T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOĞAL AFETLERDEN SONRA YÜKSEK ÖNCELİKLİ TWEET'LERİN TESBİTİ VE ÖZETLENMESİ

KADİR KEBABCI

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
DOÇ. DR. M. ELİF KARSLIGİL

İSTANBUL, 2015

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

DOĞAL AFETLERDEN SONRA YÜKSEK ÖNCELİKLİ TWEET'LERİN TESBİTİ VE ÖZETLENMESİ

Kadir KEBABCI tarafından hazırlanan tez çalışması __.__.2015 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Doç. Dr. Mine Elif KARSLIGİL

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Doç. Dr. Mine Elif KARSLIGİL

Yıldız Teknik Üniversitesi _____

Doç. Dr. Banu DİRİ

Yıldız Teknik Üniversitesi _____

Doç. Dr. Vedat COŞKUN

Işık Üniversitesi _____

ÖNSÖZ

Bu tez çalışmasında, milyonlarca kullanıcının aktif olarak paylaşım yaptıkları Twitter gibi mikroblog servislerini sosyal projelerde bilgi kaynağı olacak hale getirmek amaçlanmıştır. Bu amaç doğrultusunda sosyal projelerden en önemlisi olan, zamanın iyi kullanılması ve yerinde müdahalenin çok önemli olduğu doğal afetler sonrası yardım olduğu düşünülerek bu alanda bir çalışma gerçekleştirmeye karar verdik. Günümüzde bir çok insanın mikroblog servislerini aktif bir şekilde kullanması, paylaşımlarının bilgi kaynağı bakımından zengin ve çeşitli olması ve bu servisler arasında en çok kullanılanın Twitter'ın verilerine anlık olarak ulaşılabiliyor olması; doğal afetlerin hemen sonrasında yardım birimlerine bilgi kaynağı olabilecek bir çalışma için yeterli motivasyonu sağlamıştır. Bu doğrultuda doğal afetlerin meydana geldikten hemen sonrasında konuyla ilgili atılan tweet'lerden yüksek öncelikli olarak tanımladığımız yardım birimleri için değerli bilgi içeren tweet'lerin tespiti ve bu tweet'lerin özetlemesi gerçekleştirilmiştir.

Bu çalışma ile sosyal projeler için mikroblog servislerin kullanılabileceğini, yapılacak olan projenin amacı doğrultusunda sistemin özelleştirilmesi halinde insanların sosyal ortamlarda yaptıkları paylaşımlarının bilgi kaynağı haline getirilebileceğini gözlemledik.

Tez çalışmamın planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda ilgi ve desteğini esirgemeyen, engin bilgi ve tecrübelerinden yararlandığım, yönlendirme ve bilgilendirmeleriyle çalışmamı bilimsel temeller ışığında şekillendiren sayın hocam Doç. Dr. M. Elif KARSLIGİL'e sonsuz teşekkürlerimi sunarım.

Temmuz, 2015

Kadir KEBABCI

İÇİNDEKİLER

Sayfa

KISALTMA LİSTESİ vi

ŞEKİL LİSTESİ vii

ÇİZELGE LİSTESİ viii

ÖZET ix

ABSTRACT ix

BÖLÜM 1

GİRİŞ 1

1.1 Tezin Amacı 1
1.2 Literatür Özeti 3
1.2.1 Tweet Sınıflandırma 3
1.2.2 Tweet'lerin Özetlenmesi 4
1.2.3 Twitter Kullanılarak Yapılan Doğal Afet Çalışmaları 5
1.3 Hipotez 6

BÖLÜM 2

SİSTEM TASARIMI 7

2.1 Afet Sonrası Olaylarla ilgili Veri Toplama 7
2.2 Ön İşlemler 8
2.3 Tweet'lerin Yüksek Öncelikli ve Düşük Öncelikli Olarak Sınıflandırması . 10
2.3.1 Destek Vektör Makinesi Sınıflandırıcısı 11
2.3.2 Naive Bayes Sınıflandırıcısı 12
2.3.3 Rasgele Ormanlar (Random Forest) Sınıflandırıcısı 13
2.4 Yüksek Öncelikli Tweet'lerin Özetlenmesi 14
2.4.1 SumBasic 14
2.4.2 TF-ISF 14
2.4.3 Hibrit TF-IDF 14

BÖLÜM 3

UYGULAMA 16

3.1 Veri Toplama ve Ön İşlem 16
3.2 Yüksek Öncelikli Tweet Tespiti 18
3.2.1 Naive Bayes ile Sınıflandırma 20
3.2.2 DVM ile Sınıflandırma 21
3.2.3 Rasgele Ormanlar Yöntemi ile Sınıflandırma 22
3.2.4 Farklı Eğitim ve Test Veri Kümesi Kullanılarak Sınıflandırma 23
3.2.5 Değerlendirme 23
3.3 Yüksek Öncelikli Tweet'lerin Özetlenmesi 24
3.3.1 Veri Kümelerindeki Kullanım Sıklığı Yüksek Olan Kelimelerin Karşılaştırılması 24
3.3.2 SumBasic ile Özetleme 25
3.3.3 TF-ISF ile Özetleme 26
3.3.4 Hibrit TF-IDF ile Özetleme 27
3.3.5 Değerlendirme 29

BÖLÜM 3

SONUÇ VE ÖNERİLER 31

KAYNAKLAR 33

ÖZGEÇMİŞ 35

KISALTMA LİSTESİ

DVM	Destek Vektör Makinesi
IDF	Inverse Document Frequency
ISF	Inverse Sentence Frequency
RO	Rasgele Ormanlar
SMS	Short Messaging Service
TF	Term Frequency

ŞEKİL LİSTESİ

Sayfa

Şekil 2.1 Sistem Tasarımı 7
Şekil 2.2 Desktek Vektör Makinesi 11
Şekil 3.1 Veri Toplayıcı 16
Şekil 3.1 California Depremi Sonrası Atılan Tweet Sayısı 17
Şekil 3.2 Ön İşlemci 18

ÇİZELGE LİSTESİ

Sayfa

Çizelge 3.1 California Depremi Sonrası Saatlere Göre Tweet Sayısı 17
Çizelge 3.2 Naive Bayes ile Sınıflandırma Sonuçları 20
Çizelge 3.3 Naive Bayes Hata Matrisi 20
Çizelge 3.4 Doğrusal Çekirdek Kullanılarak Modellenen DVM Yönteminin Sınıflandırma Sonuçları 21
Çizelge 3.5 Polinomal Çekirdek Kullanılarak Modellenen DVM Yönteminin Test Sonuçları 21
Çizelge 3.6 DVM Sınıflandırıcısı Hata Matrisi 21
Çizelge 3.7 10 Ağaç Sayılı Rasgele Ormanlar Sınıflandırıcı Test Sonuçları 22
Çizelge 3.8 20 Ağaç Sayılı Rasgele Ormanlar Sınıflandırıcı Test Sonuçları 22
Çizelge 3.9 30 Ağaç Sayılı Rasgele Ormanlar Sınıflandırıcı Test Sonuçları 22
Çizelge 3.10 Rasgele Ormanlar Sınıflandırıcı Hata Matrisi 23
Çizelge 3.11 Farklı Eğitim ve Test Veri Kümesi ile Sınıflandırma Sonuçları 23
Çizelge 3.12 Genel Sınıflandırma Başarı Sonuçları 24
Çizelge 3.13 SumBasic ile Nagano Veri Kümesi Özetleme Sonuçları 25
Çizelge 3.14 SumBasic ile Nepal Veri Kümesi Özetleme Sonuçları 26
Çizelge 3.15 TF-ISF ile Nagano Veri Kümesi Özetleme Sonuçları 27
Çizelge 3.16 TF-ISF ile Nepal Veri Kümesi Özetleme Sonuçları 27
Çizelge 3.17 Hibrit TF-IDF İle Nagano Depremi Üzerinde Özetleme 28
Çizelge 3.18 Hibrit TF-IDF ile Nepal Veri Kümesi Özetleme Sonuçları 28
Çizelge 3.19 Hibrit TF-IDF ile Nepal Veri Kümesi Özetleme Sonucu 29

ÖZET

DOĞAL AFETLERDEN SONRA YÜKSEK ÖNCELİKLİ TWEET'LERİN TESBİTİ VE ÖZETLENMESİ

Kadir KEBABCI

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Doç. Dr. M. Elif KARSLIGİL

Günümüzde en çok kullanılan mikroblog servislerinden biri olan Twitter, anlık bilgi bakımından değerli bir kaynaktır. Doğal afetler sırasında kısa sürede doğru yerlere müdahalenin yapılması insan hayatı açısından büyük önem taşımaktadır. Bu çalışmada doğal afetlerden hemen sonra yazılan tweet'lerden yüksek öncelikli olanların tespit edilip özetlenmesiyle yardım birimlerine anlık doğru bilgi kaynağı sunmayı hedefleyen yeni bir sistem tasarlanmış ve gerçekleştirilmiştir. Sistemin çalışmasını değerlendirmek için doğal afetler sonrası gönderilen tweet'lerden bir veri tabanı oluşturulmuş, yaralı ve hasar durumu gibi değerli bilgiler içeren tweet'ler yüksek öncelikli, diğer tweet'ler düşük öncelikli olmak üzere iki sınıfa ayrılmıştır. Tweet'ler, ilk olarak gürültünün temizlenmesi ve sınıflandırıcıların daha başarılı şekilde değerlendirebilmesi için ön işlemden geçirilmiştir. Daha sonra Destek Vektör Makinesi yöntemi ile sınıflandırma yapılarak tweet'lerin öncelikli olup olmadıkları belirlenmiştir. Öncelikli olarak işaretlenen tweet'ler Hibrit TF-IDF yöntemi ile özetlenerek bu kümeyi en iyi temsil eden tweet'ler seçilmiştir.

Anahtar Kelimeler: Twitter, sınıflandırma, özetleme, doğal afetler

ABSTRACT

DETECTION AND SUMMARIZATION OF THE HIGH PRIORITY TWEETS AFTER NATURAL DISASTERS

Kadir KEBABCI

Department of Computer Engineering

MSc. Thesis

Adviser: Assoc. Prof. M. Elif KARSLIGİL

Twitter, being one of the most widely used micro-blog services, has been one of the best sources of instant data. It has a great significance in terms of saving lives during natural disasters by means of determining the right place for taking prompt lifesaving efforts. In this study, a system that identifies the tweets of a higher priority written immediately after natural disasters and classifies them in order to further send the right data to first responders or aid units, has been designed and realized. To evaluate how the system works, a database has been arranged containing tweets written after natural disasters and classified into two categories--tweets containing valuable information about injuries and damage being marked as those of a high priority and the other tweets marked as those of a low priority. First of all, the tweets are pre-processed to clean out the noise and evaluate the classifiers in a more successful way. Then, using the classification by means of the Support Vector Machine method, the tweets are decided on whether they are of a high priority or not. Finally, summarized by the Hybrid TF-IDF method, the high priority tweets that best represent the cluster have been selected.

Keywords: Twitter, classification, summarization, natural disasters.

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

BÖLÜM 1

GİRİŞ

1.1 Tezin Amacı

Son yıllarda, internet ortamında sosyal tabanlı içeriklerin sayısı gittikçe artış göstermektedir. Bu içeriklerin büyük bir kısmının üretildiği sosyal ağ siteleri internet ortamında düzenli kullanıcı sayılarını artırmakla kalmayıp aynı zamanda içerik zenginliği ve çeşitliliğine de katkı sağlamıştır. Sosyal tabanlı içeriklerin nispeten yeni formu olan ve normal blogların karakter sınırlaması gibi kısıtlamalar ile çok daha kısa şekli olan mikroblog servislerinin kullanımının SMS (Short Messaging Service) alt yapısına da uygun şekilde tasarlanmasıyla, kullanıcılar telefonlarından da güncellemelerini yapabilmektedir. Bu sebeple çok kısa sürede mikroblog servislerinin kullanımı popülerlik kazanmıştır. Bu servisler sayesinde kişiler istedikleri kişilerle kolayca bağlantı kurabilmekte, bağlantı kurduklarını takip edebilmekte, çok çeşitli türden bilgiler paylaşabilmektedir. Bu servisler arasında kısa sürede çok sayıda kullanıcıya ulaşan Twitter, günümüzde en aktif kullanılan mikroblog servislerinden birisidir. Twitter ile, milyonlarca kullanıcı anlık olarak bilgi ve görüş paylaşımı yapabilmektedir[1]. Sadece Türkiye'deki Twitter kullanıcı sayısı 10 milyona yaklaşmıştır. Twitter'da kullanıcılar fikirlerini "tweet" diye adlandırılan 140 karakter kısıtlamalı yazı formatında paylaşabilmektedir. Takip sisteminin mevcut olduğu bu platformda kullanıcılar takip ettiği diğer kullanıcıların tweet'lerini yeniden göndererek(retweet) kendi takipçileriyle paylaşabilmekte, beğendikleri tweet'leri favorilerine ekleyebilmektedir. Böylece bir kişinin tweet'i bir anda yüz binlerce kullanıcıya ulaşabilmektedir. Bu da twitter'ın o anki bilginin değerini ölçme konusunda sağladığı önemli faktörlerden birisidir.

Çok sayıda kullanıcının mevcut olduğu bu sosyal ağ ortamında, önemli olaylardan sonra oldukça yüksek sayıda içerik paylaşımı yapılmaktadır. Bu içeriklerde kullanıcılar olay ile ilgili anlık bilgi ve görüşlerini paylaşmaktadır. Bu durum olay ile ilgili kitle analizleri için oldukça faydalı bir platform yaratmaktadır. Doğal afetlerde bu önemli olaylar arasındadır.

Doğal afetler sonrasında yapılan acil yardımlar bir çok insanın hayatını kurtarmaktadır. Acil yardım ekiplerinin doğru zamanda doğru yerlere müdahale yapması insan hayatı açısından oldukça önemlidir. Doğru yerlere müdahale için tespitin en iyi şekilde gerçekleşmesi gerekir. Bu nedenle tespit için kullanılacak araçlar arasında insanların anlık olarak bilgi paylaşımında bulunabildiği Twitter yerini alabilir.

Doğal afetlerin hemen sonrası insanlar çevrelerinde olan bitenleri, yardıma ihtiyaç duyanları, göçük altında veya başka şekilde yardım bekleyenleri, hasar durumu vb. gibi yardım birimleri için önemli olan bilgileri Twitter üzerinden paylaşabilmektedir. Doğru yere doğru zamanda müdahale yapılabilmesi açısından paylaşılan bu bilgilerin yardım birimlerinin yararlanabileceği hale getirilmesi büyük önem arz etmektedir.

Twitter'da özellikle önemli olaylardan hemen sonra konu ile ilgili çok sayıda tweet gönderilmektedir. Bunların bir bölümü kurtarma için acil ulaşılması gereken yerler, hasar büyüklüğü, malzeme ihtiyacı gibi önemli bilgi değeri olan tweet'ler iken, çok büyük bir çoğunluğu olayla ilgili üzüntü, panik vb. duyguları, iyi dilekleri içeren mesajlar gibi bilgi içeriği açısından düşük öncelikli tweet'lerdir.

Bu çalışmada yardım ve kurtarma açısından bilgi değeri olan, doğal afet sonrası atılmış yüksek önceliğe sahip tweet'lerin makine öğrenmesi yöntemleriyle tespit edilmesine yönelik yeni bir sistem tasarlanmış ve gerçekleştirilmiştir. Twitter kullanıma sunduğu Twitter API sayesinde tweet arama ve canlı olarak dinleme olanağı sunmaktadır. Belirlenen anahtar kelimelere, kullanılan dil ve coğrafi bölgeye göre filtre oluşturulmakta ve bu filtreler ile geçmiş 8 güne kadar arama yapılabilmekte, canlı olarak o an atılan tweet'ler dinlenebilmektedir. Bu sistemde ilk olarak deprem sonrasında o depreme özel olarak belirlenen anahtar kelimeler kullanılarak kullanıcılar tarafından gönderilen tweet'ler toplanır. Ardından aralarından yüksek öncelikli tweet'ler önceden modellenmiş sınıflandırıcılar ile tespit edilir. Modelleme işlemi daha önceden toplanan deprem verileri üzerinde yüksek öncelikli tweet'leri elle etiketleyerek eğitim seti oluşturulması ve sınıflandırıcı üzerinde kullanılmasıyla yapılmaktadır.

Atılan tweet'lerin çokluğu sebebiyle çok sayıda yüksek öncelikli tweet elde edilebilmektedir. Öncelikli tweet'lerin sayısının çok olması içeriklerin hızlı değerlendirilmesini zorlaştırmaktadır. Bu nedenle, mesajlardaki önemli bilgilerin hızlı elde edilebilmesi için anlamlı tweet'lerin özeti çıkarılmıştır. Özet çıkarma, içerikten bir bölümü seçme veya otomatik özet oluşturma şeklindedir. Tweet'ler üzerinde de aynı kategoriye ait olanlar kümelenip özetleme işlemi yapılmaktadır. Genel olarak Twitter'daki "trend topic" şeklinde adlandırılan konuların anlaşılması, duygu analiziyle kategorilenen negatif veya pozitif tweet'lerin özetlenerek ana düşüncenin belirlenmesi yönünde kullanılmaktadır. Bu çalışmada ise tespit edilip kümelenen yüksek öncelikli tweet'lerin okuma kolaylığı ve zamandan tasarruf sağlamak amacıyla özet bilgi haline getirilmiştir.

1.2 Literatür Özeti

Literatür özeti Tweet Sınıflandırma, Tweet'lerin Özetlenmesi ve Twitter Kullanılarak Yapılan Doğal Afet Çalışmaları olmak üzere 3 ana başlık altında incelenmiştir.

1.2.1 Tweet Sınıflandırma

Twitter zengin bilgi kaynağı sebebiyle içerik çıkarma ile ilgili çalışmalarda oldukça popülerlik kazanmıştır. Özellikle Twitter üzerinde duygu analizi, önceden belirlenen konu hakkında kitlelerin düşüncelerini değerlendirme açısından önemli araştırma konularından birisidir. Duygu analizi, kullanıcıların attıkları tweet'leri pozitif, negatif ve nötr olarak sınıflandırmaktır. Tweet sınıflandırma, makine öğrenmesi ve doğal dil işleme yöntemleri kullanılarak üzerinde bir çok çalışma gerçekleştirilen yazı sınıflandırmanın[2] alt alanıdır. Sahip olduğu 140 karakter kısıtlaması, kendine özel terimleri ve informal dil yapısından dolayı geleneksel yazı sınıflandırma yöntemlerinin yetersiz kalabildiği ek zorluklar içermektedir.

Tweet'leri sınıflandırma üzerine geliştirilen yaklaşımlar üç ana başlık altında incelenebilir.

1.2.1.1 Sözlük Tabanlı Tweet Sınıflandırma Çalışmaları

Tweet sınıflandırma için kullanılan yöntemlerden ilki eğitmensiz öğrenim yöntemlerinden olan sözlük tabanlı yaklaşımdır[3].

Sözlük tabanlı yaklaşımda kelimelerin birbirinden bağımsız olduğu kabul edilir ve incelenen tweet bu kelimelerin birleşimi olarak ifade edilir. Duygu analizinde sıkça kullanılan bu yöntemde kelimelerin puanı vardır. Negatif diye ifade edilen kelimelerin puanu -1, pozitif diye ifade edilen kelimelerin puanı +1 olarak kabul edilirse, tweet içerisindeki kelimelerin puanlarının toplamı 0 dan büyükse pozitif, küçükse negatif ve 0 a eşitse nötr olarak sınıflandırılır[13].

Hassasiyet artırmak açısından kelimelerin puanları 0 ile 1 ve 0 ile -1 arasında değiştirilebilir. Negatifliği en çok ifade eden kelimeler -1 iken, daha az ifade eden kelimeler -1 den 0 a doğru yaklaşabilir. Bu yöntemin aynısı pozitif kelimeler içinde kullanılarak tweet in sınıflandırma aşamasında sadece kelime sayısına bağlı olmayıp kelimenin ifade ettiği duygunun kuvvetine de bağlı değerlendirilerek başarı oranı artırılabilir.

Bu yöntemin dezavantajı kelimelerin bazı yerlerde pozitif anlam ifade ederken bazı yerlerde ise negatif anlamda kullanılabiliyor olmasıdır.

İngilizce metinlerde duygu analizi için dünya genelindeki araştırmalarda sıkça kullanılan WordNet, EmotiNet gibi hazır sözlükler bulunmaktadır. Fakat farklı dillerde bu şekilde

yaygınlaşmış büyük çaplı sözlüklerin olmaması araştırmacılar açısından sözlük tabanlı yöntemi kullanmaya ek yükler getirmektedir. Kullanılacak dile ait sözlüklerin otomatik oluşturulmasına yönelik de çalışmalar bulunmaktadır

1.2.1.2 Eğiticili Makine Öğrenmesi Yöntemleri ile Tweet Sınıflandırma

Bu yöntemler eğitim verisi ve test verisi kullanır. Eğitim verisi özellik vektörleri ve onlara karşılık gelen sınıf etiketleri ile ifade edilir. Etiketleme işlemi önceden elle yada otomatik olarak yapılmaktadır[5]. Eğitim verisi kullanılarak, gelen özellik vektörüne karşılık gelen en yakın etiketi bulmaya çalışan bir sınıflandırma modeli geliştirilir. Daha sonra test verisi kullanılarak modelin başarı oranı ölçülür.

Eğiticili Makine Öğrenmesi Yöntemleri ile Tweet Sınıflandırma çalışmaları kapsamında Destek Vektör Makineleri(DVM), Naive Bayes (NB), Maksimum Entropi (ME), Rasgele Ormanlar gibi yöntemler kullanılarak duygu analizi çalışmaları yapılmıştır[3]. Duygu analizinde kullanılabilecek bazı özellikler kelime varlığı, kelime sıklığı, kelime türü, n-gram alınabilmektedir. Tweet sınıflandırma için DVM, Naive Bayes ve Rasgele Ormanlar'in ön plana çıktığı sınıflandırıcılar arasında özellik bakımından unigram tabanlı DVM diğerlerine göre daha başarılı sonuçlar vermiştir[14,15].

1.2.1.3 Tweet Sınıflandırma için Hibrit Yaklaşım

Hibrit yaklaşımda sözlük tabanlı yöntemler ve eğiticili makine öğrenmesi yöntemleri birlikte kullanılmıştır[4]. Sözlük tabanlı yöntemler, eğiticili makine öğrenmesi yöntemleri için özellik çıkarımı yapmak amacıyla kullanılmaktadır[4]. Örnek olarak, sözlük tabanlı yöntem ile tweet içerisindeki negatif ve pozitif kelimelerin sayısını tespit ederek makine öğrenmesi yöntemleri için iki özellik olarak kullanılabilmektedir. [19]'da Hibrit yaklaşım kullanılarak tweet sınıflandırma başarı oranının geleneksel yaklaşımlara göre daha yüksek olduğu gösterilmiştir.

Hibrit yaklaşım kullanılarak diğer yöntemlerin dezavantajları azaltılabilmektedir. Hibrit yaklaşımlar, makine öğrenmesi yöntemlerinin büyük problemlerinden bazıları olan ve yazı sınıflandırmada her bir kelimenin özellik olarak alınması gibi durumlarda karşılaşılan çok sayıda özellik kümesi ve hafıza karmaşıklıklarını giderebilmektedir.

1.2.2 Tweet'lerin Özetlenmesi

Mikroblog özetleme daha genel olan yazı özetlemenin bir alt alanı olarak görülebilir. Genel olarak içeriğin en önemli kısımlarını tespit etmeyi veya önemini vurgulayacak özet içeriği oluşturmayı amaçlar. Twitter'da özetleme ise aynı kümeye ait tweet'ler üzerinde o kümeyi en iyi temsil edecek öz bilgiyi çıkarmaya yöneliktir. Özetleme için kullanılan 2 ana yöntem vardır. Bunlardan ilki, tweet'ler içerisinde önemli parçaların bir araya getirilerek özet tweet'in otomatik oluşturulmasıdır[6]. İkinci yöntem ise kümeyi

en iyi temsil edebilecek tweet'in o küme içerisinden seçilmesidir[6]. Bu yönteme dahil olanlardan Rasgele Özetleyici (Random Summarizer) küme içerisinden rasgele k adet Tweet seçimine, En Son Özetleyici (Most Recent Summarizer) küme içerisinden en yeni k adet tweet seçimine, SumBasic ise kelimelerin kullanım frekanslarının toplamıyla hesaplanan en yüksek değere sahip k adet tweet'in seçimine dayanmaktadır[8]. LexRank, TextRank gibi yöntemler ise graf tabanlı özetleme yöntemleridir. [7,8]'de dikkat çeken ve diğer yöntemlere göre daha başarılı sonuç veren Hibrit TF-IDF (Kelime Frekansı – Tersine Doküman Frekansı)[8] yönteminde ise standart doküman özetleme de kullanılan TF-IDF yöntemi Twitter verisi üzerinde daha iyi sonuç verecek şekilde düzenlenmiştir.

Bu çalışmada da Hibrit TF-IDF yöntemi, doğal afet verisi üzerinde daha başarılı özetleme yapabilecek şekilde düzenlenmiştir.

1.2.3 Twitter Kullanılarak Yapılan Doğal Afet Çalışmaları

Mobil iletişim cihazlarının yaygınlaşması ile birlikte, insanlar doğal afet olayları ile ilgili bilgileri mikrobloglar üzerinden yazı, resim veya video olarak paylaşmaktadır. Bu ortamlardan en önemlisini de kullanıcıların hem web üzerinden hem de mobil cihazlar üzerinden oldukça aktif kullandığı Twitter sağlamaktadır.

Doğal afetler ile ilgili Twitter üzerinde bir çok çalışma gerçekleştirilmiştir. Twitter kullanıcıların gönderdiği tweet içerisine kullanıcının isteğine bağlı olarak coğrafi konum bilgisini de yerleştirmektedir. Çalışmaların çoğunda da tweet'lerle gelen bu coğrafi konum bilgisi kullanılmaktadır.

[9]'da yapılan çalışmada , doğal afetler sonrasında özet etiket (hashtag) ve afetin oluştuğu bölge bilgileri kullanılarak kullanıcıların attığı tweet'ler toplanmaktadır. Aynı zamanda GSIS (Geospatial Information System) olarak adlandırdıkları kaynaktan depremin oluştuğu bölge ile ilgili nüfus yoğunluğu, trafik yoğunluğu gibi bilgileri alınmaktadır. Toplanan tweet'ler üzerinden ilk olarak lokasyon bilgisi alınıp, ardından morfolojik analiz yapılarak sıkça kullanılan kelimeler ve bu kelimeleri içeren tweet'ler tespit edilmektedir. Daha sonra elde edilen bu verilerle doğal afetlerin etkileri harita üzerinde görsel olarak trafik, hava, yoğunluk vb bilgilerle gösterilmektedir.

Aynı şekilde [10]'da doğal afet sonrası o bölgeye ait kullanıcı tweet'leri toplanarak gerçek zamanlı kriz haritalama sistemi tasarlanmıştır. Doğal afetin yaşandığı bölge ile ilgili geçmişte tuttukları ve afetin olmadığı zamanlarda atılan tweet'ler üzerinden normal durum bilgisi çıkarılıp, afetin yaşandığı esnada atılan tweet'ler ile karşılaştırma yapılmaktadır. Afet sonrası toplanan ve sokak seviyesinde kriz haritalama yapılması için toplanan tweet'ler üzerinde coğrafi bilgi çıkarımı, Twitter'ın sunduğu coğrafi konum bilgi ile birlikte tweet içerisinde doğal dil işleme yöntemleri kullanılarak yapılmaktadır. Kullanılan bu yöntemler ile doğal afet olaylarının sosyal medya üzerindeki etkisini

ölçmek, doğal afetle ilgili gerçek zamanlı kriz haritalamak ve raporlama yapmak hedeflenmiştir.

[11]'de ise doğal afet sonrasında kullanıcıların Twitter üzerinde ki tweet atma, tweet'lere cevap yazma, retweet'leme, favorilerine ekleme gibi davranışları ölçülerek doğal afet bilgi ve destek sistemi oluşturmak hedeflenmiştir. Ölçümleme işlemi için her bir kullanıcının tweet yazma aktivitesini temsil edecek çok boyutlu vektör oluşturulmuştur. Bu vektör 5 farklı veri tipi içermektedir. Bunlar; yeni tweet atma zaman aralığı, tweet cevaplama zaman aralığı, başka tweet'i tekrar gönderme zaman aralığı, tweet cevaplama sayısı ve başka tweet'i tekrar gönderme sayısıdır. Doğal afet bilgi ve destek sistemi için belirli zaman aralıklarıyla hesaplanan bu vektörler üzerinden doğal afetlerin hemen sonrası kullanıcı davranışları ölçümlenmiştir.

[12]'de BMKG (The Meteorological, Climatological and Geophysical Agency of Indonesia)'nın sunduğu son 3 tsunami zamanı gönderdiği tweet'lerin kullanıcılar üzerinde etkisi (yeniden gönderme, cevaplama, favorilerine ekleme) ölçülüp analiz edilerek, tsunami erken uyarı sisteminin Twitter ile birlikte kullanımı tasarlanmıştır.

1.3 Hipotez

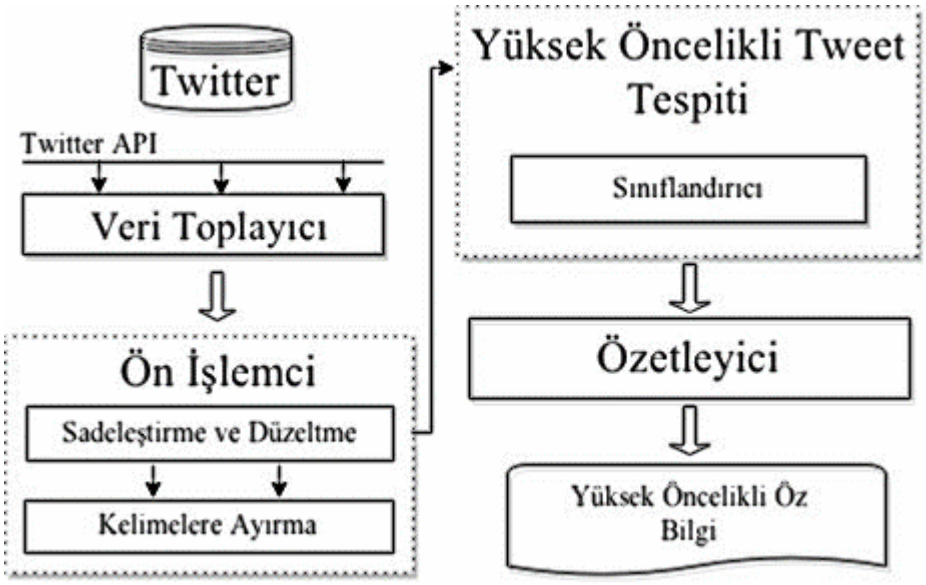
Doğal afetler sonrası Twitter kullanıcılarının doğal afet ile ilgili gönderdiği tweet'ler içerisinden yüksek öncelikli olanlarının tespiti ve tespit edilenlerin özetlenebilmesinin mümkün olduğu görülmüştür.

Çalışmalar göstermiştir ki; Twitter üzerinde yapılan diğer doğal afet çalışmalarından farklı olarak, zamanın ve acil müdahalenin insan hayatı açısından büyük önem arz ettiği doğal afetlerin hemen sonrasında, yüksek öncelikli tweet tespiti ve özetlenmesi sayesinde yardım birimleri için anlık bilgi kaynağı olabilecek hale getirmek mümkündür.

BÖLÜM 2

SİSTEM TASARIMI

Bu çalışmada, doğal afetlerin hemen sonrasında Twitter üzerinden doğal afet bölgesindeki kullanıcıların gönderdiği tweet'ler değerlendirilerek yüksek öncelikli olanların otomatik olarak tespit edildiği, tespit edilen bu yüksek öncelikli tweet'lerin özetlenerek yardım birimleri için anlık bilgi kaynağı olarak kullanabilecekleri bir sistem geliştirilmiştir. Geliştirilen sistemin blok diagramı Şekil 2.1'deki gibidir.



Şekil 2.1 Sistem Tasarımı

2.1 Afet Sonrası Olaylarla ilgili Veri Toplama

Twitter, sunduğu Twitter API ile tweet'lere ve özelliklerine erişim imkanı sağlamaktadır. Twitter verisine erişim için Twitter API iki ana yöntem içermektedir. Bunlardan ilki Stream API diye adlandırılan, canlı olarak o an atılan tweet'leri belirlenen kriterlere

göre toplayan yöntemdir. İkinci yöntem ise aynı şekilde belirlenen kriterlere göre geçmişe yönelik arama yaparak tweet'leri toplamaktır. Twitter bu iki yöntemle verilere ulaşıma belirli kısıtlamalar koymuştur. Bu kısıtlamalardan ilki gelen verinin kısıtlı şekilde sunulmasıdır. Twitter tarafından "Garden Hose" olarak adlandırılan bu kısıtlamada toplanan tweet'ler yaklaşık olarak tüm verinin 10 da 1'i olabilmektedir. Sadece Twitter ile anlaşmalı belirli şirketler "Firehose" hizmeti altında tüm tweet verisine ulaşabilmektedir. Bir diğer kısıtlama ise geçmişe yönelik yapılan aramaların en fazla 8 güne kadar olmasıdır. Daha eski verilere ulaşmak için Twitter sertifikalı şirketlerden satın alma yapılması gerekmektedir. Twitter API ile istenilen veriye ulaşmak için filtreler belirlenmesi gerekir. Anahtar kelimeler, coğrafi konum ve dil seçeneği kullanılarak belirlenen filtreler ile istenilen veriye "Garden Hose" kısıtlaması altında ulaşılabilir.

Geliştirilen sistemde ilk aşama olan Veri Toplama kısmında Twitter API kullanılarak meydana gelen California, Nagano ve Nepal depremleri hakkında kullanıcıların gönderdiği tweet'ler toplanmıştır. İngilizce doğal afet tweet'lerine yönelik 3 farklı afet ile ilgili veri seti oluşturulmuştur. Bunlardan ilki Amerika'nın California eyaletinde 24 Ağustos 2014 tarihinde meydana gelen 6.0 büyüklüğündeki deprem sonrası atılan tweet'leri içermektedir. Bu veri setinde yaklaşık olarak 150 bin tweet bulunmaktadır. İkinci veri seti Japonya'nın Nagano şehrinde 22 Kasım 2014 tarihinde meydana gelen 6.8 şiddetindeki deprem sonrası toplanan yaklaşık 2 bin tweet içermektedir. Son olarak Nepal'in Kathmondu şehrinde 25 Nisan 2015 tarihinde meydana gelen deprem sonrasında atılan 300 bin tweet veri setine eklenmiştir.

Tweet'lerin toplanılmasına yönelik ilk yapılan meydana gelen depremi tanımlayan anahtar kelimeleri kullanarak doğru filtrelerin oluşturulmasıdır. Oluşturulan filtreler kullanılarak doğal afet hakkında gönderilen tweet'leri toplamaya yönelik arama başlatılır. Yapılan arama ile JSON (JavaScript Object Notation) verisi olarak elde edilen tweet'ler sistemin diğer bölümlerinde işlenmek üzere saklanır.

2.2 Ön İşlemler

Veri Toplama ile elde edilen tweet'ler sınıflandırıcılar tarafından doğrudan kullanılmak için uygun değildir. Tweet'ler kendine özgü terimleri, 140 karakter kısıtlaması ve düzgün olmayan dil yapısıyla oldukça karışık bir veriyi temsil etmektedir. Veri içerisinde gürültü olarak ifade edilen ve sınıflandırıcı başarı oranlarını düşüren bu etmenlerden veriyi temizlemek ve başarı oranını yükseltmek için ön işlemden geçirilerek sadeleştirilmesi ve düzeltilmesi gerekir.

Sınıflandırıcılar tarafından verinin başarılı bir şekilde değerlendirilebilmesi için tweet'lere uygulanan ön işlem adımları aşağıdaki gibidir :

1. “#hashtag” kelimesi “hashtag” haline çevrilir.
   Twitter’da özel kelimelerin başına # sembölü getirilerek oluşturulan hashtag’ler bilgi değeri taşıdığı için başlarından # sembolü çıkarılarak işlenebilecek hale getirilir.

2. @username “ATUSER” ile değiştirilir.
   Twitter’da kullanıcılar isimlerinin başına @ sembolü getirilerek ifade edilir. Tweet içerisinde de bu ifadeyi kullanarak kullanıcılara bağlantı oluşturulabilmektedir. Fakat kullanıcı ismi bilgi değeri taşımadığından dolayı ATUSER ifadesi ile değiştirilir.

3. URL’ler silinir.
   Kullanıcılar tweet içerisinde bağlantılar ile web sayfaları, resimler, videolar vb. paylaşabilmektedir. Yapılacak olan analizin bu bağlantıları kapsamadığı durumlarda tweet içerisinden silinmesi gerekir. Bu nedenle tweet’ler içerisinden linkler silinmektedir.

4. Başında sayı ile başlayan kelimeler silinir.
   Başında sayı ile başlayan kelimeler sınıflandırma işlemleri için bilgi değeri taşımaması sebebiyle tweet içerisinden çıkarılır.

5. Sayı ifadeleri “ATNUMBER” ile değiştirilir.
   Tweet içerisinde depremin büyüklüğü, yaralı, ölü, hasar, mahsur sayısı gibi bilgiler bulunabilmektedir. Bu ifadelerin değerini kaybetmemek ve tekilleştirmek için sayılar “ATNUMBER” ifadesi ile değiştirilir.

6. Harf ve sayı dışındaki karakterler silinir
   Tweet içerisinde bir çok sembol ve işaret kullanılmaktadır. Geliştirilen sistemde bu ifadeler özellik olarak seçilmediği için tweet içerisinden temizlenir.

7. Retweet anlamına gelen “RT” kelimesi çıkarılır.
   Twitter API ile toplanan tweet’ler içerisinde bir kullanıcının başka bir kullanıcının tweet’ini paylaştığı yani retweet olan tweet’lerin başında RT ifadesi bulunur. Retweet’lerinde normal tweet’ler gibi işlem görmesi için başlarındaki RT ifadesi silinir.

8. Arka arkaya 2 veya daha fazla tekrar eden harfler 1’e indirilir.
   Tweet yapısal olmayan bir formata sahip olması sebebiyle içerisinde bir çok yazım yanlışı içermektedir. Örneğin “so” kelimesinin anlamını vurgulamak amacıyla kullanıcılar “soooo” şeklinde uzatarak yazabilmektedir. Eğer herhangi bir işlem yapılmaz ise sistem “so” ve “soooo” kelimelerini iki farklı kelime olarak ele alacaktır. Bu da aynı anlam taşımalarına rağmen bölünmeye sebep olarak sistemin

başarısını düşürecektir. Bu sebepten dolayı arka arkaya tekrar eden harfler 1'e indirilerek farklı yerlerde iki farklı şekilde yazılsa da tek kelime olarak ele alınacaktır. Doğru halinde arka arkaya tekrar eden harflerin mevcut olduğu "success" gibi kelimelerde ön işlem sonrası "suces" halini alacağından ve kelimenin geçtiği her durumda aynı işlem geçerli olacağından bilgi kaybı göz ardı edilebilir düzeyde olacaktır.

Bu işlemler sonucunda elde edilen sadeleştirilmiş ve düzeltilmiş tweet, kelimelerine ayrılır. Eğer bu kelimelerden her hangi biri "the", "and" gibi sıkça kullanılan ve bilgi değeri olmayan bir kelime ise listeden çıkarılır. Bu kelimelerin cümle içerisinden çıkarılması veri içerisindeki gürültüyü ve özellik sayısını azaltarak sistemin başarı oranını ve çalışma performansını arttırmaktadır. Geliştirilen sistemde İngilizce dili için oluşturulan ve yaklaşık 700 kelime içeren çok kullanılan kelime listesi kullanılmıştır.

2.3 Tweet'lerin Yüksek Öncelikli ve Düşük Öncelikli Olarak Sınıflandırması

Tweet'lerin yüksek öncelikli ve düşük öncelikli olmak üzere 2 sınıfa ayrılması hedeflenmiştir. Sınıflandırıcılar ile model oluşturmada kullanılacak eğitim kümesini oluşturmak için ön işlemden geçirilmiş tweet'ler önce "Yüksek Öncelikli" ve "Düşük Öncelikli" olarak elle etiketlenmiştir. Yüksek Öncelikli olarak etiketlenen tweet'ler deprem hakkında yaralı, ölü, hasar ve mahsur bilgisi içeren, yardım birimleri için değerli bilgi niteliği taşıyan tweet'lerdir.

Düşük öncelikli tweet örnekleri olarak aşağıdakiler verilebilir;

- *"My heart and prayers go out to those in Japan affected by the earthquake. You will be in my thoughts."*
- "*RT @taralipinski: My thoughts & love go out to the beautiful city and people of Nagano.My prayers are w all those affected from this earthq…"*
- "*RT @JRehling: BREAKING: A magnitude 6.8 earthquake has hit Nagano, Japan. http://t.co/corY3Vi6Gc*"

Yüksek öncelikli tweet örneği olarak ise aşağıdakiler verilebilir;

- *"Magnitude 6.8 earthquake hits Nagano in Japan, injuring 13 and destroying five houses"*
- "*:-/ MT @mit_obe M6.8 hits Nagano, past 10 pm Saturday, flattening 7 houses, injuring 57. http://t.co/ityTC227hC http://t.co/p2e8v4R0lo*"

- "*RT @abcnews: 21 people trapped, 13 injured and five houses destroyed as magnitude 6.8 earthquake hits Nagano in Japan http://t.co/K9m5cTEVud*"

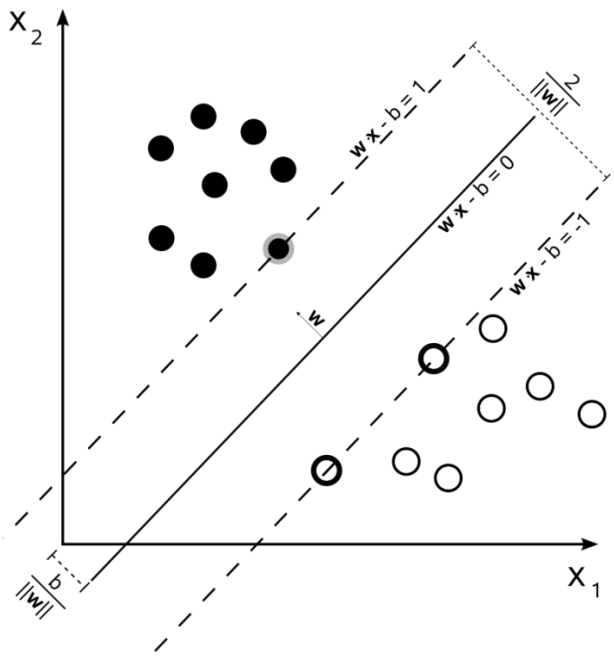
Tweet verisi içerisinde bulunan kelimeler özellik olarak seçilmiştir. Tüm özelliklerin belirlenebilmesi için etiketlenen verilerin içerdiği farklı kelimelerden sözlük oluşturulmuştur. Ardından her tweet'in özellik vektörü oluşturulurken sözlükte içerdiği kelimelerin yani özelliklerin değeri 1, içermediklerinin değeri ise 0 alınmıştır.

Tweet sınıflandırma işlemi için başarı oranları yüksek DVM, Naive Bayes ve Rasgele Ormanlar sınıflandırıcıları kullanılmıştır.

2.3.1 Destek Vektör Makinesi Sınıflandırıcısı

Destek Vektör Makinesi yöntemi, sınıflandırma ve regresyon konularında kullanılan oldukça etkili ve basit olan eğiticili öğrenme yöntemlerden birisidir. Sınıflandırma yüksek doğruluk oranlarına sahip olma, karmaşık karar sınırlarını modelleyebilme, çok sayıda bağımsız değişken ile çalışabilme, hem doğrusal olarak ayrılabilen hem de doğrusal olarak ayrılmayan verilere uygulanabilme gibi avantajlara sahiptir.

Sınıflandırma için bir düzlemde bulunan iki küme arasında bir sınır çizilerek iki sınıfa ayırmak mümkündür. Bu sınırın çizileceği yer ise iki kümeye de eşit uzaklıkta ve en uzak yer olmalıdır. DVM yöntemi bu iki küme arasındaki sınırın yerini belirler.



Şekil 2.2 Destek Vektör Makinesi

Şekil 2.2'de gösterildiği gibi iki kümeye yakın ve birbirine paralel iki sınır çizgisi çizilir ve bu sınır çizgileri birbirine yaklaştırılarak ortak sınır çizgisi üretilir.

Düzlem üzerindeki n tane noktadan her birinin tanımı Eşitlik 2.1'deki gibidir[21] ;

$$D = \{(xi, yi)|xi \in R, ci \in \{-1,1\}\}_{i-1}^{n} \quad (2.1)$$

Buradaki her x,y ikilisi için X vektör uzayındaki bir nokta ve c bu noktanın -1 veya +1 olduğunu yani sınıfını belirten değerdir. Bu düzlem içerisindeki her bir noktayı Eşitlik 2.2'deki gibi ifade etmek mümkündür.

$$wx - b = 0 \quad (2.2)$$

Eşitlik 2.2'de belirtilen w normal vektörü, x noktanın değişen parametresi ve b kayma oranıdır. Bu denkleme göre b/||w|| değeri iki grup arasındaki mesafeyi verir. Bu mesafe farkını en yüksek değere çıkarmak için Şekil 2.2'de gösterilen 0,-1 ve +1 değerlerine sahip 3 doğruyu veren denklemde 2/||w|| denklemi kullanılmıştır.

$$wx - b = -1 \quad (2.3)$$
$$wx - b = +1 \quad (2.4)$$

Eşitlik 2.3 ve Eşitlik 2.4 doğruların kaydırılması ve aradaki farkın (offset) en yüksek değeri bulunması ile elde edilir.

Bu çalışmada oluşturulan sınıflar "Yükse Öncelikli" ve "Düşük Öncelikli" şeklindedir. Bu sınırın çizileceği yer ise iki kümenin de üyelerine en uzak olan yer olmalıdır. Model oluşturulurken elimizde 2 sınıf bulunduğundan ve doğrusal ayrıma yatkın olduğundan doğrusal çekirdek kullanılmıştır. Tweet'ler içerisinde geçen ve ön işlem sonrasına kalan her bir kelime özellik olarak seçilmiş olup, bir tweet'e ait özellik vektörü oluşturulurken içerdiği kelimelerin özellik değeri 1, içermediklerinin ise 0 alınmıştır.

2.3.2 Naive Bayes Sınıflandırıcısı

Naive Bayes sınıflandırıcısı Bayes teoreminin özelliklerin birbirinden bağımsız değerlendirilmesi ile basitleştirilmiş halidir[22]. Yazı sınıflandırmada yüksek doğruluk oranlarına sahip olduğundan yaygın bir şekilde kullanılmaktadır. Bu nedenle tweet sınıflandırmada da tercih edilen yöntemler arasındadır.

Bayes teoremi Eşitlik 2.5'de gösterilmektedir.

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \quad (2.5)$$

P(A|B) olasılığı, B olayının gerçekleştiği durumda A olayının olma olasılığıdır. P(B|A) ise A olayının gerçekleştiği durumda B olayının olma olasılığıdır.

Bayes teoremi temel alınarak metin sınıflandırılması şu şekilde yapılmaktadır;

$$k = [k(1), k(2), \dots, k(L)]^T \quad (2.6)$$

Eşitlik 2.6'daki k metinlerden elde edilen kelime listesini ifade etmektedir. $P(k_i|S_i)$, ki kelimesinin Si etiketine ait olma olasılığıdır. $P(S_i)$ S_i etiketinin önsel olasılığıdır. Bu durumda k kelimelerinden oluşan d metninin S etiketli olma olasılığı Eşitlik 2.6'da görüldüğü gibidir.

$$P(S|d) = \frac{P(d|S)*P(S)}{P(d)} \quad (2.6)$$

d metni k kelimelerinden oluştuğundan dolayı Eşitlik 2.7'de görüldüğü gibi P(d|S) kelimelerin S etiketine ait olma olasılıklarının çarpımıdır.

$$P(d|S) = \prod_{i=1}^{L} P(ki|S) \quad (2.7)$$

Böylece bir metnin S etiketinde olma olasılığı Eşitlik 2.8'deki gibidir;

$$P(S|d) = \prod_{i=1}^{L} P(ki|S) * P(S) \quad (2.8)$$

2.3.3 Rasgele Ormanlar (Random Forest) Sınıflandırıcısı

Rasgele Ormanlar(RO) sınıflandırma işlemi sırasında birden fazla karar ağacının kullanıldığı bir kolektif öğrenme yöntemidir. Kolektif öğrenme yöntemleri, bir sınıflandırıcı yerine birden çok sınıflandırıcı üreten ve sonrasında onların tahminlerinden alınan oylar ile yeni veriyi sınıflandıran öğrenme algoritmalarıdır.

RO, ağaç tipi sınıflandırıcılar topluluğu olarak tanımlanabilir. RO, tüm özellikler arasından en yüksek başarıyı veren özelliği kök yaparak ağacı oluşturmak yerine, her bir düğümde rastgele olarak seçilen özellikler arasından en iyisini kullanarak her bir düğümü dallara ayırır. Her veri seti orijinal veri setinden yer değiştirmeli olarak üretilir. Sonra rastgele özellik seçimi kullanılarak ağaçlar geliştirilir. Geliştirilen ağaçlar budanmaz [20].

RO algoritmasını başlatmak için kullanıcı tarafından 2 parametre tanımlanmalıdır. Bu parametreler, en iyi bölünmeyi belirlemek için her bir düğümde kullanılan değişkenlerin sayısı (m) ve geliştirilecek ağaçların sayısı N'dir.

Rasgele Ormanlar yöntemi çok özelliğe sahip verilerde avantaj sağladığından ve yazı sınıflandırmada çok sayıda özellik mevcut olduğundan yazı sınıflandırmada kullanılan yöntemler arasındadır.

2.4 Yüksek Öncelikli Tweet'lerin Özetlenmesi

Sınıflandırma işleminden sonra tespit edilen yüksek öncelikli tweet'ler bir araya toplanarak özetlenebilecek hale getirilmiştir. Özetleme algoritmaları olarak SumBasic, TF-ISF ve Hibrit TF-IDF yöntemlerinin başarıları ayrı ayrı değerlendirilerek uygun model belirlenmiştir.

2.4.1 SumBasic

SumBasic yönteminde ilk olarak küme içerisinde kullanılan kelimelerin frekansları hesaplanır. Ardından her bir tweet içerisindeki kelimelerin frekanslarının toplamıyla o tweet'in özet puanı hesaplanır. Hesaplanan en yüksek değere sahip k adet tweet özet tweet olarak seçilir [16].

2.4.2 TF-ISF

TF-ISF(Term Frequency – Inverse Sentence Frequency) yönteminde ilk olarak SumBasic'de olduğu gibi küme içerisinde kullanılan kelimelerin frekansları hesaplanır. Ardından her bir tweet içerisindeki kelimelerin genel kullanım frekans değerleri ile birlikte kelimelerin tweet içerisindeki frekans değeri kullanılarak özet puanı hesaplanır. Hesaplanan en yüksek değere sahip k adet tweet özet tweet olarak seçilir [17].

2.4.3 Hibrit TF-IDF

Hibrit TF-IDF yöntemi Eşitlik 2.9'da gösterilen TF-IDF yönteminin tweet'ler üzerine özelleştirilmiş halidir.

$$TF_IDF = tf_{ij} * log_2 \frac{N}{df_j} \quad (2.9)$$

Eşitlik 2.9'da tf_{ij} (Term Frequency) T_j kelimesinin D_i dokümanındaki frekansı, N toplam doküman sayısı ve d_{fj} T_j kelimesini içeren doküman sayısını belirtir. Böylece IDF(Inverse Document Frequency) kelimenin o doküman içerisindeki bilgi değerini ifade eder ve ilgili kelimenin geçtiği doküman sayısının tüm doküman sayısına bölümünün logaritması alınarak bulunur. Fakat tweet standart bir doküman değildir. Eğer sınıftaki tüm tweet'ler bir doküman olarak ele alınırsa, IDF değerini kaybedecektir; çünkü tek bir doküman vardır. Diğer durumda IDF'in değerini korumak için her tweet bir doküman olarak düşünülürse TF sadece bir tweet içindeki frekansa bakacağından çok küçük ve neredeyse birbirine yakın değerlere sahip olacaktır. Bu problemin çözümüne

yönelik [8]'de TF-IDF'in tanımı hibrit dokümantasyona göre yeniden yapılmıştır. Bu tanıma göre TF değeri tüm tweet'leri bir doküman olarak ele alınıp hesaplanırken, IDF her tweet bir doküman olarak düşünülüp hesaplanır. Ek olarak uzun tweet'lerin veri içerisinde gürültü oluşturmaması için Eşitlik 2.10'daki gibi normalizasyon işlemi yapılır.

$$P(t) = \frac{Hybrid\ TF_IDF}{\max[eşik\ değer, kelime\ sayısı]} \tag{2.10}$$

Eşitlik 2.10'da t küme içerisinde ki bir tweet'i temsil ederken, P(t) o tweet'in özet puanıdır ve Hibrit TF-IDF değerinin kelime sayısı ile eşik değer arasında en yüksek değere sahip olanına bölünmesiyle elde edilir.

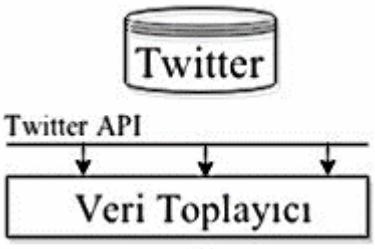
BÖLÜM 3

UYGULAMA

Yüksek öncelikli tweet tespiti ve özetlenmesine yönelik geliştirilen sistemin başarısını değerlendirmek amacı ile, meydana gelen California, Nagano ve Nepal depremleri sonrasında toplanan tweet'ler üzerinde değerlendirilerek en uygun model oluşturulmuş ve sistemin başarısı değerlendirilmiştir.

3.1 Veri Toplama ve Ön İşlem

Veri toplama işlemi için JAVA programlama dili ile Twitter API'yi kullanan bir yazılım geliştirilmiştir. Twitter API'nin sunduğu metotlar doğrultusunda belirlenen anahtar kelimeler ile geriye dönük arama veya canlı dinleme yapılabilmektedir. Deprem meydana geldikten hemen sonra o depreme yönelik anahtar kelimeler belirlenerek canlı dinleme başlatılabilmekte ve atılan tweet'ler toplanabilmektedir.



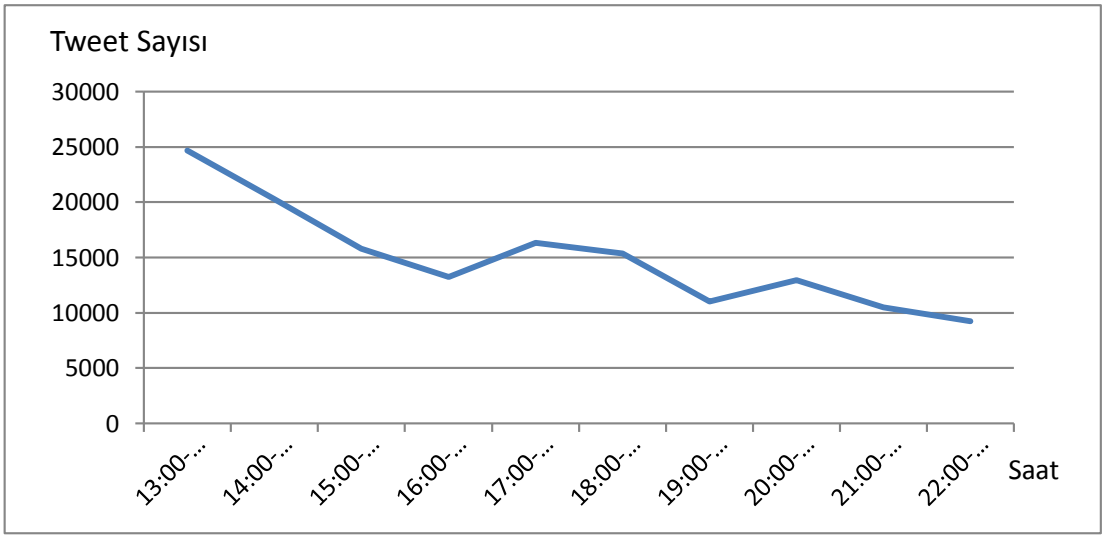
Şekil 3.1 Veri Toplayıcı

24 Ağustos 2014 tarihinde Amerika'nın California eyaletinde meydana gelen 6.0 büyüklüğündeki deprem sonrası tweet toplama işlemi başlatılmıştır. "*California*, *earhquake, quake*" anahtar kelimeleri ve İngilizce dil seçeneği ile oluşturulan filtre kullanılarak veri toplanmıştır. Türkiye saatine göre 12:20'de meydana gelen deprem sonrası toplanan tweet sayısı Çizelge 3.1'de gösterilmiştir.

Çizelge 3.1 California Depremi Sonrası Saatlere Göre Tweet Sayısı

Saat Aralığı	Tweet Sayısı
13:00-14:00	24641
14:00-15:00	20258
15:00-16:00	15804
16:00-17:00	13249
17:00-18:00	16338
18:00-19:00	15354
19:00-20:00	11001
20:00-21:00	12959
21:00-22:00	10509
22:00-23:00	9247

Deprem ile ilgili atılan tweet sayısında ilerleyen saatlerde düşüş yaşandığı gözlemlenmektedir. Grafiksel olarak gösterimi Şekil 3.1'deki gibidir.



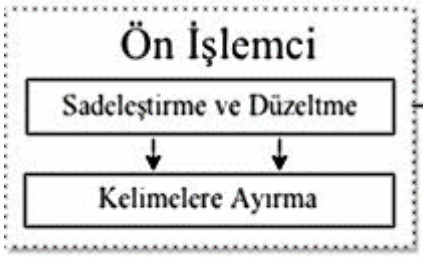
Şekil 3.1 California Depremi Sonrası Atılan Tweet Sayısı

California depreminden sonra 10 saat içerisinde deprem ile ilgili yaklaşık 150 bin tweet toplanmıştır.

Japonya'nın Nagano şehrinde 22 Kasım 2014 tarihinde meydana gelen 6.8 şiddetindeki deprem sonrasında deprem ile ilgili atılan tweet'leri toplamak amacıyla "*earthquake*", "*japan*", "*nagono*", "*quake*" şeklinde anahtar kelimeler belirlenmiştir. Geliştirilen yazılım ile bu anahtar kelimeler kullanılarak İngilizce dilindeki tweet'ler canlı dinleme yöntemiyle toplanmıştır. Toplanan yaklaşık 2000 tweet ilk olarak elle "Yüksek Öncelikli" ve "Düşük Öncelikli" olacak şekilde etiketlenmiştir. Etiketleme işleminde yaralı ve hasar durum bilgilerini ön plana çıkaran tweet'ler gibi içeriği anlamlı tweet'ler yüksek öncelikli olarak işaretlenmiştir.

Nepal'in Kathmondu şehrinde 25 Nisan 2015 tarihinde meydana gelen depremin hemen sonrasında , 12 saat sonrasında ve 24 saat sonrasında "*nepal*, *kathmondu*, *earhquake, quake*" anahtar kelimeleri ve İngilizce dil seçeneği ile oluşturulan filtre kullanılarak toplamda 300 bin tweet toplanmıştır. Bu veri üzerinde sınıflandırma ve özetleme algoritmaları test edilmiştir.

Toplanan tweet'ler örneklerde görüldüğü üzere oldukça kirli bir veriyi temsil etmektedir. Sınıflandırıcılar tarafından başarılı bir şekilde işlenebilmesi için ön işlemden geçirilmesi ve temizlenmesi gerekir.



Şekil 3.2 Ön İşlemci

Ön işlem kısmı 2 adım içerir. Bunlardan birincisi "Sadeleştirme ve Düzeltme" ikincisi ise "Kelimelere Ayırma" şeklindedir. Aşağıda yüksek öncelikli olarak etiketlenen tweet'in temizleme ve kelimelerine ayırma işlemlerinden sonraki durumu örnek olarak verilmiştir.

"*RT @asiapundits: The #Dharahara tower has collapsed, with up to 400 people suspected buried. #Nepal #Kathmandu #Earthquake. http://t.co/Xxt…"*

1- **Sadeleştirme ve Düzeltme** *: "rt atuser the dharahara tower has collapsed with up to atnumber people suspected buried nepal kathmandu earthquake"*
2- **Kelimelerine ayırma** *: rt, atuser, the, dharahara, tower, has, collapsed, with, up, to, atnumber, people, suspected, buried, nepal, kathmandu, earthquake*

Kelimelerine ayırma işleminden sonra "*rt*", "*atuser*" ifadeleri tweet'in anlamı ile ilgili herhangi bir bilgi vermediği için silinir. Aynı zamanda İngilizce dilinde sıkça kullanılan "the", "with", "up", "to", "has" kelimeleri de tweet'e her hangi bir anlam katmadığından dolayı ve özellik sayısının azaltılması için silinir.

3.2 Yüksek Öncelikli Tweet Tespiti

Yüksek öncelikli tweet tespiti için Nagano ve Nepal depremlerinden elde edilen tweet'ler kullanılmıştır. California depreminde elde edilen tweet'ler veri toplama ve ön işlem bölümünde ele alınmış olup depremin büyüklüğü ve etkisinin diğer depremlere göre daha az olmasından dolayı yüksek öncelikli tweet tespitinde kullanılmamıştır.

Japonya'nın Nagano şehrinde meydana gelen deprem sonrasında toplanan yaklaşık 2000 tweet'in 241 tanesi yüksek öncelikli, geri kalanlar ise düşük öncelikli olarak etiketlenmiştir. Ardından toplanan bu veri ön işlemden geçirilerek sınıflandırıcıların daha başarılı bir şekilde değerlendirebileceği hale getirilmiştir.

Nepal'in Kathmondu şehrinde meydana gelen deprem sonrasında belirli aralıklar ile toplamda 300 bin tweet toplanmıştır. Depremden hemen ardından 3 saat içerisinde toplanan tweet sayısı ise 180 bin'dir. Etiketleme işlemi bu veri üzerinde yapılmıştır. Yüksek öncelikli olarak 308, düşük öncelikli olarak ise 433 tweet etiketlenmiştir. Yüksek öncelikli olarak etiketlenen tweet'ler veri içerisinde tekrarları ile birlikte sayıldığında yaklaşık olarak 5 bin, düşük öncelikli olarak etiketlenen tweet'ler ise tekrarları ile birlikte yaklaşık 12 bin'dir. Etiketleme işleminin ardından tweet'ler ön işlemden geçirilerek sadeleştirilmiş ve kelimelerine ayrılmıştır.

Sınıflandırıcılar her bir veri kümesinde test edildikten sonra ek olarak bu veri kümelerinin karışımı ile de test edilmiştir. İki veri kümesinin karıştırılması sonucunda 549 yüksek öncelikli, 2179 düşük öncelikli olarak etiketlenmiş veri elde edilmiştir. Sınıflandırıcıların başarısını ölçmek amacıyla 10-katlı çapraz doğrulama yöntemi kullanılmıştır.

Sınıflandırıcıların başarıları değerlendirilirken kullanılan temel kavramlar F-Ölçütü, duyarlılık ve doğruluktur. Test sonucunda ulaşılan sonuçların başarım bilgileri hata matrisi ile ifade edilebilir. Hata matrisinde satırlar test kümesindeki örneklere ait gerçek sayıları, sütunlar ise sınıflandırıcıların tahminlerini ifade eder.

		Öngörülen Sınıf	
		Sınıf = 1	Sınıf = 0
Doğru Sınıf	Sınıf = 1	PD (Pozitif Doğru)	PY (Pozitif Yanlış)
	Sınıf = 0	NY (Negatif Yanlış)	ND (Negatif Doğru)

Hata matrisindeki değerlere göre doğruluk oranı Eşitlik 3.1'deki gibi, duyarlılık oranı Eşitlik 3.2'deki gibi ve F-Ölçütü oranı ise Eşitlik 3.4'deki gibi hesaplanır[18].

$$Doğruluk = \frac{PD+ND}{PD+NY+PY+ND} \quad (3.1)$$

$$Duyarlılık = \frac{PD}{PD+PY} \quad (3.2)$$

$$Kesinlik = \frac{PD}{PD+NY} \quad (3.3)$$

$$F - \text{Ölçütü} = \frac{2*Duyarlilik*Kesinlik}{Duyarlılık+Kesinlik} \quad (3.4)$$

Ölçümlerde doğruluk oranı en popüler ve basit yöntem olduğu için seçilmiştir. Duyarlılık oranı yüksek öncelikli tweet'lerin tespitindeki duyarlılığı verdiği için seçilmiştir. F-Ölçütü oranı tek başlarına başarı değerlendirmede yeterli olmayan duyarlılık ve keskinlik oranlarını beraber değerlendirdiği için seçilmiştir.

3.2.1 Naive Bayes ile Sınıflandırma

Nagano, Nepal ve ikisinin karışımı olan veri kümeleri üzerinde 10-katlı çapraz doğrulama yöntemi kullanılarak Naive Bayes ile yapılan sınıflandırma sonuçları Çizelge 3.2'de verilmiştir.

Çizelge 3.2 Naive Bayes ile Sınıflandırma Sonuçları

	F-Ölçütü	**Duyarlılık**	**Doğruluk**
Nagano	0.964	0.97	0.962
Nepal	0.905	0.905	0.906
Karışık	0.914	0.929	0.909

Çizelge 3.2'deki sınıflandırma sonuçları dikkate alındığında Naive Bayes sınıflandırıcısının Nagano veri kümesi üzerinde daha yüksek başarı gösterdiği, karışık veri kümesi üzerinde %90,9 oranında doğruluk değerine sahip olduğu gözlemlenmektedir.

Çizelge 3.3 Naive Bayes Hata Matrisi

	Sınıflandırılan	
	Yüksek Öncelikli	**Düşük Öncelikli**
Yüksek Öncelikli	523	26
Düşük Öncelikli	223	1956

Karışık veri kümesi üzerinde yapılan sınıflandırma sonucu ortaya çıkan hata matrisi Çizelge 3.3'de verilmiştir. Bu hata matrisine göre Naive Bayes yüksek öncelikli tweet sınıflandırmada düşük öncelikli tweet sınıflandırmaya göre daha düşük başarı sonuçları vermiştir.

3.2.2 DVM ile Sınıflandırma

Nagano, Nepal ve ikisinin karışımı olan veri kümeleri üzerinde 10-katlı çapraz doğrulama yöntemi kullanılarak DVM sınıflandırıcısı test edilmiştir. Doğrusal çekirdek ve 2. Dereceden polinomal çekirdek kullanılarak yapılan testlerin sonucu ayrı ayrı Çizelge 3.4 ve Çizelge 3.5'de verilmiştir.

Çizelge 3.4 Doğrusal Çekirdek Kullanılarak Modellenen DVM Yönteminin Sınıflandırma Sonuçları

	F-Ölçütü	**Duyarlılık**	**Doğruluk**
Nagano	0.989	0.99	0.989
Nepal	0.91	0.91	0.91
Karışık	0.964	0.965	0.964

Çizelge 3.5 Polinomal Çekirdek Kullanılarak Modellenen DVM Yönteminin Test Sonuçları

	F-Ölçütü	**Duyarlılık**	**Doğruluk**
Nagano	0.989	0.989	0.989
Nepal	0.92	0.92	0.92
Karışık	0.971	0.971	0.971

Çizelge 3.4 ve Çizelge 3.5'deki sonuçlar değerlendirildiğinde DVM, 0.07 farkla polinomal çekirdek kullanılarak modellendiğinde daha yüksek başarı oranı vermiştir. Model oluşturma süreleri karşılaştırıldığında polinomal çekirdek ile modelleme 3.96 saniye sürerken doğrusal çekirdek ile 0.91 saniye sürmüştür. Polinomal çekirdek kullanılarak modellenen DVM ile doğrusal çekirdek kullanılarak modellenen DVM arasında doğruluk oranı bakımından farkın çok az olması ve doğrusal çekirdek ile modellemenin yaklaşık 4 kat daha hızlı olması sebebiyle kullanılacak olan DVM sınıflandırıcısında çekirdek olarak doğrusal kullanılmasının daha uygun olduğu belirlenmiştir.

DVM sınıflandırıcısı Naive Bayes 'e göre Nagano, Nepal ve karışık veri setleri üzerinde daha başarılı sonuçlar vermiştir. Çizelge 3.6'da DVM sınıflandırıcısının hata matrisi gösterilmiştir.

Çizelge 3.6 DVM Sınıflandırıcısı Hata Matrisi

	Sınıflandırılan	
	Yüksek Öncelikli	**Düşük Öncelikli**
Yüksek Öncelikli	512	37
Düşük Öncelikli	43	2136

Çizelge 3.6'daki hata matrisine göre DVM sınıflandırıcısı yüksek öncelikli tweet sınıflandırmada başarı oranı %93,3 iken düşük öncelikli tweet sınıflandırmada başarı oranı %98 olmuştur. Ağırlıklı ortalama başarı oranı ise %97,6 dır.

3.2.3 Rasgele Ormanlar Yöntemi ile Sınıflandırma

Nagano, Nepal ve ikisinin karışımı olan veri kümeleri üzerinde 10-katlı çapraz doğrulama yöntemi kullanılarak Rasgele Ormanlar sınıflandırıcısı test edilmiştir. Ağaç sayısı 10, 20 ve 30 alınarak modellenen sınıflandırıcılar ile yapılan testlerin sonucu ayrı ayrı Çizelge 3.7, 3.8, 3.9'da verilmiştir.

Çizelge 3.7 10 Ağaç Sayılı Rasgele Ormanlar Sınıflandırıcı Test Sonuçları

	F-Ölçütü	**Duyarlılık**	**Doğruluk**
Nagano	0.988	0.988	0.988
Nepal	0.891	0.891	0.891
Karışık	0.963	0.963	0.963

Çizelge 3.8 20 Ağaç Sayılı Rasgele Ormanlar Sınıflandırıcı Test Sonuçları

	F-Ölçütü	**Duyarlılık**	**Doğruluk**
Nagano	0.989	0.989	0.989
Nepal	0.9	0.9	0.9
Karışık	0.966	0.966	0.966

Çizelge 3.9 30 Ağaç Sayılı Rasgele Ormanlar Sınıflandırıcı Test Sonuçları

	F-Ölçütü	**Duyarlılık**	**Doğruluk**
Nagano	0.99	0.99	0.989
Nepal	0.906	0.906	0.905
Karışık	0.966	0.966	0.966

Çizelge 3.8'de gösterilen sonuçlara göre 20 ağaç kullanılarak modellenen Rasgele Ormanlar sınıflandırıcısı 10 ağaçlı modele göre daha yüksek başarı oranları vermiştir. Fakat ağaç sayısı 30'a çıkarıldığında karışık veri kümesi üzerindeki başarı oranlarında bir artış görülmemiştir. Model oluşturma süreleri 10 ağaçlıda 3.34 saniye, 20 ağaçlıda 7.18 saniye, 30 ağaçlıda ise 10.5 saniyedir. Bu veriler göz önünde bulundurulduğunda Rasgele Ormanlar sınıflandırıcısı için 20 ağaçlı modelin en uygun olduğu tespit edilmiştir.

20 Ağaçlı Rasgele Ormanlar sınıflandırıcısının Hata Matrisi Çizelge 3.10'da gösterilmiştir. Bu matrise göre yüksek öncelikli tweet'leri doğru sınıflandırma oranı %90,7 iken düşük öncelikli tweet'leri doğru sınıflandırma oranı %98,1'dir. Ağırlıklı ortalama başarı oranı ise %96,62'dir.

Çizelge 3.10 Rasgele Ormanlar Sınıflandırıcı Hata Matrisi

	Sınıflandırılan	
	Yüksek Öncelikli	**Düşük Öncelikli**
Yüksek Öncelikli	498	51
Düşük Öncelikli	41	2138

3.2.4 Farklı Eğitim ve Test Veri Kümesi Kullanılarak Sınıflandırma

Bu bölümde yapılan uygulamanın amacı önceden eğitilmiş sınıflandırdılar kullanılarak başka bir deprem sonrası atılan tweet'lerde yüksek öncelik tespitinin başarısını ölçümlemektir. Bu sebeple Nagano veri kümesi eğitim amaçlı, Nepal veri kümesi ise test amaçlı kullanılmıştır.

Nagano depremi sonrası atılan tweet'ler ile eğitilmiş olan Naive Bayes, DVM ve Rasgele Ormanlar sınıflandırıcıları Nepal depremi sonrası atılan tweet'ler ile test edilmiştir. Önceki uygulamalardan elde edilen değerlere göre Rasgele Ormanlar sınıflandırıcısı için en uygun değer olan 20 ağaç sayısı, DVM yöntemindeki çekirdek için en uygunu olan doğrusal çekirdek seçilmiştir. Test sonuçları Çizelge 3.11'de gösterilmiştir.

Çizelge 3.11 Farklı Eğitim ve Test Veri Kümesi ile Sınıflandırma Sonuçları

	F-Ölçütü	**Duyarlılık**	**Doğruluk**
Naïve Bayes	0.79	0.81	0.773
Rasgele Ormanlar	0.806	0.920	0.765
DVM	0.842	0.927	0.811

Çizelge 3.11'deki sonuçlar değerlendirildiğinde Naive Bayes ve Rasgele Ormanlar sınıflandırıcılarının doğruluk oranlarının %80 altına düştüğü görülmektedir. Diğer sınıflandırıcılara göre daha başarılı sonuç veren DVM yöntemi ise %81 doğruluk oranı vermiştir. Böylece DVM yöntemi kullanılarak oluşturulan yüksek öncelikli tweet tespit modelinin daha sonra meydana gelen depremler sonrası atılan tweet'ler üzerinde kullanılabileceği görülmüştür.

3.2.5 Değerlendirme

Değerlendirme, Nepal ve Nagano depremlerinin karışımı olan veri kümesi üzerinde yapılan sınıflandırma sonuçları göz önünde bulundurularak yapılmıştır. Nagano veri kümesi 241 yüksek öncelikli ve 1746 düşük öncelikli tweet içerirken, Nepal veri kümesi

ise 308 yüksek öncelikli ve 433 düşük öncelikli tweet içermektedir. Bu veri kümelerinin karıştırılması ile 549 yüksek öncelikli, 2179 düşük öncelikli tweet elde edilmiştir. Karışık veri üzerinde yapılan sonuçlarda 549 yüksek öncelikli etiketli tweet'lerden 512 tanesini başarılı bir şekilde tespit eden DVM en yüksek başarıyı gösteren sınıflandırıcı olmuştur. Aynı zamanda DVM yöntemi , eğitim ve test için farklı depremlere ait veri kümeleri kullanıldığında da diğer yöntemlere göre en az etkilenen yöntem olmuştur.

Çizelge 3.12 Genel Sınıflandırma Başarı Sonuçları

	F-Ölçütü	**Duyarlılık**	**Doğruluk**
Naïve Bayes	0.914	0.929	0.909
Rasgele Ormanlar	0.966	0.966	0.966
DVM	0.971	0.971	0.971

Başarı oranlarının yüksek olmasının sebebi yüksek öncelikli olarak etiketlenen tweet'lerde "injury, callopsed, trapped, destroyed, dead" gibi kelimelerin sıklığından ötürü oldukça ayırt edici nitelikte olmasıdır.

Modelleme süresi ve yüksek öncelikli tweet sınıflandırma başarı oranları dikkate alındığında doğrusal çekirdekli DVM sınıflandırıcısının yüksek öncelikli tweet tespiti için en uygun yöntem olduğu tespit edilmiştir.

3.3 Yüksek Öncelikli Tweet'lerin Özetlenmesi

Bu bölümde Nagano ve Nepal depremleri sonrasında toplanan twitter verisi üzerinde özetleme yapılmıştır. Nepal depremiyle ilgi veri sayısı daha yüksek olduğundan ağırlıklarıyla birlikte en yüksek 3 adet özet tweet örneği verilmiştir. Her bir özetleme algoritması yüksek öncelikli tweet tespiti olmadan ve yüksek öncelikli tweet tespiti yapılarak test edilmiştir. Test işlemlerinin sonuçları verilmeden önce, her bir veri kümesi içerisindeki yüksek öncelikli tweet'ler arasında kullanım sıklığı en yüksek olan kelimelerin karşılaştırılması yapılarak yüksek önceliği belirten kelimelerin depremlere göre değişiklik gösterip gösterilmediği incelenmiştir.

3.3.1 Veri Kümelerindeki Kullanım Sıklığı Yüksek Olan Kelimelerin Karşılaştırılması

Kullanılacak olan özetleme yöntemlerinin her birinde kelimelerin kullanım sıklıkları dikkate alınmaktadır. Bu nedenle farklı veri kümelerinde yüksek öncelikli olarak etiketlenen tweet'ler içerisinde kullanım sıklığı yüksek olan kelimeler karşılaştırılmıştır.

Depremi ifade eden "*earthquake, quake*" gibi kelimeler, deprem bölgesini ifade eden "*nagano, nepal*" gibi kelimeler ve sık kullanılan fakat bilgi değeri olmayan "*in, and, a, to*" gibi kelimeler sık kullanılanlar listesinden çıkarılmıştır.

Nagano depreminde etiketlenen yüksek öncelikli tweet'ler içerisinde kullanım sıklığı en yüksek olan 10 kelime şu şekildedir : *"injuries, hits, homes, city, collapses, causes, shakes, injuring, people, injured".*

Nepal depreminde etiketlenen yüksek öncelikli tweet'ler içerisinde kullanım sıklığı en yüksek olan 10 kelime ise şu şekildedir : *"people, injured, least, killed, trapped, avalanche, everest, dead, hits, breaking".*

Nepal depreminde ölümün çok daha yüksek olması "*killed*" kelimesinin sık kullanılmasına neden olmuştur. Deprem bölgesine özel *"everest,city"* gibi kelimelerin dışında kalan *"injured, trapped, causes, collapses"* gibi diğer kelimeler veri kümeleri için yüksek önceliği belirtenler arasındadır. Depremin şiddeti ve neden olduğu hasarın büyüklüğüne göre yüksek öncelikli tweet'ler içerisinde sık geçen kelimelerden bazıları değişiklik gösterebilmektedir; fakat en önemlilerinden bir tanesi olan yaralı bilgisini veren *"injured, injuries"* gibi kelimelerin veri kümelerinde ortak olduğu görülmüştür.

3.3.2 SumBasic ile Özetleme

SumBasic özetleme yönteminde tweet özet puanı hesaplanırken içerdiği kelimelerin doküman genelindeki kullanım sıklığı toplanır. Uygulaması basit ve sık kullanılan kelimenin daha önemli olduğunu düşünen bu yöntem ile yapılan özetleme sonuçları aşağıda verilmiştir.

22 Kasım 2014 tarihinde Japonya'nın Nagano şehrinde meydana gelen deprem sonrası toplanan yaklaşık 2000 bin tweet üzerinde yüksek öncelikli tweet tespiti olmadan SumBasic ile yapılan özetleme sonuçları Çizelge 3.13'deki gibidir ;

Çizelge 3.13 SumBasic ile Nagano Veri Kümesi Özetleme Sonuçları

	SumBasic	**Sayı**
Yüksek öncelikli tweet tespiti olmadan	"*RT @HaitiNewsNet: Magnitude 6.8 Quake Hits Central Japan - A house in Hakbua, Nagano Prefecture, collapsed after a strong earthquake... htt…"*	2000
Yüksek öncelikli tweet tespiti yapılarak	**"***RT @abcnews: 21 people trapped, 13 injured and five houses destroyed as magnitude 6.8 earthquake hits Nagano in Japan http://t.co/K9m5cTEVud***"**	240

25 Nisan 2015 tarihinde meydana gelen Nepal depremi sonrasında ki 3 saat içerisinde toplanan 147 bin tweet yüksek öncelikli tweet tespiti olmadan SumBasic ile özetleme işlemi yapıldığında Çizelge 3.14'deki gibi sonuçlar alınmaktadır;

Çizelge 3.14 SumBasic ile Nepal Veri Kümesi Özetleme Sonuçları

	SumBasic		**Sayı**
	Ağırlık	**Özet Tweet**	
Yüksek öncelikli tweet tespiti olmadan	0.3059	*"Prayers for people of Nepal and India affected by the earthquake. I was in Kathmandu a couple of weeks back. Feeling lucky to be back."*	147000
	0.3040	*"People try to free a man from the rubble of a destroyed building after an earthquake [7.9 magnitude] hit Kathmandu in Nepal via /r/pics*"	
	0.2940	*"Police say at least 108 people killed after the 7.9 magnitude earthquake in Nepal. All my prayers to them*"	
Yüksek öncelikli tweet tespiti yapılarak	0.3720	"*Nepal's Home Ministry says at least 71 people have been killed in a 7.9-magnitude earthquake that hit the capital and the Kathmandu Valley*."	5550
	0.3642	"*Hundreds Killed in Nepal Earthquake, Tremors Across North India: At least 100 people are feared dead in a massive earthquake of 7.9 m...*"	
	0.3525	"*Police say at least 108 people killed after the 7.9 magnitude earthquake in Nepal. All my prayers to them*"	

3.3.3 TF-ISF ile Özetleme

TF-ISF özetleme yönteminde kelimelerin genel kullanım sıklığı ve kelime içerisindeki kullanım sıklığı kullanılarak özet ağırlık puanı hesaplanır. TF-ISF yöntemine göre hem genel olarak sık kullanılan hem de tweet içerisinde tekrar edilen kelimeler daha önemlidir. Bu yöntem kullanılarak yapılan özetleme sonuçları Çizelge 3.15 ve Çizelge 3.16'daki gibidir;

Çizelge 3.15 TF-ISF ile Nagano Veri Kümesi Özetleme Sonuçları

	TF-ISF	Sayı
Yüksek öncelikli tweet tespiti olmadan	*"6.8 quake rattles Nagano: NETBALL RESULTS: BWU Linda Brooks... The magnitude-6.8 earthquake hit ... http://t.co/jHAkN8t4G8"*	2000
Yüksek öncelikli tweet tespiti yapılarak	***"**Yesterday, earthquake tremors 6 has occurred in Nagano. The ground was shaking!I was scared!! http://t.co/CvQmvUMiJP**"***	240

TF-ISF ile Nepal depremi veri kümesi üzerinde yüksek öncelikli tweet tespiti yapılarak özetleme yapıldığı durumda Çizelge 3.16'daki sonuçlar alınmaktadır;

Çizelge 3.16 TF-ISF ile Nepal Veri Kümesi Özetleme Sonuçları

TF-ISF			Sayı
Ağırlık		**Özet Tweet**	
Yüksek öncelikli tweet tespiti yapılarak	6.1164	*"NEPAL STILLS 2 - STILLS show earthquake damage, injured treated. 0830GMT / No.2056362"*	5550
	5.7608	*"NEPAL STILLS - STILLS show people in street, quake injured. 0830GMT / No.2056360"*	
	5.7062	*"RT @ANI_news: Everyone is very panicked, several house walls and buildings have collapsed: Arun Joshi (Assignment Head, News24 Nepal) #eart…"*	

3.3.4 Hibrit TF-IDF ile Özetleme

Hibrit TF-IDF yönteminde kelimelerin genel kullanım sıklığı ile tweet bazında kulanım sıklığını hesaba katarak özet ağırlık puanı hesaplanır. Hibrit TF-IDF yöntemine göre sık kullanılan ve olabildiğince fazla sayıda tweet içerisinde kullanılan kelimeler daha değerlidir. Bu yöntem kullanılarak yapılan özetleme sonuçları Çizelge 3.17 ve Çizelge 3.18'de verilmiştir.

Nagano depremi veri kümesi üzerinde yüksek öncelikli tweet tespiti olmadan yapılan özetleme sonuçları Çizelge 3.17'de gösterildiği gibidir;

Çizelge 3.17 Hybrid TF-IDF İle Nagano Depremi Üzerinde Özetleme

	Hibrit TF-IDF	**Sayı**
Yüksek öncelikli tweet tespiti olmadan	"http://t.co/Dx62zsu3Fu Magnitude 6.8 earthquake hits Nagano, Japan: A magnitude 6.8 earthquake has struck centr... http://t.co/IYRflj21oe"	2000
Yüksek öncelikli tweet tespiti yapılarak	**"***21 people trapped, 13 injured and five houses destroyed as magnitude 6.8 earthquake hits Nagano in Japan http://t.co/K9m5cTEVud***"**	240

Nepal depremi sonrasında ki 2 saat içerisinde toplanan 147 bin tweet yüksek öncelikli tweet tespiti olmadan ve yüksek öncelikli tweet tespiti yapılarak Hibrit TF-IDF ile özetleme işlemi yapıldığında Çizelge 3.18'deki sonuçlar alınmaktadır;

Çizelge 3.18 Hybrid TF-IDF ile Nepal Veri Kümesi Özetleme Sonuçları

	Hybrid TF-IDF		**Sayı**
	Ağırlık	**Özet Tweet**	
Yüksek öncelikli tweet tespiti olmadan	0.0227	"*7.9-magnitude earthquake strikes Nepal: 7.9-magnitude earthquake strikes Nepal http://t.co/ZO1B9Ru6z6... http://t.co/SiCFNUBpQv*"	147000
	0.0220	*"#Earthquake of 7.9 magnitude shakes Nepal, tremors felt in India: A 7.9 magnitude earthquake…http://t.co/sxy7EitcUR*"	
	0.0193	"*Buildings Collapse, At Least 2 Killed as 7.9 Earthquake Hits Nepal, India: Witnesses: An earthquake measuring 7.9 magnitude, and only...*"	
Yüksek öncelikli tweet tespiti yapılarak	0.0273	"*7.9-magnitude quake hits Nepal near Kathmandu: More than 150 people have died after a powerful 7.9 magnitude e... http://t.co/xSi0pSncbo*"	5550
	0.0250	"Buildings Collapse, At Least 2 Killed as 7.9 Earthquake Hits Nepal, India: Witnesses: An earthquake measuring 7.9 magnitude, and only..."	
	0.0214	"*50 people trapped in nepal and three people have died in NEPAL 7.9 earthquake .. My thoughts are with everyone*"	

Nepal depreminden 24 saat sonra yaklaşık 2-3 saatlik zaman diliminde toplanan 120 bin üzerinde yüksek öncelikli tweet tespiti yapılarak özetleme yapıldığı durumda ise Çizelge 3.19'daki sonuçlar elde edilmiştir.

Çizelge 3.19 Hybrid TF-IDF ile Nepal Veri Kümesi Özetleme Sonucu

	Hibrit TF-IDF	**Sayı**
Yüksek öncelikli tweet tespiti yapılarak	*"2,000 confirmed dead in Nepal quake: At least 2,152 people are now confirmed dead in the massive earthquake that… http://t.co/Tt4vdZJ1hR"*	2000

3.3.5 Değerlendirme

Yapılan özetlemelerde , özetleme öncesi yüksek öncelikli tweet tespiti yapılmadığı takdirde kullanılan yönteme bağlı olarak daha çok depremin büyüklüğü ve meydana geldiği bölge hakkında bilgi veren tweet'ler ön plana çıkmaktadır. Özetleme öncesi yüksek öncelikli tweet tespiti kullanıldığı takdirde özetleme sonucunda deprem hakkında ölü, yaralı, mahsur ve hasar bilgisi veren bilgi açısından değerli tweet'ler ön plana çıkmıştır.

Kullanılan özetleme yöntemlerinden SumBasic'de tweet özet puanı, içerdiği kelimelerin kullanım sıklıkları toplanarak hesaplandığından uzun tweet'ler daha ön plana çıkmaktadır.

TF-ISF yönteminde ise hesaba kelimenin tweet içerisindeki kullanım sıklığının (Inverse Sentence Frequency) da katılması sebebiyle içerisinde tekrar eden kelimeleri içeren tweet'ler ön plana çıkmaktadır.

Hybrid TF-IDF yönteminin SumBasic'ten farklı olarak normalizasyon işlemini içermesi ve TF-ISF'ten farklı olarak tweet'lerin genelini (Inverse Document Frequency) hesaba katması daha başarılı sonuçlar vermiştir. Ek olarak anlamlı bilgilerin kaybını engellemeye yönelik, tweet'ler içerisinden "earthquake" ve depremin ait olduğu bölgeye özel olarak "japan, nagano, nepal" gibi kelimeler filtrelenerek özetlemeye dahil edilmemiştir.

Özetleme sonuçlarının başarılarını ölçerken, sonuçların göreceli olması sebebiyle, özet olarak nitelendirilen tweet'leri yaralı ve hasar durumu hakkında net ve öz bilgi vermesi açısından elle değerlendirilmiştir.

Hybrid TF-IDF ile Nagano verisi üzerinde yüksek öncelikli tweet tespiti ile birlikte yapılan özetleme sonucu gelen tweet ; *"21 people trapped, 13 injured and five houses*

destroyed as #magnitude 6.8 #earthquake hits #Nagano in #Japan http://t.co/Xp95pJKKUI" şeklindedir. Özet tweet'te görüleceği gibi amaçlanan şekilde mahsur, yaralı ve hasar durumu hakkında bilgi edinilmektedir.

BÖLÜM 4

SONUÇ VE ÖNERİLER

Bu çalışmada, milyonlarca kullanıcının aktif olarak paylaşım yaptıkları Twitter gibi mikroblog servislerini sosyal projelerde bilgi kaynağı olacak hale getirmeyi genel hedef edindik.

Günümüzde sosyal medya kullanımının gittikçe yaygınlaşması ile birlikte, kullanıcılar aktif bir şekilde paylaşımlarda bulunmakta ve bu paylaşımlar çeşitli türlerde bir çok bilgiyi kendi bünyesinde barındırmaktadır. Doğru analiz yöntemleri kullanılarak bu türden bilgiler çeşitli alanlarda kaynak haline getirilebilir. Internet üzerinde kullanıcı paylaşımlarının analizi üzerine yapılan ilk araştırmalardan biri olan filmler hakkında yapılan yorumların değerlendirilerek izleyici tepkisinin ölçülmesi bu duruma örnek olarak verilebilir.

Artan sosyal medya kullanımı ile birlikte çeşitlenen kullanıcı paylaşımları, veri madenciliğinin felsefesi olan "kullanılabilir veriden faydalı bilgi çıkarma" ilkesi doğrultusunda popülerlik kazanan bir araştırma alanı olmuştur.

Bu çalışmanın hedefi doğrultusunda sosyal projelere bilgi kaynağı olabilecek nitelikte olan, aynı anda çok sayıda kullanıcı tarafından paylaşım yapıldığı ve mikroblog servisler arasında günümüzde en yaygın kullanımda olan Twitter seçildi. Aynı zamanda Twitter'ın sağladığı API sayesinde istenilen anahtar kelimelere yönelik veri toplama imkanı elde edilmesi kaynağı oluşturabilmek adına önemli bir kriter olmuştur.

Sosyal projelerin en önemlilerinden birisi olan, zamanın iyi kullanılmasının ve yerinde müdahalenin çok önemli olduğu doğal afetler sonrası yardım olduğu düşünülerek bu alanda bir çalışma gerçekleştirilmiştir. Doğal afetler sonrasında kullanıcılar konuyla ilgili bir çok tweet atmaktadır. Bu tweet'lerden çok büyük bir kısmı olay ile ilgili üzüntü ve iyi dilekleri belirtirken bir kısmı ise yaralı, hasar ve mahsur bilgisini verebilmektedir. Sayısı az fakat yardım birimleri için oldukça değerli olabilecek bu tür bilgilerin tespit edilmesi ve sunulması bu çalışma kapsamında edinilen temel amaç olmuştur. Yüksek öncelikli

olarak tabir ettiğimiz bu türden bilgileri veren tweet'lerin tespiti için makine öğrenmesi yöntemleri kullanılmıştır. Bu yöntemler ile yüksek öncelikli tweet'ler tespit edildikten sonra yazı sınıflandırma alanında da kullanılan ve doküman içerisinden öz bilgiyi çıkarmayı hedefleyen özetleme yöntemleri kullanılmıştır. Özetleme yapılarak yüksek öncelikli olanları en iyi temsil edebilecek nitelikte olan tweet'ler seçilerek yardım birimleri için hem zamandan tasarruf sağlamak hem de öz bilgiye ulaşmada kolaylık sağlamak hedeflenmiştir.

Geliştirilen sistemin başarısı, Twitter'ın geriye dönük en fazla sekiz günlük aramaya izin vermesi ve daha önceki tarihlere ait veriye ulaşmanın ücretli olması sebebiyle bu çalışma esnasında gerçekleşmiş olan deprem sonrasında atılan tweet'ler üzerinde değerlendirilmiştir. Bu tweet'ler Amerika'nın California eyaletinde, Japonya'nın Nagano şehrinde ve Nepal'in Kathmandu şehrinde meydana gelen depremler sonrası toplanmıştır.

Uygulamalardan elde edilen sonuçlara göre yüksek öncelikli tweet'lerin tespiti ve özetlenebilmesi sayesinde, yardım birimlerinin doğru yere doğru zamanda müdahale edebilmesi için kullandığı kaynaklar arasına kullanıcıların anlık bilgi paylaştığı Twitter'ın eklenmesinin de mümkün olduğu ve önemli bir katkı sağlayacağı görülmüştür.

İleriki çalışmalarda, diğer doğal afetlerde ve aynı zamanda sosyal projelerde kullanılabilecek yardım ve ihtiyaç tespitinin yapılması, bu tespitlerin gerekli birimlerin faydalanabileceği hale getirilmesi hedeflenmektedir.

KAYNAKLAR

[1] Java, A., Song, X., Finin, T. ve Tseng, B. (2007). "Why we Twitter: Understanding Microblogging Usage and Communities", Proceeding of the 9th Web KDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, 56-65

[2] Niharika, S., Sneha, V.L. ve LavanyA, D.R. (2012). "A SURVEY ON TEXT CATEGORIZATION", International Journal of Computer Trends and Technology, 3:39-45

[3] Neethu, M.S. ve Rajasree R., (2013)."Sentiment Analysis in Twitter using Machine Learning Techniques", 4th ICCNT, 2013 July 4-6, Tiruchengode, India

[4] Bahrainian, S. ve Andreas, D., (2013). "Sentiment Analysis Using Sentimen Feautures", IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), 3:26-29

[5] Lima, A. ve De Castro, L.N., (2012). "Automatic Sentiment Analysis of Twitter Messages", Fourth International Conference on Computational Aspects Of Social Networks, 52-57

[6] Bahrainian, S. ve Dengel, A., (2013). "Sentiment Analysis and Summarization of Twitter Data ", Proceedings of the 2013 IEEE 16th International Conference on Computational Science and Engineering, 227-234

[7] Sharifi, B., Hutton, M.A. ve Kalita, J.K., (2010). "Experiments in Microblog Summarization", In Proc. of IEEE Second International Conference on Social Computing, 49-56

[8] Inouye, D. ve Jugal, K., (2011). "Comparing Twitter Summarization Algorithms for Multiple Post Summaries", IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, 298 - 306

[9] Yamamoto, Y., Tomita, M., Uchida, O. ve Kajita, Y., (2014). "Disaster mitigation support system using Twitter and GIS", ICT and Knowledge Engineering (ICT and Knowledge Engineering), 18-23

[10] Middleton, S.E., Middleton, L. ve Modafferi, S., (2014). “Real-Time Crisis Mapping of Natural Disasters Using Social Media”, IEEE Computer Society Social Intelligance and Technology, 29(2):9-17

[11] Matsumoto, S. ve Toriumi, F., (2014). “A Method to Quantify Twitter User's Posting Activities for Constructing Disaster İnformation Support System”,Computational Intelligence and Applications (IWCIA), IEEE 7th International Workshop, 137-140

[12] Chatfield, A.T. ve Brajawidagda, U., (2013). “Twitter Early Tsunami Warning System: A Case Study in Indonesia’s Natural Disaster Management”, System Sciences (HICSS), 46th Hawaii International Conference, 2050 - 2060

[13] Bhuta, S., Doshi, A., Doshi, U. ve Narvekar, M., (2014). “A Review of for Sentiment Analysis Of Twitter Data”, InternationalConference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 583 - 591

[14] Go, A., Bhayani, R. ve Huang, L., (2009). “Twitter Sentiment Classification using distant supervision”, technical report, Standford

[15] Pang, B., Lee, L. ve Vaithyanathan, S., (2012). “Thumbs up Sentiment Classification using Machine Learning Techniques.” In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, University of Pennsylvania, 79-86

[16] Vanderwende, L., Suzuki, H., Brockett, C. ve Nenkova, A., (2007). “Beyond Sumbasic: Task-Focused Summarization with Sentence Simplification and Lexical Expansion”, Information Processing & Management, 43(6):1606–1618

[17] Wadhvani, R., Pateriya, R.K. ve Roy, D., (2013). “A Topic-driven Summarization using K-mean Clustering and Tf-IsfSentece Ranking”, International Journal of Computer Applications (0975 8887), 79(8):39-45

[18] Sokolova, M., Japkowicz, N. ve Szpakowicz, S., (2006). “Beyond Accuary, FScore and ROC: a Family of Discriminant Measures for Performance Evaluation”, Advances in Artificial Intelligence

[19] Bahrainian, S. ve Dengel, A., (2013). “Sentiment Analysis and Summarization of Twitter Data”, IEEE 16th International Conference on Computational Science and Engineering, 227 - 234

[20] Breiman, L. ve Adele, C., (2001). “Random Forests”, Machine Learning, 45(1):5-32

[21] Durgesh, K. S., ve Lekha, B., (2009), “Data Classification Using Support Vector Machine”, Journal of Theoretical and Applied Information Technology, 12(1):1-7

[22] Lowd, D. ve Domingos, P., (2005) “Naive Bayes Models for Probability Estimation”, International Conference on Machine Learning, 2005, Bonn

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Kadir KEBABCI

Doğum Tarihi ve Yeri : 17.11.1990 / Bakırköy

Yabancı Dili : İngilizce

E-posta : kadir.bmh@gmail.com

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Lisans	Bilgisayar Mühendisliği	İstanbul Üniversitesi	2012
Lise	Fen Bilimleri	Kırklareli Anadolu Lisesi	2008

İŞ TECRÜBESİ

Yıl	Firma/Kurum	Görevi
2014	Forum Medya Eklam Organızasyon Ve Tıc. Ltd. Ştı.	Proje Yazılım ve Geliştirme Uzmanı
2013	Defne Telekomünikasyon A.Ş.	Yazılım Geliştirici
2012	Innova Bilişim Çözümleri A.Ş	Yazılım Geliştirici
2011	Nodeser	Yazılım Geliştirici

YAYINLAR

Bildiri

Kebabcı, K. ve Karslıgil, M.E., (2015), “High priority tweet detection and summarization in natural disasters”, Signal Processing and Communications Applications Conference (SIU), 2015 May 23th, Malatya