelmektedir. Böylece diğer kullanıcılarda bu ürünle ilgili yorumları görebilmekte ve ön bilgiye sahip olmaktadır. İlgili kurumlar ise kısa sürede geri bildirim sağlayabilmektedir. Ayrıca, sosyal medya kullanıcılarının paylaşımlarından nasıl bir duyguya sahip olduğu da ortaya

çıkarılmaktadır. İnsanın sahip olduğu duyguya göre karşısına ilgisini çekebilecek reklamlar ve öneriler çıkarılmaktadır. Ek olarak ruh hali kötü olan bir kişinin tespit edilip, ilgili yerlere haber sağlanması açısından duygu çıkarımında bulunabilmek önemlidir. Her türlü paylaşımlar sonucunda verilerin analiz edilmesi gereksinimi kaçınılmaz hale gelmektedir. Böylece verileri inceleyen kişiler veya yerler sosyal medya uzmanlarıyla çalışmak durumundadır.

Sosyal medya platformlarından birisi olan Twitter 2006 yılında kurulmuş bir mikro blog sitesidir. Kullanıcıların duygusunu, düşüncesini ve paylaşımlarını 280 karakterle anlatmasına izin vermektedir. Her platform gibi her gün kullanımı artan Twitter yeni reklam ve pazarlama platformu olarak da görülmektedir. Kullanıcıların özgürce fikirlerini söyleyebildiği Twitter'daki kullanıcı sayısı 2017 yılı itibariyle yaklaşık 328 milyona ulaşmıştır. Günlük atılan tivit sayısı ise yaklaşık olarak 7 milyonu bulmaktadır. Bu tezde duygu analizi yapabilmek için Gizli Dirichlet Ayırımı yönteminden yararlanılmış ve veri seti olarak Twitter verileri ve haber başlıkları içeren veri seti kullanılmıştır.

1.1 Literatür Özeti

Hasan vd. [1], Twitter mesajlarının ifade ettikleri duyguları bulmak için Emotex isimli yeni bir yaklaşım önermişlerdir. Duygusal durumları Circumplex modeli ile elde etmiş ve Twitter mesajlarının etiketlerini Twitter karma etiketini kullanılarak otomatik bir şekilde etiketlemişlerdir. Twitter mesajlarındaki duyguları sınıflandırmak için Destek Vektör Makineleri (DVM), K-En Yakın Komşu (k-NN), Karar Ağacı ve Naive Bayes (NB) gibi makine öğrenmesi algoritmalarını kullanmışlardır. Önerilen Emotex yaklaşımı, çok sayıda kısa metni % 90 doğrulukla sınıflandırmıştır.

Ekinci vd. [2], yaptıkları çalışmada özellik tabanlı duygu analizi kapsamında bir otel ile ilgili Türkçe kullanıcı yorumları içerisinden gizli olarak bulunan ürün özelliklerini çıkarmak için popüler konu modelleme yöntemlerinden biri olan Gizli Dirichlet Ayırımı (GDA) algoritmasını kullanmışlardır. Otel ile ilgili Türkçe kullanıcı yorumlarından ürün özelliklerini konu modelleme yöntemi ile belirlemişlerdir. Yapılan çalışmada duygu analizi ve konu modelleme yöntemleri birleştirilerek başarılı bir özellik çıkarımı yapılmıştır.

Mohammad ve Kirtchenko [3], duygu-kelime hashtaglarını kullanarak Hashtag Emotion Corpus veri setini ve bu veri setiyle duygu terimleri sözlüğü olan Hashtag Emotion Lexicon'ı oluşturmuşlardır. Hashtag Emotion Lexicon sözlüğünü kullanarak metinlerden kişilik çözümleme ile ilgili çalışmalar yapmışlardır. Heyecan, suçluluk duyma, özlem ve hayranlık gibi duyguların karşılığı olan kelimelerin kişilik algılanmasında önemli özellikler olduğunu tespit etmişlerdir.

Chaffar ve Inkpen [4], haber başlıklarını, masalları ve blog yazılarını birleştirerek oluşturdukları heterojen veri kümesiyle altı farklı duyguyu (sinirli, iğrenç, korku, mutlu, üzgün ve şaşkın) tespit edebilmek için denetimli bir makine öğrenmesi yaklaşımı önermişlerdir. Bu veri setlerini oluştururken kelime torbası (Bag of Word, BOW) ve n-gram gibi çeşitli özellikleri kullanmışlardır. DVM sınıflandırıcısının diğer sınıflandırma algoritmalarından daha başarılı olduğunu göstermişlerdir.

Bollen vd. [5], İngilizce tivitleri kullanarak bir duygu analizi gerçekleştirmişlerdir. Tivit içeriklerinden duyguları (gergin, bunalmış, öfkeli, enerjik, yorgun, karışık) bulmak için bir zaman çizelgesi kullanarak her gün için altı boyutlu bir ruh hali vektörü hesaplamışlardır. Medyadan ve kaynaklardan toplanan popüler olaylar ile sistemlerini test etmişlerdir. Toplumsal, siyasal, kültürel ve ekonomik alanlardaki olayların, toplumun çeşitli boyutları üzerinde önemli, doğrudan ve son derece belirgin bir etkiye sahip olduklarını tespit etmişlerdir.

Pak ve Paroubek [6], duygu sınıflandırması yapmak için kullanılacak olan veri setinin otomatik bir şekilde toplanması için bir yöntem geliştirmişlerdir. Duygu sınıflandırmasında, duygusal metinlerin güçlü göstergelerinden biri olan Sözcük Türü (Part of Speech) etiketlemeyi yapabilmek için ağaç etiketleyicisini kullanılarak pozitif, negatif ve nötr durumlar arasındaki farkları gözlemlemişlerdir. Bu gözlemler ile yazarların duygularını tanımlamak için söz dizimsel yapılar kullandıklarını raporlamışlardır.

Colace vd. [7], Gizli Dirichlet Ayırımı yaklaşımını belgelerin sınıflandırılmasında kullanarak elde edilen terimlerin grafiği ile ağırlık-kelime çiftlerine dayalı yeni bir yaklaşım önermişlerdir. Bir dokümanın olumlu ya da olumsuz olduğunu bulmak için standart ve gerçek veri setleri (gerçek zamanlı tivitler) kullanmışlardır. Her iki durumda

da, elde edilen sonuçların diğer yaklaşımlarla elde edilen sonuçlardan daha iyi olduğunu göstermişlerdir.

Çoban ve Özyer [8], Türkçe tivitlerden oluşan veri setini metin sınıflandırma yöntemleri ile analiz ederek olumlu veya olumsuz olup olmadığını incelemişlerdir. Deneysel sonuçlar DVM, NB, Multinomimal Naive Bayes (MNB) ve k-NN algoritmalarıyla test edilmiştir. Vektör Uzay model ile temsil edilen öznitelikleri, BoW ve n-Gram modeli olmak üzere iki farklı şekilde oluşturmuşlar ve sınıflandırma sonuçlarına olan etkisini göstermişlerdir.

Onan [9], pozitif ve negatif yorumlar içeren Türkçe tivitlere GDA modelini uygulayarak makine öğrenmesi sınıflandırıcıları ile duygu analizi gerçekleştirmiştir. Beş farklı makine öğrenmesi algoritması (NB, DVM, k-NN, Rastgele Orman ve Lojistik Regresyon) ve 4 farklı konu sayısı (k=50, k=100, k=150, k=200) kullanmıştır. Deneysel çalışmalarda en yüksek başarı konu sayısı 50 olduğunda %78.34 ile Naive Bayes algoritması vermiştir. Gizli Dirichlet Ayırımı modelinin Türkçe bir metin belgesinin sınıflandırılmasında da uygun bir yöntem olabileceğini önermiştir.

Çoban ve Özyer [10], yaptıkları çalışma ile mesaj içeriklerini otomatik olarak analiz ederek; pozitif, nötr ve negatif olduğunu tespit etmek için bir sistem önermişlerdir. Duygu analizinde önemli bir problem olan duygu sınıflandırmasını konu modelleme ile gerçekleştirmişlerdir ve bu modellemenin sonuçlar üzerindeki etkilerini incelemişlerdir. GDA algoritmasını kullanarak konu modeli oluşturmuşlardır. Önerdikleri sistemi farklı öznitelik çıkarım modelleri ve sınıflandırıcılar kullanarak test etmişlerdir. Konu modelinin daha önce kullanılan yöntemlerden daha başarılı olduğunu bulmuşlardır. Sonuç olarak, geliştirdikleri otomatik analiz yapan yöntem ile duygu sınıflandırmasında başarının %26 oranında arttığı tespit edilmiştir.

Bolelli vd. [11], dokümanları zamansal ve yinelemeli bir şekilde sıralayabilmek için GDA'ya dayalı bir model önermişlerdir. Dokümanların her bir bölümünde keşfedilen konuların daha sonraki süreçlerde konu bulma faaliyetlerine girebilmesi için konular zaman dilimlerine ayrılmıştır. Bilimsel yayınların zaman sıralamasını kullanarak dokümanlar için yazar-konu ve konu-kelime sistemini öğrenen bir doküman modeli olan S-ATM'yi önermişlerdir. S-ATM modeli, CiteSeer veri deposundan alınan bir örnek veri

kümesine uygulanması ile bilimsel konuların ve en çok ilgi çeken yazarların zaman içerisinde gösterdikleri gelişmeleri etkin bir şekilde keşfedilebileceği gösterilmiştir.

Beyhan [12], verilerin çekilmesi ve çekilen verilere duygu analizi yapabilmek için Botego firmasını kullanmıştır. Botego firmasının desteği ile belirlenen tarih aralıklarında GSM sektöründeki üç firmanın Twitter'da adının geçtiği tüm tivitler alınmıştır. Alınan tivitler farklı modeller kullanılarak kümelenmiştir. Veriler anlamlı ve değerlendirilebilir olduğunda kümeleme işlemi sonlandırılıp elde edilen sonuçlar ile duygu analizi sonuçları birleştirilmiştir. Firmalara ait paylaşımların duygu analizleri ve modelleri uygulamanın sonucunu oluşturmuştur. Uygulama sonucu firmalara ait satış verileri, reklam ve pazarlama stratejileri ile birlikte değerlendirilmiştir.

Evirgen [13], duyguların kısa ve anlamlı bir şekilde ifade edildiği ve kelimelerin doğrudan işlenebilecek formatta olmayan Türkçe tivitleri, R programlama dili kullanarak işlenebilmesi için bir sistem önermiştir. Türkçe ile ilgili yapılan çalışmalar analiz edilerek veri seti hazırlanmıştır. Makine öğrenmesi algoritmalarından DVM, Rastgele Orman (RO) Karar Ağaçları, Boosting, Maksimum Entropi, Yapay Sinir Ağları ile geliştirilen sistem karşılaştırılmıştır.

Al-Bndi [14], Twitter veya web sitelerinde bulunan mikro bloglama fonksiyonu ile bu sitelerden duyguları çıkarabilmek ve duygu analizi yapabilmek için bir sistem geliştirmiştir. Çalışmanın ilk amacı tivitlerdeki duyguları, kelime bulma veya kelime frekansı özellikleri ile unigram öznitelik çıkarım faktörünü kullanarak duyguların sınıflandırılmasını yapmaktır. Kelime bulma özelliği ile yüksek bir doğruluk oranına ulaşmışlardır. Çalışmanın ikinci amacı da tivitlerdeki duyguları n-gram ($1<n<4$) öznitelik çıkarımını kullanıp, çıkarmak ve sınıflandırmaktır. Eğitim setinde bulunan tivit sayıları arttırıldığında ve eşit sayıda pozitif, negatif tivitler kullanıldığında yapılan testler sonucunda sistem yüksek bir doğruluk oranı ile çalışmıştır. DVM yöntemlerinden olan Sıralı Minimal Optimizasyon ve MNB yöntemi ile birlikte kullanılmıştır. Tivitlerin duygularını sınıflandırma da kullanılan iki makine öğrenmesi algoritmalarının yüksek başarı oranına sahip olduğu görülmüştür.

Giriş [15], ürünler ile ilgili müşterilerin yorumlarını ve kullandıkları dil özelliklerini dikkate alarak yorumların polaritesini tespit etmiştir. Yorumlar negatif ve pozitif olarak iki farklı

web sitesi kullanarak toplanmıştır. Çalışmanın asıl amacı, müşterilerin ürün yorumlarından insanların tutumları ve düşünceleri ile ilgili özellikleri çıkartmaktır. Bu çalışmada 3 farklı sınıflandırma yöntemi ve 3 farklı makine öğrenme algoritması kullanılmıştır. Sınıflandırma yöntemleri olarak İkili Sınıflandırma, Frekans Sınıflandırması ve Ağırlıklı Frekans Sınıflandırması; makine öğrenme algoritmaları olarak da BayesNet, VotedPerceptron ve J48 karar ağacı kullanmıştır. Çalışma sonunda en iyi sonuçlar ikili sınıflandırma kullanılarak BayesNet algoritmasında elde edilmiştir.

Çoban [16], Türkçe tivitlerden iki sınıflı duygu analizi yapmıştır. Duygu analizini metin sınıflandırma gibi düşünerek duygu analizi teknikleri ile metin sınıflandırma tekniklerini birlikte kullanmıştır. Tivitlerdeki etkin duyguları otomatik bir şekilde bulabilmek için makine öğrenme algoritmaları kullanılmıştır. Çalışmanın asıl amacı duygu analizi başarı oranını arttırmaktır. Bundan dolayı Türkçe tivitlerin duygu analizi için farklı ön işleme, etiketleme, sınıflandırma ve benzerlik yöntemlerini kullanarak bunların etkisi incelenmiştir. Konu etiketleme yöntemi önerilerek %92,50 oranında başarıya ulaşılmıştır. Önceki çalışmalara göre başarının daha yüksek olduğu gözlemlenmiştir. Metin sınıflandırmayı ve duygu analizi süreçlerini yapmak için Türkçe ve İngilizce metinleri işleyebilen bir yazılım prototipi geliştirmiştir.

Kama [17], forumlardaki yorum ve incelemelerle oluşturmuş olduğu Türkçe veri setine İngilizce için yapılan işlemleri uygulamıştır. Huawei hakkındaki donanimhaber.com üzerinden toplanan yorumlar ile veri seti oluşturulmuştur. Yorumlar hakkında öznitelik çıkarımı için frekans tabanlı ve duygusal kelime destekli görüş çıkarımı kullanılmıştır. Türkçe için bir özellik tabanlı duygu analizi yöntemi geliştirmiştir.

Demirci [18], mikro-bloglardaki duyguların otomatik bir şekilde analizini gerçekleştirebilmek için bir çalışma yapmıştır. Mikro-blog metinlerindeki duyguyu önemli ölçüde etkileyen özel ifadeler, semboller ve bazı kolaylıklar vardır. Mikro-bloglar metin madenciliğindeki metinler gibi işlenebilecek formata sahip olmadığından duygu analizde kullanmak için Türkçe tivitlerden yeni bir veri seti oluşturmuştur ve genel bir yapı önermiştir. Sınıflandırma için sırasıyla NB, DVM ve k-NN kullanılmıştır.

Boynukalın [19], Türkçede ulaşılabilecek veri seti olmadığından iki farklı kaynak ile tivitlerden oluşan bir veri seti oluşturmuştur. Veri setini mutlu, üzgün, korku ve sinirli

olmak üzere dört duygu ile etiketlemiştir. Türkçenin İngilizceden farklı olan özelliklerinden dolayı kullanılan yöntemlere çeşitli eklemeler yaparak Türkçe metinlerin duygu analizi başarısını arttırmıştır. DVM, NB ve Tümleyen NB makine öğrenme yöntemlerini kullanarak sonuçları karşılaştırmıştır.

Özgirgin [20], sosyal medya verilerine duygu analizi yaparak sınıflandırmayı ve bu sınıflandırma işleminin gerçek nedeni ile niteliklerini belirlemeyi amaçlamıştır. Veriler telefon operatörleri ile alakalı yorumlar toplanarak olumlu ve olumsuz olarak iki gruba ayrılmıştır. Duygu analizinde veriler ön işlemeden geçirilip, metinler düzeltildikten sonra makine öğrenmesi yöntemlerini kullanarak analizleri yapılmıştır. Niteliklerin belirlenmesinde kelimelerin yapısal olarak incelenmeleri, terim varlığı ve frekans analizleri ile çıkarımlar elde edilmiştir. Sınıflandırıcı olarak DVM, Lojistik Regresyon, Çok Katmanlı Yapay Sinir Ağları kullanmıştır ve sistemin başarısını incelemiştir.

Agarwall vd. [21], Twitter verileri üzerinde duygu analizi gerçekleştirmişlerdir. Geliştirdikleri sistem, sözcük türü özel ön polarite özelliklerini tanıtmada ve çekirdek kullanımını yaygınlaştırmada katkı sağlamıştır. Duygu analizi yönteminde daha önceden de önerilen unigram modeli temel alınarak yapılan sınıflandırma işleminde %4'lük bir başarı artışı sağlanmıştır. Başlangıçta heyecan duygusu için yapılan analizin diğer duygular için yapılan analizden çok farklı olmadığı sonucuna ulaşılmıştır.

Strapparava ve Mihalcea [22], geliştirdikleri "Affective Text" yapısı ile duyguları ve değerli bilgileri haber başlıklarını sınıflandırmada kullanmıştır. Geliştirdikleri yapı duygular ile sözcüklerin anlamları arasındaki bağlantıları bulmada kullanılmıştır.

Wang vd. [23], duyguları tanıma konusunda yapılan çalışmalar ve tivitlerdeki duygu ile ilgili hashtagları kullanarak duygusal durumlar ile ilgili kapsamlı bir duygu etiketli veri seti oluşturmuşlardır. Duyguları otomatik olarak tanımlayabilmek için 7 duygu kategorisine ait yaklaşık 2 milyon tivit toplanmıştır. Duygu analizi için iki farklı makine öğrenmesi tekniği uygulamışlardır. Unigram, bigram, duygu ifadeleri ve konuşma bölümlerinin duyguların toplanmasında etkili olduğunu gözlemlemişlerdir. Yaklaşık 2 milyon tivit içeren eğitim verisi kullanarak en yüksek doğruluk oranına yaklaşık %65 ile ulaşılmıştır.

Liu vd. [24], yapmış oldukları çalışma da beş farklı dışsal faktörü inceleyerek, yapılan analizler ile kullanıcıların duygusal durumlarını incelemişlerdir. Kullanıcıların geçmişte

paylaştıkları tivitlere bakarak hava, haber maruziyeti, sosyal medyaya katılım zamanı, sosyal ağ duyarlılığı ücreti ve duygu yatkınlığı gibi beş özellik türünü araştırmışlardır. Bu çalışma ile günlük yaşantımızın bazı yönlerinin duyguları temellendirmede kullanıldığını ve belirli duygusal tepkileri etkileyen unsurları tespit etmeye çalışmışlardır. Kullanıcıların duygusal durumlarını %67 doğruluk oranı ile tahmin etmişlerdir. Bireysel özellikler incelendiğinde, kullanıcının önceki metin içeriğinin on iki saat sonra yaşamış olduğu duygusal tepki ile arasında yüksek bir korelasyon ilişkisi olduğunu ve kullanıcıların duygusal durumlarını analiz etmede tutarlı olduğu görülmüştür.

Roberts vd. [25], son zamanlarda duygu içerikli metinlere erişimin artış gösterdiği mikro-bloglamada duygu analizi çalışmışlardır. Mikro-blog veya Twitter ile 14 konudan özetlenen yedi duyguyu (öfkeli, nefret, korku, neşeli, sevgi, üzgün ve sürpriz) içeren bir yapı sunmuşlardır. Oluşturdukları yapıyı birkaç duygu seti ile karşılaştırarak konuların duygusal yapılarının analizini, her bir konunun dilsel tarzını ve yapıda kullanılan her bir duyguyu karakterize etmişlerdir. Yapmış oldukları bu analiz, yeni denetlenen ve denetlenmeyen duygularını belirlemede yeni tekniklerinin tasarımına yol açmıştır.

Çelikyılmaz vd. [26], GDA modelinin soru cevaplama sistemine uygulanmasını incelemiştir. Kullanıcının sorduğu soru ile aday cevaplar arasındaki benzerlik ölçütlerinin bulunup sıralanması GDA ile yapılmıştır. Sınırlı miktarda bilgilerle, soru cevaplarla veri kümelerindeki konuyla ilgili yapıların yeni gösterimiyle, soru cevap sıralama sisteminin performansı üzerinde iyileştirmeler göstermiştir.

Li ve Zhang [27], metin sınıflandırma için GDA'yı kullanmışlardır. Bu çalışmada bir dokümandaki kelime sıklığına (tf), kelimenin birden fazla dokümanda geçme sıklığına (idf) göre sözlüğe eklenmesini kontrol eden kelime özellik modelleriyle GDA'nın karşılaştırması yapılmıştır.

Çelikyılmaz vd. [28], konuşma anlama üzerine semantik bir işlem uygulamışlar ve konuşma anlama sisteminde semantik yapıyı öğrenmek için gizli n-gram kümeleme ve yarı denetlenmiş GDA kullanmışlardır. Ürün özelliklerinin çıkarılmasında da konu modelleri kullanılmaktadır. Gizli anlamsal modelleme, sözdizimsel ayrıştırma veya konu izleme gibi birçok doğal dil işleme görevini iyileştirmek için araştırılmıştır. Anlam bilim temellerinden edinilmiş olan ön bilgiyle GDA'nın genişletilmesini önermişlerdir.

Geliştirilen GDA yöntemiyle elde edilen konu semantik yapı için öğrenme modeline ek kısıtlama getirmişlerdir. En yüksek anlamsal başarı artışı Şartlı Rastgele Alan'da (CRF) görülmüştür.

Titov vd. [29], geliştirdikleri Çok Parçalı GDA ile çıkarılan yerel konuları oylanan özellikleri, global konuları ise ürün özelliklerini çıkartmak için kullanmışlardır. Çok parçalı konuları başlatmak için GDA ve Olasılıksal Gizli Anlamsal Analiz gibi standart konu modelleme yöntemlerine dayanarak Çok Parçalı GDA modeli geliştirmişlerdir. Çok parçalı modellerin çalışma için daha uygun olduğunu iddia etmişlerdir. Geliştirdikleri model ile sadece kabul edilebilir nitelikler çıkarmakla kalmamışlar, ayrıca bunları tutarlı konulara kümelemişlerdir.

Lin vd. [30], metinden aynı anda hem duyguyu hem de konuyu tespit eden Ortak Duygu / Konuşma Modeli (JST) olarak adlandırılan GDA temelli yeni bir olasılıkçı modelleme çerçevesi önermişlerdir. Önerilen JST modeli tamamen denetlenmemektedir. Model ile sinema yorumlarından ürün özelliklerini ve duygu ifadelerini eş zamanlı olarak çıkarmışlardır. JST tarafından elde edilen sonuçlar olumlu olarak gösterilmiştir.

Lee vd. [31], tüketici yorumlarından otomatik olarak algılama haritaları ve radar çizelgeleri oluşturmak için Mining Conceptual Map adlı bir yöntem önermişlerdir. Sanal dokümanlar oluşturularak, ağırlıklı GDA ile ürünlerin özellikleri çıkarılmıştır. Algılanan haritalar ve radar çizelgeleri çok sayıda tüketici incelemesinden elde edildiğinden, önerilen yöntem öznel kişisel eğilimleri azaltabileceği öngörülmüştür. Akıllı telefonların tüketici yorumlarından elde edilen analiz sonuçları, önerilen yöntemin akıllı telefon şirketleri için pratik bilgiler sunabileceğini göstermiştir. Ayrıca, şirketlerin yeni ürünleri konumlandırmalarına ve etkili pazarlama ve rekabetçi stratejileri formüle etmelerine yardımcı olabileceği öne sürülmüştür.

Chatterjee vd. [32], esas olarak Twitter verilerine odaklanan konu ve görüş sınıflandırma yöntemlerinden faydalanarak bilgilerin güvenilirliğini değerlendirmek için bir yöntem önermişlerdir. Geliştirilmiş GDA algoritması tivit konularının sınıflandırılması için; yarı eğiticili olarak kullanılan DVM ise duygu analizi için kullanılmıştır. Bir konudaki karşı görüşlerin sayısını karşılaştırarak çoğunluk kararı güvenilirlik analizi için uygulanmıştır.

Bununla birlikte Twitter API'si alınan postaların dilini belirlemeye izin verdiği için geliştirilen yöntemin diğer dillere kolayca adapte edilebileceği öngörülmüştür.

Poria vd. [33], GDA'ya kavramlar arası ilişki bilgisini de dahil ederek sözdizimsel bir GDA'dan anlamsal bir GDA'ya geçiş önermişlerdir. Bunun sonucunda istatistiksel bir yöntem yerine kelimeler arası anlamsal ilişkiden yararlanarak başarılı bir kümeleme işlemi yapmışlardır.

Feuerriegel vd. [34], GDA'yı finans haberlerindeki konuları çıkartarak bu konuların Alman borsasını nasıl etkilediğini belirlemek için kullanmışlardır. Amaca özel konular GDA kullanılarak belirlenmiş olup etkili bir şekilde, 40 konu çıkarmayı başarmışlardır. Konu gruplarının ise etkisinin birbirinden büyük ölçüde farklılık gösterdiği belirlenmiştir.

Onan vd. [35], yapmış oldukları çalışmada GDA'nın metin duygularını sınıflandırma üzerindeki etkisini incelemişlerdir. Çalışmada NB, DVM, Lojistik Regresyon, Radyal Temel İşlevi Ağı ve k-NN olmak üzere beş farklı makine öğrenme algoritması; ayrıca Bagging, AdaBoost, Rastgele Altuzay, Oylama ve İstifleme gibi beş farklı yöntem kullanılarak dört adet duygu veri setiyle değerlendirmede bulunmuşlardır. Sonuçta topluluk öğreniminde kullanılan veri setlerindeki sınıflandırma algoritmalarının performanslarında artış gözlemlenmiştir.

Literatürler incelendiğinde her iki dildeki çalışmalarda algoritma üzerine yeni yöntemler geliştirmişlerdir. Geliştirilen yöntemlere göre makine öğrenme yöntemlerinin başarılarını hesaplamışlardır. Literatürde Türkçe ile alakalı çalışmalar çok fazla bulunmamaktadır. Olan çalışmalar da incelendiğinde ise genelde veri seti olumlu, olumsuz ve nötrden oluşmaktadır. Yapılan Türkçe çalışmalarda veri setlerine ulaşım olmadığından herhangi bir karşılaştırma imkanı olmamaktadır. İngilizce de ise konu modelleme ile ilgili çalışmalar daha fazla bulunmuştur. Veri setlerinde sınıf sayısı dört ile yedi arası değişmektedir. İngilizce de veri setlerinin bazılarına ulaşım sağlanmaktadır. Her iki dilde ki çalışmalar incelendiğinde önerdikleri yöntemler başarıyı arttırmıştır.

1.2 Tezin Amacı

Gerek haber başlıklarının türü gerekse Twitter'da yapılan paylaşımların duygu durumlarının sınıflandırılması gelişen teknoloji ile beraber medya gibi ilgili alanlarda

kullanım açısından büyük önem taşımaktadır. Twitter, kullanıcıların ve her gün atılan tivitlerin çokluğu ile verileri anlamlı olan bir sosyal medya platformudur. Bu sebeple kullanıcıların fikirlerini paylaştığı tivitlerin önemi de gittikçe artmaktadır. Herkesin erişimine açık olan API'ler aracılığıyla tivitler kolayca toplanabildiğinden kullanımı tercih edilmektedir.

Bu tez çalışması ile klasik GDA yöntemi üzerinde bir geliştirme yapılarak önerilen sistemin başarısı, haber başlıklarının konularına göre ayrılması ve Twitter verilerindeki duygu durumunun tespit edilmesinde kullanılmıştır. Çalışmanın ilk aşamasında haber başlıklarının hangi türe ait olduğunu belirlemek için konu modelleme yöntemi kullanılarak analiz edilmiştir. Haberler için veriler yedi sınıfa ayrılmaktadır. İkinci aşamada kullanıcıların paylaştığı tivitler ise duygusal açıdan beş sınıfa ayrılmıştır. Duygu sınıfları konu modelleme yöntemiyle kendi içinde analiz edilerek, hangi duyguya ait olduğunu belirleyen kelimelerin dağılımı incelenmiştir. Kelimelerin ağırlıklarına ve bulunduğu konuya bakılması sonucunda tivitin hangi duyguyu içerdiğini sistem öngörmektedir.

1.3 Hipotez

Gelişen teknoloji ile beraber büyük verilerin analiz edilmesi kolaylaşmıştır. Böylece kullanıcıların paylaştığı yorumlar ve haberler yüksek hızda işlenerek analiz edilmektedir. Gelişmeler ile filmler hakkındaki yorumlar, ürün yorumları kullanıcıların yorumları üzerinden analiz edilebilmekte ve bir tahmin çıkarılabilmektedir.

Tezin hipotezi, haberlerin başlıklarından türünün ve Twitter'daki kullanıcıların tivitlerinden duygusunun tespit edilebilmesi varsayımıdır. Haber başlıklarının sınıflandırılması gelişen teknoloji ile beraber medya gibi ilgili alanlarda kullanım açısından büyük önem taşımaktadır. Bir haber başlığının hangi türe ait olduğu otomatik olarak kolayca atanabilecektir. Ayrıca, paylaşılan tivitlerin içerdiği duygu bulunarak kullanıcıya özgü reklamların ve içeriklerin yönlendirilmesine yardım sağlayacağı varsayılmıştır. Kullanıcıların attığı tivitlerin duygusu mutlu, üzgün, sinirli, korku ve sürpriz olarak belirlenmesi tivitin aslında hangi aşamada değerlendirileceği için önemlidir. Tezde öncelikle verilerin konu modelleme yöntemi olan GDA ile etiketlenmesi sağlanarak doğruluğu araştırılmak istenmiştir. Aynı işlem ön aşama gibi düşünülerek ilk önce haber

başlıklarından türünü yöntemle etiketleyip doğruluğunu inceleme içinde yapılması öngörülmüştür. Bu doğrultuda yapılan hipotez, kullanıcının paylaştığı tivitlerden kişi hakkında nasıl duyguya sahip olduğu hakkında çıkarım yapabileceği varsayımına dayanmaktadır.

BÖLÜM 2

MATERYALLER

Bu bölümde kullanılmış olan veri setleri, kullanılan programlama dilleri, kök bulma araçları ve metotlar ayrıntılı olarak açıklanmaktadır.

2.1 Veri Setleri

Sistemin eğitim ve test aşamalarında Türkçe, İngilizce haber başlıkları ve Türkçe tivitlerden oluşan veri setleri kullanılmıştır.

2.1.1 Türkçe Haber Veri Seti

Türkçe metinlerden oluşan haber başlıkları için Milliyet[1], Mynet[2] gibi sitelerden faydalanarak veri seti oluşturulmuştur. Her bir haber başlığı maksimum 30 kelimeden oluşmaktadır. Haber başlığının genişletilmiş olmasının sebebi, bazı haber başlıklarının çok az kelimeden oluşması ve türünün belirlenmesini engellemesidir. Bu yüzden html'de tanımlama etiketindeki alt başlığı içeren bölümde başlık olarak alınmıştır. Veri seti; ekonomi, magazin, siyaset, spor, sağlık, teknoloji ve yaşam olmak üzere 7 farklı türde haberden oluşmaktadır. Her haber türüne ait 600 adet haber başlığı toplanmıştır. Veri seti 3 haber sınıfı (ekonomi, yaşam, spor) için 1800, 5 haber sınıfı (ekonomi, yaşam, spor, siyaset, magazin) için 3000 ve 7 haber sınıfı (ekonomi, magazin, siyaset, spor, sağlık,

[1] http://www.milliyet.com.tr/

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

teknoloji, yaşam) için 4200 adet haber başlığından elde edilen denemelerde kullanılmak üzere 3, 5 ve 7 sınıf etiketine sahip 3 farklı veri seti hazırlanmıştır.

2.1.2 İngilizce Haber Veri Seti

İngilizce veri seti için haber başlıklarından oluşan Uci-news[1] veri setinden ve spor sitelerinden yararlanarak yeni bir veri seti oluşturulmuştur. Veri seti; ekonomi, magazin, sağlık, teknoloji ve spor olmak üzere her birinden 1000 adet haber başlığına sahip olan 5 farklı türden oluşmaktadır. Yine denemeler için veri seti 3 (magazin, sağlık, teknoloji) ve 5 (ekonomi, magazin, sağlık, teknoloji, spor) sınıf olmak üzere sırası ile 3000 ve 5000 veriden oluşan 2 farklı veri seti oluşturulmuştur.

2.1.3 Türkçe Tivit Veri Seti

Twitter aracılığıyla Türkçe tivitlerden oluşan veri seti oluşturulmuştur. Tivitlerde duygunun anlaşılabilmesi için duyguyu ifade edebilecek en az bir kelime olmasına dikkat edilmiştir (mutluyum, şaşırdım vs.). Veri seti; mutlu, üzgün, şaşkın, korku ve kızgın olmak üzere 5 farklı duygudan oluşmaktadır. Her duyguya ait 800 adet tivit toplanmıştır. Üç duygu sınıfı (kızgın, korku, mutlu) için 2400, 5 duygu sınıfı (mutlu, üzgün, şaşkın, korku, kızgın) içinse 4000 tivitten oluşan ve eğitimde kullanılmak üzere, her biri 3 ve 5 sınıf etiketine sahip 2 farklı veri seti hazırlanmıştır.

Her veri setinin %80'i eğitim, %20'si de test için kullanılmıştır.

2.2 Kök Bulma Araçları

Veri setlerindeki kelimelerin köklerini elde etmek için dört tür araç kullanılmıştır.

2.2.1 Zemberek

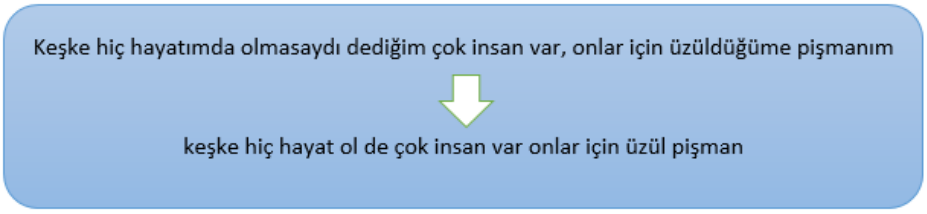
Zemberek[2], açık kaynak kodlu Türkçe doğal dil işleme kütüphanesidir. OpenOffice, LibreOffice eklentisi olarak kullanılmaktadır. Tamamen Java ile geliştirilmiştir. Kütüphanede hatalı kelime için düzeltme önerisi, yazım denetimi, kelimelerin köklerini

[1] https://www.kaggle.com/uciml/news-aggregator-dataset/data

[2] https://tr.0wikipedia.org/wiki/Zemberek_(yaz%C4%B1l%C4%B1m)

tespit etme, kelimeyi eklerine ayrıştırma, heceleme, morfolojik analizini yapma gibi işlemler uygulanabilmektedir. Zemberek'in ikinci sürümünde Mozilla Kamu Lisansı'na geçilmiştir, Türkçe'de bir doğal dil işleme altyapısı oluşturulması için gerekli mimari değişiklikler yapılmıştır. Ayrıca, kütüphane java dilinde yazıldığından platform bağımsızdır.

Çalışmada kütüphane aracılığıyla Türkçe veri setlerindeki kelimelerin biçimbirimsel analizi yapılmıştır. İlgili biçimbirimsel yapıya sahip kelimelerin kökleri bulanarak veri seti güncellenmiştir. Şekil 2.1'de Türkçe tivit veri setindeki bir cümleye ait kelimelerin köklerinin bulunmasına örnek verilmiştir.



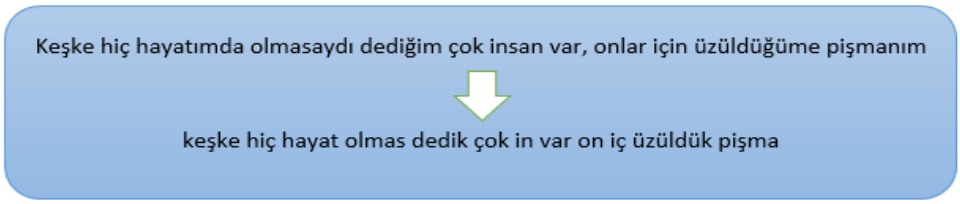
Şekil 2. 1 Türkçe tivit veri setinde zemberek ile kök bulma örneği

2.2.2 Snowball Stemmer

Snowball[1], kelimeler hakkında bilgi edinmek için kökleri çıkarma algoritmaları oluşturmakta kullanılan tasarlanmış bir dizi işleme dilidir. Snowball derleyicisi bir Snowball komut dosyasını başka bir dile dönüştürmektedir. Şu anda ISO C, Java, Python, Rust ve Go programlama dillerinde desteklenmektedir. İngilizce, Fransızca, İspanyolca, Portekizce, Almanca, Türkçe ve daha birçok dilde kelimelerin köklerini bulma işlemini gerçekleştirmektedir.

Çalışmada kütüphane kullanılarak veri setlerindeki kelimelerin köklerini bulma işlemi yapılmıştır. Kelimelerin kökleri ile veri seti güncellenmiştir. Şekil 2.2'de Türkçe tivit veri setindeki bir cümleye ait kelimelerin köklerinin bulunmasına örnek verilmiştir.

[1] http://snowballstem.org/

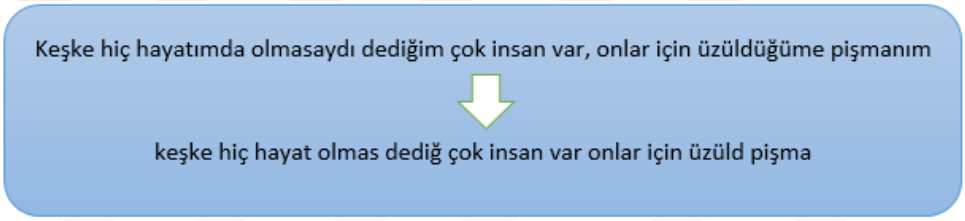


Şekil 2. 2 Türkçe tivit veri setinde snowball ile kök bulma örneği

2.2.3 Kelimenin İlk Beş Harfinin Kök Kabul Edilmesi

Türkçe çalışmalarda genel olarak kullanıldığı tespit edilen bir yöntemdir. Bu yöntem kelimenin ilk beş harfinin kök olarak alınıp, kabul edilmesine dayanmaktadır. Beş harften az olan kelimelerin tamamı kök olarak alınır.

Anlatılan yöntem çalışmamızda baz alınarak veri seti güncellenmiştir. Şekil 2.3'de Türkçe tivit veri setindeki bir cümleye ait kelimelerin köklerinin bulunmasına örnek verilmiştir.



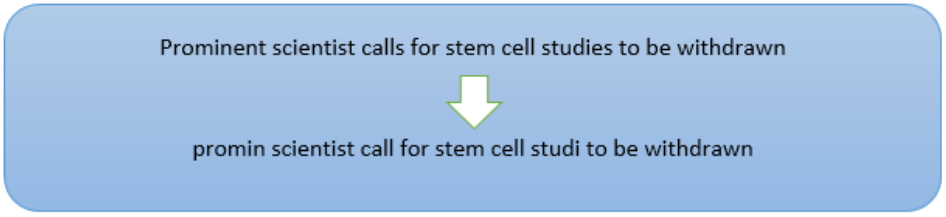
Şekil 2. 3 Türkçe tivit veri setinde ilk beş harf ile kök bulma örneği

2.2.4 Porter Stemmer

Python'a nltk kütüphanesiyle beraber yüklenen kök bulma aracıdır. İngilizce dilini desteklemektedir. Porter kök bulma kütüphanesi, İngilizce'deki kelimelerin biçimbirimsel analizini yaparak kelime kökünü bulan bir yöntemdir[1].

Tezdeki çalışmada İngilizce olarak bulunan haberler veri seti için kök bulma aracı olarak kullanılmıştır. Sonuç olarak bulunan kökler ile veri setinin güncellenmesi sağlanmıştır. Şekil 2.4'de İngilizce haber veri setindeki bir cümleye ait kelimelerin köklerinin bulunmasına örnek verilmiştir.

[1] https://tartarus.org/martin/PorterStemmer/



Şekil 2. 4 İngilizce haberler veri setinde porter stemmer ile kök bulma örneği

2.3 Programlama Dilleri

Bu tez kapsamında geliştirilen uygulamanın modülleri iki farklı programlama dili kullanılarak geliştirilmiştir.

2.3.1 Java

Java[1], Sun Microsystems şirketinin geliştirmiş olduğu bir programlama dilidir. Bu dil, C ve C++'dan birçok sözdizim türetse de daha basit nesne modeli ve daha az düşük seviye olanaklar içermektedir. Java uygulamaları bilgisayar mimarisine bağlı olmaksızın herhangi bir Java Virtual Machine'de çalışabilmektedir.

Yazılan Java uygulaması; Zemberek ile kelimelerin morfolojik analizine göre köklerini tespit etmek ve veri seti dosyasını tekrar oluşturmak için kullanılmıştır. Ayrıca, Weka'da kullanmak üzere, oluşturulmuş olan "csv" uzantılı dosyayı, "arff" uzantıya çevirmek için de uygulama yazılmıştır.

2.3.2 Python

Python, genel amaçlı programlama için yorumlanmış bir üst düzey programlama dilidir. Kodun okunabilirliğini vurgulayan bir tasarım felsefesine sahiptir. Ayrıca, programcıların fikirlerini daha az kod satırında ifade etmesini sağlayan bir sözdizimi vardır. Küçük ve büyük ölçeklerde açık programlamaya yapı sağlamaktadır. Dinamik bir sistem olup otomatik bellek yönetimi özelliği bulunmaktadır. Nesne yönelimli, zorunlu, işlevsel ve işlemsel olmak üzere çoklu programlama paradigmalarını desteklemektedir. Geniş ve

[1] https://en.0wikipedia.org/wiki/Java_(programming_language)

kapsamlı standart bir kütüphaneye sahip olan Python'un, birçok işletim sisteminde yorumlayıcıları da mevcuttur[1].

Tezde yazılan uygulamanın ana dili olarak kullanılmıştır. Konu modelleme kütüphanesi Python'a yüklenerek kodsal olarak uygulanmıştır. Dosya okuma işlemleri, veri setine uygulanan ön işlemler, konu modelleme algoritması olan GDA, ilgili değer olan tutarlık (coherence) değeri hesaplama, sistemin konularını bulma, konuları etiketleme ve sistemin doğruluğunu tespit etme, kelime-ağırlık grafiklerini çıkarma işlemleri gibi çoğu işlem Python uygulamasıyla beraber yazılmıştır.

2.4 Weka

Weka, veri madenciliği işlem adımları için geliştirilmiş bir makine öğrenme algoritmasıdır. Algoritmalar doğrudan veri kümesine uygulanabilmektedir. Ayrıca, Java kodundan çağrılabilme özelliği mevuttur. Weka, veri ön işleme, sınıflandırma, regresyon, kümeleme, ilişki kuralları ve görselleştirme yöntemlerini içermektedir. Aynı zamanda yeni makine öğrenme tasarısı geliştirmek için de uygundur. Weka, GNU Genel Kamu Lisansı altında yayınlanan açık kaynaklı bir yazılımdır [36].

Weka aracı çalışmada GDA algoritması ile veri setinin konu ağırlığına göre etiket bilgisiyle oluşturulan dosyanın sınıflandırıcılar ile başarısını ölçmek için kullanılmıştır.

2.5 Metotlar ve Algoritmalar

Geliştirilen sistemin temel algoritması konu modelleme için kullanılan GDA'dır. Diğer metotlar GDA'dan faydalanılarak çıkarılan özellik bilgisi ile sınıflandırma işlemi yapmak için kullanılmıştır.

Naive Bayes (NB), Multinomial Naive Bayes (MNB), Destek Vektör Makineleri (DVM), Rastgele Orman (RO) ve Multilayer Perceptron (MP) sınıflandırıcılar; GDA algoritması ile veri setinin konu ağırlığına göre etiketli bilgisiyle oluşturulan dosyanın Weka ile çalıştırılması sonucu kullanılmıştır.

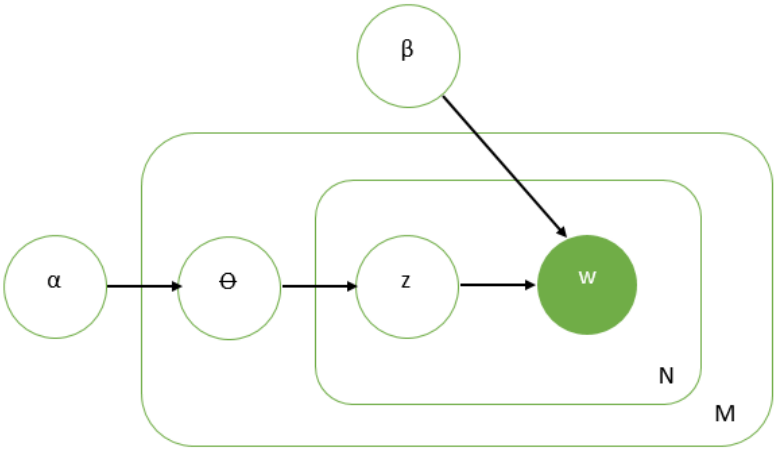
[1] https://en.0wikipedia.org/wiki/Python_(programming_language)

2.5.1 Gizli Dirichlet Ayırımı Algoritması

GDA, olasılık tabanlı bir konu modelleme yöntemi olup, bir dizi dokümandan belirlenmiş sayıdaki konulara ait kelimeleri ve ağırlıklarını oluşturmaktadır. GDA yönteminde, metin belgesi konuların birleştirilmiş şekli olarak tanımlanmaktadır. Yöntemin temelinde, konular kelimeler üzerinde bir olasılık dağılımına, metin belgeleri de konular üzerinde bir olasılık dağılımına sahiptir. Her konunun ise sabit kelime dizisi üzerinde bir dağılımı olmaktadır [37]. Model, gözlemlenen veri setiyle oluşan kelimeler ve ağırlık değerleriyle temel konu yapısını belirlemeyi amaçlar. Dokümanlardaki kelimeler, sistemde gözlemlenen verilerdir.

GDA metin belgelerindeki konuları bulmak için kullanılan etkili bir denetimsiz öğrenme yöntemidir. Yöntem her dokümanı, kelimeler üzerinde çok terimli bir dağılıma karşılık gelen her bir konunun karışımı olarak modelleyen üretken bir işlemdir. GDA tarafından öğrenilen dokümanın konusu ve konu-kelime dağılımları, dokümanlar için en iyi konuları ve her bir konu için en açıklayıcı kelimeleri tanımlar [11].

Şekil 2.5'de GDA süreci grafiksel olarak gösterilmiştir. Rastgele olan değişkenler düğümler ile belirtilmiştir. Düğümler arasındaki muhtemel bağlantılar ise kenarlar kullanılarak temsil edilmektedir.



Şekil 2. 5 Gizli dirichlet ayırımı süreci [38]

Gösterimdeki parametreleri belirtecek olursak;

- α doküman başına konu dağılımını verir.
- β konu başına kelime dağılımını verir.

● Θ belli doküman için konu dağılımını gösterir.

● z her bir kelime için atanan konulardır.

● w gözlemlenen kelimelerdir.

Şekil 2.5'de ki yapıdan da anlaşılacağı gibi α ve β parametreleri, sistem oluşturulurken bir kez örneklenmektedir. Θ parametresi ise sistemdeki her bir doküman için örneklenmektedir [38].

GDA algoritmasında dokümanlardaki tüm kelimelere rastgele konu atanmaktadır. Konu atama işlemi dokümanlara yapıldıktan sonra bu bilgiyle çeşitli istatistikler çıkartılır. Yerel istatistik, her dokümandaki konulara kaç adet kelime atandığını gösterirken, global istatistik ise tüm doküman için her kelimenin her konuya kaç kere atandığını göstermektedir. İstatistiksel bilgiler elde edildikten sonra her doküman için her kelimenin yeniden konu ataması gerçekleştirilir. Bunun için mevcut kelime bilgileri de güncellenmektedir.

$$\frac{n_{ik} + \alpha}{N_i - 1 + K\alpha} \tag{2.1}$$

Kelimeler, konulara atanırken ilk önce mevcut dokümanın konular ile ne kadar ilişkili olduğu hesaplanır (2.1). n_{ik}, *i.* haberde *k*. konuya atanan kelime sayısını göstermektedir. N_i ise dokümanda yer alan toplam kelime sayısıdır. Değerden 1 çıkartılmasının nedeni kullanılan kelimenin yok sayılmasıdır. α değeri; konuların dokümanlardaki dağılımını vermektedir. *K* değeri de belirlenen konu sayısıdır.

K konu sayısı sistemde; konu modelleme ölçütü olan tutarlık (coherence) değeriyle belirlenmektedir. Tutarlık değeri; kelimelerin birbirine benzerliğini ölçer ve seçilecek olan konu sayısı hakkında bize bilgi vermektedir. Belirtilen konu sayıları için hesaplanan tutarlık değerleri arasından en yüksek çıkana ait konu değeri olan *k* değeri, konu sayısı olarak seçilmektedir.

$$\frac{n_{word,k} + \beta}{\sum_{w \in V} n_{w,k} + V\beta} \tag{2.2}$$

Yöntemde ikinci olarak, her kelimenin konular ile ne kadar ilişkili olduğu hesaplanır. Hesaplama sonucunda kelimenin, verilen konu ile ilgili ağırlığı hakkında bilgi çıkarılır. (2.2)'de; $n_{word,k}$ geçerli kelimenin *k*. konuya tüm dokümanda kaç kere atandığını gösterir.

β değeri; kelimelerin konulardaki dağılımını verir. *V* ise veri setindeki tüm kelimelerden oluşturulan sözlüğün boyutudur. (2.1) ve (2.2)'den elde edilen sonuçlar çarpılarak geçerli kelimenin *k*. konuya atanma olasılığı hesaplanmaktadır. Tüm doküman sayısı boyunca değerler tekrar hesaplanmaktadır. En yüksek değere ait olan konu, kelimenin yeni konusu olarak belirlenir. Veri setindeki tüm dokümanlara ait her kelime için aynı işlemler uygulanarak dokümanların konuları bulunur. Sistemde belirlenen iterasyon sayısına kadar konuları güncelleme sürmektedir. Kelimelerin konu dağılımı ataması gerçekleştikten sonra sistemin modelini çıkarmak için doküman-terim matrisi oluşturulur. Bu matris ile kelime ağırlıkları hesaplanarak, kelimelerin konulardaki ağırlıkları ortaya çıkarılır [39].

2.5.2 Naive Bayes

Naive Bayes algoritması, belirli bir veri setindeki değerlerin frekansları ve değer birleşimlerini sayarak bir olasılık setini hesaplayan basit bir olasılıkçı sınıflandırıcıdır. Algoritma, Bayes teoremini kullanır ve sınıf değişkeninin değeri göz önüne alındığında tüm özniteliklerin bağımsız olduğunu varsaymaktadır. Bu koşullu bağımsızlık varsayımı nadiren gerçek dünyadaki uygulamalarda geçerlidir, dolayısıyla niteleme Naive olarak geçmektedir. Ancak algoritma çeşitli denetlenen sınıflandırma problemlerinde iyi performans göstermekte ve hızla öğrenme eğilimindedir [40].

Naive Bayes sınıflandırıcısı, Bayes teoremine ve toplam olasılığın teoremine dayanmaktadır. Bir x = (f_1, f_2, … f_m) öğesine sahip olduğumuzu varsayarsak ve f_i, m özelliklerinden birinin değeridir. Sonra x'in sınıf c_i olabilme olasılığı (2.3)'deki gibi ifade edilmektedir;

$$P(c_i|f_m) = \frac{P(f_m|c_i) * P(c_i)}{P(f_m)} \tag{2.3}$$

(2.3)'de, $P(c_i)$ yani c_i sınıfının olasılığı, c_i sınıfındaki örnek sayısının toplam örnek sayısına bölünmesiyle hesaplanmaktadır. $P(f_m)$ sınıfa bağımlı olmadığı için, payda atılabilmektedir. Ardından formül, zincir kuralı kullanılarak yeniden yazılabilir; bir sınıfın bir özelliğinin varlığı, sınıf göz önüne alındığında diğer özelliklerin varlığından bağımsız olduğu varsayımına sahip olur;

$$P(c_i|f_m) = P(c_i)\prod_{i=1}^{m} P(f_m|c_i) \quad (2.4)$$

Böylece, her sınıf için test örneğinin sonraki (posterior) olasılığı (2.4) ile hesaplanmaktadır. En yüksek değere sahip sınıf (2.5)'te olduğu gibi tahmin edilerek seçilmektedir [19].

$$class = argmax_{class} P(class) * \prod_{i} P(feature_i \mid class) \quad (2.5)$$

2.5.3 Multinomial Naive Bayes

Multinomial Naive Bayes (MNB), bir dokümandaki kelimelerin çok terimli olarak dağılımını modellemektedir. Bir doküman bir dizi sözcük olarak ele alınır ve her bir sözcük konumunun birbirinden bağımsız olarak oluşturulduğu kabul edilmektedir.

Sınıflama için sabit sayıda sınıf olduğu varsayılmaktadır ve her bir sınıf sabit çoklu terim parametresine sahiptir. Sınıflar kümesi C = $\{c_1, c_2, \ldots c_m\}$, her bir c sınıfı için ise parametre vektörü $V(\theta_c) = \{\theta_{c1}, \theta_{c2}, \ldots \theta_{cn}\}$ ve kelime sözlüğünün boyutu n olarak belirtilmiştir. $\sum_i \theta_{ci} = 1$ işlemi sonucu ortaya çıkan θ_{ci}, i kelimesinin sınıfta olma olasılığıdır. Bir dokümanın olasılığı, dokümanda belirlenen kelime parametrelerinin bir sonucudur;

$$P(d|\theta_c) = \frac{\sum_i f_i\,!}{\prod_i f_i\,!} * \prod_i (\theta_{ci})^{f_i}, \quad (2.6)$$

f_i, doküman d'deki i kelimesinin frekans sayısını vermektedir. Sınıf kümesinin üzerinde öncelikli bir dağıtım yaparak, sınıfı en büyük sonraki (posterior) olasılıkla seçen minimum hata sınıflandırma kuralına ulaşabilmektedir [41].

2.5.4 Destek Vektör Makineleri

Destek Vektör Makineleri (DVM), verileri sınıflandırmak için çok boyutlu düzlemleri kullanan denetimli öğrenme algoritmasıdır. Çok boyutlu düzlemler, farklı sınıfların en yakın noktalarını ayıran kenar boşluklarını bulmak amacıyla oluşturulmuştur. Sınırlar boyunca uzanan noktaları destek vektörleri göstermektedir. Basit problemler için çok boyutlu düzlemler sınıfları kolayca ayırmaktadır. Daha karmaşık problemler için, çok

boyutlu düzlem, kenar boşluğunun yanlış tarafındaki noktalar ile tamamen ayrılamamaktadır. Çok boyutlu düzlem,

$$h(x) = w^T x \tag{2.7}$$

(2.7)'de w, bir düzlem üzerindeki h(x) = 0 boyunca olan noktaların ağırlık vektörüdür (destek vektörü).

$x_i \in R$ ve y ∈ {1, -1} olan bir iki sınıflı bir eğitim veri seti (x_i, y_i) için, DVM en uygun hale getirmektedir;

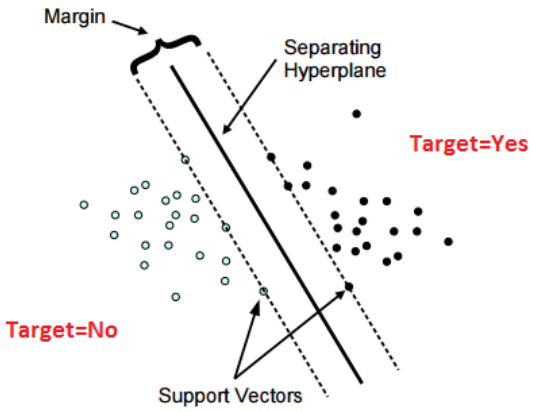
$$min_{w,\varepsilon} = \frac{1}{2} w^T w + C \sum_i \varepsilon_i \tag{2.8}$$

$$y_i(w^T \varphi x_i) \geq 1 - \varepsilon_i \, , \varepsilon_i \geq 0$$

(2.8)'de C, marjinin dayanıklılığını kontrol eden bir maliyet faktörüdür; ε, bir kenarın yanlış tarafından beklenen kenarına olan veri noktası mesafesidir. φ ise bir kernel fonksiyonu tarafından tanımlanan girdi verisinin yüksek boyutlu haritalanmasıdır. φ için, doğrusal olmayan çok boyutlu düzlem ile ayrılan ortak bir kernel fonksiyonu radyal taban fonksiyonu (2.9)'da belirtilmiştir;

$$k(x, x') = \exp\left(e^{-\gamma \left|\left|x - x'\right|\right|^2}\right) \tag{2.9}$$

γ, radyal taban fonksiyonu için kernel parametresidir.



Şekil 2. 6 Destek vektör makineleri [42]

Şekil 2.6'da DVM yapısının çalışmasına örnek gösterilmiştir. Sınıflandırma, DVM'nin yüksek boyutlu alanı belirli bölgelere ayırma becerisine dayanır. Bir bölge içindeki üye noktalar, veri noktasının nihai sınıflandırmasını belirlemektedir [13].

2.5.5 Rastgele Orman Algoritması

Rastgele orman (RO) algoritması, bir yanıt değişkeni üzerinde toplu olarak bildirmek için bir grup model kullanan denetimli bir öğrenme algoritmasıdır. Grup içindeki bir model, çoklu sınıflandırma ve regresyon ağacıdır (CART). CART algoritması; bir yanıt değişkeniyle bir dizi açıklayıcı değişken arasındaki ilişkiyi, her parçadan gelen sınıflandırma hatalarının en aza indirgenmesi için bir dizi bölümleme kuralıyla modelleyen bir algoritmadır. Parçaların dizisi, sonuç olarak tüm veri noktalarının sınıflandırılmasına neden olur. Her bölünme sırasında, bir veri noktasının doğru olarak sınıflandırmanın frekansı (f_i) veya yanlış sınıflandırmanın frekansı (1-f_i) hesaplanabilmektedir. Sınıflandırma performansının genel seviyesi Gini indeksinin azaltılmasına dayanır. Gini indeksi şu şekilde hesaplanır;

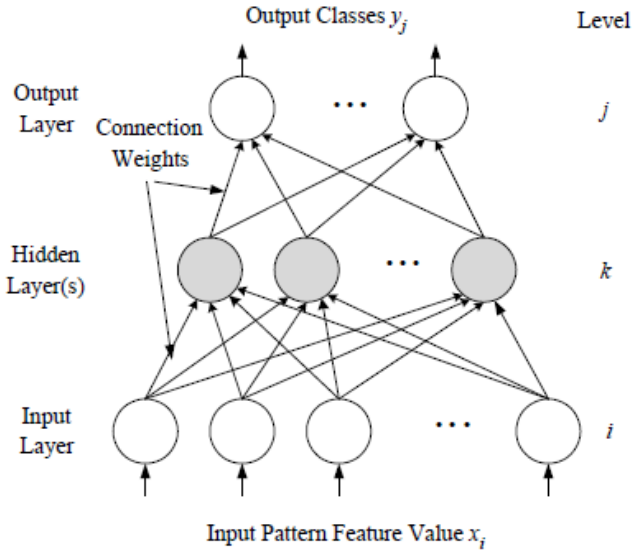
$$I_G = \sum_i^n f_i * (1 - f_i) \tag{2.10}$$

CART'lar ve bir RO algoritması arasındaki iki önemli farklılık, RO'da birçok ağacın oluşturulması, tüm veri noktalarına karşın orijinal verinin önyüklenmiş (bootstrapping) örneğine dayanmaktadır. Ayrıca, RO algoritması, her bölünmede tüm değişkenler yerine açıklayıcı değişkenlerin alt kümesini kullanmaktadır. Tek bir ağaç için önyükleme sırasında, veri noktalarının yaklaşık üçte ikisi model geliştirme için kullanılır. Her bir bölünmede, değişken sayısının karekökü kullanılır. Hem önyükleme yapmak hem de açıklayıcı değişkenlerin bir alt kümesini kullanmak, ağaçların birbirleriyle ilişkilendirilmemesini sağlamaktadır. Tek bir ağaç, aşırı uyma (overfitting) ve daha düşük algoritma performansı sonucu oluşan veri setindeki gürültüye aşırı duyarlı olma eğilimi göstermektedir. RO'da birden fazla ilişkisiz ağaç kullanılmasıyla, gürültü ve varyans azaltılarak daha iyi algoritma performansı elde edilir.

Tüm RO algoritması, verileri sınıflandırmak için rastgele orman grubundaki tüm ağaçları kullanır. Her ağaç, sınıflandırmayla ilgili tek bir kararı temsil eder ve tüm ağaçların çoğunluk kararı bir veri noktasının sınıfını belirlemektedir [13].

2.5.6 Multilayer Perceptron

Multilayer Perceptron (MP), Pazarlama Karar Destek Sistemi'nde (MDSS) en sık kullanılan sinir ağı mimarilerinden biridir ve denetimli sinir ağları sınıfına aittir. MP, katmanlarda sıralanan bir düğüm ağı (işlem elemanları) içerir. Tipik bir MP ağı, üç veya daha fazla işlem düğümü katmanından oluşur. Bunlar girdi katmanı, bir veya daha fazla gizli katman ve sınıflandırma sonucunu üreten çıktı katmanıdır. Diğer katmanların aksine giriş katmanında hesap yapılmamaktadır. Ağın prensibi, veriler giriş katmanında bulunduğunda, ağ düğümlerinin çıktı düğümlerinin her birinden bir çıktı değeri elde edilinceye kadar ardışık katmanlarda hesaplamalar gerçekleştirmesine dayanmaktadır. Bu çıktı sinyali, giriş verileri için uygun sınıfı gösterebilmelidir. Diğer bir deyişle, doğru sınıf düğümünde yüksek çıkış değeri ve diğer tümünde düşük çıkış değerleri olması beklenebilir.

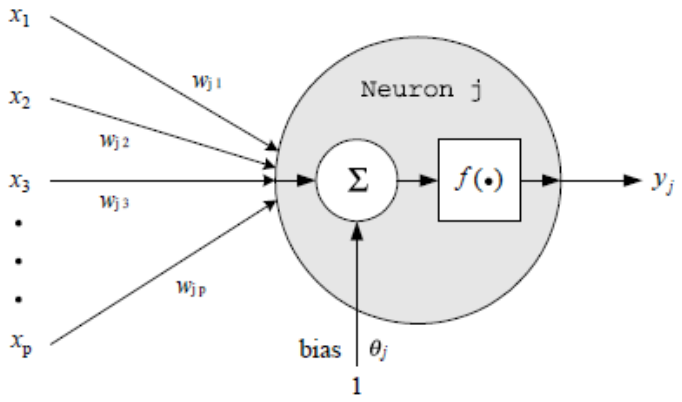


Şekil 2. 7 Multilayer perceptron yapısı [43]

MP'deki bir düğüm, sapmanın varlığında girdilerin ağırlıklı toplamını hesaplayan yapay bir nöron (Şekil 2.8) olarak modellenebilir ve etkinleştirme fonksiyonu sayesinde bu toplamı geçirir. Bütün süreç şu şekilde tanımlanmaktadır:

$$v_j = \sum_{i=1}^{p} w_{ji} x_i + (\theta_j), \qquad y_j = f_j(v_j) \tag{2.11}$$

(2.11)'de v_j girişlerin (x_1, x_2, .. x_p) doğrusal birleşimi, θ_j sapma değeri, w_{ji} x_i girişi ile nöron j arasındaki ilişki ağırlığıdır. f_j, j. nöronun aktivasyon fonksiyonudur. y_j ise sistemin çıktısıdır.



Şekil 2. 8 MP'nin bir düğümü; yapay nöron [43]

Sigmoid fonksiyonu, aktivasyon fonksiyonun yaygın olarak tercih edilenidir (2.12).

$$f(a) = \frac{1}{1 + e^{-a}} \tag{2.12}$$

Sapma terimi θ_j, θ_j'nin pozitif veya negatif bir değer alıp almayacağına bağlı olarak, sigmoid aktivasyon fonksiyonun sol veya sağa kaymasına katkıda bulunmaktadır [43].

2.5.7 k-Katlı Çaprazlı Doğrulama

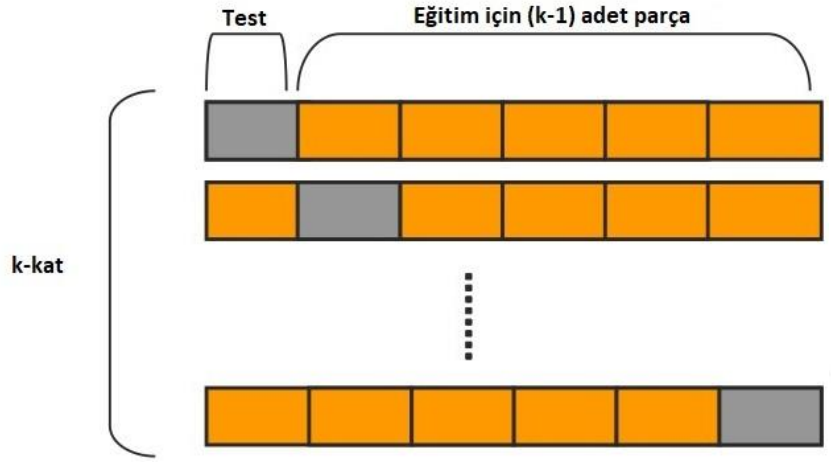
Çapraz Doğrulama (ÇD), makine öğrenimi modellerinin performansını değerlendirmek için çok kullanışlı bir tekniktir. Makine öğrenme modelinin, bağımsız veri setinde nasıl genelleneceğini bilmeye yardımcı olmaktadır. Modelin eğitimde vereceği tahminlerin ne kadar doğru olduğunu tahmin etmek için bu teknik kullanılmaktadır.

Bir makine öğrenimi problemi olduğunda, eğitim ve doğrulama (test) olmak üzere iki tip veri seti kullanılmaktadır. Çapraz doğrulamayı kullanarak, aşırı eğitim olup olmadığını kontrol etmek ve makine öğrenme modelinin problem için verilen doğrulama veri seti olan bağımsız verileri nasıl genelleneceği hakkında bir fikir edinmek için "eğitim" aşamasında makine öğrenme modeli "test" edilebilmektedir.

k-katlı Çapraz Doğrulama (k-KÇD), makine öğrenmesinde yaygın olarak kullanılan yaygın bir çapraz doğrulama türüdür. k-KÇD, aşama olarak açıklamak gerekirse;

- Orijinal eğitim veri seti k eşit alt kümelere ayrılmaktadır. Her alt kümeye bir kat denir. Katlar $f_1, f_2, \ldots f_k$olarak adlandırılmaktadır.
- i = 1 'den k'ya kadar;

  Doğrulama veri seti olarak f_i katı alındığında, geri kalan k-1 tane kat eğitim veri seti olarak kullanılmaktadır. Çapraz doğrulama, eğitim setini kullanarak makine öğrenmesi modelini eğitmektedir. Ardından tahmin edilen sonuçları doğrulama setine karşı doğrulayarak modelin doğruluğunu hesaplamaktadır.
- Tüm k çapraz doğrulama durumlarında elde edilen doğrulukların ortalaması alınarak, makine öğrenim modelinin doğruluğu tahmin edilmektedir.



Şekil 2. 9 K-katlı çapraz doğrulama çalışma yapısı [44]

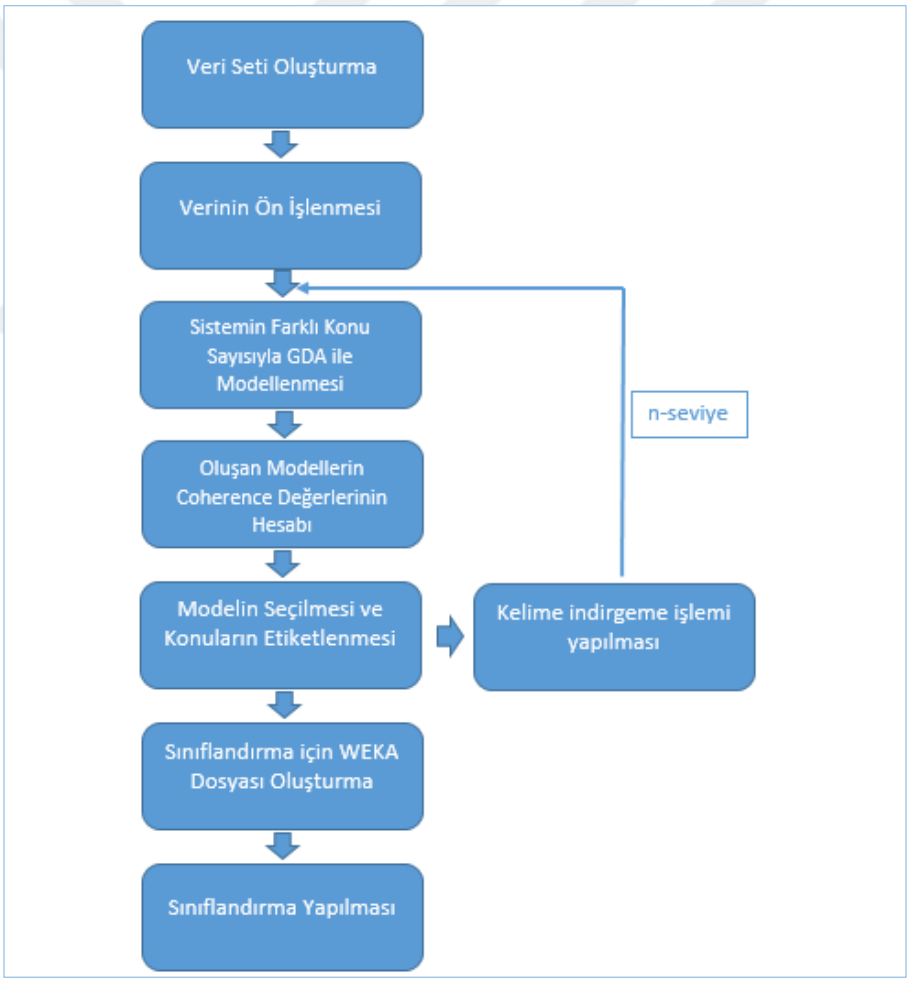
k-katlı çapraz doğrulama yönteminde, orijinal eğitim veri setindeki tüm girdiler hem eğitim hem de doğrulama yapabilmek için kullanılmaktadır. Ayrıca, her girdi yalnızca bir kez doğrulama için kullanılır [44].

Çapraz doğrulama tekniği, farklı makine öğrenim modellerinin performansını aynı veri seti üzerinde karşılaştırmak için kullanılabilir. Çalışmada önceki bölümlerde açıklanan makine öğrenme yöntemleri için bu yöntem kullanılmıştır. k değeri genel olarak 10 alınmıştır. Veri seti tamamen bir yerde eğitim seti olarak toplanarak sistem Weka'da çalıştırılmıştır.

BÖLÜM 3

SİSTEM TASARIMI VE UYGULANMASI

Tez kapsamında geliştirilen uygulamanın işlem adımları bu bölümde anlatılmaktadır. Şekil 3.1'de kurgulanan sistemin tasarımı verilmektedir.



Şekil 3. 1 Sistemin tasarımı

Çalışmada ilk olarak veri setinin oluşturması gerçekleştirilmiştir. Sonrasında toplanan veri seti üzerinde ön işlemler uygulanarak anlaşılır hale gelmesi amaçlanmıştır. Ardından

ön işlem uygulanan veri setleri, belirli farklı konu sayılarıyla konu modelleme algoritması olan GDA ile eğitilmiştir. Konu sayılarıyla beraber oluşan modellerin ayrı ayrı tutarlık (coherence) değerleri hesaplanmıştır. Alınan sonuca göre model seçilmiştir. Seçilen modelin konularına ilgili etiketin atanma işlemi gerçekleştirilmiştir ve sistemin doğruluğu hesaplanmıştır. Ardından başarı artışı sağlayabilmek için sözlükteki kelime sayısını azaltabilecek yöntem geliştirilmiştir. Bu yöntem GDA'nın N-seviyeli olarak kullanılabilmesini sağlamıştır. Sistemin başarısı tekrar ölçülerek ve konuya göre kelime-ağırlıklarından Weka'ya uygun formatta bir dosya oluşturulmuştur. Sonrasında sınıflandırmada alınan başarı ölçülmüştür.

3.1 Veri Seti Oluşturma

Veri seti materyaller bölümünde anlatıldığı gibi oluşturulmuştur. Kısaca hatırlatmak gerekirse 3 çeşit veri seti kullanılmıştır;

- Türkçe haber veri seti: Türkçe metinlerden oluşan haber başlıkları için haber sitelerinden faydalanarak veri seti oluşturulmuştur. Veri seti; ekonomi, magazin, siyaset, spor, sağlık, teknoloji ve yaşam olmak üzere 7 farklı etikete sahip haberlerden oluşmaktadır. Her haber türüne ait 600 adet haber başlığı toplanmıştır. Üç haber türü (ekonomi, spor, yaşam) için 1800, 5 haber türü (ekonomi, magazin, siyaset, spor, yaşam) için 3000 ve 7 haber türü (ekonomi, magazin, siyaset, spor, sağlık, teknoloji, yaşam) için 4200 adet haber başlığından oluşan 3 farklı veri seti hazırlanmıştır.
- İngilizce haber veri seti: Haber başlıklarından oluşan hazır UCI-news veri seti kullanılmıştır. Veri seti; ekonomi, magazin, sağlık, teknoloji ve spor olmak üzere her 5 farklı haber türünden oluşmaktadır. Üç haber türü (magazin, sağlık, teknoloji) için 3000, 5 haber türü (ekonomi, magazin, sağlık, teknoloji, spor) için 5000 haber başlığından oluşan 2 farklı veri seti hazırlanmıştır.
- Türkçe Tivit veri seti: Twitter sitesi aracılığıyla Web scraping yapılan Import.io[1] sitesi yardımıyla Türkçe tivitlerden veri seti oluşturulmuştur. Seçilen tivitlerde duyguyu ifade eden en az bir kelimenin olmasına dikkat edilmiştir. Veri seti; mutlu, üzgün,

[1] https://www.import.io/

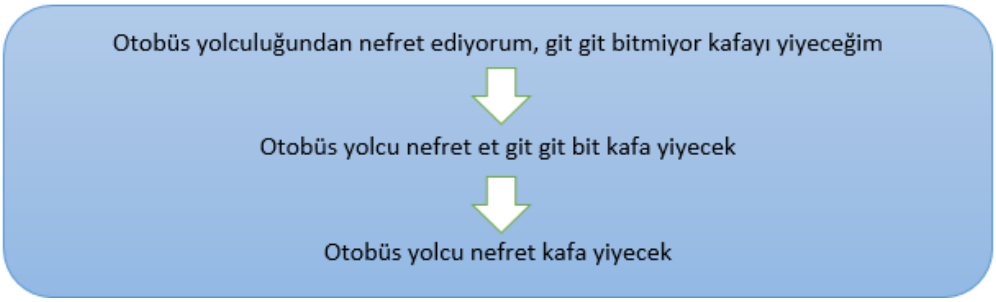
şaşkın, korku ve kızgın olmak üzere 5 duygudan oluşmaktadır. Üç duygu türü (mutlu, kızgın, korku) için 2400, 5 duygu türü (mutlu, üzgün, şaşkın, korku, kızgın) için 4000 tivitten oluşan 2 farklı veri seti kullanılmıştır.

3.2 Veri Ön İşleme

3.2.1 Türkçe Veri Setinde Ön İşlem

Türkçe haberler ve tivitlerden oluşan iki ayrı veri setinde aynı ön işlem adımları uygulanmıştır. Sırayla ön işlem adımlarını belirtmek gerekirse;

- Cümlelerdeki noktalama işaretleri ayrıştırılmıştır.
- Tüm veri seti küçük harfe dönüştürülmüştür. Türkçe karakterler okunurken kodsal olarak sıkıntılar olduğu için İ, Ö, Ç gibi İngilizcede olmayan harfler kod içerisinde küçük harfe çevrilmiştir.
- Cümle içinde sınıflandırma için önemi olmayan etkisiz kelimeler (stopwords) cümlenin içerisinden temizlenmiştir.
- Veri seti için duygu ve haberi belirlemede anlam taşımayan kelimelerden her biri için liste oluşturulmuştur. Bu listedeki kelimeleri içeren cümlelerde ayrıştırma işlemi Şekil 3.2'deki gibi uygulanmıştır.



Şekil 3. 2 Örnek bir aşama olarak ön işleme

- Kelimelerin kökünü bulmak için 3 farklı yöntem kullanılmış ve her biri için VS-Z, VS-S ve VS-5 ismi verilen ayrı veri setleri oluşturulmuştur;
  - VS-Z: Zemberek kütüphanesi kullanılarak kelimelerin kökleri elde edilerek haberler veri seti için isim, fiil ve kısaltma içeren kelimelerden oluşan veri seti

oluşturulmuştur. Tivit veri seti için ise isim, sıfat, fiil ve tepkileri içeren kelimelerden veri seti elde edilmiştir.

- VS-S: Snowball stemmer kütüphanesi kullanılarak kelimelerin kökleri elde edilmiştir. Kök uzunluğu 8 karakterden uzun olanlar için ilk 5 harf kök olarak kabul edilmiştir. Diğer kelimelerin aynısı alınarak veri seti oluşturulmuştur.
- VS-5: Veri seti içerisindeki kelimelerin ilk 5 harfi kök kabul edilerek veri seti oluşturulmuştur.

3.2.2 İngilizce Veri Setinde Ön İşlem

İngilizce haber veri setinde;

- Noktalama işaretlerinin ayrıştırılması gerçekleştirilmiştir.
- İngilizcede sınıflandırma için önemli olmayan etkisiz kelimeler (stopwords) çıkarılmıştır.
- Kök bulma işlemi için de Porter Stemmer[1] (UCI-P) kütüphanesi kullanılmıştır. Sonucunda köklerden oluşan bir veri seti oluşturulmuştur.

3.3 Sistemin GDA ile Modellenmesi

Konu modelleme algoritması olan GDA ile modelleme 3 veri seti içinde yapılmıştır. GDA eğitimsiz öğrenme algoritması olduğu için önceden tanımlanmış kelimelere ihtiyaç duymamaktadır. Modelde konu sayısını belirleme işleminden sonra, sınıflara göre konulara etiket atanmaktadır.

Model oluşturma sayıları veri setlerinde farklılık göstermektedir. Veri setleri için maddeler halinde açıklama yapacak olursak;

- Türkçe haber veri seti: Ön işlem adımları uygulandıktan sonra 3 farklı kök bulma yöntemi kullanılarak oluşturulmuş olan 3, 5 ve 7 farklı sınıf için VS-Z, VS-S ve VS-5 veri setleri kullanılmıştır. 3, 5 ve 7 sınıflı veri setlerinde farklı konu sayıları kullanarak GDA algoritmasıyla modeller oluşturulmuştur. Üç sınıf için 6 ile 30 arası (6, … 30), 5 sınıf

[1] https://tartarus.org/martin/PorterStemmer/

için 10 ile 50 arası (10, … 50), 7 sınıf için ise 14 ile 49 arası (14, … 49) konu sayısı olan k parametresi kullanılmıştır.

- İngilizce haber veri seti: : Ön işlem adımları uygulandıktan sonra kök bulma aracı UCI-P ile güncellenen 3 ve 5 sınıflı veri setleri kullanılmıştır. 3 ve 5 sınıflı veri setlerinde farklı konu sayıları kullanarak GDA algoritmasıyla modeller oluşturulmuştur. 3 sınıf için 6 ile 30 arası (6, … 30) ve 5 sınıf için 10 ile 50 arası (10, … 50) konu sayıları kullanılmıştır.
- Türkçe Tivit veri seti: Ön işlemler veri setine uygulandıktan sonra farklı kök bulma yöntemleri kullanılarak oluşturulmuş olan 3 ve 5 sınıfa sahip VS-Z, VS-S ve VS-5 veri setleri kullanılmıştır. 3 ve 5 sınıflı veri setlerinde farklı konu sayıları kullanarak GDA algoritmasıyla modeller oluşturulmuştur. 3 sınıf için 6 ile 30 arası (6, … 30) ve 5 sınıf için 10 ile 50 arası (10, … 50) konu sayıları kullanılmıştır.

3.4 Uygun Konu Sayısının Bulunması

Çalışmada konu değerinin bulunması için iki adet parametrik değerin hesaplanması araştırılmıştır. Bu değerler modelin sapma (perplexity) ve tutarlık (coherence) değerleridir. Sistemde iki değerin karşılaştırılmasından sonra daha iyi anlaşılır olan ve başarılı sonuçlar veren tutarlık değeri kullanılmıştır.

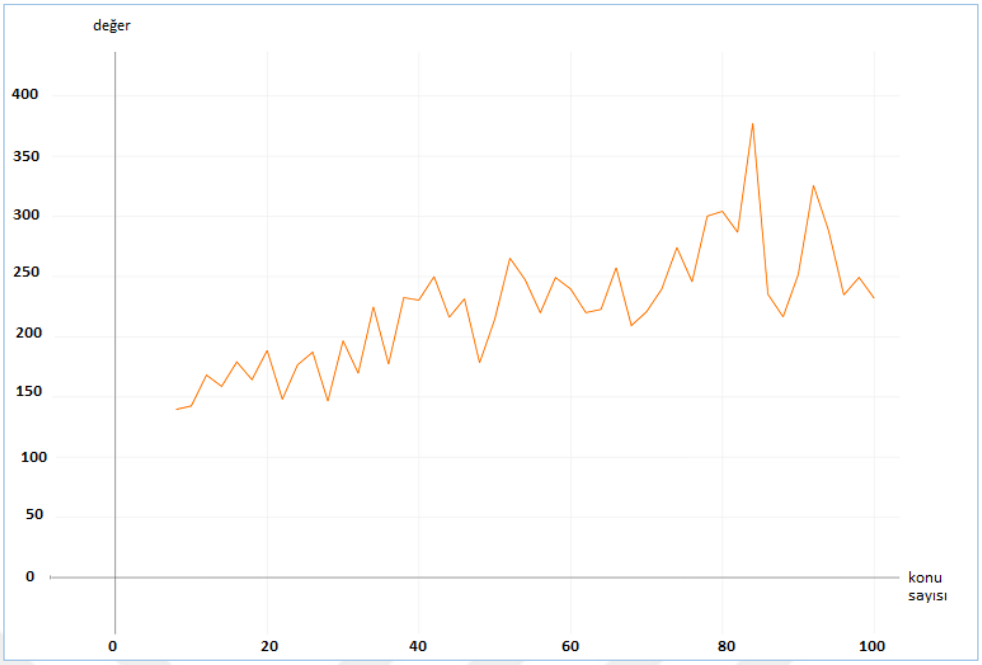
3.4.1 Perplexity Değeri Hesaplanması

Modelleri değerlendirmek ve konu sayısını belirlemek için dışarıdan bir test setinin sapma değeri hesaplanmaktadır. Dil modellemesinde kullanılan değer, test verilerinin olasılığı bakımından monoton olarak azalmaktadır. Sapma değeri, matematiksel olarak kelime başına olasılığın geometrik ortalamasının tersine eşittir. Daha düşük bir sapma skoru, daha iyi genelleştirme performansını işaret etmektedir [37].

M dokümanının bir test seti için sapma değeri:

$$perplexity(x_{test}) = \exp\left(-\frac{log\ p(x_{test})}{N_{test}}\right) \quad (3.1)$$

(3.1)'de x_{test}, test kümesini ve N_{test} test kümesinin boyutunu göstermektedir. Daha düşük bir sapma değeri daha iyi bir nitelik bildirmektedir.



Şekil 3. 3 Örnek perplexity grafiği

Şekil 3.3'te plotly[1] aracılığıyla çalışmada uygulanan örnek bir sapma grafiği çizme işlemi gösterilmiştir. Burada Türkçe haberler veri setinin VS-Z ile oluşturulmuş hali kullanılmıştır. 10'dan 100'e kadar (10, 15, …. 100) modelin sapma değerleri hesaplanarak plotly sitesi aracılığıyla çizimi elde edilmiştir. Sapma değerinin düşük olduğu konu sayısına göre sistem eğitilmiştir. Sonrasında tutarlık değeriyle kıyaslanması yapılmıştır.

3.4.2 Coherence Değeri Hesaplanması

Konu tutarlık (coherence) ölçütleri, konu içindeki yüksek ağırlıklı kelimeler arasındaki semantik benzerlik derecesini ölçerek tek bir konuyu puanlandırmaktadır. Bu ölçümler, semantik olarak yorumlanabilen konulardan ve istatistiksel çıkarım yapılacak olan konular arasında ayrım yapmaya yardımcı olur. GDA yöntemi için tasarlanmış UCI ve UMass iki yeni tutarlık ölçütü göz önüne alınmaktadır. Ölçütlerin her ikisinin de konu niteliğindeki insani kararlarla uyuştuğu gösterilmiştir.

Her iki ölçüt, bir konunun tutarlılığını, V konu kelime dizileri üzerindeki çift yönlü dağılım benzerlik skorlarının toplamı olarak hesaplamaktadır. Genelleştirmek gerekirse;

[1] https://plot.ly/python/getting-started/#initialization-for-offline-plotting

$$coherence(V) = \sum_{(v_i, v_j)\ \epsilon\ V} score(v_i, v_j, \varepsilon) \quad (3.2)$$

(3.2)'de V, konuyu açıklayan bir kelime dizisidir ve ε ağırlığın reel sayıları döndürmesini garantileyen düzeltme faktörünü belirtir ($\varepsilon = 1$).

- UCI ölçütü, bir sözcük çiftinin skorunu iki sözcük arasındaki noktasal karşılıklı bilgi (PMI) olarak tanımlamaktadır.

$$score(v_i, v_j, \varepsilon) = \log \frac{p(v_i, v_j) + \varepsilon}{p(v_i) * p(v_j)} \quad (3.3)$$

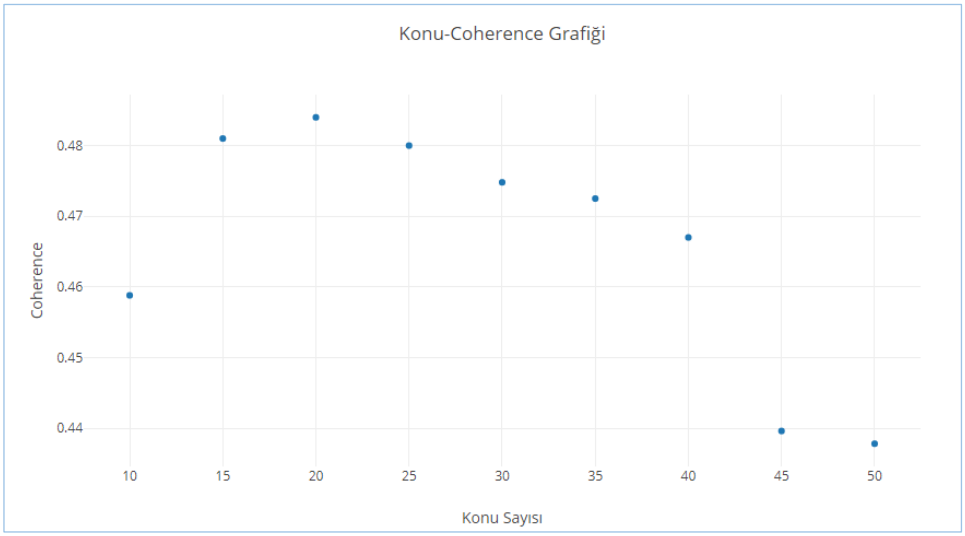
  Kelime olasılıkları, harici bir derlem üzerinde kayan bir pencerede kelime ortak oluşum frekanslarını sayarak hesaplanmaktadır.

- UMass ölçütü, ortak oluşan dokümana dayanarak elde edilen skor ile tanımlanmaktadır.

$$score(v_i, v_j, \varepsilon) = \log \frac{D(v_i, v_j) + \varepsilon}{D(v_j)} \quad (3.4)$$

  (3.4)'te D(x, y), x ve y sözcüklerini içeren dokümanların sayısını, D(x) ise x içeren dokümanların sayısını saymaktadır. UMass metriği, bu sayıları harici bir derlem yerine konu modellerini eğitmek için kullanılan orijinal derlem üzerinden hesaplar. Bu ölçüt doğada daha gerekçidir. Modellerin, derlemta olduğu bilinen verileri öğrendiğini doğrulamaya çalışmaktadır [44].

Hesaplama yapılırken kaydırma penceresinin boyutu UCI için 10 iken, Umass için önemsizdir. Tutarlık değeri; kelimelerin birbirine benzerliğini ölçer ve seçilecek olan konu sayısı hakkında bize bilgi vermektedir. Belirtilen konu sayıları için hesaplanan tutarlık değerleri arasından en yüksek çıkana ait konu değeri olan *k* değeri, konu sayısı olarak seçilmektedir. Sistemde bir önceki bölümde bahsedilmiş olan her konu sayısı için tutarlık değeri GDA kütüphanesindeki modeller aracılığıyla hesaplanmıştır. Pencerenin boyutu kütüphanede varsayılan değer olan 110 seçilmiştir. Örnek kullanılmak üzere Türkçe Tivitlere ait VS-Z veri setinde 5 sınıf için tutarlık değeri sıralanması Şekil 3.4'te gösterilmektedir.

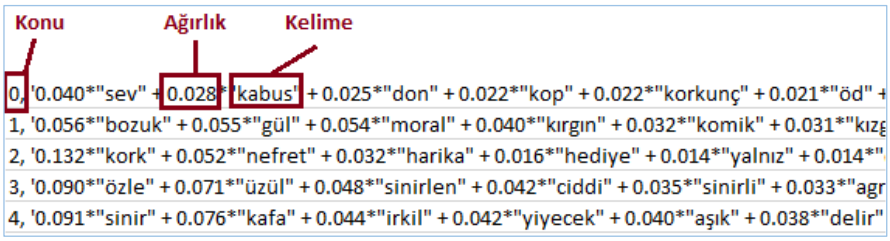


Şekil 3. 4 Tutarlık değeri örnek sıralama

Şekil 3.4'te görüldüğü üzere en yüksek değer 20 konu sayısına aittir. Tutarlık değeri, sistemin 20 konu sayısı için daha iyi sonuç vereceğini öngörmektedir.

3.5 Uygun Modelin Seçilmesi

Konu sayısı, tutarlık değerine göre seçildikten sonra model o konu sayısında eğitilmektedir. GDA modeli, bize kelime ve ağırlık değerinden oluşan sonuçları göstermektedir.



Şekil 3. 5 Türkçe tivit veri seti için örnek konu gösterimi

Şekil 3.5'te duygu veri setinden 20 konu sayısı ile oluşturulan modelden örnek alınmıştır. Görüldüğü gibi konu ve o konuya ait kelimeler ağırlıklarıyla beraber gösterilmektedir. Bu bilgilerden yararlanarak konuları belirlenen duygu sınıflarınca etiketleme işlemi yapılmaktadır. Konulara uygun duygu etiketi verme işleminde içerisinde farklı duyguları ifade eden kelimeler de yer alabilmektedir. Böyle durumlarda konudaki kelimelerin ağırlık değerlerine bağlı olarak hangi duyguyu içerdiğine önem verilmektedir.

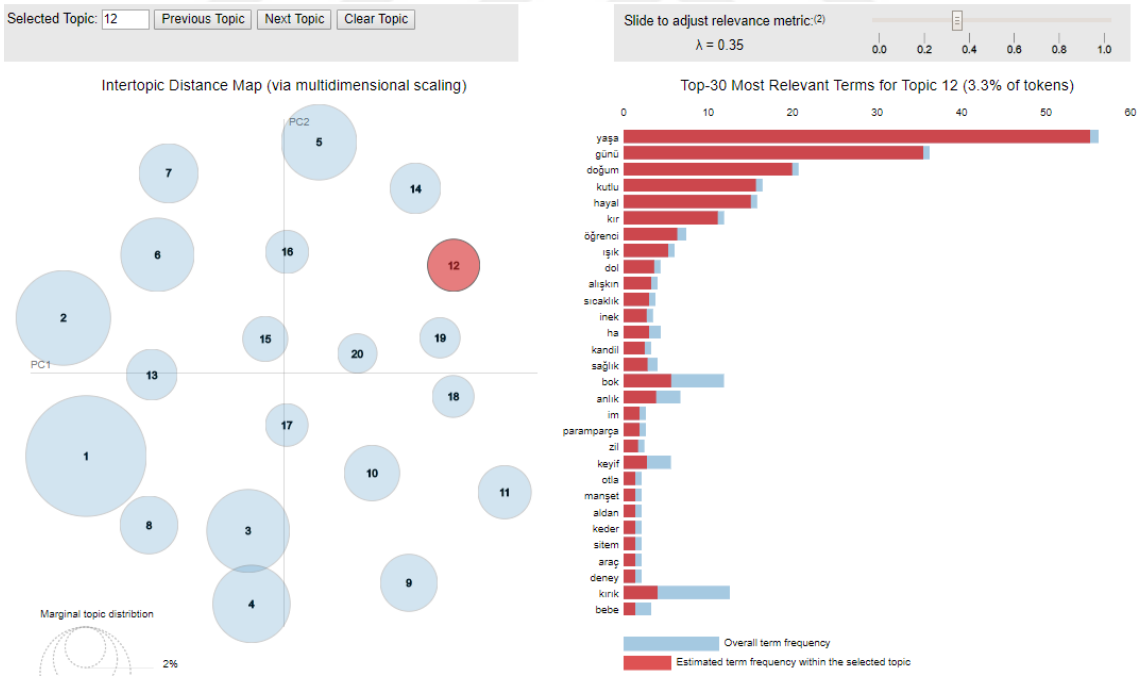
Çizelge 3. 1 Türkçe tivit VS-Z verileri üzerinde örnek modelin etiketlenmesi

KONU SAYISI	KELİMELER VE AĞIRLIKLARI	ATANAN DUYGU ETİKETİ
2	'0.132*"kork" + 0.052*"nefret" + 0.032*"harika" + 0.016*"hediye" + 0.014*"yalnız" + …	Korku
7	'0.091*"film" + 0.063*"korku" + 0.055*"düş" + 0.055*"heyecan" + 0.039*"umar" + …	Korku
4	'0.091*"sinir" + 0.076*"kafa" + 0.044*"irkil" + 0.042*"yiyecek" + 0.040*"aşık" + …	Kızgın
8	'0.094*"boz" + 0.053*"sinir" + 0.051*"kalbi" + 0.032*"bozul" + 0.027*"dal" + …	Kızgın
6	'0.235*"mutsuz" + 0.113*"hüzün" + 0.031*"hasta" + 0.023*"tatlı" + 0.019*"halim" + …	Üzgün
9	'0.119*"yor" + 0.115*"tüken" + 0.109*"bık" + 0.074*"can" + 0.073*"yıpran" + …	Üzgün
14	'0.161*"yaşa" + 0.103*"günü" + 0.058*"doğum" + 0.046*"kutlu" + 0.044*"hayal" + …	Mutlu
18	'0.121*"beğen" + 0.103*"zor" + 0.066*"hadi" + 0.044*"laf" + 0.032*"manyak" + …	Mutlu
17	'0.201*"hayret" + 0.188*"şaşır" + 0.162*"şaşkın" + 0.051*"aaa" + 0.035*"şok" + …	Sürpriz
19	'0.178*"inan" + 0.172*"şok" + 0.099*"şaş" + 0.086*"oha" + 0.055*"vay" + …	Sürpriz

Örnek olması açısından Çizelge 3.1'de 5 sınıflı duygu veri setine ait 20 konu sayısı olan modelde etiketlenen her duygu sınıfı için ikişer tane örnek verilmiştir. İncelendiğinde 2 numaralı konuya korku etiketi atanmıştır (Çizelge 3.1). Konudaki kelimeler ve ağırlıklara bakıldığında korkmayı ifade eden "kork" ve "yalnız" kelimesi gösterilebilir. Diğer kelimeler (nefret, harika, hediye) farklı duyguları ifade etse de korku duygusuna ait kelimelerin ağırlığı konuyu belirlemiştir.

Tüm veri setlerinde (Türkçe tivit, Türkçe haber, İngilizce haber) oluşturulan modellere de konuları atama işlemleri aynı şekilde değerlendirmeyle yapılmıştır.

Şekil 3.6'da gösterilen pyLDAvis[1] aracının konulara atama yapılması açısından da görsel olarak katkısı olmuştur. Oluşturulan modele göre çizdirilen grafik, konulara ait çemberlerden, kelimelere ait barlardan oluşmaktadır. Barlardaki açık renkli kısım kelimenin tüm dokümandaki terim frekansını verirken; koyu renkli kısım ise seçilen konuda hesaplanan terim frekansıdır. λ değeri ise; barlardaki hesaplanan değerlerin değişimini sağlamaktadır.

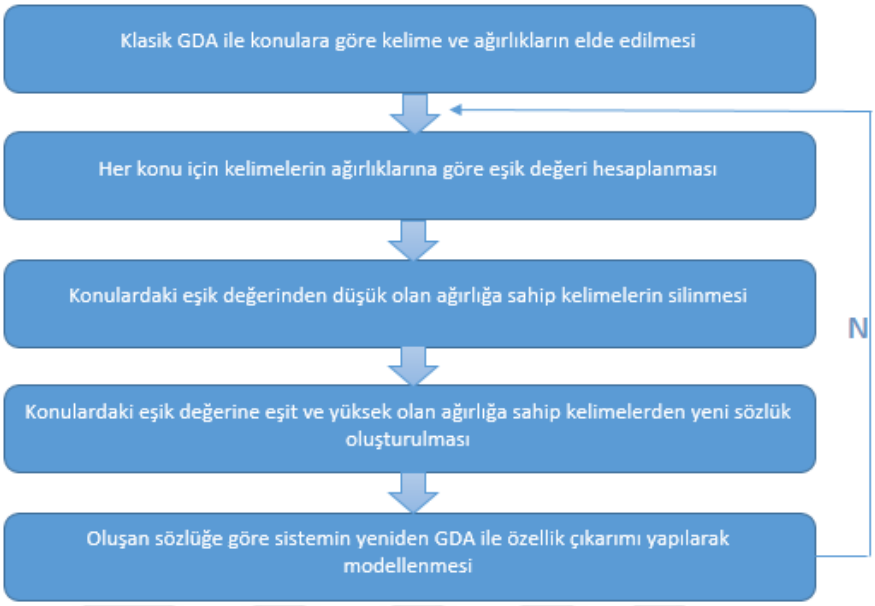


Şekil 3. 6 pyLDAvis ile çizilien Türkçe tivit veri setinde örnek grafik

[1] http://pyldavis.readthedocs.io/en/latest/index.html

3.6 N-seviyeli GDA Yönteminin Geliştirilmesi

Geliştirilen N-seviyeli GDA yönteminin işleyişi Şekil 3.7'de aşama aşama gösterilmiştir. Yöntem aynı işleyişte N defa uygulanabilmektedir.



Şekil 3. 7 N-seviyeli yöntemin işleyişi

Çalışmada oluşan modelin başarısını artırmak için yöntem üzerinde yeni geliştirmeler yapılmıştır. Yöntem ile beraber kelimeler ve ağırlıkları kullanılarak hesaplanan eşik değeriyle, tüm dokümana ait sözlükteki kelime sayısının azaltılması amaçlanır. Sözlükteki kelime sayısını azaltmanın nedeni, modelde etiketlenen konulardaki ağırlığı düşük olan ve yanlış sınıflandırmaya neden olan kelimelerin olmasıdır. N seviyede gerçekleştirilebilen bu yöntem, her aşamada daha az kelimeyle ağırlıklandırma yaparak, başarıyı artırmayı amaçlamaktadır. N değeri; sistemdeki veri setinin büyüklüğüne göre değişebilir.

$$ed(k_i) = \frac{\sum_{j=1}^{m} w_j}{n_i} \qquad i = 1,2,\ldots,k \; adet \; konu \tag{3.5}$$

$$w_j \geq ed(k_i)$$

Yöntemi açıklamak gerekirse;

- Bir konu içindeki bulunan kelimelerin toplam ağırlık değerinin ($\sum_j w_j$), toplam kelime sayısına (n_i) oranlanmasıyla bir eşik değeri elde edilir. (3.5)'deki $ed(k_i)$, i. konuya ait eşik değerini ifade etmektedir.

- Konuya ait eşik değerden eşit veya yüksek olan ağırlığa sahip kelimelerle (w_j) sözlük yeniden oluşturulmaktadır.

Geliştirilen yöntemin uygulanmasıyla kelime sözlüğünün boyutunda gözle görülür şekilde bir düşüş gözlenmektedir. Yeni kelime sözlüğü kullanılarak sistemde modeller tekrar oluşturulmuştur. Yeni modellerin tutarlık değeri hesaplanarak konu sayısı tekrardan belirlenir. Belirlenen yeni konu sayısıyla etiketleme işlemi tekrardan yapılır. Kelimelerin ağırlık değeri; geliştirilen modelde ki N değeri artış göstermektedir. Bunun nedeni kelimelerden oluşan sözlüğün boyutunun azalmasıdır.

Çizelge 3. 2 Türkçe tivit veri setinde kök bulma araçlarına göre N-GDA için sözlükteki kelime sayısı (5 sınıf için)

AŞAMA \ KÖK BULMA ARAÇLARI	VS-Z	VS-5	VS-S
N = 1-aşamalı	2208	4043	4291
N = 2-aşamalı	359	593	443
N = 3-aşamalı	309	168	149

Çizelge 3.2 incelendiğinde kök araçlarına göre 5 sınıflı veri setindeki sözlükteki kelime sayıları aşama olarak gösterilmiştir. Geliştirdiğimiz N-seviyeli GDA modeline göre dokümanda geçen kelime sayısı, aşama değeri arttıkça azalmaktadır. Aşama değeri arttıkça sözlükteki kelime sayısındaki düşüş ile çok sayıda gereksiz ve ağırlığı düşük kelimelerin silinmesi gerçekleşmiştir.

Çizelge 3. 3 Türkçe tivit VS-Z veri setinde “kork” kelimesinin aşamalardaki ağırlıkları

AŞAMA SAYISI (N)	AĞIRLIK DEĞERİ
1-aşamalı	0.132
2-aşamalı	0.343
3-aşamalı	0.688

Türkçe tivitlerin VS-Z'ye kök alma aracına göre oluşturulan veri setindeki "kork" kelimesinin ağırlığının artması Çizelge 3.3'te örnek olması açısından gösterilmiştir. Böylece veri setindeki ilgili konuların, test aşamasındaki verilere daha doğru olarak atanacağı öngörülmektedir.

3.7 Weka'da Sınıflandırma Yapabilmek için Dosya Oluşturma

GDA ile özellik çıkarım işlemi yapılıp, elle etiketlenen sistemin başarısı ölçüldükten sonra, Weka'daki sınıflandırma araçları da denenerek sistemin başarısının ölçülmesi istenmiştir. Weka'da sınıflandırma yapabilmek için GDA ile oluşturulan modeldeki konuların kelime ve ağırlıklarının kullanılması amaçlanmıştır. Dosya oluşturma işlemini adım adım açıklamak gerekirse;

- GDA ile uygun konu sayısında model oluşturulmuştur.
- Oluşan GDA modeline göre konuların kelime ve ağırlıklarının incelenmesi amaçlanmıştır.
- Eğitim ve test veri setlerinin birleştirilip, tek veri seti oluşturulmuştur. Veri setinde cümlenin hangi sınıfa ait olduğu da belirtilmiştir.
- Yeni veri setindeki her cümleye ait kelimelerin GDA modelindeki konularda bulunan kelime-ağırlıklarına bakarak her konu için ağırlıklarının bulunması sağlanmıştır.
- Cümledeki kelimelerin tek tek konulara ait ağırlıkları hesaplandıktan sonra cümlenin konulara ait toplam ağırlık değerleri hesaplanmıştır.
- Her cümlenin konu sayısı kadar ağırlıkları ve sınıf bilgileri Python programlama dilinde yazılan kod ile txt uzantılı olarak dosyaya kaydedilmiştir. Ardından txt uzantılı dosyadaki veriler csv uzantılı dosyaya aktarılmıştır.
- Csv uzantılı dosyada arff uzantılı dosyaya ikuz[1] web sitesi aracılığıyla dönüştürülmüştür ve Weka kullanımı için uygun hale getirilmiştir.

[1] http://ikuz.eu/csv2arff/

3.8 Sınıflandırma İşlemi

Weka içerisinden eğitici yöntemlerden olan Naive Bayes, Multinomial Naive Bayes, Rastgele Orman, Destek Vektör Makineleri ve Multilayer Perceptron algoritmaları ile sistemin başarısı ölçülmüştür. Sınıflandırmada k-katlı çapraz doğrulama kullanılmıştır. k değeri olarak 10 seçilmiştir.

BÖLÜM 4

DENEYSEL ÇALIŞMALAR

Çalışmada 3 farklı veri seti kullanılmış ve deneysel sonuçlar bu veri setlerinden elde edilmiştir. Geliştirilen yöntemin ilk deneysel çalışması Türkçe ve İngilizce haber veri setleri kullanılarak gerçekleştirilmiştir. Sonrasında sistemin ana çalışma konusu olan Türkçe tivit veri seti ile duygu tanıma yapılmıştır.

4.1 Türkçe Haber Veri Seti için Sonuçlar

Veri setlerine ön işlem adımları uygulandıktan sonra farklı kök bulma yöntemleri kullanılarak oluşturulmuş olan VS-Z, VS-S ve VS-5 veri setinin 3, 5 ve 7 sınıf için ayrı ayrı tutarlık değerleri hesaplanmıştır. Her veri seti için tutarlık değeri en yüksek çıkan değer konu sayısı olarak alınmış ve bizim sistemimizi eğitmek için kullanacağımız değer olarak kabul edilmiştir. Örnek olarak VS-Z için sınıflara göre tutarlık değeri ve bulunan konu sayıları Çizelge 4.1'de belirtilmiştir.

Çizelge 4. 1 Sınıf sayılarında tutarlık değerine göre konu sayıları (VS-Z için)

SINIF SAYISI	TUTARLIK DEĞERİ	KONU SAYISI
3	0.5207	12
5	0.5068	20
7	0.4618	21

Tüm kök alma araçlarıyla yeniden elde edilen veri setleri için bulunan konu sayıları üzerinde GDA algoritmasıyla elde edilen sistemde başarı oranı, sınıf sayısı arttıkça

beklenildiği gibi azalmaktadır. Sistemin 3, 5 ve 7 sınıf için GDA modeli ile elde edilen başarısı Çizelge 4.2'de gösterilmektedir.

Çizelge 4. 2 Kök bulma araçlarına göre sınıflarda GDA'nın başarısı (%)

SINIF \ ARAÇ	VS-5	VS-Z	VS-S
3	73.5	**81.38**	67.78
5	65.3	**69**	67
7	**53**	52.3	48

GDA ile oluşan model sonucunda haber başlıkları içerisinde yer alan sözlükteki kelime sayısını azaltmak için geliştirilen iki aşamalı bir yöntem ile sistem yeniden modellenmiştir. Bu N-seviyeli yöntemin (N=2) sonucunda sözlükteki toplam kelime sayısı yaklaşık olarak 1/4 oranına düşmektedir. VS-Z, VS-5 ve VS-S kök bulma araçlarıyla güncellenen her veri seti için yeni elde edilen kelime sözlüğü ile tüm sınıflar için tutarlık değerleri yeniden hesaplanarak uygun konu sayıları belirlenmiştir. Sistem iki aşamalı GDA (2-GDA) ile modellendiğinde alınan başarı sonuçları Çizelge 4.3'te gösterilmektedir.

Çizelge 4. 3 Geliştirilen 2-GDA yöntemi sonucunda sistemin başarısı (%)

SINIF \ ARAÇ	VS-5	VS-Z	VS-S
3	**91.66**	90.28	81.38
5	74.83	**76.5**	70.33
7	57.61	**57.62**	50.35

Modele geliştirilen 2-GDA yönteminin etkisi incelendiğinde başarı oranlarında %3 ile %18 arasında artış olmuştur.

2-GDA ile oluşturulan modelden sınıf etiketine sahip dosya oluşturularak Weka'da sistemin başarısı ölçülmek istenmiştir. Genel olarak en iyi sonuçları veren VS-Z veri seti seçilerek dosya oluşturulmuştur. Her sınıf için oluşturulan dosyada, en uygun konu sayısı ve bir adet sınıf etiketi olmak üzere özellik bulunmaktadır. 2-GDA ile özel oluşturulan

“arff” uzantılı dosyanın sınıflandırma algoritmalarındaki başarısı Çizelge 4.4’te gösterilmiştir. Sınıflandırma algoritmalarında 10-katlı çapraz doğrulama yöntemi seçilerek F-measure başarısı ölçülmüştür.

Çizelge 4. 4 2-GDA ile oluşturulan dosyanın sınıflandırma algoritmalarında başarısı (%)

ALGORİTMA \ SINIF	3	5	7
NB	88.55	79.46	66.26
MNB	89.33	80.3	67.59
RO	**95.22**	**88.06**	**83.19**
DVM	88.77	79.96	70.5
MP	92.61	82.73	75.33

4.2 İngilizce Haber Veri Seti için Sonuçlar

İngilizce haber veri setine ön işlem adımları uygulandıktan sonra UCI-P kök bulma aracıyla 3 ve 5 farklı haber sınıfına sahip güncellenmiş veri setleriyle tutarlık değerleri yeniden hesaplanmıştır ve konu sayıları GDA modeli için seçilmiştir. Seçilen konu sayısına göre sistemin başarısı Çizelge 4.5’te gösterilmiştir.

Çizelge 4. 5 Sınıf ve bulunan konu sayıları göre GDA’nın başarısı

SINIF	TUTARLIK DEĞERİ	KONU SAYISI	BAŞARI (%)
3	0.487	15	67.16
5	0.462	15	59.4

İngilizce veri setinde de sistem için geliştirilen 2-GDA algoritması uygulanmıştır. Elde edilen yeni tutarlık değerleri ve konu sayılarına göre sistem yeniden modellenerek sistemin başarısı ölçülmüştür. Sistemin başarısı Çizelge 4.6’da gösterilmiştir.

Çizelge 4. 6 GDA ve 2-GDA başarısının karşılaştırılması

SINIF	GDA (%)	2-GDA (%)
3	67.16	**71.83**
5	59.4	**63.7**

Çizelge 4.6 incelendiğinde 2-GDA'nın sistem üzerindeki olumlu etkisi olduğu gözükmektedir. GDA yöntemine göre %3-%7 arası başarı artışı sağlanmıştır.

2-GDA ile sistemin başarısı hesaplandıktan sonra 2-GDA ile çıkarılan kelimeler özellik olarak alınarak Weka'da bulunan sınıflandırma algoritmaları ile başarı ölçülmüştür. 2-GDA ile oluşturulan arff uzantılı dosyanın sınıflandırma algoritmalarındaki F-measure sonuçları Çizelge 4.7'de gösterilmiştir.

Çizelge 4. 7 2-GDA ile oluşturulan dosyanın sınıflandırmadaki başarısı (%)

ALGORİTMA \ SINIF	3	5
NB	77.56	77.18
MNB	75.8	75.48
RO	**92.63**	**92.74**
DVM	82.16	80.04
MP	86.8	83.68

Her sınıf için oluşturulan dosyada, konu sayısı kadar ve bir de sınıf etiketi olmak üzere özellik bulunmaktadır. Sınıflandırma algoritmalarında 10-katlı çapraz doğrulama yöntemi seçilerek başarı incelenmiştir. En başarılı yöntem Rastgele Orman olmuştur. Sınıf sayısı arttıkça başarının azaldığı da görülmektedir.

4.3 Türkçe Tivit Veri Seti için Sonuçlar

Türkçe tivit verileri üzerinde GDA ile özellik çıkarımı yapılarak gerçekleştirilmiş duygu sınıflandırması yapan çalışma sayısı yok denecek kadar azdır. İncelenen çalışmalarda da etiketler çoğunlukla olumlu, olumsuz ve nötrden oluşmaktadır. Ayrıca, literatürde

rastlanan çalışmalarda kullanılmış olan veri setlerine erişim de mümkün olmadığından özellik çıkarımı geliştirilen model klasik GDA ile karşılaştırılmıştır.

Ön işlem adımları veri setine uygulandıktan sonra farklı kök bulma yöntemleriyle oluşturulmuş olan VS-Z, VS-S ve VS-5 veri setinin 3 ve 5 sınıf için tutarlık değerleri hesaplanmıştır. 3 ve 5 sınıflı veri setlerinde konu sayısını bulmak için sınıf sayısı kadar artacak şekilde her veri seti için 10 tutarlık değeri hesaplanmıştır. Tutarlık değeri en yüksek olan konu sayısı, sistemi eğitmek için kullanılacak konu değeridir. Örnek olarak VS-Z ile güncellenen veri seti için sınıflara göre tutarlık değeri ve bulunan konu sayıları Çizelge 4.8'de belirtilmiştir.

Çizelge 4. 8 Sınıf sayısına göre elde edilen konu sayıları (VS-Z için)

SINIF SAYISI	TUTARLIK DEĞERİ	KONU SAYISI
3	0.4998	9
5	0.484	20

Sınıf sayılarına bakıldığında 3 sınıf için tutarlık değeri 9 konu sayısı için en yüksek değerde bulunurken, 5 sınıf için tutarlık değeri 20 konu sayısında en yüksek değere çıkmıştır (Çizelge 4.8).

Sınıf sayılarında belirlenen konu değerine göre sistem GDA modeliyle eğitilmiştir. Sistemin 3 ve 5 sınıf için GDA ile elde edilen başarısı Çizelge 4.9'da gösterilmektedir.

Çizelge 4. 9 Kullanılan kök bulma araçlarına göre GDA'nın başarısı (%)

SINIF \ ARAÇ	VS-5	VS-Z	VS-S
3	58	**65.83**	51
5	48.75	**60.375**	47

Kök bulma araçlarının tümüyle güncellenen veri setlerinde başarıları incelendiğinde 3 ve 5 sınıf içinde en iyi başarıyı Zemberek (VS-Z) vermiştir.

Türkçe tivit veri setinde yer alan sözlükteki kelime sayısını azaltmak amacıyla GDA algoritması için önerilen n aşamalı yöntem kullanılmıştır. Yeni oluşturulan sözlükteki toplam kelime sayısı eskisine oranla yaklaşık olarak 1/4 oranına düşmüştür. VS-Z, VS-5 ve VS-S kök bulma araçlarıyla güncellenen her veri seti için yeni elde edilen kelime sözlüğü ile tüm sınıflar için tutarlık değerleri yeniden hesaplanarak uygun konu sayıları belirlenmiştir. Sistem iki aşamalı GDA (2-GDA) ile modellendiğinde alınan başarı sonuçları Çizelge 4.10'da gösterilmektedir.

Çizelge 4. 10 Geliştirilen 2-GDA yönteminin başarısı (%)

SINIF \ ARAÇ	VS-5	VS-Z	VS-S
3	68.5	**80.83**	74.375
5	67.375	**70.5**	56.875

Sistemin başarısı incelendiğinde 2-GDA ile başarı GDA'ya göre artış göstermektedir. Yine geliştirilen modelde en yüksek başarı VS-Z veri setinde elde edilmiştir (Çizelge 4.10).

Modele iki seviyeli GDA (2-GDA) işlemi uygulandıktan sonra, en iyi sonucu veren VS-Z veri seti seçilerek, üç seviyeli GDA (3-GDA) ile modellenmesi sağlanmıştır. 3-GDA ile sözlükteki kelime sayısı 2-GDA'dan daha da azalarak gereksiz kelimeler atılmıştır. Ardından tekrardan hesaplanan tutarlık değerleriyle konu sayıları belirlenmiştir. Çizelge 4.11'de sistemin 3-GDA ile modellenmesi sonucundaki başarısı ve diğer modeller ile karşılaştırılması gösterilmiştir.

Çizelge 4. 11 Oluşturulan modellerdeki başarı oranları (%)

METOT \ SINIF	3	5
GDA	65.83	60.375
2-GDA	80.83	70.5
3-GDA	**81.5**	**76.375**

Geliştirilen n aşamalı GDA yönteminin olumlu etkisi Çizelge 4.11'de gözükmektedir. n değerinin artmasıyla beraber başarı oranı da doğrusal olarak artmıştır. 3-GDA'da 2-GDA'ya göre 3 sınıf için yaklaşık %1 başarı artışı sağlamış iken, 5 sınıf için bu oran yaklaşık %6 olarak gözükmektedir.

Son olarak Weka'daki sınıflandırma araçları ile denenerek sistemin başarısının ölçülmesi amaçlanmıştır. Weka'da sınıflandırma yapabilmek için 3-GDA ile oluşturulan modeldeki konuların kelime ve ağırlıkları kullanılarak arff uzantılı dosya elde edilmiştir. Dosya, eğitim ve test veri setinden oluşan toplam tivit sayısında konu sayısı ve sınıf etiketi özelliklerinden oluşmaktadır. 3-GDA için 5 sınıflı veri setinde tutarlık değeri en yüksek çıkan konu sayısı 50 olduğundan, dosyada 50 konu sayısı ve 1 sınıf etiketi olmak üzere 51 özellik bulunmaktadır. 3 sınıflı veri setine ait dosya ise konu sayısı 27 ve 1 sınıf etiketi olmak üzere 28 özellik içermektedir. 3-GDA ile oluşturulan arff uzantılı dosyanın sınıflandırma algoritmalarındaki F-measure başarısı Çizelge 4.12'de gösterilmiştir.

Çizelge 4. 12 3-GDA'dan elde edilen dosyanın sınıflandırmaki başarısı (%)

ALGORİTMA \ SINIF	**3**	**5**
NB	86.625	80.8
MNB	87.416	83.175
RO	**97.79**	**95.925**
DVM	87.58	85.925
MP	94.54	89.975

Çizelge 4.12'deki sınıflandırma algoritmalarının hepsi 10-katlı çapraz doğrulama yöntemiyle çalıştırılmıştır. Sistemdeki başarı oranı yüksek bir başarı oranı sağlamıştır. Yine en başarılı sınıflandırıcı Rastgele Orman'dır. Sınıf sayısı arttıkça sistemin başarısı da düşmektedir.

BÖLÜM 5

SONUÇ VE ÖNERİLER

Yapılan bu tez aşamasında konu modelleme yöntemi olarak kullanılan GDA üzerinde geliştirmeler yapılarak, klasik GDA'nın sınıflandırmaya olan etkisi arttırılmıştır.

İlk olarak Türkçe haber başlığı içeren veri setinden yola çıkarak haberlerin hangi türe ait olduğu konu modelleme algoritması olan GDA ile tespit edilmiştir. İki aşamalı olarak geliştirilen GDA ile klasik GDA yöntemi arasında %5 ile %10 arası bir başarı artışı gözlemlenmiştir. Bunun en önemli nedeni tüm belgede kullanılan kelime sayısının, az ağırlığa sahip kelimelerin silinmesinden dolayı azalmasıdır. Ardından 2-GDA'nın kelime-ağırlık değeriyle elde edilen sınıf etiketli dosyanın sınıflandırma algoritmalarına uygulanmasıyla %4-%19 arası başarı artışı olmuştur. Sınıflandırma algoritmalarında en yüksek başarı Rastgele Orman algoritması ile elde edilmiştir.

İkinci olarak İngilizce haber başlığı içeren veri setinden yola çıkarak haberlerin hangi türe ait olduğu konu modelleme algoritması olan GDA ile tespit edilmiştir. İki aşamalı olarak geliştirilen GDA ile klasik GDA yöntemi arasında %3 ile %7 arası bir başarı artışı gözlemlenmiştir. 2-GDA'nın kelime-ağırlık değeriyle elde edilen sınıf etiketli dosyanın sınıflandırma algoritmalarına uygulanmasıyla %6-%22 arası başarı artışı olmuştur. Sınıflandırma algoritmalarında en yüksek başarı Rastgele Orman algoritması ile elde edilmiştir.

Son olarak sosyal medyada atılan tivitlerden yola çıkarak hangi duygu türüne ait olduğu konu modelleme algoritması olan GDA ile tespit edilmiştir. Sistemde n (2 ve 3 için) aşamalı olarak geliştirilen GDA ile klasik GDA yöntemi arasında başarı farkı gözlemlenmiştir. Klasik GDA yönteminin başarısı 2-GDA yöntemi uygulandığında %10 ile %15 arasında artış gösterdiği gözlemlenmiştir. 3-GDA yöntemi sisteme uygulandığında

ise 2-GDA'ya göre %1 ile %6 arası başarı artışı sağlanmıştır. Ardından 3-GDA'nın kelime-ağırlık değeriyle elde edilen sınıf etiketli dosyanın sınıflandırma algoritmalarına uygulanarak %5-%18 arası başarı artışı olmuştur. Sınıflandırma algoritmalarında en yüksek başarı yine Rastgele Orman algoritması ile elde edilmiştir. Sistemdeki n değeri veri setinin büyüklüğüne göre arttırılabilir. Ayrıca, veri setindeki tivitlerde bağlantılı kelime az olması durumunda n seviye sayısını artırmak daha mantıklı olacaktır.

Gelecek konu modelleme çalışmalarında N-GDA algoritmasını; müzik türünü, yazılan metinin hangi yazar tarafından yazıldığını, soru cevap sistemlerinde doğru cevabı tespit etmek için kullanabileceği öngörülmektedir. Geliştirilen yönteme ek olarak konulardaki kelime-ağırlık değerlerinden sınıf etiketine sahip veri seti elde edilerek sınıflandırma algoritmalarındaki başarıyı da ölçebiliriz. Geliştirdiğimiz yöntem ile kelimelere her aşamada daha doğru bir ağırlıklandırma yapıldığından sisteme olumlu yönde bir katkı sağlayacağını düşünmekteyiz.

KAYNAKLAR

[1] Hasan, M., Rundensteiner, E. ve Agu, E., (2014). "EMOTEX: Detecting Emotions in Twitter Messages", Ase Bigdata/Socialcom/Cybersecurity Conference, 27-31 Mayıs 2014, Stanford.

[2] Ekinci, E. ve Omurca, S.I., (2016). "Ürün Özelliklerinin Konu Modelleme Yöntemi ile Çıkartılması", Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 9(1): 51-58.

[3] Mohammad, S. ve Kiritchenko, S., (2014). "Using Hashtags to Capture Fine Emotion Categories from Tweets", Computational Intelligence, 31(2): 301-326.

[4] Chaffar, S. ve Inkpen, D., (2011). "Using a Heterogeneous Dataset for Emotion Analysis in Text", Canadian Artifical Intelligence, 25-27 Mayıs 2011, Vancouver.

[5] Bollen, J., Mao, H. ve Pepe, A., (2011). "Modeling Public Moodand Emotion: Twitter Sentimentand Socio-Economic Phenomena" ,Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 17-21 Temmuz 2011, Barselona.

[6] Paroubek, P. ve Pak, A., (2010). "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Proceedings of the International Conference on Language Resourcesand Evaluation, 17-23 Mayıs 2010, Malta.

[7] Colace, F., Santo, M. ve Greco, L., (2013). "A Probabilistic Approach to Tweets' Sentiment Classification", Humaine Association Conference on Affective Computing and Intelligent Interaction, 2-5 Eylül 2013, Cenevre.

[8] Coban, O., Ozyer, B. ve Ozyer, G.T., (2015). "Sentiment Analysis for Turkish Twitter Feeds", In Signal Processingand Communications Applications Conference (SIU), 16-29 Mayıs 2015, Malatya.

[9] Onan, A., (2017). "Türkçe Twitter Mesajlarında Gizli Dirichlet Tahsisine Dayalı Duygu Analizi", Akademik Bilişim, 8-10 Şubat 2017, Aksaray.

[10] Çoban, Ö., ve Ozyer, G. T., (2016). "Sentiment Classification for Turkish Twitter Feeds using LDA", In Signal Processingand Communications Applications Conference (SIU), 16-19 Mayıs 2016, Zonguldak.

[11] Bolelli, L., Giles, L. ve Ertekin, Ş., (2009). "Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation", Proceedings of the 31th

European Conference on IR Researchon Advances in Information Retrieval, 6-9 Nisan 2009, Toulouse.

[12] Beyhan, H.L., (2014). Sosyal Medya Üzerinden Metin Madenciliği ve Duygu Analizi ile Pazar Değerlendirme, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[13] Evirgen, E., (2016). Sentiment Analysis of Turkish Tweets, Yüksek Lisans Tezi, Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[14] Al-Bndi, M.S., (2015). Sentiment Analysis and Opinion Mining via MicroBlogging in Social Media Like: Twitter, Yüksek Lisans Tezi, Çankaya Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

[15] Giriş, S.F., (2014). Sentimedia: Opinion Mining and Sentiment Analysis on Social Media, Yüksek Lisans Tezi, Fatih Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[16] Çoban, Ö., (2016). Metin Sınıflandırma Teknikleri ile Türkçe Twitter Duygu Analizi, Yüksek Lisans Tezi, Atatürk Üniversitesi Fen Bilimleri Enstitüsü, Erzurum.

[17] Kama, B., (2016). Feature Based Sentiment Analysis on Informal Turkish Texts, Yüksek Lisans Tezi, ODTÜ Fen Bilimleri Enstitüsü, Ankara.

[18] Demirci, S., (2014). Emotion Analysis on Turkish Tweets, Yüksek Lisans Tezi, ODTÜ Fen Bilimleri Enstitüsü, Ankara.

[19] Boynukalın, Z., (2012). Emotion Analysis of Turkish Texts by Using Machine Learning Methods, Yüksek Lisans Tezi, ODTÜ Fen Bilimleri Enstitüsü, Ankara.

[20] Özgirgin, T., (2016). Sosyal Medyada Duygu Analizi ve Nitelik Çıkarımı, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[21] Agarwal, A., Xie, B., Vovsha, I., Rambow, O. ve Passonneau, R., (2011). "Sentiment Analysis of TwitterData", Proceedings of the Workshop on Language in Social Media, 23 Temmuz 2011, Portland.

[22] Strapparava, C. ve Mihalcea, R., (2007). "SemEval-2007 Task 14: Affective Text", SemEval '07 Proceedings of the 4th International Workshop on Semantic Evaluations, 23-24 Temmuz 2007, Prag.

[23] Wang, W., Chen, L., Thirunarayan, K. ve Sheth, A.P., (2012). "Harnessing Twitter 'Big Data' for Automatic Emotion Identification", 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, 3-5 Eylül 2012, Washington.

[24] Liu, V., Banea, C. ve Mihalcea, R., (2017). "Grounded Emotions", Seventh International Conference on Affective Computing and Intelligent Interaction, 23-26 Ekim 2017, Teksas.

[25] Roberts, K., Roach, M., Johnson, J., Guthrie, J. ve Harabagiu, S., (2012). "EmpaTweet: Annotating and Detecting Emotions on Twitter", In Proceedings of the 8th International Conference on Language Resourcesand Evaluation, 23-25 Mayıs 2012, İstanbul.

[26] Çelikyılmaz, A., Tur, G. ve Tur, D., (2010). “LDA Based Similarity Modeling for Question Answering”, Proceedings of the NAACL HLT 2010 Workshop on Semantic Search, 1-9.

[27] Li, L. ve Zhang, Y., An Empirical Study of Text Classification Using Latent Dirichlet Allocation, http://www.cs.cmu.edu/~yimengz/papers/MLReport.pdf, 20 Kasım 2017.

[28] Çelikyılmaz, A., Tur, G. ve Tur, D., (2013). “Latent Semantic Modeling for Slot Filling in Conversational Understanding”, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 26-31 Mayıs 2013, Vancouver.

[29] Titov, I. ve McDonald, R., (2008). “Modeling Online Reviews with Multigrain Topic Models”, In Proceedings of International Conference on World Wide Web, 21-25 Nisan 2008, Pekin.

[30] Lin, C. ve He, Y., (2009). “Joint Sentiment/Topic Model for Sentiment Analysis”, In Proceedings of ACM International Conference on Information and Knowledge Management, 2-6 Kasım 2009, Hong Kong.

[31] Lee, A.J.T., Yang, F., Chen, C., Wang, C. ve Sun, C., (2016). “Mining Perceptual Maps from Consumer Reviews”, Decision Support Systems, 82: 12-25.

[32] Chatterjee, R. ve Agarwal, S., (2016). “Twitter Truths: Authenticating Analysis of Information Credibility”, In 2016 International Conference on Computing for Sustainable Global Development, 16-18 Mart 2016, Yeni Delhi.

[33] Poria, S., Chaturvedi, I., Cambria, E. ve Bisio, F., (2016). “Sentic LDA: Improving on LDA with Semantic Similarity for Aspect-Based Sentiment Analysis”, International Conference on Neural Networks, 24-29 Temmuz 2016, Vancouver.

[34] Feuerriegel, S., Ratku, A. ve Neumann, D., (2016). “Analysis of How Underlying Topics in Financial News Affect Stock Prices Using Latent Dirichlet Allocation”, 49th Hawaii International Conference on System Sciences, 5-8 Ocak 2016, Koloa.

[35] Onan, A., Korukoglu, S. ve Bulut, H., (2016). “LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis”, International Journal of Computational Linguistics and Applications, 10(55): 1051-1055.

[36] Weka, Weka 3: Data Mining Software in Java, https://www.cs.waikato.ac.nz/ml/weka/index.html, 10 Şubat 2018.

[37] Blei, D.M., Ng, A.Y. ve Jordan, M.I., (2003). “Latent Dirichlet Allocation”, Journal of Machine Learning Research, 3: 993-1022.

[38] Wikipedia, Latent Dirichlet Allocation, https://en.0wikipedia.org/wiki/Latent_Dirichlet_allocation, 12 Eylül 2017.

[39] Amazon AWS, Latent Dirichlet Allocation (LDA) with Python, https://rstudio-pubs-static.s3.amazonaws.com/79360_850b2a69980c4488b1db95987a24867a.html, 15 Eylül 2017.

[40] Tina, R.P. ve Sherekar, S.S., (2013). “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification”, International Journal Of Computer Science And Applications.

[41] Rannie, J.D.M., Shih, L., Teevan, J. ve Karger, D.R., (2003). “Tackling the Poor Assumptions of Naive Bayes Text Classifiers”, Proceedings of the Twentieth International Conference on International Conference on Machine Learning, 21-24 Ağustos 2003, Washington.

[42] DnI Institute, Building Predictive Model using SVM and R, http://dni-institute.in/blogs/building-predictive-model-using-svm-and-r/, 15 Aralık 2017.

[43] Yan, H., Jiang, Y., Zheng, J., Peng, C. ve Li, Q., (2006). “A Multilayer Perceptron-Based Medical Decision Support System for Heart Disease Diagnosis", Expert Systems with Applications, 30(2):272-281.

[44] Magoosh, What Is K-Fold Cross Validation, https://magoosh.com/data-science/k-fold-cross-validation/, 12 Aralık 2017.

[45] Stevens, K., Kegelmeyer, P., Andrzejewski, D. ve Buttler, D., (2012). “Exploring Topic Coherence Over Many Models and Many Topics”, Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 12-14 Temmuz 2012, Jeju Adası.

EK-A

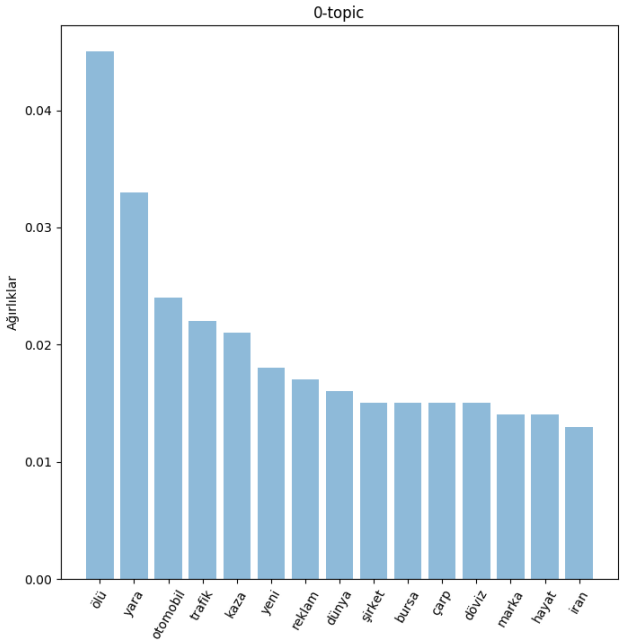
KONULARIN GRAFİKLERİ VE ETİKETLERİ

Tüm veri setleri için geliştirilen N-seviyeli GDA yönteminde en iyi sonucu veren konuların grafikleri ve konuya atanan sınıf etiketi eklerde belirtilmiştir. Her veri seti için sınıfı ifade eden 2 konu grafik örneği gösterilmiştir.

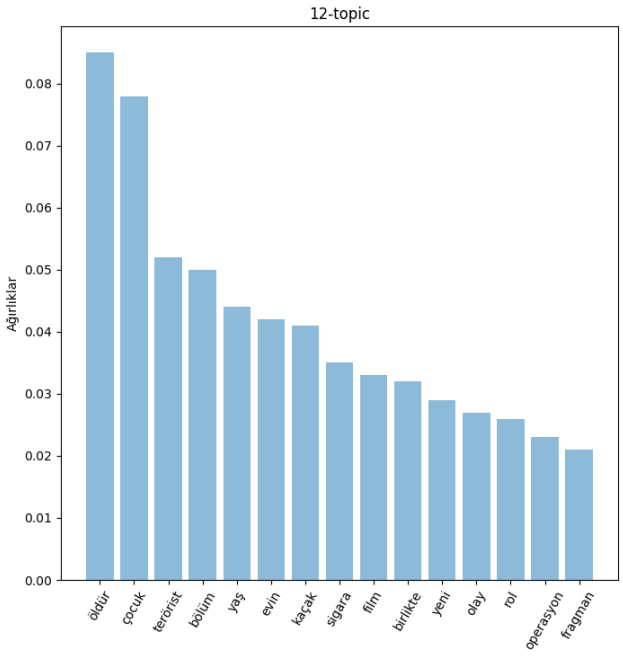
A-1 Türkçe Haber Veri Seti

"Yaşam" haberleri olarak etiketlenen konuları incelersek;

- Model sonucunda oluşan 0. konu, "ölü, yaralı, kaza" gibi kelimelerin ağırlığı fazla olduğundan yaşam haber sınıfına atanmıştır.

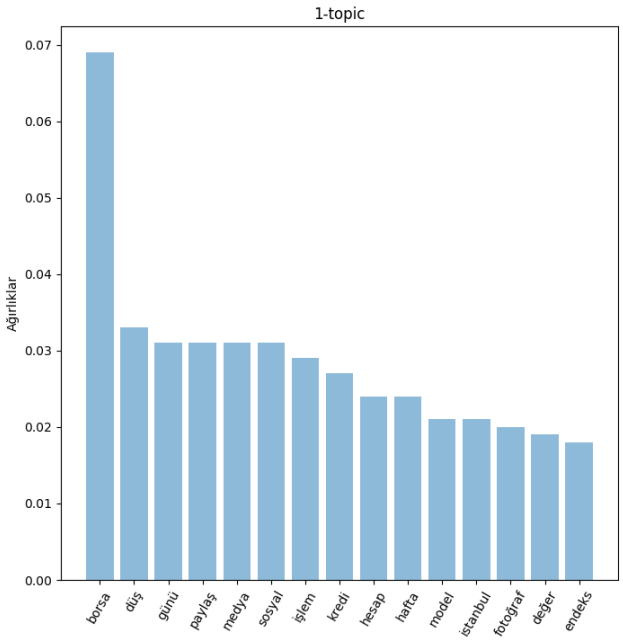


- 12. konu, “öldür, terörist, kaçak” gibi kelimelerin ağırlığı fazla olduğundan yaşam haber sınıfına atanmıştır.

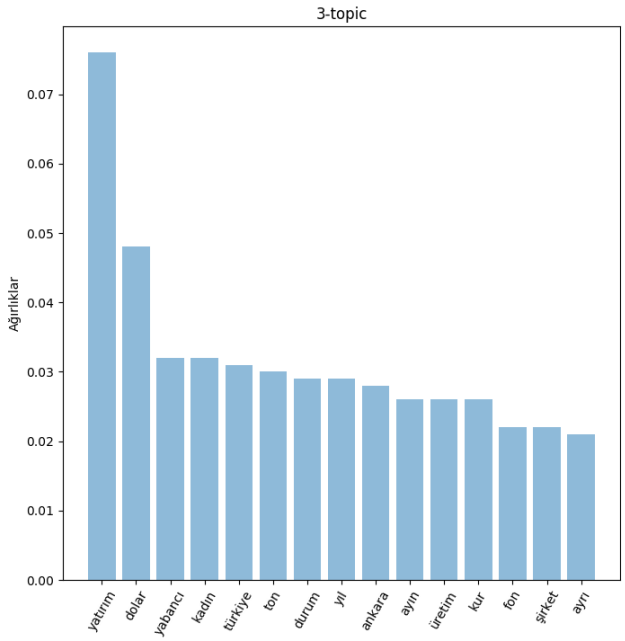


“Ekonomi” haberleri olarak etiketlenen konular;

- 1. konu, “borsa, düş, kredi” gibi kelimelerin ağırlığı fazla olduğundan ekonomi haber sınıfına atanmıştır.

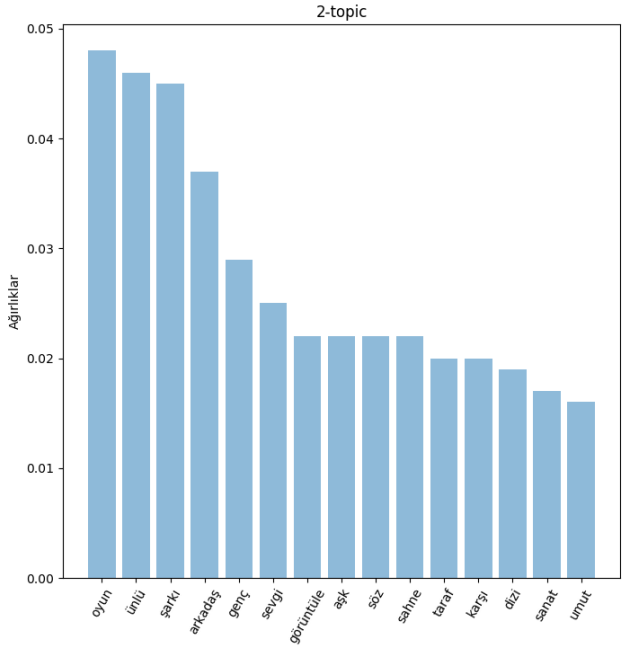


- 3. konu, “yatırım, dolar” gibi kelimelerin ağırlığı fazla olduğundan ekonomi haber sınıfına atanmıştır.

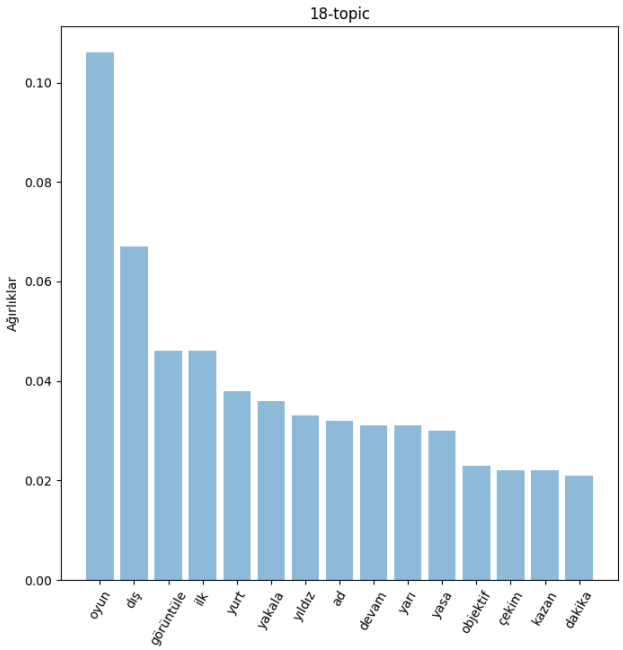


“Magazin” haberleri olarak etiketlenen konular;

- 2. konu, “oyun, ünlü, şarkıcı” gibi kelimelerin ağırlığı fazla olduğundan magazin haber sınıfına atanmıştır.

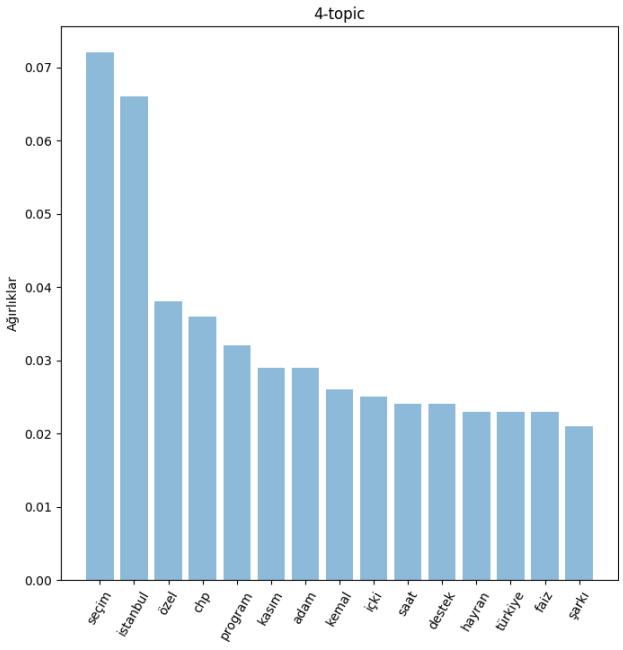


- 19. konu, “oyun, görüntüle, objektif” gibi kelimelerin ağırlığı fazla olduğundan magazin haber sınıfına atanmıştır.

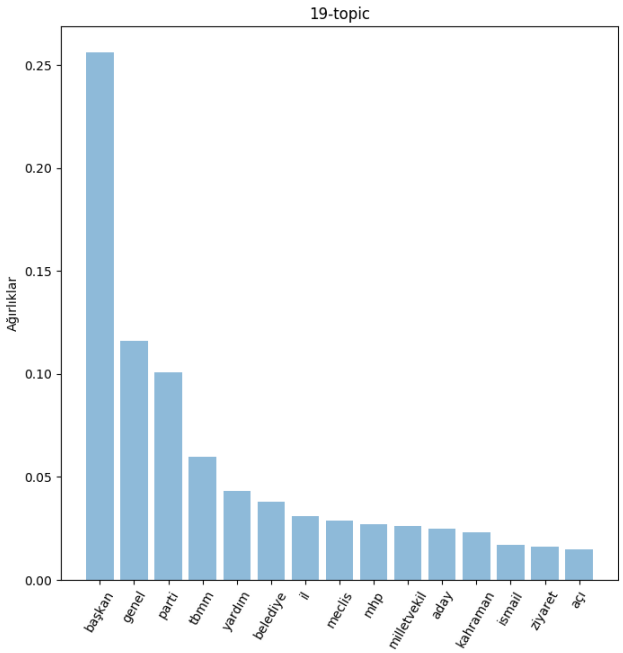


“Siyaset” haberleri olarak etiketlenen konular;

- 4. konu, “seçim, program” gibi kelimelerin ağırlığı fazla olduğundan siyaset haber sınıfına atanmıştır.

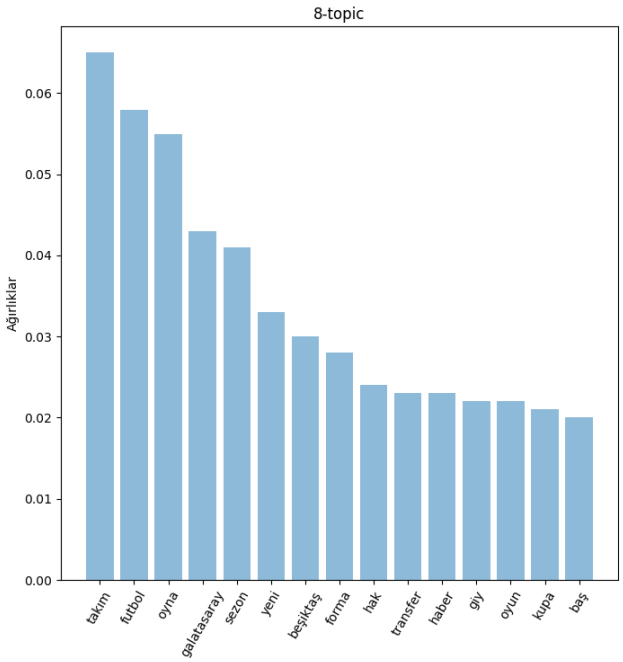


- 19. konu, “başkan, parti” gibi kelimelerin ağırlığı fazla olduğundan siyaset haber sınıfına atanmıştır.

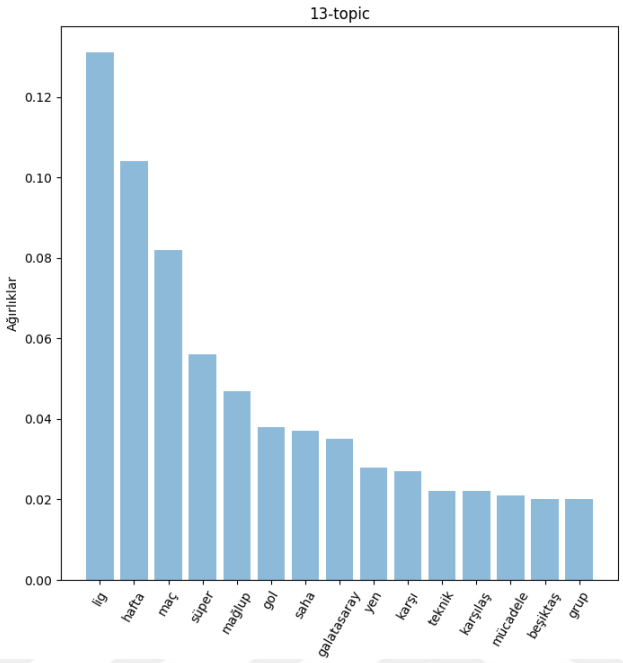


“Spor” haberleri olarak etiketlenen konular;

- 8. konu, “takım, futbol, galatasaray” gibi kelimelerin ağırlığı fazla olduğundan spor haber sınıfına atanmıştır.

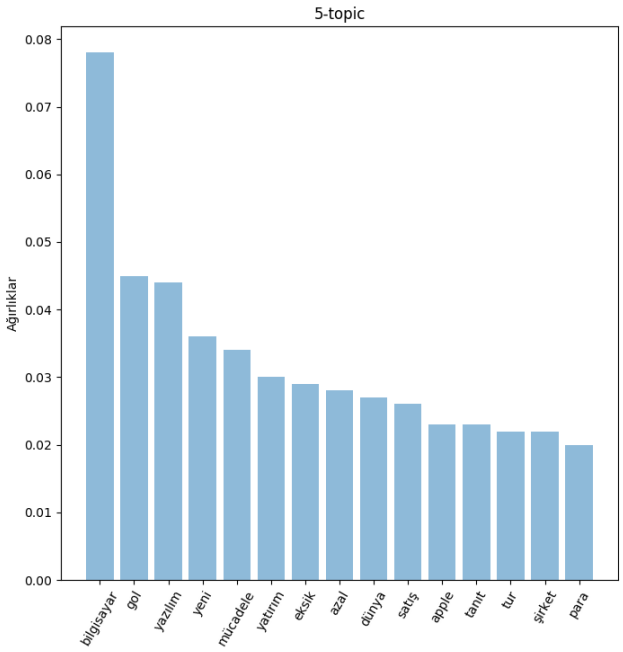


- 13. konu, “lig, maç, mağlup” gibi kelimelerin ağırlığı fazla olduğundan spor haber sınıfına atanmıştır.

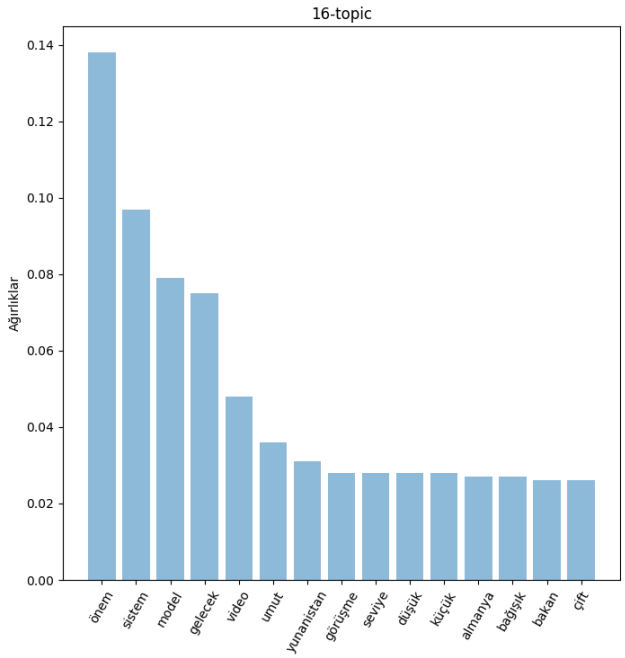


“Teknoloji” haberleri olarak etiketlenen konular;

- 5. konu, “bilgisayar, yazılım” gibi kelimelerin ağırlığı fazla olduğundan teknoloji haber sınıfına atanmıştır.

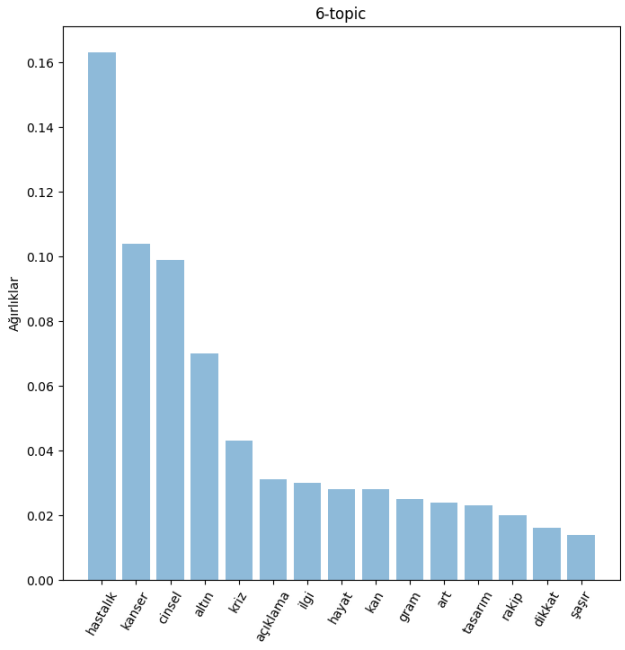


- 16. konu, “sistem, model” gibi kelimelerin ağırlığı fazla olduğundan teknoloji haber sınıfına atanmıştır.

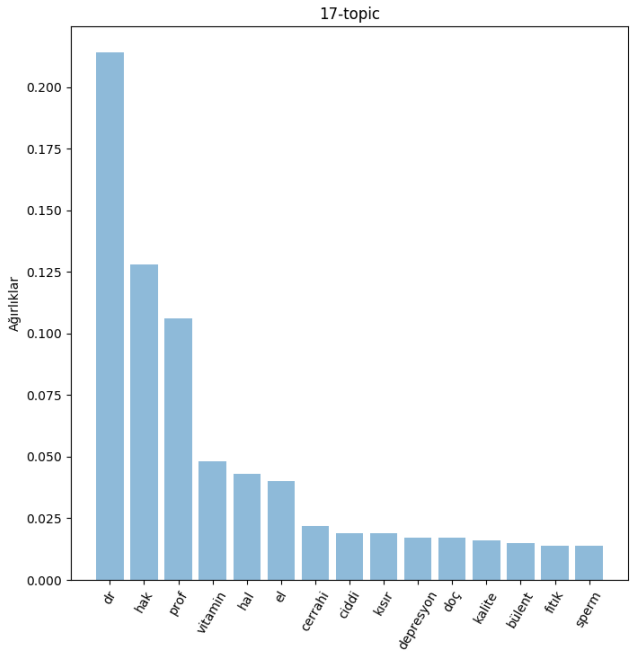


“Sağlık” haberleri olarak etiketlenen konular;

- 6. konu, “hastalık, kanser” gibi kelimelerin ağırlığı fazla olduğundan sağlık haber sınıfına atanmıştır.



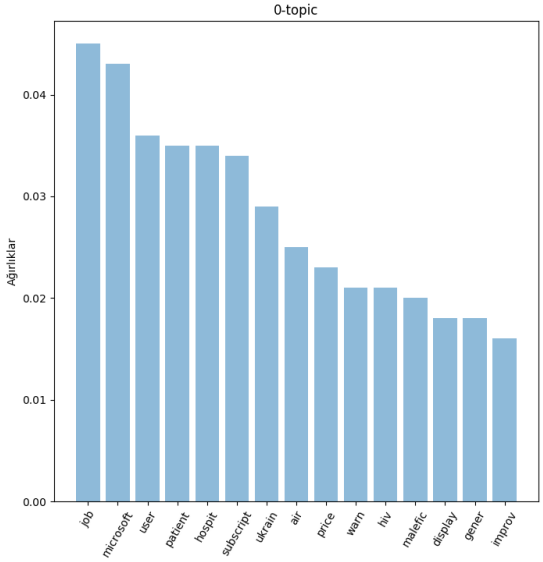
- 17. konu, “dr, prof, vitamin” gibi kelimelerin ağırlığı fazla olduğundan sağlık haber sınıfına atanmıştır.



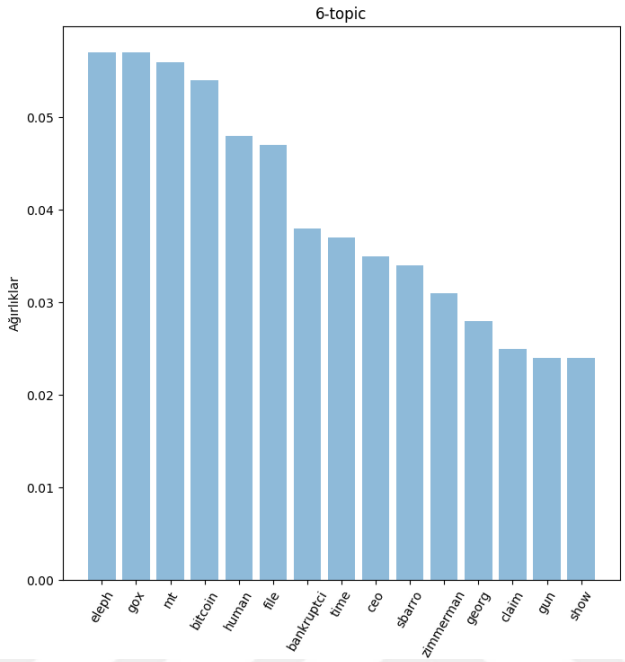
A-2 İngilizce Haber Veri Seti

“Business” sınıf etiketi olarak atanan konular;

- 0. konu, “job, microsoft, user” gibi kelimelerin ağırlığı fazla olduğundan business haber sınıfına atanmıştır.

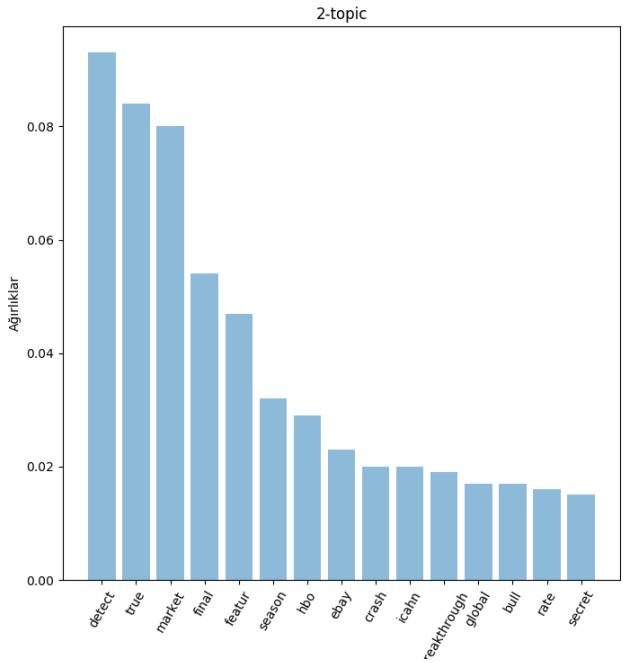


- 6. konu, “bitcoin, gox, bankrupticy” gibi kelimelerin ağırlığı fazla olduğundan business haber sınıfına atanmıştır.

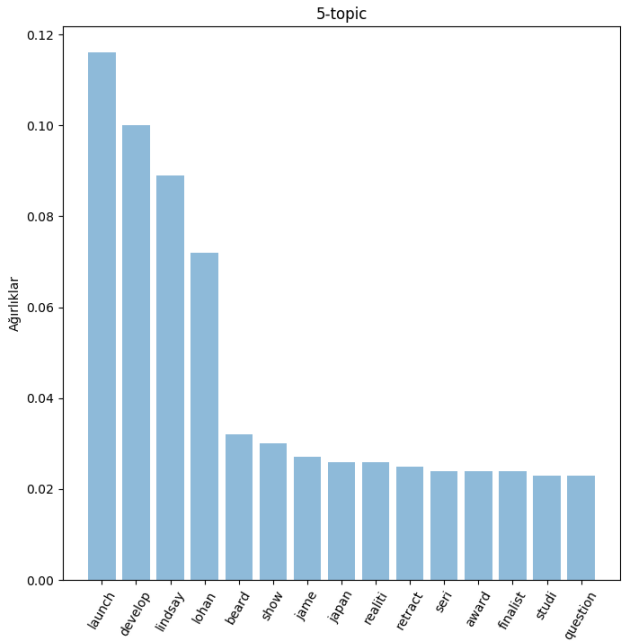


“Entertainment” sınıf etiketi olarak atanan konular;

- 2. konu, “season, detect” gibi kelimelerin ağırlığı fazla olduğundan entertainment haber sınıfına atanmıştır.

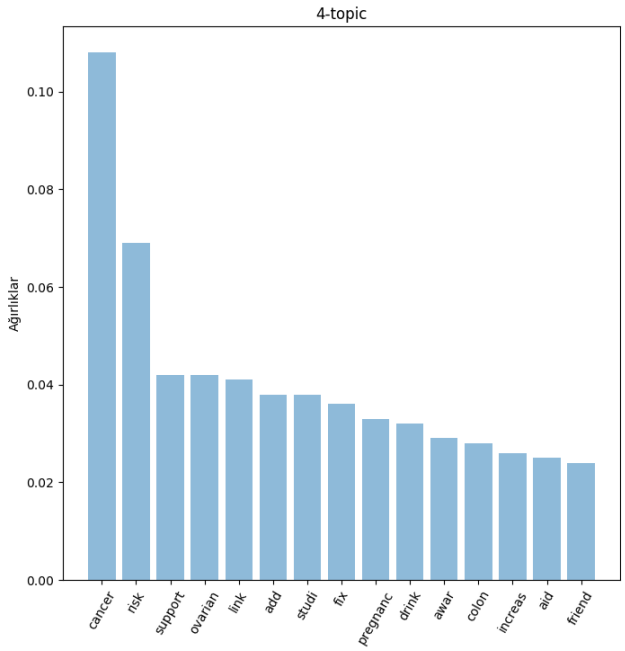


- 5. konu, “launch, show” gibi kelimelerin ağırlığı fazla olduğundan entertainment haber sınıfına atanmıştır.

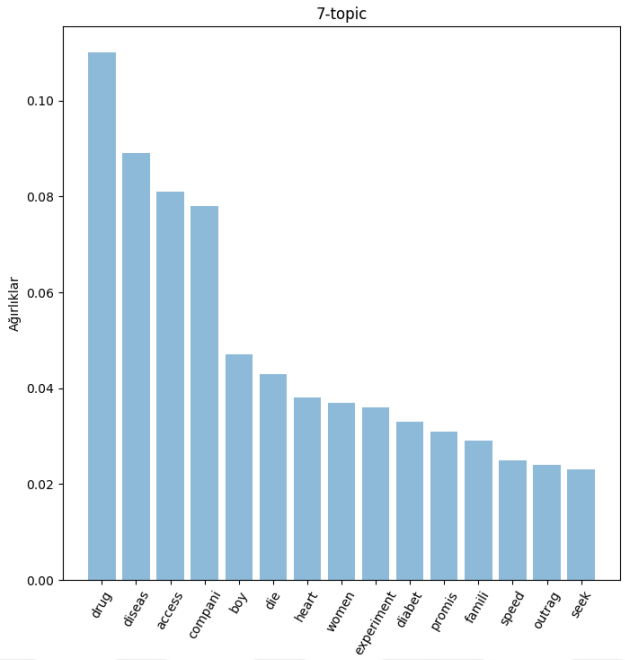


“Medicine” sınıf etiketi olarak atanan konular;

- 4. konu, “cancer, risk” gibi kelimelerin ağırlığı fazla olduğundan medicine haber sınıfına atanmıştır.

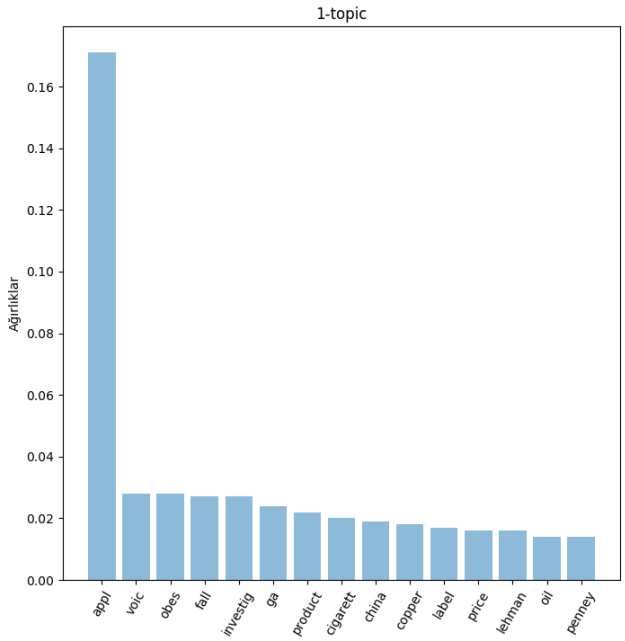


- 7. konu, “drug, disease” gibi kelimelerin ağırlığı fazla olduğundan medicine haber sınıfına atanmıştır.

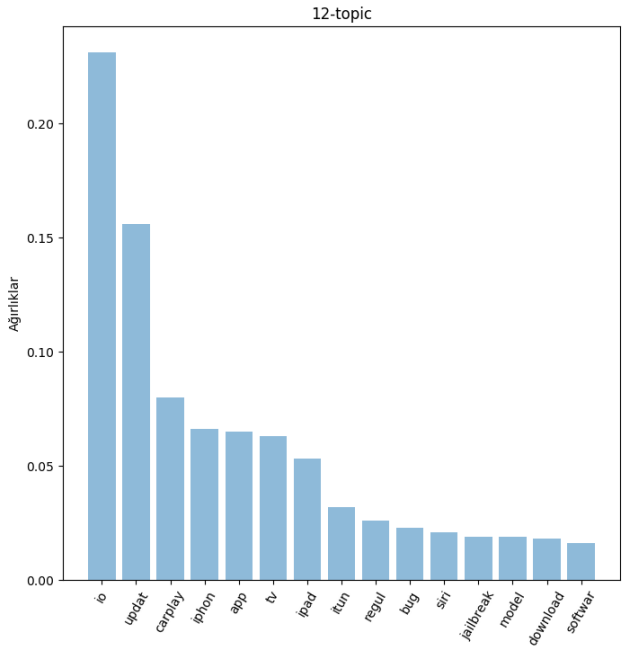


“Technology” sınıf etiketi olarak atanan konular;

- 1. konu, “apple, voice” gibi kelimelerin ağırlığı fazla olduğundan technology haber sınıfına atanmıştır.

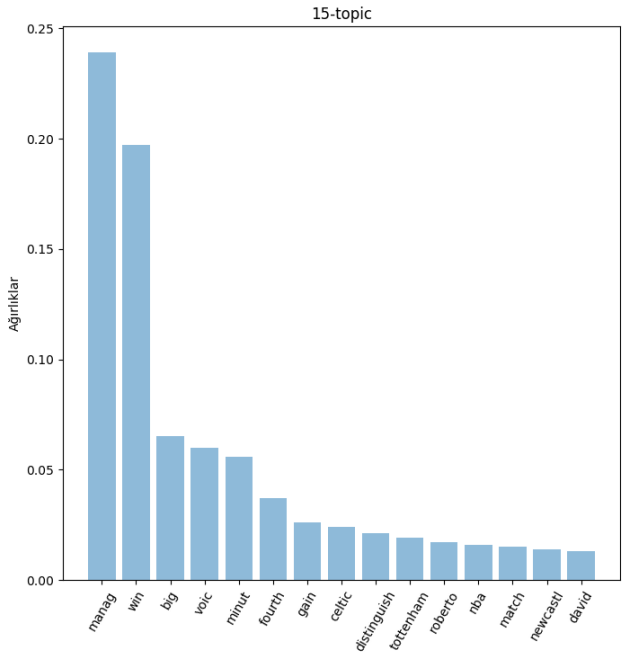


- 12. konu, “ios, update” gibi kelimelerin ağırlığı fazla olduğundan technology haber sınıfına atanmıştır.

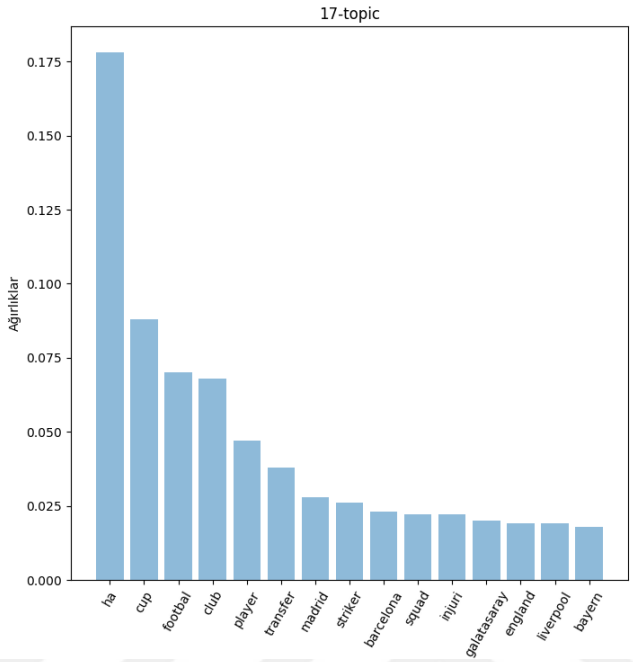


“Sport” sınıf etiketi olarak atanan konular;

- 15. konu, “manage, win, match” gibi kelimelerin ağırlığı fazla olduğundan sport haber sınıfına atanmıştır.



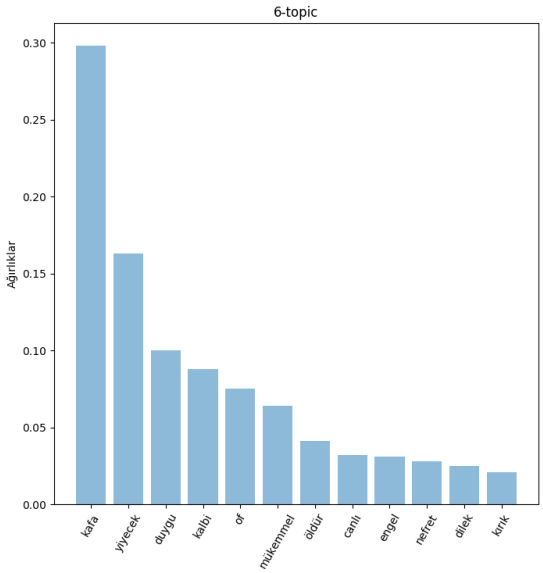
- 17. konu, “cup, football, club” gibi kelimelerin ağırlığı fazla olduğundan sport haber sınıfına atanmıştır.



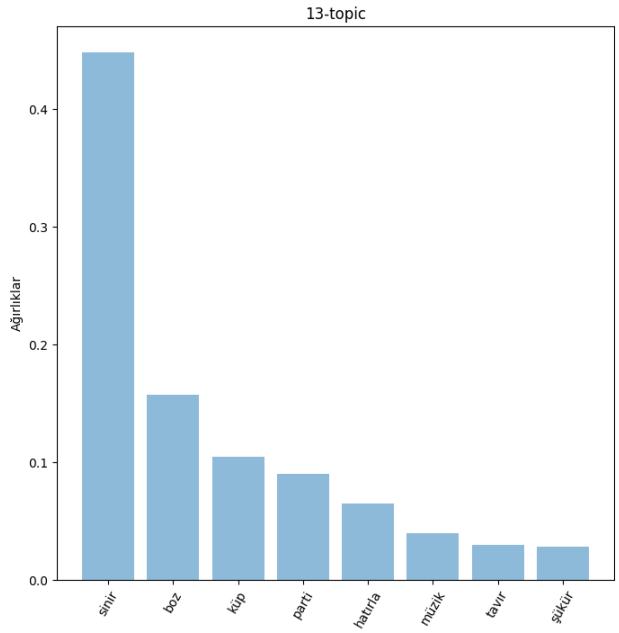
A-3 Türkçe Tivit Veri Seti

“Kızgın” duygu etiketi olarak atanan konular;

- 6. konu, “kafa, yiyecek, nefret” gibi kelimelerin ağırlığı fazla olduğundan kızgın haber sınıfına atanmıştır.

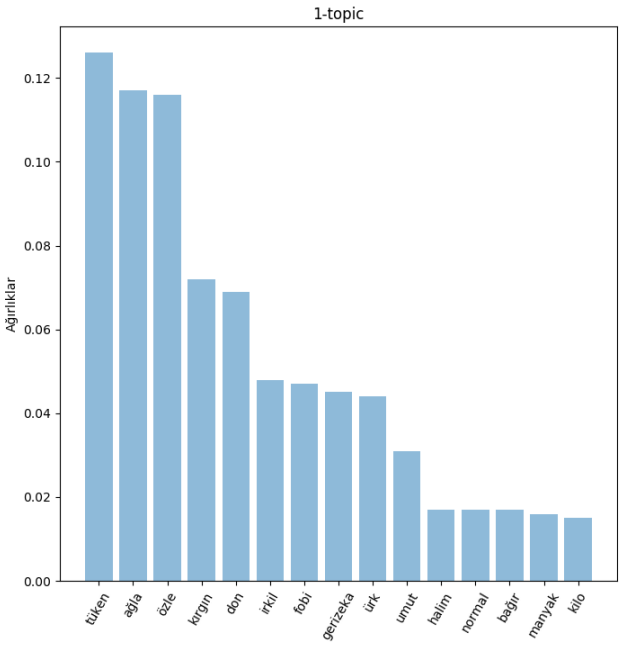


- 13. konu, “sinir, boz” gibi kelimelerin ağırlığı fazla olduğundan kızgın haber sınıfına atanmıştır.

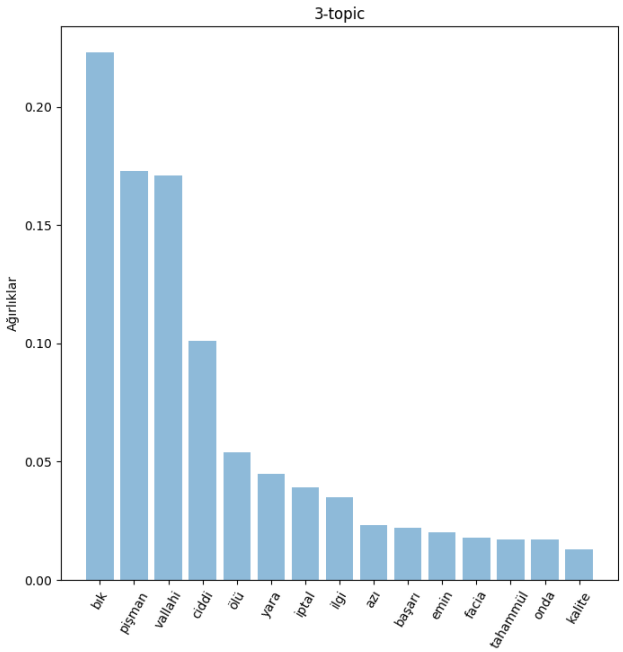


“Üzgün” duygu etiketi olarak atanan konular;

- 1. konu, “tüken, ağla, özle” gibi kelimelerin ağırlığı fazla olduğundan üzgün haber sınıfına atanmıştır.

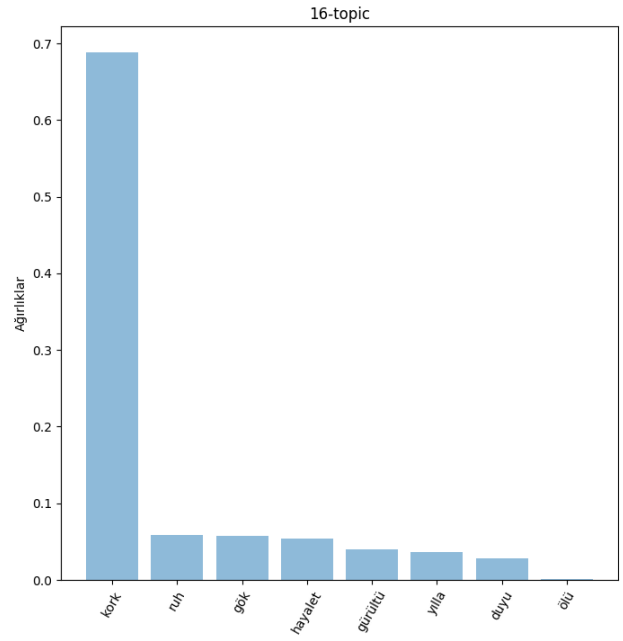


- 3. konu, “bık, pişman” gibi kelimelerin ağırlığı fazla olduğundan üzgün haber sınıfına atanmıştır.

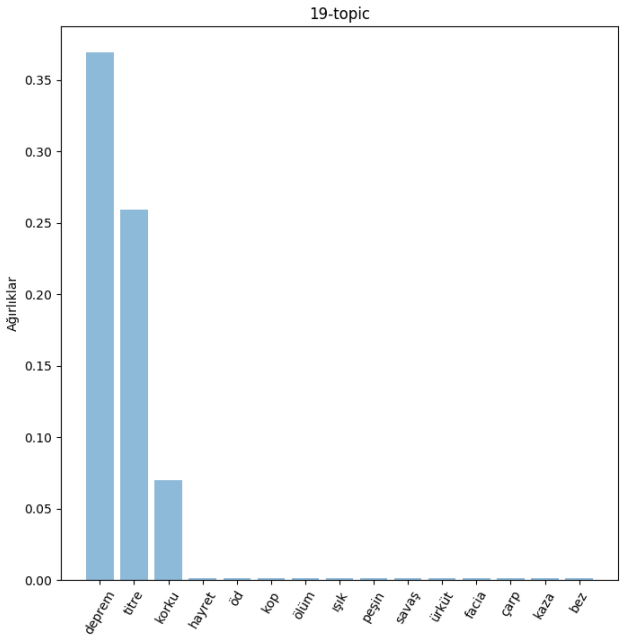


“Korku” duygu etiketi olarak atanan konular;

- 16. konu, “kork, ruh” gibi kelimelerin ağırlığı fazla olduğundan korku haber sınıfına atanmıştır.

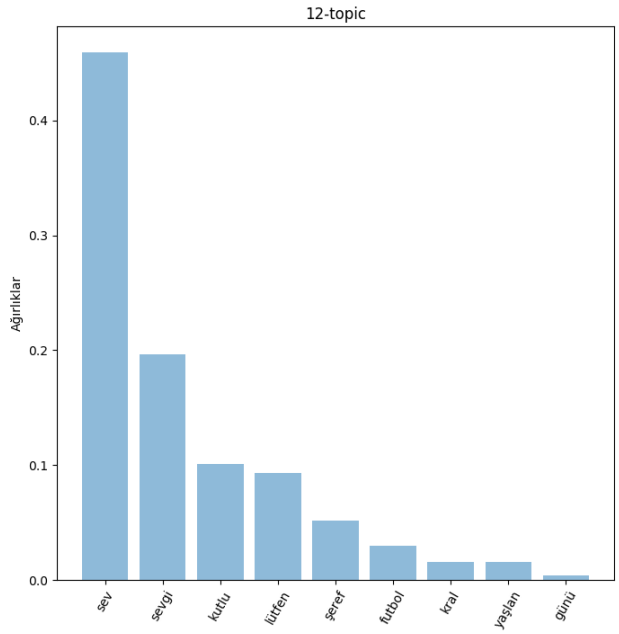


- 19. konu, “deprem, korku, titre” gibi kelimelerin ağırlığı fazla olduğundan korku haber sınıfına atanmıştır.

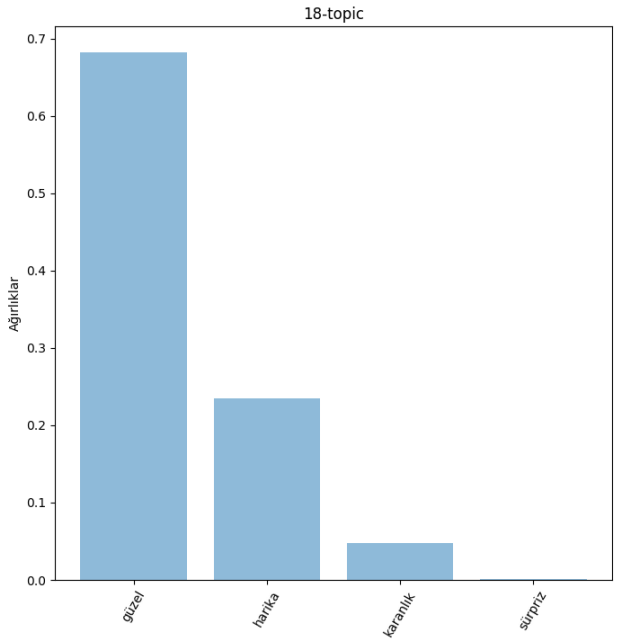


“Mutlu” duygu etiketi olarak atanan konular;

- 12. konu, “sev, sevgi, kutlu” gibi kelimelerin ağırlığı fazla olduğundan mutlu haber sınıfına atanmıştır.

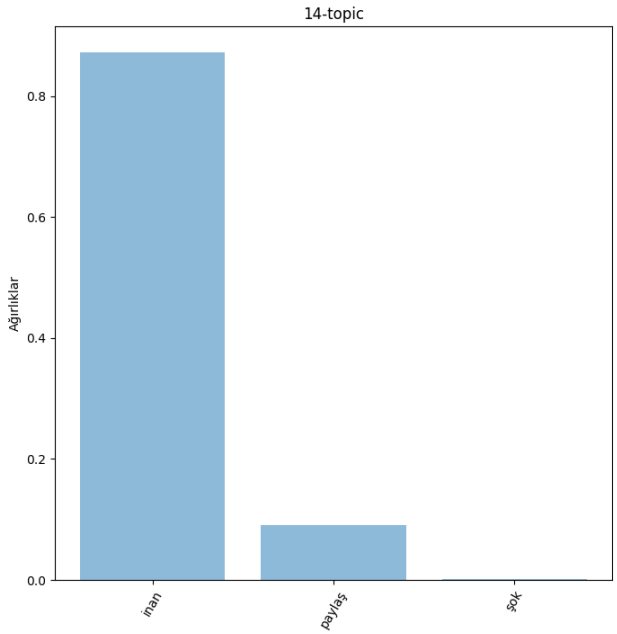


- 18. konu, “güzel, harika” gibi kelimelerin ağırlığı fazla olduğundan mutlu haber sınıfına atanmıştır.

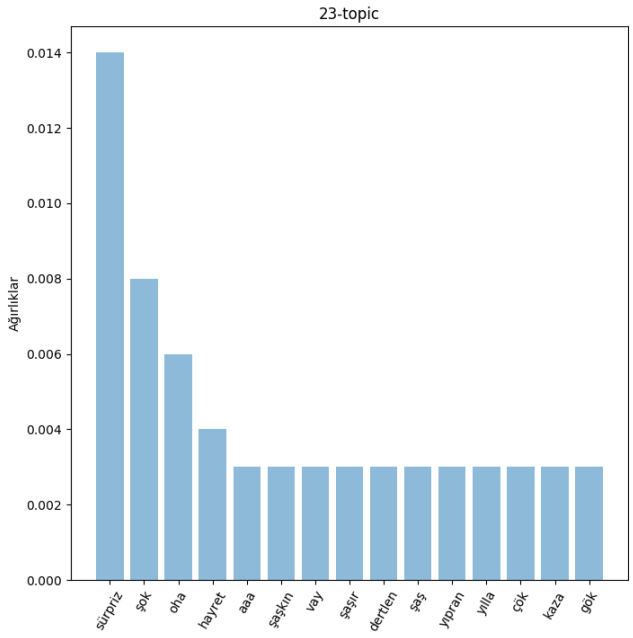


“Şaşkın” duygu etiketi olarak atanan konular;

- 14. konu, “inan (-mıyorum!)” gibi kelimenin ağırlığı fazla olduğundan şaşkın haber sınıfına atanmıştır.



- 23. konu, “sürpriz, şok, oha” gibi kelimelerin ağırlığı fazla olduğundan şaşkın haber sınıfına atanmıştır.



ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Zekeriya Anıl GÜVEN

Doğum Tarihi ve Yeri : 12/01/1992 – Havza

Yabancı Dili : İngilizce

E-posta : anilguven1055@gmail.com

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Y. Lisans	Bilgisayar Mühendisliği	Yıldız Teknik Üniversitesi	
Lisans	Bilgisayar Mühendisliği	Kocaeli Üniversitesi	2015
Lise	Fen Bilimleri	Amasya Anadolu Öğretmen Lisesi	2010

İŞ TECRÜBESİ

Yıl	**Firma/Kurum**	**Görevi**
2017	Recep Tayyip Erdoğan Üniversitesi	Araştırma Görevlisi
2016	Terapi Yazılım	Yazılım Uzmanı
2015	Finnet Elektronik	Yazılım Uzmanı

YAYINLARI

Bildiri

1. Güven, Z.A., Diri, B. ve Çakaloğlu, T., “n-seviyeli Gizli Dirichlet Ayırımı ile Türkçe Tivit Duygularının Sınıflandırılması”, EBBT 2018, 18-19 Nisan 2018, İstanbul.

2. Güven, Z.A., Diri, B. ve Çakaloğlu, T., “İki Aşamalı Gizli Dirichlet Ayırımı ile Haber Başlıklarının Sınıflandırılması”, IDAP 2018, 28-30 Eylül 2018, Malatya. (gönderildi)