T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE TWITTER'DA SORU ALGILAMA

ZEYNEP BANU ÖZGER

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
DOÇ. DR. BANU DİRİ

İSTANBUL, 2014

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE TWITTER'DA SORU ALGILAMA

Mustafa Özgür CİNGİZ tarafından hazırlanan tez çalışması 30.01.2014 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Doç. Dr. BANU DİRİ

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Doç. Dr. BANU DİRİ

Yıldız Teknik Üniversitesi _____

Yrd. Doç. Dr. AYSUN GÜRAN

İstanbul Kültür Üniversitesi _____

Yrd. Doç. Dr. M. FATİH AMASYALI

Yıldız Teknik Üniversitesi _____

ÖNSÖZ

Teknolojinin gelişmesi neticesinde sosyal ağlara günümüzde mobil cihazlardan da erişilebilmektedir. Bu durum; internet kullanımını yaygınlaştırmakta ve sosyal ağ kullanıcılarının sayısını hızla arttırmaktadır. Son yıllarda mikro bloglar sadece bir eğlence aracı olmaktan çıkmış, bilgi paylaşımı yapılan interaktif bir ortam haline de gelmiştir.

Soru algılama, Doğal Dil İşleme' nin bir alt çalışma alanı olup, verilen bir derlemden soru içeren ifadelerin bulunabilmesini amaçlar. Bu çalışmada ilk olarak Türkçe tvitlerden soru içerenlerin tespit edilmesi ardından da yedi farklı soru türünden hem soruyu hem de cevabı içeren tvitlerin tespit edilmesi amaçlanmıştır. Her iki aşamada da Şartlı Rastgele Alanlar kullanılarak algılama işlemi gerçeklenmiştir.

Bilgi birikimi, görüşleri ve pozitif yaklaşımıyla çalışmamın her aşamasında destek olan, değerli hocam Doç. Dr. Banu Diri'ye, çalışma süresince yardımını esirgemeyen sevgili meslektaşım Canan Girgin'e ve çalışmamı bitirebilmem için motivasyonumu daima yüksek tutan değerli aileme teşekkür ederim.

Ocak, 2014

Zeynep Banu ÖZGER

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ

$p(X, Y)$	Olasılık dağılımı
y^*	Etiket dizisi
x^*	Gözlem Dizisi
(V, E)	Yönlü bir Grafik
G	Grafik
λ_i	Çalışma verilerinden tahmin edilebilecek parametre
$Z(x)$	Normalleştirme faktörü
s_k	Durum özellik işlevi

KISALTMA LİSTESİ

LDA	Linear Discriminant Analysis
POS	Part of Speech
5N1K	Ne, ne zaman, nerede, nasıl, neden, kim
CRF	Conditional Random Fields
GMM	Gizli Markov Model
MEMM	Maksimum Entropi Markov Model
OAuth	Open Authorization
API	Application Interface
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
MUC-6	Sixth Message Understanding Conference
MUC-7	Seventh Message Understanding Conference
QA	Question With Answer
NQ	Not a Question
RQ	Rhetorical Question
FK	Factual Knowledge

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

TÜRKÇE TWITTER'DA SORU ALGILAMA

Zeynep Banu ÖZGER

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Doç. Dr. Banu Diri

Twitter gibi mikro-blog servislerinin kullanımının son yıllarda katlanarak arttığı görülmektedir. Her gün tvit adı verilen, 140 karakterden oluşan, kullanıcıların günlük aktiviteleri, görüşleri ve ilgi alanlarından oluşan milyonlarca mesaj gönderilmektedir. Bununla birlikte Twitter; kullanıcılara birbirlerine doğrudan mesaj göndermek yoluyla kendisini takip edenler ile bir sosyal ağ kurma imkânı da sağlamaktadır. Günümüzde kullanıcılar her yerden erişebildikleri için sosyal medya ağlarını bilgi paylaşmak ve sorularına cevap alabilmek için de kullanmaktadırlar. Kullanıcıların Twitter üzerinde oluşturduğu büyük miktardaki ilişkisel ve metinsel veri, araştırmacıları bu alanda çalışmalar yapmaya teşvik etmektedir.

Soru algılama Doğal Dil İşleme' nin bilgi çıkarımı alanının bir alt dalıdır. Dilin yapısal kurallarına uyan veya uymayan derlemlerden soru içeren cümleleri tespit etmeyi amaçlar. Soru algılama ile ilgili yapılan ilk çalışmalar kurallı metinler üzerinde olup İnternetin yaygınlaşması ile birlikte forum siteleri gibi düzensiz verilere yönelmiştir. Çalışmalar genellikle, derlemin yazıldığı dile bağlı kural tabanlı olarak tasarlanmış olup eğitim aşamasında çeşitli makine öğrenmesi yöntemlerinden yararlanılmıştır.

Tez kapsamında Türkçe tvitlerden oluşan bir veri seti için, Şartlı Rastgele Alanlar metodu kullanılarak geliştirilmiş bir soru algılama sistemi geliştirilmiştir. Çalışma genel olarak dört adımdan oluşmaktadır. İlk olarak Türkçe tvitleri içeren bir veri seti oluşturulmuş ve bir ön-işleme metodu ile tvitler retvit, kullanıcı adı gibi sistem için

anlamlı olmayan veriden arındırılmıştır. Çalışmanın ikinci aşamasında, veri setinden kural tabanlı bir yöntem ile soru içermeye aday tvitler belirlenmiştir. Ardından Türkçe için soru kalıpları tanımlanarak, Şartlı Rastgele Alanlar metodu ile soru olmaya aday tvitlerden soru içerenler tespit edilmiştir. Çalışmanın son aşamasında ise veri setindeki yedi farklı soru türünden birini algılamaya yönelik bir sistem yine Şartlı Rastgele Alanlar metodu kullanılarak geliştirilmiştir.

Performans değerlendirme sonuçlarına göre, örüntüleri desteklemek için tanımlanan küçük boyuttaki sözlüklerin başarıyı artırdığı gözlemlenmiştir. Ayrıca, özellik olarak tanımlanan örüntülerin hassaslaştırılması; soruların tespiti aşamasındaki başarıyı artırırken, tvitlerin kuralsız veri olmasından dolayı soru olmayan tvitlerin soru olarak etiketlenmesindeki hata oranını artırmaktadır. Bu nedenle, örüntüler her iki taraftaki hatayı dengede tutacak şekilde tanımlanmıştır.

Anahtar Kelimeler: Soru algılama, Doğal Dil İşleme, Şartlı Rastgele Alanlar, Twitter, Sosyal ağlar

ABSTRACT

QUESTION IDENTIFICATION ON TURKISH TWITTER

Zeynep Banu ÖZGER

Department of Computer Engineering

MSc. Thesis

Advisor: Assoc. Prof. Dr. Banu DİRİ

The use of micro-blogging services such as Twitter has increased exponentially in recent years. Twitter users are sent to millions of 140-character messages every day which called Tweet. Tweet texts are composed of users' daily activities, opinions and interests. However, Twitter also provides an opportunity for users to establish social networks of people who follow one another's Tweets via send direct messages to each other. Today users are using Twitter to share information and to get answers to their questions so they can access from anywhere. Large amounts of textual and relational data on Twitter, has spurred to researchers working in this field.

Question detection is a sub-task of information extraction field of Natural Language Processing. It aims to detect sentences which contain question statements from a regular or an irregular corpus. Earlier studies about question identification are on regular texts. Yet, with the popularization of using the Internet, researchers have turned to irregular data such as online forums and micro-blogging services. Studies about this area are designed as rule based, in training phase various machine learning methods were used.

In this study, Conditional Random Fields method is used to identify questions from a data set which consists of Turkish tweets. The study consists of four steps. Initially, it is formed a data set including Turkish tweets and pre-processing method tweets in data set is cleaned from unnecessary data for the system like retweet and user name. In the second stage of study, candidate tweets were identified via a rule-based system.

Following, question tweets were detected from candidate tweets using Conditional Random Field method. In the last step, a system has been developed that can detect one of seven different question types.

According to the performance evaluation results, it was observed that the small-size dictionaries for using in patterns increased the success. Additionally, sensitizing the patterns that identified as feature increased the success in question detection phase, because of the irregular structure of tweets, labeling non-question tweets as question tweets error rate also increased. Therefore, patterns are identified to hold in balance both of two sides.

Key words: Question detection, Natural Language Processing, Conditional Random Fields, Twitter, Social network.

BÖLÜM 1

GİRİŞ

Bilgi teknolojilerindeki ilerlemeler neticesinde kullanıcılar artık sadece bir okuyucu değil aynı zamanda içerik sağlayıcısı konumuna gelmiştir. Web 2.0 teknolojisi ile ortaya çıkan ve hızla yaygınlaşan sosyal ağ ve mikro-bloglarda kullanıcılar düşüncelerini, aktivitelerini, fotoğraf ve videolarını paylaşabilmektedir. En yaygın mikro-blog sitelerinden biri olan Twitter' da kullanıcılar kısa içeriklerini paylaşmaktadırlar. Tvit olarak adlandırılan bu kısa içerikler araştırmacılara büyük kitlelerin düşüncelerine ulaşabilme imkânı verdiği için son yıllarda mevcut tvit verilerini incelemek adına pek çok çalışma yapılmaktadır.

Soru algılama Doğal Dil İşleme ve Bilgi Çıkarımı disiplinlerinin bir alt çalışma alanıdır. Soru algılama çalışmalarının amacı bir metinde geçen soru cümlelerini tespit edebilmektir. Günümüzde Twitter sanal ortamda gayri resmi bir bilgi etkileşim ortamı haline gelmiştir. Soru sormak ise bu etkileşimin önemli bir parçasıdır [1].

Bu çalışma kapsamında Türkçe tvitlerden, soru içeren tvitlerin belirlenmesi işlemi gerçekleştirilmiştir. [1, 2, 3, 4 ve 5] deki çalışmalarda kullanılan soru türleri incelenmiş bunlar arasından yedi soru türü belirlenmiş ve tvitin soru içerip içermediği tespit edildikten sonra bu soru türlerinden seçilen bir türe ait soruların bulunması işlemi gerçekleştirilmiştir.

1.1 Literatür Özeti

Bireylerin sınırları belirlenmiş bir sistem içinde halka yarı/açık profil oluşturmasına, bağlantıda olduğu diğer kullanıcıların listesini açıkça vermesine ve bu kullanıcıların

sistemdeki listelenmiş bağlantılarını görmesine ve aralarında gezinmesine izin veren web tabanlı hizmetlerin tümüne Sosyal Ağlar denmektedir [6].

Web 2.0 teknolojisinin gelişmesi ile birlikte amacı bilgi alışverişi yapmak, aynı sektörde çalıştıkları veya ilgi alanı aynı olan insanlarla tanışmak fikir alış verişi yapmak isteyen kullanıcılar sosyal ağ sitelerini kullanmaya başladılar. Bu sitelerde oluşan ilişkiler gelişerek, gerçek hayatta da buluşan insanların oluşturdukları sosyal medyayı ortaya çıkardı.

Sosyal medyada insanlar zaman ve mekân sınırlaması olmadan paylaşım ve tartışma yapabilmektedir. Gayri resmi bir eğitim alanı olarak da görülen sosyal medya günümüzün en etkili pazarlama alanlarından biri olarak kullanılmaktadır.

Günümüzdeki en yaygın sosyal ağlardan biri olan Twitter 2006 yılında kurulmuştur. Kullanımının ve erişiminin kolay olması Twitter' ın kullanımını artırmıştır. Twitter ile kullanıcılar duygu ve düşüncelerini paylaşabilmekte, fotoğraf ve videolarını yayınlayabilmektedir [8].

2013 yılında Twitter' a ait istatistikler aşağıdaki gibidir [9]:

- Aylık 230+ milyon aktif kullanıcıya sahiptir.
- Günde 500 milyon tvit gönderilmektedir.
- Aktif kullanıcıların %76'sı mobil cihazlardan erişim yapmaktadır.
- Hesapların %77'si ABD dışındandır.
- Twitter 35'den fazla sayıda dili desteklemektedir.

Kullanıcıların gönderdikleri iletilere tvit denir. Her bir tvit 140 karakter ile sınırlandırılmış metin, video, fotoğraftan oluşabilir. Kullanıcılar 140 karakter sınırlamasını daha verimli kullanabilmek için kelimeleri kısaltmakta ve sesli harflerin kullanımını azaltmaktadır. Bu durum ise tvitlerin analiz edilmesini zorlaştırmaktadır. Paylaşımlar belirli bir grupla sınırlandırılabileceği gibi herkesin görebileceği şekilde de yapılabilir. Kullanıcılar diğer kullanıcıların tvitlerine üye olabilir, üye oldukları kişi veya kurumların kimleri takip ettiğini görebilirler [10].

Twitter' da çeşitli kısaltmalar ve semboller kullanılmaktadır. Bu kısaltma ve semboller [3,5]:

- **Trending Topic (TT):** Sol taraftaki en çok konuşulan 10 konunun listelendiği alana Trending Topic denir. Her ülkenin kendi TT listesi vardır. Ayrıca dünya genelinde en çok konuşulan 10 konuyu gösteren bir liste mevcuttur.
- **Friday Follow (FF):** Bu kısaltma takip ettiğiniz kişilerin sizi de takip etmesi içindir. Sadece cuma günleri yapılır.
- **ReTweet (RT):** Başka bir Twitter kullanıcısının tvitini sizin takipçileriniz de görsün diye retvit tuşuna basıp sizin profilinizde görünmesidir.
- **Direct Message (DM):** Başka bir twitter kullanıcısının size atmış olduğu postayı işaret eder, bir nevi e-mail işlevi görür.
- **@kullanıcıAdı:** Bir kullanıcının başka bir kullanıcıyı referans göstermek veya başka bir kullanıcıya cevap vermek için cevap verilmek istenen kullanıcı adının önüne '@' sembolü eklemesidir. Ör:@zeynep
- **#etiket(tag, hashtag):** '#' sembolünden sonra gelen ifade etiketleri oluşturur. Belirli konu, kişi, olay vs. için atılan tvitler bir başlık altında toplanabilmektedir. Ör: #YildizTeknikUniversitesi

Twitter özellikle son yıllarda medya, ünlüler, siyasi partiler tarafından sıkça kullanılmaktadır. Ayrıca, oldukça büyük bir reklam kaynağı olan Twitter ünlülere para da kazandırmaktadır.

Son yıllarda kullanım ve erişim kolaylığı nedeniyle Twitter kullanıcılarının sayısı hızla artmaktadır. Twitter üzerinde yer alan verinin büyüklüğü, anlamlı bilgi çıkarımı yapmak isteyen araştırmacılara büyük bir kitlenin görüşlerine ulaşabilme imkânı sağlayarak çalışmalarını sosyal ağlar üzerine yönlendirmelerine neden olmuştur.

[11] deki çalışmada Go ve arkadaşları tvitlerden düşünce analizi yapmıştır. Bu çalışma tvitlerden düşünce analizine yönelik yapılan çalışmaların ilkidir. Tvitler olumlu veya olumsuz olarak etiketlenerek tüketicilerin almak istedikleri ürünler veya firmalar ile ilgili toplumun görüşleri hakkında fikir edinebilmelerini amaçlamışlardır. Başarılarını artırmak için gürültülü veri olan tvitlere bir önişlem uygulanmıştır. Makine öğrenmesi

yöntemlerinden Naive Bayes, Maksimum Entropi ve Destek Vektör Makineleri' nin etiketlemedeki başarılarını Uzak Denetim (Distant Supervision) yöntemini kullanarak ölçmüşlerdir. Twitter'dan düşünce analizine yönelik bir başka çalışmada [12] ise meta seviye özellikleri gibi farklı düşünce boyutları kullanarak Twitter'dan düşünce analizindeki başarının artırılması hedeflenmiştir. Görüş gücü, düşünce ve polarite göstergeleri gibi özellikler bir araya getirilerek kullanılmıştır. Çalışmanın sonucunda meta seviye özelliklerinin birleşimlerinin kullanılması düşünce analizindeki başarıyı artırdığı görülmüştür.

[13] deki çalışmada Twitter'dan tartışmalı olayların belirlenmesi amaçlanmıştır. Bir olay eğer halk tarafından çok konuşuluyorsa bu olayın tartışmalı bir olay olduğu varsayılarak; o olayın merkezinde yer alan kişiler, durumlar veya kurumlar hedef olarak belirlenmiş, belirli bir zaman diliminde ilgili hedefe yönelik tüm tvitler toplanarak analiz edilmiştir. Çıkarılan özelliklere göre hedefler için puanlar elde edilmiştir. Eğer hedefler belirli bir puan seviyesinde ise mevcut olay tartışmalı olay olarak etiketlenmiştir. [14] deki çalışmada tvitlerden acil durumların tespit edilmesi hedeflenmiştir. Tvitlerin gürültülü veri olduğu ve gerçek olmayan durumları da içerdiği göz önüne alınarak tvitler gerçek zamanlı olarak analiz edilmiştir. Doğal Dil İşleme yöntemleri ile gerçek olayların tespit edilmesi amaçlanmış ve kurtarma operasyonlarını desteklemek için olay hakkında önemli bilgiler elde edilebileceği gösterilmiştir.

[15]'de Michelson ve arkadaşları Twitter kullanıcılarının ilgi duydukları konuları tespit etmeye yönelik ilk çalışmalarını sunmuşlardır. Çalışmada tvitlerde bahsi geçen varlıklar incelenmiştir. Böylece kullanıcıların ilgi alanlarına yönelik arama yapabilmelerini ve konu profillerinin oluşturulması amaçlanmıştır. İlk olarak tvitlerde geçen varlık isimleri çıkarılmış ardından çıkarılan varlık isimlerinin dâhil olabileceği kategoriyi bulabilmek için Wikipedia'dan faydalanılmıştır. Böylece varlıkların dâhil olduğu kategoriler bir grafik yapısında temsil edilmiştir.

[16] da ise Twitter verisi kullanılarak popüler aktiviteler tahmin edilmektedir. Çalışma üç alt işlevden oluşmaktadır. Gelecekteki aktiviteleri tanımlayabilmek için aktivitelerin sınıflandırılması, tvitlerden aktivitelerin çıkarılması ve bir aktivitenin popülaritesinin belirlenmesi gerekmektedir. Sınıflandırma işlemi için Destek Vektör Makineleri ve

Naive Bayes sınıflandırıcıları kullanılmıştır. Aktiviteleri tanımlamak için LDA (Linear Discriminant Analysis) ve POS (Part of Speech) etiketi yaklaşımından faydalanılmıştır. Son olarak Adapte Edilmiş Frekans Sayıları (Adapted Frequenty Counts) kullanılarak popüler aktiviteler sıralanmıştır.

Bir sinema filmi ne kadar çok konuşuluyorsa o kadar çok izlenmiştir hipotezinden yola çıkılarak [17] de vizyona giren filmlerin gişe gelirleri Doğrusal Regresyon yöntemiyle tahmin edilmiştir. Twitter' da yapımcıların açmış olduğu film başlıklarına atılan tvitlerin oranları, filmlere ait url, video ve fotoğraf paylaşımlarından yararlanılmıştır. Gişe gelirlerini tahmin edebilmek için tvitlerden düşünce analizi yapılarak filme gidenlerin sayısı belirlenmiş ve çok konuşulan bir filmin izleyicisinin çok olacağına dair hipotez doğrulanmıştır.

[18] de Twitter tabanlı suç olayı tahminine yönelik bir çalışma sunulmuştur. Yazarlar otomatik duygu analizi ve tvitlerin anlaşılması tabanlı bir yaklaşım geliştirmiştir. Gizli Dirichlet Tahsisi (Latent Dirichlet Allocation) yoluyla boyut indirgeme, doğrusal model ile de tahmin yapmışlardır. Model gelecekteki vur-kaç kazalarını tahmin etme görevi ile test edilmiştir. Geliştirilen model gün boyunca meydana gelen vur-kaç kazalarını tahmin etmeye yönelik bir temel modeli oluşturmuştur.

Twitter' da kullanıcılar ilgi duydukları alanlara yönelik içerik paylaşan diğer kullanıcıları takip edebilmektedir. [10]'da ki çalışmada kategori bilgisi bilinen kullanıcıların ait oldukları kategoriyi ne ölçüde temsil ettiği gözlemlenmiştir. Dört farklı kategorideki haberler eğitim verisi olarak kullanılmış ve eğitim modeli Destek Vektör Makinesi ve çok değişkenli Naive Bayes ile oluşturulmuştur. Normal kullanıcı ve haber içeriği giren kullanıcıların içerikleri (tvitleri) ise test verisi olarak kullanılarak, haber içeriği giren kullanıcıların (botlar) içerikleri, normal kullanıcı içeriklerine oranla daha başarılı kategorize edildiği gözlemlenmiştir.

Güncel olaylar ve bunların Twitter ile ilişkisi üzerinde bir çalışma yapan Shamma [19] 2008 Amerika başkanlık seçimi sürecindeki haber trafiğini tvitler üzerinden incelemiş ve gündelik olayların tvitleri nasıl etkilediğine dair bir araştırma yapmıştır.

Ece çalışmasında [20] etiket, tvit ve kullanıcı arasındaki ilişkiyi belirlemeye çalışmıştır. Bu işlem için tvitlerdeki kelime ve etiketlerden kelime, etiket ve kullanıcı listesi

oluşturmuştur. Böylece ilgi alanları birbirine yakın olan kullanıcılar belirlenmeye çalışılmıştır. Her bir etiket için, o etikete ait tvitlerdeki kelimeler ve Google, Wikipedia gibi kaynaklar kullanılarak elde edilen kelimelerden bir özellik kümesi tanımlanmıştır.

[5]'de İngilizce tvitlerden otomatik soru çıkarmayı amaçlayan bir çalışma gerçeklenmiştir. Yazarlar soru algılama işlemini iki adımda gerçekleştirmiştir. İlk olarak soru içeren tvitler belirlenmiş ardından da bu tvitlerden gerçek bilgi veya yardım isteyenler çıkarılmıştır. Soru içeren tvitleri belirlemek için kural tabanlı ve öğrenme tabanlı metotlar kullanılırken, gerçek bilgi içeren tvitlerin çıkarılmasında tvitlerin içerik özelliklerinden yararlanılmıştır. İlgili çalışmadan esinlenilerek bu çalışma kapsamında Türkçe tvitler için bir soru algılama sistemi gerçeklenmiştir.

Doğal Dil İşleme'nin alt çalışma alanlarından biri olan soru algılama verilen bir doküman koleksiyonunda geçen soru cümlelerinin Doğal Dil İşleme yöntemleri ile otomatik olarak çıkarılmasıdır [21]. Soru algılama işlemi kurallı metinlerde gerçeklenebileceği gibi son yıllarda artan internet kullanımı ve forum siteleri, elektronik postalar, interaktif sohbet siteleri, sosyal medya gibi ortamlarda kullanıcıların öğrenmek istediği konularda çeşitli sorular sorup cevap beklemelerinden dolayı düzensiz verilere yönelmiştir.

Dent ve Paul [22] dilbilimsel bir ayrıştırıcı ile Twitter'dan otomatik soru çıkarımı yapan bir araç geliştirmeyi hedeflemişlerdir. Yazarlar tvitlerdeki 140 karakter sınırlamasından kaynaklanan kelimelerin kısaltılması veya eksik yazılması, kendine has bir dilinin olması, noktalama işareti kullanılmaması gibi problemler ile karşılaşmıştırlar. Bu nedenle tvitlere adapte olmuş bir simgeleştirici (tokenizer), sözlük ve ayrıştırıcı geliştirmişler ve İngilizce için yerel dilbilgisi özelliklerinden yararlanmışlardır.

Bir başka çalışmada [23] forum sitelerinden soru algılamak ve soruların cevaplarını bulmak üzerine bir sistem geliştirilmiştir. Soruları algılamak için ardışıl örüntü özellikleri, cevap adaylarını tanımlamak için grafik tabanlı bir metot geliştirilmiştir. Forum sitelerinden soru algılama yapan bir başka çalışmada [24]'de sunulmuştur. Wang et al. (2010) tarafından önerilen Prefixspan Algoritması ile yapılan ardışıl örüntü çıkarımının yanı sıra söz dizimsel örüntüler ile de sözlüksel özellikler elde edilerek özellik kümesi oluşturulmuştur. Eğitim modeli olarak da Destek Vektör Makinelerini

kullanmışlardır. Forum sitelerinden soru algılamada özellik olarak yaygın kullanılan soru işareti ve 5N1K kelimelerinin yanı sıra [25]'de sitede soru adayı cümleyi yazan yazarlara ait bir takım bilgileri ve n-gramları da özellik olarak tanımlayarak sınıflandırma tabanlı bir soru algılama ve cevaplama sistemi geliştirmişlerdir. Ding ve arkadaşları forum sitesinden soru ve cevap algılama içerikli çalışmalarında [26] Conditinal Random Fields (Şartlı Rastgele Alanlar) yöntemini kullanmışlardır.

Elektronik postaları özetlemek için elektronik postalardan soru ve cevap algılamaya yönelik bir çalışma gerçeklenmiştir [27]. Yazarlar özellik kümelerini pos etiketlerinden oluşturmuşlardır. Soru algılamaya yönelik kuralları öğrenmek için Ripper sınıflandırıcısını kullanmışlardır.

1.2 Tezin Amacı

Özellikle mobil cihazlarda da kullanılabilir olmasıyla birlikte Twitter insanların sadece boş zamanlarında oturum açtıkları bir sanal ortam olmaktan çıkmış, aynı zamanda bilgi paylaşılan ve soru sorulan bir ortam haline gelmiştir. [2]'de insanların Twitter'daki soru sorma alışkanlıkları incelenmiştir. [3]'deki çalışmada ise insanların Twitter'ı bilgi ihtiyaçlarını karşılamak için de kullandıkları ve sorulan soruların genellikle arama motorları ile bulunması zor olan öznel sorular oldukları sonucuna varılmıştır. Bu nedenle sosyal medya tabanlı soru algılama ve cevaplama sistemleri popüler olmaya başlamıştır [4]. Tez kapsamında Türkçe tvitlerden soru içerenlerin bulunması amaçlanmıştır.

Soru algılama için yapılan çalışmalar çoğunlukla forum siteleri, e-mail metinleri üzerinedir. Twitter üzerine pek fazla çalışma bulunmamaktadır. Var olanlar ise İngilizce tvitler için hazırlanmış çalışmalardır.

Bu çalışmada, konudan bağımsız Türkçe tvitler için, Conditinal Random Fields (CRF-Türkçe karşılığı Şartlı Rastgele Alanlar olup literatürdeki yaygın kullanımı nedeniyle tez kapsamında kısaltması olan CRF ifadesi kullanılarak bahsedilecektir) kullanılarak, soru algılama yapılmıştır. Çalışma dört adımdan oluşmaktadır. İlk olarak twitter4j ile konu ve kullanıcıdan bağımsız olacak şekilde tvit toplanmıştır. İkinci adımda toplanan tvitlere bir önişleme uygulanarak tvitlerdeki etiket, link gibi bilgi içermeyen veriler temizlenmiştir. Üçüncü adımda tanımlanan bir takım kurallarla aday sorular belirlenmiştir. Aday soru

veri setinde soru içeren tvitler olduğu gibi soru içermeyen tvitlerde bulunabilmektedir. Ve son olarak gerçekten soru içeren tvitleri tespit edebilmek için CRF kullanılarak tvitler etiketlenmiştir.

1.3 Hipotez

Literatürde Türkçe tvitlerden soru algılamaya yönelik bir çalışmaya rastlanmamıştır. Bu çalışmada ilk aşamada Türkçe tvtilerlerden soru içerenler, ardından ise tanımlanan soru türlerinden biri olan, soruyu ve cevabı içeren tivler otomatik olarak çıkarılmıştır.

Tezin ikinci bölümünde sistemin geliştirilmesi sırasında kullanılan yöntemlere değinilmiştir. Üçüncü bölümde veri setinin nasıl oluşturulduğu ve verileri sınıflandırmak için kullanılan etiketlerden bahsedilmiştir. Dördüncü bölümde uygulamanın nasıl geliştirildiğine dair bilgiler verilmiştir. Beşinci bölümde ise uygulamaya dair başarı ölçümleri verilmiştir.

BÖLÜM 2

KULLANILAN MEVCUT YÖNTEMLER

Tez kapsamında yararlanılan kaynaklar bu bölümde sunulmuştur. Eğitim kümesinin özelliklerini tanımlamak için Bölüm 2.1'de soru tanımada kullanılan Markov Model tabanlı istatistiksel sınıflandırma yöntemi olan Şartlı Rastgele Alanlar'dan (İngilizce karşılığı Conditional Random Fields - CRF olup literatürdeki yaygın kullanımı nedeniyle tez kapsamında CRF olarak geçecektir) bahsedilmektedir. Bölüm 2.2' de Java tabanlı istatistiksel bir Doğal Dil İşleme paketi olan Mallet'den bahsedilmiştir. Mallet paketi aracılığıyla kullanılan CRF yönteminde kuralları tanımlamak için, kural tabanlı yöntemlerde yaygın olarak kullanılan 'Düzenli İfadeler' den (Regular Expression) yararlanılmıştır. Bölüm 2.3' de ise Düzenli İfadeler' e yer verilmiştir. Son olarak sistemin başarısını test etmekte yararlanılan değerlendirme metrikleri Bölüm 2.4' de anlatılmıştır.

2.1 Şartlı Rastgele Alanlar

CRF makine öğrenmesi ve örüntü tanımada, yapılandırılmış veride kullanılan istatistiksel bir sınıflandırma yöntemidir. Doğrusal zincirli CRF Doğal Dil İşleme problemlerinde, giriş dizileri için etiket dizilerini tahmin etmede sıklıkla kullanılır [28]. Tez kapsamında, soru algılamada eğiticili bir sistem olan CRF yöntemi kullanılmıştır.

2.1.1 Sıralı Verinin Etiketlenmesi

Bir gözlem dizisi kümesine etiket dizisi atanması işlemi konuşma tanıma, varlık ismi tanıma gibi birçok alanda ortaya çıkmaktadır [29,30]. Doğal dil işlemede etiketleme işlemi her bir kelime ve sözcük türüne göre analiz edilerek etiketlenebilmektedir. Örneğin;

<İsim>Sibel</İsim><DiğerZarflar>ile</DiğerZarflar><İsim>arkadaş</İsim><TamlananEki>ları</TamlananEki>,<İsim>tatil</İsim><DolaylıTümleçEki>den</DolaylıTümleçEki><DiğerZarflar>dün</DiğerZarflar><Yüklem>dön</Yüklem><ZamanEki>dü</ZamanEki>

Cümlelerin söz dizimsel analizi yapılarak etiketlenmesi, sözcüklerin içerdiği bilgileri artırarak doğal dil işleme ile ilgili çalışmalarda kolaylık sağlamaktadır.

Olasılıksal sonlu durum otomatları veya Gizli Markov Modelleri (GMM) bir cümledeki sözcüklerin olasılığı en yüksek olan etiketi tanımlamak için kullanılan yaygın yöntemlerdendir [31, 32].

GMM, zamana bağlı değişkenlerin bulunduğu problemlerde sıklıkla kullanılan bağlantı tabanlı üretken bir model şeklidir. Temelde ardışık düğümlerin olasılığının bulunmasında kullanılan Markov Model teorisine dayanır. GMM'yi Markov Modelden ayıran durum ise Markov Modelde düğümler sadece gözlemlenen varlıkları ifade ederken diğerinde gözlemlenemeyen ancak, varsayılan bağlantılar için saklı düğümler oluşturulur. Markov model denklemin matematiksel modelini, GMM'nin çıkardığı model ise denkleme yakınsayan bir modeldir [33].

X ve Y gözlem dizileri üzerinde değişen rastgele değişkenler olmak üzere p(X,Y) ortak bir olasılık dağılımını tanımlar. Bu tür ortak bir olasılık dağılımını tanımlamak için üretimsel model tüm olası gözlem modellerini sıralamalıdır. Gözlem elemanları izole birimler olarak temsil edilmediği sürece bu zor bir işlemdir. Herhangi bir anda gözlem elemanları sadece etiket veya duruma doğrudan bağlı olabilir. Bu basit birkaç veri seti için uygun bir varsayımdır. Ancak, gerçek dünya uygulamalarında gözlem dizileri en iyi çoklu etkileşim özellikleri ve gözlem elemanları arasındaki uzun menzilli bağımlılıklar açısından temsil edilebilir [34].

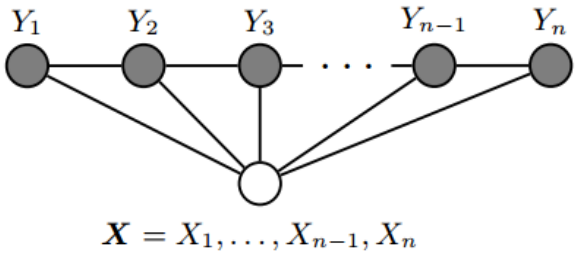
Bu temsil sorunu ardışıl verileri etiketlerken karşılaşılan en temel problemlerden biridir ve izlenebilir çıkarımı destekleyen ve nedensiz bağımsız varsayımlar yapmadan veriyi temsil edebilen bir modele ihtiyaç duyar. Bu iki şartı da sağlayan bir yol; gözlem dizisi ve etiketin her ikisi üzerinde de ortak bir dağılım yerine verilen belirli bir x gözlem dizisi de etiket dizilerinin şartlı bir olasılığını p(y/x) tanımlayabilen bir model kullanılmasıdır.

Şartlı modeller p(y*|x*) şartlı olasılığını en üst düzeye çıkaran y* etiket dizisini seçerek yeni bir x* gözlem dizisini etiketlemek için kullanır. Bu tür modeller koşullu doğaları gereği gözlemleri modellemede boşa çaba harcamazlar ve nedensiz bağımsız varsayımlar yapmak zorunda kalmazlar. Böyle bir modelle gözlem verisinin rastgele özellikleri elde edilebilir.

CRF [34] ardışıl veriyi etiketlemek ve bölmek için şartlı yaklaşıma dayalı olasılıksal bir yöntemdir. CRF verilen bir gözlem dizisindeki etiket dizileri üzerinde tek bir log-lineer dağılım tanımlayan bir yönsüz grafik modeli biçimidir. CRF'in GMM'den birincil üstünlüğü şartlı yapısıdır. Bu yapı sayesinde GMM'lerin içerdiği bağımsız varsayımlardan kurtulmuş olur. Ayrıca CRF ile Markov modellerinin bir zayıf noktası olan etiket önyargı problemi önlenir. Yani CRF Maksimum Entropi Markov Model (MEMM) [30,36] ve GMM'den çok sayıda dizi etiketlemeyi gerektiren görevlerde daha başarılıdır [30,35].

2.1.2 Yönsüz Grafik Modelleri

CRF; Markov Rastgele Alan [37] veya bir rastgele değişken olan ve gözlem dizilerini temsil eden X üzerinde global olarak devam eden yönsüz bir grafik modeli olarak da tanımlanabilir. Genelde yönsüz bir grafik G=(V,E) olarak gösterilir. Burada v ϵ V, Y nin bir Y_v elemanını her bir rastgele değişkene karşılık gelen bir düğümdür. Eğer her bir Y_v rastgele değişkeni G açısından Markov özelliğine uyuyorsa (Y,X) bir şartlı rastgele alandır. Teoride G grafiğinin yapısı rastgele olabilir. Ancak diziler modellenirken karşılaşılan en basit ve yaygın olan grafik yapısı Şekil 2.1 de görüldüğü üzere Y elemanına karşılık gelen düğümün basit birinci derece zincir oluşturduğu durumdur [34].



Şekil 2. 1 Diziler için zincir yapılı bir grafik modeli. Renklendirilmemiş düğümlere bağlı değişkenler model tarafından üretilmemiştir.

2.1.3 Potansiyel Fonksiyonlar

CRF'deki grafik modeli Y nin Y_v elemanları üzerindeki koşullu bağımsızlık kavramından elde edilen, gerçek değerli potansiyel işlevlerin normalize edilmiş ortak dağılımını çarpanlarına ayırmak için kullanılabilir. Her bir potansiyel fonksiyon G'nin köşelerinin temsil ettiği rastgele değişkenlerin bir alt kümesini kullanır. Yönsüz grafik modellerin şartlı bağımsızlık tanımına göre, G'deki iki köşe arasında bir kenar yoksa bu köşenin temsil ettiği rastgele değişken modeldeki tüm diğer rastgele değişkenlerden şartlı olarak bağımsızdır. Bu nedenle potansiyel fonksiyonlar şartlı bağımsız rastgele değişkenlerin aynı potansiyel fonksiyon içinde bulunmayacak şekilde ortak olasılığı çarpanlarına ayırmanın mümkün olduğunu sağlamalıdır. Her bir potansiyel fonksiyonun G içinde maksimal grubu oluşturan noktalara karşılık gelen rastgele değişkenler dizisinde işlev görmesini gerektirmek ilgili gerekliliği sağlamanın en kolay yoludur. Bu köşeleri doğrudan bağlı olmayan ve herhangi bir rastgele değişken çiftinin hiçbir potansiyel fonksiyona işaret etmemesi gerektiğini gösterir. Eğer iki köşe aynı grup içinde bir arada bulunuyorsa bu ilişkinin açık olmasını sağlar. Şekil 2.1 de de görüldüğü üzere her bir potansiyel fonksiyon Y_i ve Y_{i+1} bitişik etiket değişkeni çiftlerinde çalışacaktır.

İzole edilmiş potansiyel bir fonksiyonun doğrudan olasılıksal bir yorumunun bulunmaması dikkat çekicidir. Ancak fonksiyondaki rastgele değişkenler kısıtlamaları temsil eder. Bu, global konfigürasyonun olasılığını etkiler. Yüksek olasılıklı global bir konfigürasyon düşük olasılıklı global bir konfigürasyon ile karşılaştırıldığında yüksek olasılıklı olanın ilgili kısıtlamalardan birçoğunu yerine getirmesi daha mümkündür [34].

2.1.4 Şartlı Rastgele Alanlar

Lafferty ve arkadaşları [28] eşitlik 2.1 de x gözlem dizisindeki belirli bir y etiket dizisinin olasılığını potansiyel fonksiyonun normalleştirilmiş bileşkesi şeklinde tanımlamıştır.

$$\exp\left(\sum_j \lambda_j t_j\,(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i)\right) \qquad (2.1)$$

Burada $t_j(y_{i-1}, y_i, x, i)$ tüm gözlem dizisinin, i ve i-1 etiket dizisinde yer alan etiketlerin geçiş dizisidir. $s_k(y_i, x, i)$ ise i konumundaki etiketin ve gözlem dizisinin durum özellik fonksiyonudur. λ_i ve μ_k ise eğitim verisinden tahmin edilen parametrelerdir.

Özellik fonksiyonları tanımlanacağı zaman eğitim verisindeki deneysel dağılımın bir takım karakteristiklerini gösteren gözlemin gerçek değerli bir kümesi olan $b(x, i)$ üretilmelidir. Bu özellik şu şekilde ifade edilebilir.

$$b(\boldsymbol{x}, i) = \begin{cases} 1 & \text{İ durumundaki gözlem eğer 'Türkiye' ise} \\ 0 & \text{diğer durumlar} \end{cases}$$

Eğer mevcut durum (durum fonksiyonu durumunda) veya önceki ve mevcut durumlar (geçiş fonksiyonu durumunda) belirli değerleri aldıysa her özellik fonksiyonu gerçek değerli b(x,i) gözlem özelliklerinden birinin değerini alır. Tüm özellik fonksiyonları bu nedenle gerçek değerlidir. Örneğin;

$$t_j(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} b(\boldsymbol{x}, i) & Y_{i-1} \text{ sıfat} \rightarrow y_i \text{ isimdir.} \\ 0 & \text{Diğer durumlar} \end{cases}$$

Biraz daha basitleştirilirse

$$s(y_i, \boldsymbol{x}, i) = s(y_{i-1}, y_i, \boldsymbol{x}, i)$$

ve

$$F_j(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \boldsymbol{x}, i),$$

elde edilir.

Her bir $f_j(y_{i-1}, y_i, x, i)$ fonksiyonu ya bir $s(y_{i-1}, y_i, x, i)$ durum işlevi veya $t(y_{i-1}, y_i, x, i)$ geçiş işlevidir.

Bu durum bilinen bir x gözlem dizisindeki y etiket dizisi olasılığının Eşitlik 2.2 deki gibi formülize edilebilmesini sağlar.

$$p(y|x, \lambda) = \frac{1}{Z(x)} \exp(\sum_j \lambda_j F_j(\boldsymbol{y}, \boldsymbol{x})) \quad (2.2)$$

Burada $Z(x)$ normalizasyon faktörüdür.

2.1.5 Maksimum Entropi

CRF sıralı veriyi işaretleme ve bölümleme işlemlerinde MEMM ve GMM nin genel yapısını yansıtarak, örüntü tanımada sıklıkla kullanılan bir istatistiksel modelleme yöntemidir.

CRF, [38] de gösterildiği gibi maksimum entropi prensibine dayalı olarak bir eğitim verisi kümesinden olasılık dağılımlarını tahmin eden bir yöntemdir. Bir olasılık dağılımının entropisi [39] bir belirsizlik ölçüsüdür ve ancak dağılımın homojenliği mümkün olabilecek en yüksek noktaya çıktığında maksimize edilebilir. Maksimum entropi ilkesi [40] sonlu eğitim verisi gibi yetersiz bilgiden oluşturulan olasılık dağılımında dahi mevcut sınırlamaları temsil eden bir maksimum entropi olduğunu ve diğer tüm durumların yersiz varsayımlar içereceğini ileri sürmektedir.

Eğer bilgi bir özellik fonksiyonu kümesi kullanılarak temsil edilen eğitim verisinde özetlenirse, mümkün olduğunca homojen olan maksimum entropi modeli dağılımı her bir özellik fonksiyonunun beklentisinin eğitim verisinin deneysel dağılımının model dağılımına göre özellik fonksiyonunun beklenen değerine eşit olmasını sağlar.

2.2 Mallet

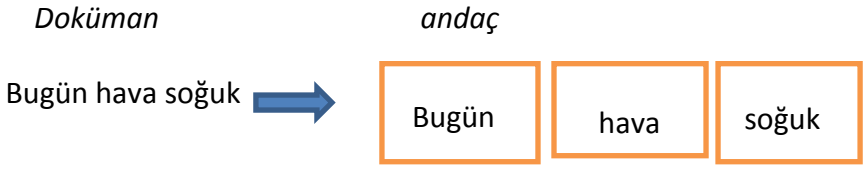
Soru algılama için geliştirilen sistemi CRF metodu ile eğitmek için ‘Mallet’ ara yüzü kullanılmıştır.

Mallet; doküman sınıflandırma, bilgi çıkarımı ve metinlere uygulanan diğer makine öğrenmesi teknikleri için kullanılan Java tabanlı istatistiksel bir Doğal Dil İşleme paketidir. Mallet:

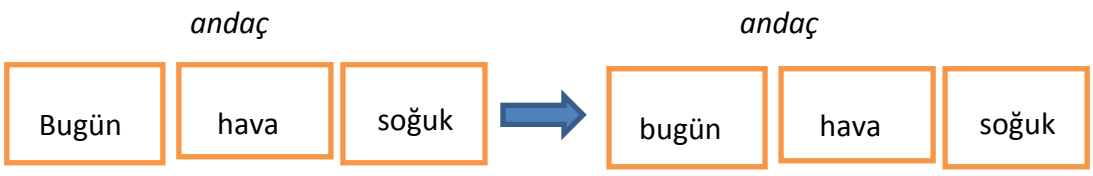
- Doküman sınıflandırma için metinlerden özellikler çıkaran yordamlar, Naive Bayes, Karar Ağaçları gibi algoritmaları ve performans ölçüm metriklerini,
- Dizi etiketleme işlemi için ise Saklı Markov Model, Maksimum Entropi Markov Model ve Şartlı Rastgele Alanlar algoritmalarını,
- Konu modelleme için Gizli Dirichlet Tahsisi, Pachinko Tahsisi ve Hiyerarşik Gizli Dirichlet Tahsisinin örnekleme tabanlı uygulamaları ve
- Metin belgelerini sayısal olarak ifade edebilmek için dizgeleri alt dizgelere ayırma (tokenizing), engellenecek kelimeleri (stop words) kaldırma ve dizilerin sayı vektörlerine dönüştürülmesi yordamlarını içerir [41].

Metin dokümanları, Mallet ile işlenmeden önce sayısal vektörlere dönüştürülür. Özellik vektörlerinin elemanları 'özellik değeri' (feature value) olarak adlandırılır. Özellik vektöründe satırlardaki tüm elemanlar kelimelerin dokümanda geçme sayısını temsil eder. Özellik vektörü çıkarma işlem adımları:

1- Dokümanlar andaçlara(token) dönüştürülür



2- Andaçlar büyük harflerin küçük harfe çevrilmesi, gereksiz kelimelerin atılması gibi ön işlemlerden geçirilerek standart bir hale getirilir.



3- Andaçlar özellik dizilerine dönüştürülür. Andaç dizisindeki indise karşılık gelen özellik dizisindeki sayısal değerler, ilgili andacın tüm dokümanda kaç kere geçtiğinin bilgisidir.

4- Özellik dizileri özellik torbalarına (bag) dönüştürülür. Burada '1. 0' ifadesi ilgili andacın etiketli olduğunu göstermektedir.

Soru algılama için geliştirilen sistemde Mallet'in kendi bünyesindeki simgeleştirici (tokenizer) kullanılmamıştır. Çünkü soru algılama için tvitler Mallet'in yaptığı gibi kelime olarak değil cümle olarak bir örneği temsil etmektedir. Bu nedenle dokümanları cümle cümle ayıran bir yordam sisteme entegre edilmiştir. Böylece özellik vektöründeki her bir özellik kelimeyi değil, cümleyi temsil etmiştir.

Mallet' de örneklerin 4 parametresi vardır.

1- **İsim:** Örneği temsil edecek bir dizgi

2- **Etiket:** Sonlu bir etiket kümesindeki bir değerdir. Mallet' de 'LabelAlphabet' dizisinin içinde tutulmaktadır.

3- **Data:** Genellikle bir özellik vektörü veya bir özellik dizisidir.

4- **Source:** Örneğin simgeleştirilmemiş bir temsilidir.

Örnekler 'InstanceList' olarak adlandırılan listelerde tutulur. Veri yükleneceği zaman bir takım işleme adımları gerçekleşir. Mallet' de bu adımlar 'Pipe' ara yüzüyle temsil edilir. Tipik bir pipeline yineleyiciler (Iterator) ile başlar. Örneğin dosya yineleyici (FileIterator)

her örneğin bir dosyada olduğu durumlarda, CSVIterator ise tüm örneklerin tek bir dosyada olduğu durumlarda kullanılır. Geliştirilen sistemde tüm tvitler tek bir dosyada olduğundan CSVIterator kullanılmıştır. CSVIterator'da düzenli ifadeler kullanılarak kurallar tanımlanabileceği gibi kullanıcı kendi pipe'ını yazabilir. Geliştirilen sistemde ise düzenli ifadelerden faydalanılmıştır.

Pipe işlemleri sonucunda eğitim setindeki her bir veri ağırlıklandırılır. Eğer sistem doğru etiketi tahmin ettiyse bu ağırlık 1.0, aksi halde 0.0'dır.

Bir InstanceList nesnesi Pipe, dataAlphabet(veriler) ve LabelAlphabet(etiketler) sınıflarından oluşur.

2.2.1 Mallet ile Sıralı Verinin Etiketlenmesi

Veri dizilerden meydana gelir ve her durum için birbiri ile ilişkili etiketler vardır. Bölüm 2.1' de bahsedildiği gibi CRF Saklı Markov Model'den türetilmiş istatistiksel bir yöntemdir. Bir verinin her etikete ait olma olasılığını bulur. Gizli Markov Modelden farklı olarak özellik dizileri ile değil, özellik vektör dizileri ile çalışır. Veri yüklenirken her bir satırda bir kelime olacak şekilde yüklenir ve örnekler birbirinden boş satır ile ayrılır. Soru algılama işleminin cümle bazında çalışması gerektiği için geliştirilen sistemde veri her bir cümle yani her tvit bir satırda olacak şekilde yüklenmiştir. Veri yüklendikten sonra veriler Mallet'in yordamı (SimpleTaggerSentence2TokenSequence()) ile token dizilerine, token dizileri başka bir yordam (TokenSequence2FeatureVectorSesuence()) ile özellik vektör dizilerine dönüştürülür. 'Pipe' altında tanımlanan kurallar örneklere uygulanarak kuralların eğitim verisindeki örneklerde görülme istatistiklerine ve bu örneklerin etiket bilgilerine göre bir model oluşturur. Oluşturulan modele göre de yeni gelen etiketsiz veri sınıflandırılır [42].

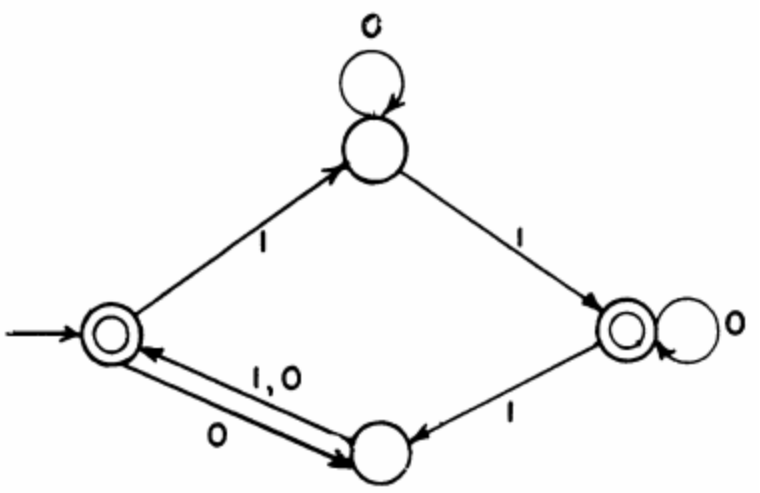
2.3 Örüntü Eşleştirme

Bilgisayar bilimlerinde örüntü en geniş anlamıyla verileri dönüştürmek için kullanılan kurallardır. Algılanan simgeler dizisini, bir takım örüntülerin var olup olmaması açısından kontrol etmenin bir yoludur. Kullanıcı tanımlı örüntülerin metin belgesi içerisinde aranması metin düzenleme ve bilgi çıkarımı uygulamalarında ortak bir

gereksinimdir [43]. Örüntüler genellikle dizi veya ağaç yapısındadır. Dizi yapısındaki örüntüler genellikle düzenli ifadeler ile açıklanmaktadır ve geriye doğru izleme (backtracking) gibi yöntemlerle eşleştirme yapılır.

2.3.1 Düzenli İfadeler

Düzenli İfadeler Algoritması temelde bir otomat mantığıyla çalışır, sıfır ve bir dizilerinin sınıflarını belirtir. Sıfır ve bir dizileri bir otomatın giriş dizisi olarak da düşünülebilir. Düzenli İfade Algoritmasının işleyiş biçimi Şekil 2.2' deki gibi bir otomat teorisi üzerine kurulmuştur [44].



Şekil 2. 2 Sonlu Otomat

Düzenli ifadeler bir karakter dizgesinden istenilen özel bir dizgenin seçilmesini sağlayan, Doğal Dil İşleme uygulamalarında sıklıkla kullanılan bir tekniktir. Metin içinde aranacak her bir karakter, sıradaki tüm olası karakterlerin bir listesine karşı bir dizi içinde incelenir. İnceleme süresince gelebilecek tüm sonraki karakterlerin yeni bir listesi oluşturulur. Şimdiki listenin sonuna gelindiğinde, oluşturulan yeni liste şimdiki liste olur, sonraki karakter elde edilir ve işlem böylece devam eder [45]. Zamanla geliştirilen algoritma belirli bir ifadeyi belirli bir kalıba göre bulmayı, değiştirmeyi veya parçalamayı sağlayan küçük bir yazılım dili haline gelmiştir. Birden fazla kalıbı tek bir kuralda toparlayarak daha esnek kurallar tanımlayabilmek için çalışma kapsamında CRF'in özelliklerini belirlemede Düzenli İfadeler' den yararlanılmıştır.

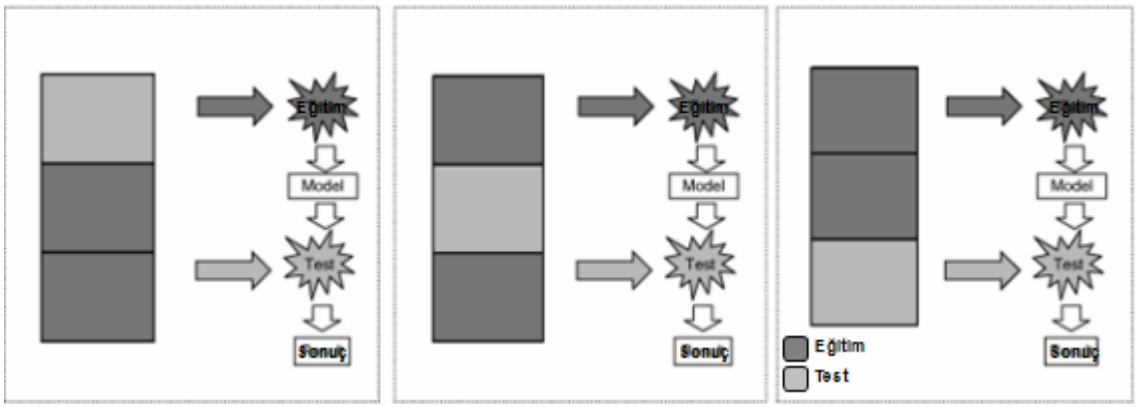
2.4 Değerlendirme Metrikleri

Geliştirilen sistem 5 kat çapraz geçerleme yöntemi ile test edilmiştir. Sistemin başarısını ölçmek için sınıf hata matrisi elde edilerek tutturma(precision), bulma (recall) ve F1-Puanı değerleri hesaplanmıştır.

2.4.1 K-Kat Çapraz Geçerleme

Çapraz geçerleme biri eğitimde diğerleri testte kullanılmak üzere veriyi k parçaya bölerek öğrenme algoritmalarının değerlendirilmesi ve karşılaştırılması veya aynı öğrenme algoritmasının farklı verilerle test edilebilmesi için geliştirilen istatistiksel bir yöntemdir. Çapraz geçerlemenin en temel şekli k-kat çapraz geçerlemedir.

K-kat çapraz geçerlemede tüm veri k eşit parçaya bölünür ve k adet iterasyon yapılır. Her bir iterasyonda farklı bir bölüm test için ayrılarak kalan k-1 adet bölüm eğitim için kullanılır ve başarı hesaplanır. İterasyonlar sonlandığında k adet başarı sonucunun ortalaması alınarak sistemin gerçek başarısı elde edilmiş olur. Şekil 2.3' de 3-kat Çapraz Geçerleme modellenmiştir [46].



Şekil 2. 3 3-kat çapraz geçerleme

2.4.2 Hata Matrisi

Makine öğrenmesinde hata matrisi, algoritmanın performansını gözlemlemek için kullanılan özel bir tablo biçimidir. Genellikle eğiticili öğrenme metotlarında kullanımı yaygındır, eğiticisiz metotlardaki karşılığı 'eşleştirme matrisi' dir. Hata matrisinde sütunlar algoritmanın tahmin ettiği sınıf bilgisini, satırlar ise gerçek sınıf bilgisini temsil

eder. Çizelge 2.1’ de 2 sınıflı bir veri seti için oluşturulmuş sınıf karışıklık matrisi verilmiştir.

Çizelge 2. 1 Sınıf karışıklık matrisi

		Gerçek	
		Soru	$\overline{\text{Soru}}$
Tahmin	Soru	TP	FP
	$\overline{\text{Soru}}$	FN	TN

Çizelge 2.1’ de;

- TP (True Positive): Sistemin ‘soru’ olarak etiketleyip gerçekte de soru olan örnek sayısı,
- FP (False Positive): Gerçekte soru olmayıp sistemin ‘soru’ olarak tahmin ettiği örnek sayısı,
- FN (False Negative): Gerçek sınıfı ‘soru’ olup sistemin sınıfını ‘soru değil’ olarak tahmin ettiği örnek sayısı
- TN (True Negative): Gerçek sınıfı ‘soru değil’ olup sisteminde sınıfını ‘soru değil’ olarak tahmin ettiği örnek sayısıdır [47].

Doğruluk (Accuracy): Geliştirilen sistemin doğru tahmin ettiği örnek sayısının tüm örnek sayısına oranıdır. Çizelge 2.1’ deki değişkenler için doğruluk değeri Eşitlik 2.3’ de verilmiştir.

$$Doğruluk = (TP + TN)/(TP + FP + FN + TN) \quad (2.3)$$

Tutturma (Precision): Doğru tahmin edilen pozitif sınıfındaki örneklerin tüm örneklere oranıdır. Çizelge 2.1’ deki değerler için hassasiyet Eşitlik 2.4 deki gibidir.

$$P = TP/(TP + FP) \quad (2.4)$$

Bulma (Recall): Doğru tahmin edilen ‘pozitif’ sınıfındaki örneklerin ‘pozitif’ sınıfındaki tüm örneklere oranıdır. Çizelge 2.1’deki değerleri için Eşitlik 2.5’de gösterildiği gibi hesaplanır.

$$R = TP/(TP + FN) \tag{2.5}$$

F1-Puanı (F1-Score): Doğruluk değeri, veri setinde etiketlerin dağılımı arasındaki fark fazla ise sistemin gerçek başarısını yansıtmaz. Bu nedenle sistemin gerçek başarısını bulmak için hassasiyet ve duyarlılığın harmonik ortalamasını kullanan F1-Puanı değeri alınır. F1-Puanı değerinin formülü Eşitlik 2.6 da verilmiştir.

$$\mathrm{F1-Puanı} = \frac{2*TP}{(2*TP)+FP+FN} \tag{2.6}$$

BÖLÜM 3

VERİ SETİ

Bu bölümde geliştirilen soru algılama sisteminin eğitimi için kullanılan veri setine dair bilgiler sunulmuştur. Bölüm 3.1'de eğitim setinin nasıl oluşturulduğu ve bu safhada yararlanılan twitter4j kütüphanesinden ve tvitlerin etiketlenme formatından bahsedilmiştir. Bölüm 3.2'de tvitlerin nasıl etiketlendiğinden ve etiket türlerinden bahsedilmiştir.

3.1 Eğitim Setinin Oluşturulması

Geliştirilen sistemin eğitilmesi için, bir Java kütüphanesi olan twitter4j ile 22.05.2013 ile 20.06.2013 tarihleri arasında 1.000.000 tvit toplanmıştır. Tvitlerde herhangi bir kullanıcı adı, konu, yer, vs. kısıtlaması yoktur.

Twitter4j; Twitter API'si için gayri resmi bir Java kütüphanesidir. Tvitleri çekebilmek için önce Twitter'ın [48] deki sitesinden hesap oluşturmak gerekmektedir. Kimlik doğrulama için 'OAuth' protokolünü kullanmaktadır. OAuth (Open Authorization) yetkilendirme için kullanılan açık bir standarttır. Kullanıcıların video, fotoğraf gibi özel kaynaklarını paylaşabilmelerine imkân verir. Hesap oluşturduktan sonra tvitlere erişim yapılabilmektedir. Twitter4j ile;

- Bir tvitin tvit metni, alıcı ve göndericisi, gönderildiği tarih ve saat, konu başlığı, konum, dil bilgileri ayrı ayrı elde edilebilmektedir.
- Toplanacak tvitler belirli bir kullanıcı, konu, zaman, konum bilgilerine göre filtrelenebilmektedir.

- Belirli bir kullanıcıya için ilgili kullanıcının takipçileri, takip ettikleri bilgileri de elde edilebilmektedir.

Tvitlere sadece Türkçe dilinde yazılmış olma şartı konulmuş olup bunun dışında her hangi bir kısıtlama konulmamıştır [49].

Eğitim setinin etiketlenmesinde MUC-6 [50] konferansında belirlenen, MUC-7 [51] konferansında ise kapsamı genişletilen 'ENAMEX' formatı kullanılmıştır. ENAMEX; varlık adı ifadesi (entity name expression) kavramından oluşturulmuştur. Doğal Dil İşleme' nin bir çalışma alanı olan Varlık İsmi Tanıma çalışmalarında kullanımı yaygındır.

Geliştirilen sistemin ilk aşamasında yapılan soru algılama işleminde tvit soru içeriyorsa 'ENAMEX' etiketleri arasına alınmıştır.

Örneğin; <ENAMEX_TYPE="soru"> nerde açıklanmış notlar? </ENAMEX> →Soru

<ENAMEX_TYPE="\0"> sonra komaya girdi ?:/</ENAMEX> → Soru Değil

Şekil 3.1' de soru algılama için oluşturulan eğitim setinin bir kesiti bulunmaktadır. Yaklaşık 13 KB büyüklüğünde bir eğitim seti oluşmuştur. Gerçekte soru olan tvitler 'Enamex' etiketlerinin içerisine alınmış olup; tipi 'soru' olarak belirlenmiştir. Soru olmayanlar ise olduğu gibi bırakılmıştır ve programın çalıştırılması sırasında yukarıdaki örekteki gibi tipi '\0' olarak atanmaktadır.

```
<ENAMEX_TYPE="soru"> ve sonra döndüm dedimki LAN BEN NİYE ESRA EROL İZLİYORUM? </ENAMEX>
<ENAMEX_TYPE="soru"> Ily suan tirnak arasinda kalan cig kofteleri yiyor sizce kac unf ? </ENAMEX>
<ENAMEX_TYPE="soru"> Ne yapıyosunuz lo? </ENAMEX>
Senin adam dediklerini çağır bakıyım?
<ENAMEX_TYPE="soru"> nerden biliyosun? :( </ENAMEX>
<ENAMEX_TYPE="soru"> Bugün kazanacağız değil mi abi? ! :) </ENAMEX>
```

Şekil 3. 1 Soru algılama için eğitim seti

Çalışmanın ikinci aşamasında ise, soru olan tvitlerin içerisinden QA (Question With Answer) türündeki tvitler tespit edilmeye çalışılmıştır. Eğitim seti tüm veri setindeki gerçekten soru olan tvitlerden oluşmaktadır. Benzer şekilde bulunmak istenen tür olan QA etiketli tvitler 'enamex' etiketleri arasına alınmış olup diğerleri olduğu gibi bırakılmıştır. Şekil 3.2' de QA algılama için kullanılan eğitim setinin bir kesiti bulunmaktadır.

<ENAMEX_TYPE="soru"> Mutluluk ? Zeki insan işi! </ENAMEX>
ohaaa ıy pislik ama simdi haklı kadın nereye sıcacak bu millet !?
<ENAMEX_TYPE="soru"> Saglikli olmak ne demektir? Enis:Hasta olmamaktir addfghhhhg </ENAMEX>
KONUŞMUYORUZ Kİ HOCAM DERS ÇALIŞIYORUZ. (Ee kanka sonra ne olduuu?)

Şekil 3. 2 QA algılama için eğitim seti

3.2 Eğitim Veri Seti Sınıfları

Twitter'da soru algılama veya ne tür soruların sorulduğunu tespit etmeye yönelik çalışmalarda soru türlerini gruplandırmak için çeşitli etiket türleri kullanılmıştır.

[5] deki İngilizce tvitlerden soru algılamaya yönelik yapılan çalışmada altı soru türü belirlenmiştir.

- Reklam; Okuyucuya soru sormak ve reklam sunmak için yazılan tvitler,
- Web üzerinde, içerisinde soru barındıran, makale veya haber başlığı içeren tvitler,
- Yanıtı ile soru,
- Tırnak içinde soru,
- Retorik soru; bir cevap beklemeyen sorular içeren tvitler,
- Qweet; gerçek bilgi veya yardım isteyen türde soru içeren tvitler.

[3] deki çalışmada ise yazarlar insanların Twitter'da soru sorma alışkanlıklarına dair bir anket çalışması yapmıştırlar. Kullanıcıları hangi türde daha çok sorular sorduklarını belirlemek için 8 soru türü belirlemişlerdir.

- Öneri (Recommendation),
- Görüş (Opinion),
- Gerçek Bilgi (Factual Knowledge),
- Retoric (Rhetorical),
- Davet (Invitation),
- Yardım (Favor),

- Sosyal Bağlantı (Social Connection),
- Teklif (Offer)

[1]de insanların micro-bloglarda sorduğu soruların karakteristiği, Twitterda soruların oynadığı rol ve ne tür sorular sorulduğuna dair bir çalışma yapılmıştır. Yazarlar soru türlerini:

- Bir soru-cevap çifti önermek, Reklam; Kullanıcıyı bir soru üzerinde düşünmeye teşvik eder ve cevap vermesi ya da soruyu anlaması için bir yol sunar bu da genellikle başka bir web sitesine yönlendirmekle olur. Kullanıcının arkadaşlarının birinden gönderilmez. Bu tvitlerin genellikle spam veya başka bir istem olması muhtemeldir,
- Bir soru-cevap çifti önermek – Sosyal; Bir önceki soru türü gibidir farklı olarak kullanıcının arkadaşlarının birinden gönderilir,
- Bilgi Talep Eden Sorular: Açıklama; Bir cevap veya somut kaynak bekleyen bilgi arama amaçlı sorulan sorulardır,
- Bilgi Talep Eden Sorular: Bireysel veya grup görüşü; Topluluğun görüşlerini içeren bilgi için sorulur. Bu görüşlerin bireyin veya bir başlıkla ilgili bir grubun görüşü olabilir,
- Bir aktivite daveti,
- Bir görüş ya da mevcut durumun ifade edilmesi; Retorik sorulardır. Bildirim ifade etmek için soru içerir. Doğrudan bir bilgi arayışı yoktur,
- Bir faaliyetin koordinasyonu,
- Açık değil; Soru içeren fakat içeriği itibariyle ayırt edilemeyen veya mevcut kategorilerden birine dâhil edilemeyen tvistleri içerir,
- Soru değil

olarak belirlemişlerdir.

Twitter'da soru ve cevap alışkanlıklarını inceleyen bir başka çalışmada [2], yazarlar soru türü olarak Morris'in [3] deki çalışmasında belirlediği etiket türlerini kullanmışlardır.

Geliştirilen çalışmanın ilk aşamasında soru içeren tvitlerin tespit edilmesi amaçlandığından tvitler 2 tür etikete sahiptir: Soru veya soru değil.

Çalışmanın ikinci aşamasında ise bir soru türünün belirlenmesi amaçlanmıştır. Ayrıca Türkçe tvitlerdeki soru cevap alışkanlıkları da gözlemlenmiştir. Bu amaçla [1, 3, 5] deki çalışmalardan esinlenilerek sekiz soru türü belirlenmiştir.

1. RQ (Rhetorical Question): Retorik sorular. Tvit soru ifadesi içeriyor ancak bir cevap almak için sorulmamış sorulardır. Daha çok felsefi cümleler, şarkı sözleri içermektedir. Bir başkasından nakledilen ve sorulan sorunun cevabını içerisinde barındırmayan sorularda bu gruba dâhil edilmiştir.

   Ör: “Hayatın önemli kurallarından birisi "beklentiye girmemek" iken, diğer bir önemli kuralı "umudunu yitirme" nasıl olabiliyor?”

2. FK (Factual Knowledge) : Gerçek sorular. Kullanıcının bir cevap almak amacıyla sorduğu sorulardır.

   Ör: “nerde açıklanmış notlar?”

3. IA (Invitation): Davet ve organizasyon içeren sorulardır.

   Ör: “hafta sonu yaza kaçalım mi? geliyo musun???”

4. RN (Recommendation): Öneri almak için sorulan sorulardır.

   Ör: “Oje sürcem ne renk süreyiim??”

5. QA (Question With Answer): Tvit hem soruyu hem cevabı içerir. Genellikle kullanıcının kendisi sorup kendisi cevapladığı veya bir başkasıyla arasında geçen bir soru cevap diyaloğunu naklettiği sorulardır.

   Ör: “+KANADI VAR UÇAMAZ PETEĞİ VAR BAL YAPAMAZ NEDİR BU? - ÖZÜRLÜ ARI :Dd:D:dD:Dd”

6. NQ (Not a Question): Bir soru içermeyen ancak yapısı veya içerdiği kelimeler itibariyle soru gibi görünen tvitlerdir.

   Ör: “ben de merak içindeyim ????”

7. RT (Request): İzin, rica anlamı taşıyan sorulardır.

Ör: “Mertcan, bana adresini mesaj ile atar mısın? sana son kitabımdan göndereyim..”

8. ON (Opinion): Herhangi bir konuda fikir almak için sorulan sorulardır.

   Ör: “Sence akşam ki maç ne olur”.

Twitter4j kütüphanesi kullanılarak çekilen tvitler Bölüm 4’de ayrıntılı olarak anlatılacağı şekilde bir ön temizleme işleminden geçirilmiştir. Ardından tanımlanan kurallar ile ‘Aday Soru Havuzu’ oluşturulmuştur. Çekilen 1.000.000 tvitten Aday Soru Havuzuna giren 136.449 tanesi için etiketleme işlemi el ile yapılmıştır. Veri setinin büyüklüğünden dolayı etiketleme işlemi beş kişi tarafından yapılmış olup, her bir tvit bir kişi tarafından etiketlenmiştir.

Çizelge 3.1’de tvitlerin soru türlerine göre dağılımları gösterilmiştir.

Çizelge 3. 1 Soru türü dağılımı

NQ	RQ	FK	QA	RT	ON	IA	RN
63.389	47.382	17.567	4511	1831	1244	315	210

Çizelge 3.1'de de görüldüğü üzere en çok tvit NQ sınıfındadır. Aday havuzunda NQ sınıfındaki verinin çok olmasının nedeni; aday sorular çıkarılırken gerçekten soru olan tvitleri de kaçırmamak için kuralların esnek tutulması ve kullanıcıların tvit yazarken yazım kurallarına uymamaları, kelimeleri kısaltmaları veya sesli harfleri art arda kullanmaları gibi sebeplerden oluşan hatalardan kaynaklanmaktadır. RQ sınıfındaki tvitlerin sayısının çok olma nedeni ise felsefik cümleleri, alay içeren ifadeleri, şarkı-şiir sözlerini sıklıkla içermesinden kaynaklanmaktadır. Soru türü dağılımına dair kısıtlar Bölüm 4'de ayrıntılı olarak ele alınmıştır.

BÖLÜM 4

SİSTEM TASARIMI

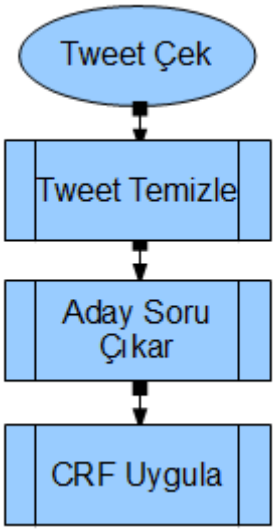
Bu bölümde geliştirilen soru algılama sisteminin nasıl gerçeklendiğinden bahsedilmiştir. Bölüm 4.1'de uygulamanın genel işleyişi, tvitlerin çekilmesi, temizlenmesi işlemleri anlatılmıştır. Bölüm 4.2'de soru algılama için geliştirilen uygulama, tanımlanan özellikler ve kısıtlardan bahsedilmiştir. Bölüm 4.3'de ise QA algılama için geliştirilen uygulama anlatılmıştır.

4.1 Sistemin İşleyişi

Geliştirilen uygulama temelde 4 adımdan oluşmaktadır.

1. Twitter4j kütüphanesi ile tvit veri tabanı oluşturulur.
2. Tvitlerdeki gereksiz veriler bir ön işlem fonksiyonu ile temizlenir.
3. Tanımlanan kurallar ile 'aday soru havuzu' oluşturulur.
4. Aday Soru Havuzundaki tvitlere 'Mallet' kütüphanesi [52] ile CRF uygulanarak soru algılama işlemi gerçekleştirilir.

Şekil 4.1’ de sistemin genel işleyişi özetlenmiştir.



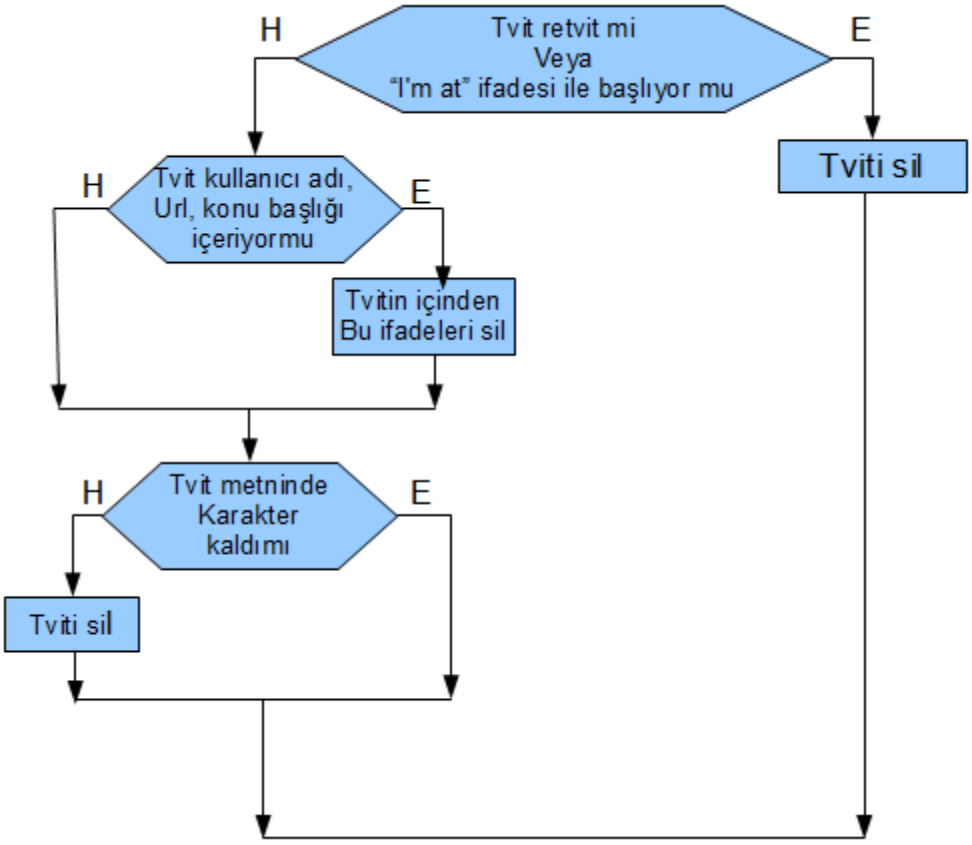
Şekil 4. 1 Sistemin genel işleyişi

4.1.1 Tvitlerin Temizlenmesi

Tvitlere, sistemin kullanmayacağı bilgilerden arındırılmak için bir ön işleme fonksiyonu uygulanmıştır. Önişleme adımları aşağıdaki şekilde sıralanmaktadır.

1. Twitter’da ‘retvit’ olarak adlandırılan bir başka kullanıcının tvitinin gönderilmesi anlamına gelen bir özellik mevcuttur. Veri setinin eğitiminde aynı tvitin birden fazla bulunması istenmediğinden veri seti retvitlerden arındırılmıştır.
2. Yer bildirimi içeren tvitlerde soru içermeyeceği için, “I’m at” ile başlayan, yer bildiriminde bulunan tvitler elenmiştir.
3. Tviterin içerisinde geçebilen url, kullanıcı adı, konu başlığı bilgileri de sistem için anlamlı olmadığından, bu ifadeler tvitlerin içerisinden çıkarılmıştır.
4. Bazı tvitler sadece url, kullanıcı adı veya konu başlığından ibaret olduğu için temizleme sırasında bu ifadeler silindiğinde, tvit, içeriği boş bir kayıt haline gelebilmektedir. Bu nedenle temizleme işleminin en sonunda içeriği boş olan kayıtlar veri setinden silinmiştir.

Şekil 4.2’de tvit temizleme işlemi modellenmiştir.



Şekil 4. 2 Tvit temizleme

4.1.2 Aday Soru Havuzunun Oluşturulması

Uygulamada kullanılmak üzere, twitter4j kütüphanesi ile bir milyon tvit çekilmiştir. Sistemin doğruluğunun test edilebilmesi için tvitlerin etiketlenmesi gerekmektedir. Etiketleme işlemi Bölüm 3'de anlatıldığı üzere el ile yapılmıştır. 1 milyon tvitin hepsinin okunup el ile etiketlenmesi maliyetli olacağından, tanımlanan kurallar ile bir aday soru havuzu oluşturulup, soru olma ihtimali yüksek olan tvitler bu havuzda toplanmıştır. Gerçekte soru olan tvitlerin aday soru havuzunun dışında kalmasını önlemek için tanımlanan kurallar esnek tutulmuştur.

Türkçe' de soru cümleleri; soru ekleri, soru zarfları, soru zamirleri veya soru sıfatları ile oluşturulurlar ve soru cümlelerinin sonuna soru işareti (?) konulur [53]. Türkçe' deki bu soru cümlesi için geçerli ifadeler ile kurallar tanımlanmış ve aday soru havuzu oluşturulmuştur.

Kurallar aşağıda sıralanan 4 gruptan oluşmaktadır. Bir tvit bu kurallardaki ifadelerin her hangi birini içeriyor ise aday soru havuzuna dâhil edilmekte ve diğer kurallar aranmamaktadır. Kurallar düzenli ifadeler kullanılarak tanımlanmıştır.

1. Soru İşareti
2. Soru Eki; mı, mısın, mısınız, mıyım, mıyız, mıydın.
3. Soru Kelimesi; ne, neden, niçin, niye, nasıl, kim, kaç, hangi.
4. Özel Kelime; acaba, naber, napica(n/z/m/k), noluyo, napabil.., demi, dimi, napıyo, noldu

Özel kelimeler grubu tvitlerde sıklıkla geçebilen, bulunduğu tvite soru anlamı katabilen kelimelerden oluşmaktadır. 'napabil..' ifadesi napabilir, napabilecek, napabiliyor gibi bu kelimeden türetilebilecek diğer kelimeleri de içersin diye bu şekilde tanımlanmıştır.

4.1.2.1 Aday Soru Havuzu İçin Kısıtlar

Aday soru havuzunu oluşturma işleminde, tvitlerin düzensiz veri formatında olmasından kaynaklanan problemler ile karşılaşılmıştır.

1. Tvitler düzensiz veri olduğu için imla kurallarına uyulmamaktadır. Türkçe yazım kurallarına göre kelimeden ayrı yazılması gereken soru ekleri bitişik olarak da yazıldığı için, soru eki için tanımlanan düzenli ifadelerde kelimenin yukarıdaki soru eklerinden biriyle bitip bitmediği kontrol edilmiş olup, kelimeden ayrı yazılma şartı aranmamıştır. Bu durum özellikle 'mi' soru eki için yanlış etiketlemelere yol açmıştır. Örneğin: 'özlemi' şeklinde bir ifade 'mi' ile bittiği için soru olmamasına rağmen sistem tarafından aday soru havuzuna dâhil edilmiştir. Bu tarz durumların çokluğu aday soru havuzundaki NQ sınıfına dâhil tvitlerin sayısını, dolayısıyla da hataları yüksek oranda artırmıştır. Ancak eklere kelimeden ayrı yazılma kuralı eklendiğinde birçok soru içeren tvitin havuza dâhil edilmediği tespit edildiğinden, kural uygulanmıştır.
2. Soru kelimelerinden birini içeren tvitler aday soru havuzuna dâhil edilmiş olup, kelimenin sonrasında boşluk şartı aranmamıştır. Çünkü tanımlanan soru kelimeleri birçok formatta yazılabilmektedir. Örneğin 'kim' soru kelimesinin

'kime, kimde, kimden vs.' gibi birçok şekli vardır. Kuralın Soru kelimesi+boşluk şeklinde tanımlanması halinde bu kelimelerden türetilenler kaybedilecektir. Kural uygulandığı zaman ise 'kimya' gibi aslında soru kelimesi olmayan kelimeleri içeren tvitlerde aday soru havuzuna alınması problemi oluşmuştur.

3. Soru kelimeleri soru anlamı vermeyebilir. Örneğin: 'nasıl güzel bir hava' şeklindeki bir tvitte 'nasıl' soru anlamı taşımamaktadır. Ancak 'nasıl' kelimesi soru kelimeleri listesinde bulunduğundan tvitin bu kelimeyi içermesi durumunda aday soru havuzuna dâhil edilmiştir.
4. Veri setinde soru olmadığı halde soru işareti içeren birçok tvit bulunmaktadır. Sistem soru işareti içeren tvitleri aday soru havuzuna dâhil ettiği için soru olmayan tvitlerde havuza girmiştir.

CRF ile sistemi eğitmek için mevcut kurallar hassaslaştırılarak ve çeşitli örüntüler tanımlanarak aday soru havuzunda oluşan hatalar azaltılmıştır.

4.1.3 Şartlı Rastgele Alanlar'ın Uygulanması

CRF, uygulamada hem soru içeren tvitlerin tespit edilmesinde hem de QA sınıfındaki tvitlerin tespit edilmesinde kullanılmıştır. Soru içeren tvitlerin tespitinde aday soru havuzundaki tüm tvitler kullanılırken, QA soru tipi algılama da aday soru havuzundaki gerçekte soru olan tvitler eğitim setinde kullanılmıştır. Her iki uygulama içinde eğitimde kullanılmak üzere, çeşitli özellikler tanımlanmıştır.

4.2 Soru Algılama İçin Geliştirilen Uygulama

Tüm tvitlerin içerisinden soru içeren tvitlerin algılanması için CRF ile bir uygulama geliştirilmiştir. Veri setini eğitmek için 73 özellik tanımlanmıştır. Tanımlanan özellikler aday soru havuzunu oluşturmak için tanımlanmış özelliklerle benzer olup, aday soru havuzunu oluştururken karşılaşılan problemler çözümlenmeye çalışılmıştır. Özellikler aday soru havuzu oluşturmada olduğu gibi kelimelerden değil örüntülerden oluşturulmuştur ve tvitlerde yoğunlukla kullanılan günlük konuşma diline ait örüntüler de özellik olarak eklenmiştir.

4.2.1 Soru Algılama İçin Tanımlanan Özellikler

Soru içeren tvitleri tespit edebilmek için aday soru havuzunu oluşturmada kullanılan özellikler temel alınarak, aday havuzunu oluştururken karşılaşılan hataları gidermeye yönelik özellikler geliştirilmiştir. Tanımlanan özellikler dört grupta birleştirilmiştir. Tvitler kendilerine özgü bir konuşma dili içermektedir. Bu nedenle sıklıkla geçen örüntüler tvitlerin incelenmesi ile tespit edilmiştir. Özellik grupları aşağıda sıralanmaktadır.

I. Soru İşareti: Bu grupta tek bir özellik bulunmaktadır. Tvitin soru işareti içerip içermediği bir düzenli ifade ile kontrol edilir.

II. Soru Eki: Soru ekleri için 17 özellik tanımlanmıştır. Kontrolü yapılan soru ekleri : -mı, -mısın, -mısınız, -mıyım, -mıyız, -mıydık, -mıydım, -mıydın, -mıydınız, -mıymış, -mıdır, -bilirmi, -bilirlermi, -limmi, -sekmi, -dıkmı, -larmı şeklinde sıralanmaktadır. Tvitlerin kuralsız yapısı nedeniyle eklerin ayrı yazılması kuralının çoğu zaman ihlal edilmesi ekleri tespit etmeyi zorlaştırmıştır. Eklerin kelimeden ayrı olması durumu tamamen göz ardı edilip, ilgili eklerden her hangi biriyle biten kelimelerin tvitlerde aranması ise bölüm 4.1.2.1'de bahsedilen sorunlara sebep olmaktadır. Yapılan analizler sonucu –mi eki hariç diğer soru eklerinin ayrı yazılma şartı aranmaması uygun görülmüştür. Ancak, –mi ekinin oluşturduğu problemi çözmek için ilgili eke yönelik dört ayrı özellik tanımlanmıştır. Özellikle çok yaygın kullanılan fiiller (al, bak, gör vs.) ve kelimeler (mümkünmü, doğrumu vs.) tvitlerden tespit edilerek küçük bir sözlük oluşturulmuştur. Sözlükteki kelimeler için ayrı olma şartı aranmamıştır.

   Kelimeden ayrı yazılma kısıdı göz ardı edilen ekler için; bir kelimenin ilgili ekle bittiği halde soru eki olmamasından kaynaklanan hatalar oluşmuştur.

   Örneğin: O ikisi sınıfta çok samimiydi. -->-miydi eki

   Favorimsin Fenerbahçe. -->-mısın eki

   Alarmı kurmayı unutmuşum. -->-larmı eki

   Unutmiyim diye not aldım. -->-miyim eki

   Kankam varsa bende varımdır. -->-midir eki

Bu tarz hataları en aza indirgeyebilmek adına tvitler incelenerek bu duruma neden olan ve sıklıkla geçen ifadeler tespit edilmiş ve her ek için ayrı küçük sözlükler oluşturulmuştur. İlgili eklerin kendileri için tanımlanan sözlükteki ifadelerden birinin ardından gelmemesi, özelliklere kısıt olarak eklenmiştir.

III. Soru Kelimesi: Aday soru havuzunu oluşturmada kullanılan soru kelimeleri, oluşan hataları önleyebilmek için hassaslaştırılmış ve sık geçen örüntüler tespit edilerek tvitte örüntünün bulunması veya bulunmamasına göre özellikler tanımlanmıştır.

Soru kelimesi içeren tvitleri tespit edebilmek için tespit edilen örüntüler için 33 özellik oluşturulmuştur. Kontrolü yapılan soru kelimeleri; niye, kaç, nerde, neler, nedir, kim, ney, nasıl, neden, nolmuş, noldu, ne, napıyo, niçin, naber, nedersin, ..neymiş..miş, neyin var, niye .. çünkü, ne diye sorsa, kaçıncı, napıcan, naptı, napmış, noluyo, ne kadar, neli, neci, nesi, ne zaman, ne demek şeklinde sıralanmaktadır.

Yukarda sıralanan ifadeler soru kelimesi olarak geçebildiği gibi sıfat, zarf vs. olarak da bulunabilmektedir.

Örneğin: ‘Nasıl geldin buraya kadar’ bir soru iken

‘Nasıl eğlendik ama yaa’ soru değildir.

Bu tarz durumlardan oluşabilecek hataları azaltabilmek için, tvitlerin analiz edilmesi ile çıkarılan kalıplar örüntü olarak tanımlanmıştır. Mesela, yukarıdaki örnek için ‘nasıl’ kelimesinin cümle içerisinde hangi kalıplarla birlikteyken soru kelimesi olduğu ve olmadığı tespit edilmiş, bu duruma yönelik örüntüler oluşturulmuştur. Örüntüleri tanımlayabilmek için soru kelimeleri için küçük sözlükler oluşturulmuştur.

IV. Özel Kelime: Tvitlerde sık geçtiği tespit edilen, kelime ve kelime grupları ‘özel kelime’ grubunu oluşturmaktadır. Bu kelimeler geçtiği cümleye soru anlamı katabilen kelimeler olabileceği gibi soru kelimelerinin günlük konuşma dilindeki halleri de olabilmektedir.

Acaba, dimi, demi, sence, sizce, rica etse.., hayırdır, napabil.., nesin, değilmi, varmı, olurmu, öylemi, yokmu, napalım, tamammı, nen var, neymiş o, ..dedi..dedim, diye sordum ..dedi, diye sorsalar ..derim, eee şeklinde sıralanmaktadır. Burada 'napabil..' napabilir, napabilirim, napabiliriz gibi farklı ekler ile oluşturulmuş hallerini de içerebilsin diye kısaltılarak tanımlanmıştır. '..dedi .. dedim' şeklindeki kalıplarda '..' olan yerde herhangi bir şey olabilir anlamındadır.

İlgili özellikler her zaman soru içeren tvitlerde bulunmamaktadır. Zaman zaman soru olmayan tvitlerde de bulunabilmektedir. Ancak, uygulamanın amacı soru olanların tespitine yönelik olduğu için soru bulmadaki hataların azaltılması öncelikli hedef sayılmış, özellikler ve örüntüler buna göre düzenlenmiştir.

Ayrıca, tvitlerde 140 karakter sınırlaması olduğundan zaman zaman sesli harflerin yazılmadığı tespit edildiğinden örüntülerdeki kelimelere sesli harflere bulunma veya bulunmama esnekliği verilmiştir. Kullanıcılar ağırlıklı olarak sesli harfler olmak üzere bir kelimedeki bir harfi birden fazla yazabildikleri için özelliklerdeki birçok harfe birden fazla bulunma esnekliği de verilmiştir.

4.2.2 Soru Algılama İçin Kısıtlar

Soru algılamaya yönelik uygulamanın geliştirilmesi safhasında karşılaşılan en büyük problem; aynı örüntünün her iki sınıfı da temsil eden örnekler içermesinden kaynaklanan çakışık sınıf problemi olmuştur. Uygulamanın soru algılama aşamasında iki sınıf vardır; sorudur veya soru değildir. Bu durumu içeren her özellik sınıflardan birinin başarısını artırırken diğerininkini azaltmıştır. Aşağıda her iki sınıf içinde örnek içeren örüntüler sıralanmaktadır.

1. Kaçıncı

   Örnek: Bidaha Philips mi asla!! Kaçıncı bu ya →soru değil

   Kaçıncı seviyedesin →soru

2. Kim; ilgili soru kelimesini içeren özellik tanımlanırken, örüntü içerisinde 'kim' ile başlayan bir kelime geçip geçmediğini kontrol edecek şekilde tanımlanmıştır.

Çünkü, ‘kim’ birçok farklı ek alıp farklı hallerde (kime, kiminle, kimin gibi) soru ifade edebilmektedir. Bu durumda bazı ‘kim’ ile başlayan ve aslında soru amaçlı kullanılmayan bir takım kelimeleri içeren tvitlerde bu özellik tarafından tanınır hale gelmiştir. Bu durumdan kaynaklanan hatalar örüntülere eklenen küçük sözlükler ile azaltılmıştır. Ancak tamamen yok edilememiştir.

Örnek: Kimine göre lazım değil aşk→kimine, soru değil

3. Ne... ne

Örnek: ne yalnızlığım geliyor aklıma ne mutsuzluğum →Soru değil

ne bu şimdi mayhemle ne alakası var ? →Soru

4. ..ne demek..

Örnek: ne demek her zaman :)) → soru değil

Şaltolon ne demek ? →soru

5. ..mi…mi

Örnek: Kitap mı okuyorum komedi film mi izliyorum belli değil →Soru değil

Ece mi ben mi ? both deme — ikiniz →Soru

6. Neden olacak

Örnek: Seni biraz daha dinlemem biraz daha yarım kalmama neden olacak→soru değil

Bu adam başka hangi ölümlere neden olacak? →soru

7. Ne olur

Örnek: Ne olur ara sıra haberdar et →soru değil

Sizce hükümet düşerse ne olur ? →soru

8. Ne işin var

Örnek: Ne isin var sokakta yaa →soru değil

hayırdır ne isin var oralarda →soru

9. Neden oldu

Örnek: Ötüken'de açık bırakılan ütü yangına neden oldu →soru değil

Yüzündeki yara neden oldu →soru

10. Kimin

Örnek: Şimdi kim bilir kimin gülüşlerini hayatına katıyorsun →soru değil

kimin konseri →soru

11. Ne var

Örnek: Hızlı yazınca harf hatası yapıyorum bunda ne var yani→soru değil

ne var besiktasta? →soru

12. Nerede

Örnek: nerde olursan ol, gün doğumu daima güzeldir. →soru değil

nerde açıklanmış notlar? →soru

13. Neden

Örnek: İstediğini yapmam için tek neden yok →soru değil

Neden gitmedin okula →soru

14. ..mi ; soru eki –mi zaman zaman tvitlerde, soru eki olmadığı halde soru eki gibi ayrı yazılabilmektedir.

Örnek: parmağı mı kırdım→soru değil

15. Kaç; Soru kelimesi olarak sıkılıkla kullanılan bu kelime fiil olarak da kullanılabilmektedir. Çalışmada her hangi bir ayrıştırıcı kullanılmadığından veya kelimeleri kök ve eklerine ayıracak bir yapı geliştirilmediğinden kelimenin soru kelimesi mi yoksa fiil mi olduğu ayırt edilememektedir.

Örnek: dayak gördünmü kaç abicim →soru değil

16. Hangi

Örnek: Hangi yöne gideceğimizi seçme şansımız kalmadı →soru değil

Hangi renk oje süreyim →soru

17. ..msn; diğer birçok özellikte de olduğu gibi –mısın eki için de sesli harflerin yazılmaması esnekliği sağlanmıştır. Ancak bu durum beraberinde bir takım hatalar da getirmiştir.

    Örnek: msn e gelsene →soru değil

    Gelir msn artık →soru

Özellikle 'msn' kelimesi içeren tvitlerin özellik kapsamına alındığı görülmüştür. Burada 'msn', 'massenger' kelimesinin kısaltması olarak bulunmaktadır. Herhangi bir soru anlamı yoktur. İkinci durumda ise soru eki olarak bulunmaktadır anca sesli harfleri kullanıcı tarafından yazılmamıştır.

Bahsedilen kısıtlar uygulamanın başarısını etkilemektedir. Kısıt oluşturan özelliklerin tanımlanması durumunda soru bulmadaki başarı artarken, soru olmayan tvitlerin soru olarak etiketlenmesindeki hatada artmaktadır. Bu durum ise genel başarıyı etkilemektedir. Durumu örneklendirmek için, yukarda 1. kısıt olarak bahsedilen özelliğin başarı üzerindeki etkisi ölçülmüştür. Bölüm 4.2.1'de anlatılan özelliklerin hepsinin eğitimde kullanılması durumunda gerçekte soru olan 73.060 tvitin 69.584 tanesi tespit edilebilmektedir ve soru olmadığı halde soru olarak etiketlenen tvit sayısı 16.533'dür. 'Kaçıncı' soru kelimesini içeren örüntünün aynı zamanda soru olmayan tvitleri de modellemesi nedeniyle özelliklerden kaldırılması durumunda ise 69.545 tanesi tespit edilebilmekte olup, soru olmadığı halde soru olarak etiketlenen tvit sayısı ise 16.503'e düşmektedir. İlk durumda veri setindeki soru olan 73.060 tviti %95,24 doğruluk ile tespit edebilirken, özelliğin kaldırılması durumunda soru algılama doğruluk yüzdesi % 95,18'e düşmektedir. Bu oran bazı kısıtlar için daha fazla bazıları için ise daha az olmaktadır. Tüm kısıtların neden olduğu hatalar göz önüne alındığında doğruluk yüzdesi ciddi oranda düşmektedir. Bu nedenle soru olmayan tvitlerin soru olarak etiketlenme hata sayısında artışa neden olmasına rağmen kısıtlarda bahsedilen özellikler veri setinin eğitiminde kullanılmıştır.

Uygulamanın amacı soru olan tvitleri tespit etmek olduğu için soru olmayan tvitlerin soru olarak etiketlenmesini önlemeye yönelik özellikler tanımlamış olsa da, soru olan tvitlerin tespit edilebilmesine yönelik özellik tanımlamaya ağırlık verildiğinden kısıtların oluşturduğu hatalar göz ardı edilmiştir.

4.3 QA Algılama İçin Geliştirilen Uygulama

Geliştirilen uygulamanın ikinci aşamasında yedi farklı soru tipi içerisinden QA sınıfındaki tvitler ayırt edilmeye çalışılmıştır. QA algılama için de CRF yönteminden yararlanılmıştır. Eğitim için kullanılan veri seti, soru algılama için kullanılan veri setinden türetilmiş olup, NQ sınıfı haricindeki tvitlerden oluşmaktadır. Eğitim setindeki 73.060 tvitin 4.511 tanesi QA sınıfına aittir. Tvitlerin sınıflara göre dağılımı Çizelge 3. 1' de verilmektedir. QA algılama için 43 özellik tanımlanmıştır.

4.3.1 QA Algılama İçin Tanımlanan Özellikler

Veri setindeki QA sınıfındaki tvitler incelendiğinde ilgili sınıftaki tvitlerin daha çok karşılıklı konuşma şeklinde olduğu gözlemlenmiştir. Bu nedenle tanımlanan örüntüler çoğunlukla karşılıklı konuşmaları modelleyebilecek şekilde tanımlanmıştır. Yedi farklı türdeki sorulardan QA sınıfındaki tvitleri algılamak için tanımlanan özellikler aşağıdaki şekilde sıralanmaktadır.

1. ....neymiş...miş

   Örnek: Neymiş efendim öyle birden sevemezmiş. Vade farksız 12 taksit yapıyoruz

2. ....napıyormuşuz... muşuz

   Örnek: Demek ki napıyomuşuz ? Kabaran saçları sevmiyomuşuz

3. ....dediklerinde/dediğinde...demek istediğim...

   Örnek: “Neyin var?” dediklerinde “Zaten anlamayacaksın uğraştırma beni.” demek istediğim insanlar var.

4. ....ne diye sorsa...

   Örnek: Hayattaki en kötü şey ne diye sorsalardı, onlara canınızdan çok sevdiğiniz insanın gözlerinizin önünde size karşı yabancılaşması derdim.

5. dedi....dedim/dedik/cevap....

   Örnek: Okeyde kıza elin nasıl? dedim. Ojeli dedi. Ben şoka girdim o migrosa.

6. ....diye sordum/ sordular/ soruyorlar/ soruyorda/ sorarsa...dedi/ dedim/ cevap/ derim/ diyom...

Örnek: Tanımadığım kızın biri eklemiş Faceden. Sen kimsin diye sordum aldığım cevap; tanımadığım birine bütün dertlerimi anlatmak oldu.

7. ....diye sorsalar/ deseler/ desem/ desen/ dediğinde/ sorduğumda/ sorsalardı... derdim/ deyip/ dersiniz/ derim

Örnek: Sevmek mi daha zor? Unutmak mı? deseler. Sabah erken uyanmak derim.

8. .... ne demektir/ nedir .... dir/ tir.

Örnek: Marksizm nedir diye sorsan cevap veremeyecek insanların aynı battaniyeyi, yemeğini ağacını paylaşmasıdır.

9. ....dediklerinde/dediğinde .... demiş

Örnek: Yildirim Aktunaya sormuşlar efendim arşivler doldu evrakları napalim? o da demis hepsinin fotokopisini çekin asillari imha edin! bürokrasi !

10. ....var/ yok mı/ mıdır var/ yok

Örnek: Yarın sınavım varmı? Var

11. .... yapar mı yapar...

Örnek: İstediğinde yapar mı yapar.

12. .... demedik mi dedik...

Örnek: olmaz günah :)) kadın sadece evde olacak demedik mi dedikk :))

13. .... diyor/ diye soruyorum/ diyorum/ diyolar/ soruyorum... diyorum/ diyor/ diyolar

Örnek: Babam paran var mı diyo, var diyorum. Ne kadar var diyo masum bi şekilde her zaman 10 diyorum. Hemen 50 çıkarıp veriyo aşkım iki günde bir :D

14. .... mı/ mıyım/ mıyız/ mıdır/ mısın/ mısınız/ mıydım/ dimi... evet/hayır

Örnek: Kafede otururken masanın üzerindeki peçete, kürdan, şeker gibi şeyleri sömürmeyen bizden midir arkadaslar? HAYIIIIIIIIIIRRRRRR.

15. neden/ niye/ ne için/ niçin.. mi...

Örnek: 1000 tanede diş fırçası olsa benımkı tanılıyomus neden mı Hep ruj lekesi varmıs.

16. neden/ niye/ ne için/ niçin ... için/ diye/ yoo/ çünkü/ meğer...

Örnek: Neden savaşlar oluyor biliyor musun ? Çünkü bu dünya insansiz başladi ve onsuz bitecek ...

17. naber...iyi

Örnek: naber ??? — İYİ SENDEN :D

18. ....ne olur ... cevap/ olur

Örnek: SBS'ye girersem ne olur, dumur olur.

19. ....mıyım ... değilim

Örnek: bilmem iyiyim galiba. iyi miyim? yok değilim

20. ....napıyorum/ nedir/ kim/ mi/ miyim... tabi/ tabiki/ tabiki de...

Örnek: Hersey belgeleniyor. Kim tarafindan tabiki redhack tarafindan. Muhalefet ne yapiyor. Bakiyor. Aciyorum onlara.

21. ....sorusuna/ derse/ sorulunca/ sorsalar/ sorduklarında/ sorusunun ... dedim/ dedi/ cevabını/ derim/ diyeceğim/ yanıtı/ cevabım...

Örnek: Bana ne zaman "Özgürlük nedir?" diye sorsalar, hep "Temiz bir vicdan" diye cevaplarım.

22. ....olur/ olmaz/ olabilir mi... olur/ olmaz/ olabilir

Örnek: Olmaz mı olabilir.;)

23. ....ne denir ... denir

Örnek: kanka bısey sorucam çıplak ördeğe ne denir ? Ne denir ? cıbılduck :F:F.d..d:D...Sfa

24. ....mısın... evet/hayır/yoo

Örnek: Mesaj atmış yaşıyor musun demiş, yok öldüm cennetten yazıyorum

25. ....ne demek... demek

Örnek: Biz dostça ayrıldık ne demek. aşkım ben ayrılmak istiyorum . tamam kardeşim ne demek. sağol kanka. lafı mı olur

26. .... ne kadar... kadar

Örnek: Ne kadar sevdin diye soruyorlar bazen Hep aynı cevabı veriyorum ; Hak Etmediği Kadar !!

27. .... ne zaman... zaman

Örnek: Zaten aşk ne zaman doğru zaman kollar hiçbiir zamaan

28. ....cevap: ...

Örnek: kendimi sorguladim ktüyü seçtiğimemi pişmanım iktisatı mı cevap:ikisinide!!!

29. ...sordum... dedi

Örnek: Gecen bi kiza "suan kafam patlasa napardin?" diye sordum "ogretmenleri cagirirdim dedi, CAGIR NOLUR

30. ...eee ne demişler...

Örnek: eee ne demişler hayvanın hatrı yoksa bile sahibinin vardır....

31. .... deyince ... demesi

Örnek: Ya bir kizaa ne ismarlim ne istersin deyince onundaaa hic bisey istemiyorum demesii beni sinir ediyo yaaa :/

32. .... derse/ diyenlere/ soranlara ... cevap/ cevabım...

Örnek: Neden ve nasıl bu kadar mutlusun, enerjiksin diyenlere tek bir cevabım var. Mutluluğum kimseye değil bir tek kendime baglı muahhhh :)

33. .... ne/ nedir biliyor musun...

Örnek: İnsan nedir biliyor musun ? Ağaçları kesip kağıt yapan sonrada kağıda ağaçları koruyalım yazan

34. ....ne/ nedir bilir misin...

Örnek: Aşk'ın en adaletsiz yanı nedir bilir misin? Ağızdan çıkmaya cesareti olmayan sözlerin yürekte fırtınalar kopartmasıdır.

35. ....noldu ... oldu...

Örnek: ben çabuk tuttum da noldu, efes zengin oldu :(((

36. +...-...

Örnek: +Moralini kim bozdu? -MURAT BOZdu :D:D:d:Dasdfhj

37. -...+...

Örnek: - Karnede asker var mı + Ne askeri tim geliyor timm

38. ...=....=

Örnek: Harry=Sevgilin var mı? Sen=Hayır. Harry=Neden? Sen=Çünkü hazır değilim. Harry=Hm. Sen=Peki senin neden yok? Harry=Çünkü hazır degilsin..

39. -...-...

Örnek: -Yastık savaşı yapalım mı -Neyle -Sen masayı al, ben kapıyı söküp geliyorum.

40. ... : .... : ...

Örnek: Yaşlı adam: sana dayımın selamı var Genç adam : hangi dayın Yaşlı adam : namazdayım gdik:DsJdSjH:hDGj

41. ... ; .... ; ...

Örnek: Hasan hoca; irem bu gömlek ne? Culsuz musun??? Ben; hocam 220 irem dinç sözlü olarak ;);)))

42. ...?-...

Örnek: Neden sevgiliniz yok? -Tercih meselesi

43. ...? —...

Örnek: En sevdiğin kahvaltı nedir? — normal bir kahvaltı işte yeter bana :)

4.3.2 Soru Algılama İçin Kısıtlar

Bölüm 4.3.1'de bahsedildiği üzere, QA algılama için geliştirilen uygulamada özellikler yoğunlukla karşılıklı konuşma kalıplarına yöneliktir. Tanımlanan örüntüler tvitler analiz edilerek oluşturulmuştur. Karşılıklı konuşma kalıpları aynı zamanda QA sınıfında olmayan tvitleri de modelleyebilmektedir. Örneğin '...dedi ... dedim' şeklindeki bir örüntü içerisinde iki cümle barındırmaktadır. QA sınıfındaki tvitlerde ilk cümle bir soru ikinci cümle ise cevap şeklinde olmaktadır. Benzer şekilde QA sınıfında olmayan tvitlerde de karşılıklı konuşmalar yer almaktadır.

Örnek: Dünya yuvarlak mıdır dedi evet dedim.

Dünya yuvarlaktır dedi yenimi farkettin dedim.

Yukardaki örnekte her iki tvitte aynı kalıp ile modellenebilmektedir. Ancak, ilk örnek QA sınıfında iken ikinci örnek değildir. Tanımlanan özelliklerde, bu şekilde karşılıklı konuşmayı modelleyen tüm özellikler, QA sınıfında olmayan tvitleri de bu sınıfa dâhil etmektedir. Bu durum uygulamanın hatalarını oluşturmaktadır. İlgili özellikler kaldırıldığında ise, tahmin edileceği üzere QA sınıfındaki birçok tvit tespit edilememektedir. Örneğin, hatalar göz ardı edilerek, tüm özelliklerin QA algılama için kullanılması durumunda gerçekte QA sınıfında olup uygulamanın tespit edemediği tvit sayısı 983 iken kısıtlarda belirtilen 5. örüntüyü kapsayan özelliğin QA olmayan tvitleri QA olarak etiketlemesini azaltmak amacıyla kaldırılması durumunda, tespit edilemeyen QA sınıfındaki tvit sayısı 1.057 olmaktadır. Amaç, QA sınıfındaki tvitleri tespit edebilmek olduğu için ilgili özellikler kullanılarak oluşturduğu QA sınıfında olmayan tvitlerin QA sınıfına dâhil edilmesine yönelik hatalar göz ardı edilmiştir.

BÖLÜM 5

DENEYSEL SONUÇLAR

Bu bölümde soru içeren ve QA sınıfındaki tvitlerin tespitine yönelik geliştirilen uygulamanın deneysel sonuçları verilmiştir. Bölüm 5.1'de uygulamanın ilk aşaması olan soru algılama kısmının test ve başarı sonuçları verilirken Bölüm 5.2'de, uygulamanın QA sınıfındaki tvitlerin algılanmasına yönelik kısmının test ve başarı sonuçları verilmiştir.

5.1 Soru Algılama

Bu bölümde, tüm tvitleri içeren veri setinden soru olan tvitlerin tespit edilmesine yönelik uygulamanın 5-kat çapraz geçerleme ile sistemin test edilmesi sonucu elde edilen sistem başarı değerleri verilmektedir.

Veri setinde toplam 136.449 tvit bulunmaktadır. Tvitler, Bölüm 3'de bahsedildiği üzere sekiz sınıftan oluşmaktadır. Bu sınıflardan biri NQ olup, gerçekte soru olmayan tvitleri temsil etmektedir. Diğer yedi sınıf ise soru türlerini temsil etmektedir. Uygulamanın soru algılama aşaması; gerçekte soru olan yedi sınıfa ait tvitler tespit edebilmektir. 136.449 tvitin 73.060 tanesi soru içermekte olup, 63.389 tanesi de NQ sınıfına dâhildir.

5.1.1 Uygulamanın Doğruluğu

Soru algılamaya yönelik geliştirilen uygulamanın doğruluğunu ölçümleyebilmek için tüm veri seti hem eğitim hem de test için kullanılmıştır. Sistemin özellik gruplarından her biriyle tek başlarına modellenmesi sonucunda elde edilen doğruluk, tutturma, bulma ve F1-Puanı değerleri Çizelge 5.1'de gösterilmektedir.

Çizelge 5. 1 Özellik gruplarının tekli başarı ölçümleri (Sİ: Soru İşareti, SE: Soru Ekleri, SK: Soru Kelimeleri, ÖK: Özel Kelimeler)

Özellik	Doğruluk	Tutturma	Bulma	F1-Puanı
Sİ	0,698	0,926	0,474	0,627
SE	0,686	0,922	0,453	0,607
SK	0,651	0,767	0,502	0,607
ÖK	0,535	0,535	1	1

Soruları tespit etmeye yönelik çıkardığımız özellik gruplarının sisteme eklenmesi ile oluşan sistemdeki başarı değişimine yönelik değerler Çizelge 5.2'de özetlenmektedir.

Çizelge 5. 2 Özellik gruplarının eklenmesi sonucu başarıdaki değişim

Özellik	Doğruluk	Tutturma	Bulma	F1-Puanı
Sİ	0,698	0,926	0,474	0,627
Sİ+SE	0,786	0,900	0,677	0,773
Sİ+SE+SK	0,844	0,807	0,931	0,864
Sİ+SE+SK+ÖK	0,875	0,808	0,952	0,874

Her bir örüntü uygulamaya özellik olarak eklenmeden önce, eklenmeden önceki ve eklendikten sonraki doğruluk yüzdeleri incelenmiştir. Çizelge 5.2' den de görüldüğü üzere, özellik grupları eklendikçe, tutturma hariç tüm değerler artış göstermiştir. Tutturma değerinin düşmesinin nedeni ise Bölüm 4'de değinilen aynı örüntünün iki sınıfı da modellemesinden kaynaklanan hatalardır. Her yeni örüntü soru olan ve algılanabilen tvitlerin sayısını artırmış, aynı zamanda soru olmadığı halde soru olarak etiketlenen tvit sayısını da artırmıştır. Bu durum tutturma değerinin düşmesine neden olmuştur.

Özelliklere eklenen sözlüklerin başarıdaki etkisini gözlemleyebilmek için, her bir sözlükten iki adet kelime çıkarılarak sistem tüm veri setiyle yeniden eğitilmiştir.

Eğitimde kullanılan tüm veri setinin test için de kullanılması sonucunda elde edilen F1-Puanı değeri 0,839 çıkmıştır. Çizelge 5.2'den de görüldüğü üzere sözlüklerden kelime çıkarmadan önce ise F1-Puanı değeri 0,874'dür. Bu durum sözlüklerin sistemin başarısındaki etkisini göstermektedir.

Çizelge 5.3' de ise tüm veri setinin hem eğitim hem de test için kullanılması durumunda oluşan sınıf karışıklık matrisi verilmiştir.

Çizelge 5. 3 Soru algılama için sınıf karışıklık matrisi

		Gerçek	
		Soru	$\overline{\text{Soru}}$
Tahmin	Soru	69.584	16.533
	$\overline{\text{Soru}}$	3.476	49.856

5.1.2 5-Kat Çapraz Geçerleme

Soru algılama için geliştirilen uygulamanın başarısını ölçmek için veri seti beşe bölünerek 5-kat çapraz geçerleme yapılmıştır. Eğitim ve test kümelerindeki veriler rastgele seçilmemiş olup, her biri sırayla tüm veri setinin beşte birlik kısmından oluşturulmuştur. Çizelge 5.4' de soru algılama uygulamasının başarısına yönelik sonuçlar verilmektedir.

Çizelge 5. 4 Soru algılama başarı ölçümü

K	Doğruluk	Tutturma	Bulma	F1-Puanı
1	0,855	0,811	0,953	0,876
2	0,851	0,806	0,949	0,872
3	0,855	0,810	0,953	0,876
4	0,853	0,807	0,954	0,874
5	0,854	0,806	0,953	0,873
Ort.	0,8536	0,808	0,9524	0,8742

Geliştirilen uygulamanın 5-kat çapraz geçerleme yapıldıktan sonraki başarı sonuçları, Çizelge 5.2'de tüm özellik gruplarının kullanılması sonucunda elde edilen başarı sonuçları ile çok yakın çıkmıştır.

Çizelge 5.4'de görüldüğü üzere tutturma değeri diğerlerine göre düşük çıkmıştır. Bu durumun nedeni Bölüm 4'de kısıtlarda anlatıldığı üzere aynı örüntünün iki sınıfı da modelleyebiliyor olmasındandır.

Tvitlerin kendine özgü bir yazım dili vardır. Çoğunlukla imla ve yazım kurallarından tamamen bağımsız, günlük konuşma dilinde yazılmaktadırlar. Bu durum gerçekte soru olan bazı tvitlerin her hangi bir kalıpla modellenebilmesini engelleyerek bulma değerindeki hatayı oluşturmuştur.

5.2 QA Algılama

Veri setinde bulunan toplam 136.449 tvitin 73.060 tanesi gerçekte soru olan tvitlerden oluşmaktadır. QA algılama için mevcut veri setindeki, gerçekte soru olan 73.060 tviti içeren bir veri seti türetilmiştir. 73.060 tvitin 4.511 tanesi QA sınıfındadır. Uygulamanın doğruluğu tüm veri setinin hem eğitim hem de test için kullanılması ile test edilmiştir. Başarıyı ölçmek için ise 5-kat çapraz geçerleme yapılmıştır.

5.2.1 Uygulamanın Doğruluğu

Tüm veri setinin hem eğitimde hem de testte kullanılması sonucu elde edilen sınıf karışıklık matrisi Çizelge 5.5'de verilmektedir.

Çizelge 5. 5 QA algılama için sınıf karışıklık matrisi

		Gerçek	
		QA	$\overline{QA}$
Tahmin	QA	3.528	2.372
	$\overline{QA}$	983	66.177

Çizelge 5.6'de ise tüm veri setinin hem eğitim hem de test için kullanılması sonucu elde edilen başarı ölçümleri bulunmaktadır.

Çizelge 5. 6 Tüm veri seti için başarı

Doğruluk	Tutturma	Bulma	F1-Puanı
0,954	0,598	0,782	0,678

5.2.2 5-Kat Çapraz Geçerleme

QA algılama için geliştirilen uygulamanın başarısı 5- kat çapraz geçerleme yapılarak ölçülmüştür. Eğitim ve test kümesindeki veriler rastgele seçilmemiş olup, her birinin tekrarsız olacağı şekilde bölümleme yapılmıştır. Ayrıca, test kümesindeki verilerin eğitim kümesinde bulunması da önlenmiştir.

Çizelge 5.7' da 5-kat çapraz geçerleme sonuçları verilmektedir.

Çizelge 5. 7 QA algılama başarı ölçümü

K	Doğruluk	Tutturma	Bulma	F1-Puanı
1	0,953	0,589	0,777	0,670
2	0,955	0,604	0,758	0,672
3	0,953	0,596	0,79	0,679
4	0,955	0,599	0,789	0,681
5	0,956	0,609	0,785	0,686
Ort.	0,9544	0,5994	0,7798	0,6776

Tutturma değeri, QA olarak etiketlenen tvitlerin yüzde kaçının gerçekten QA sınıfına dâhil olduğunu göstermektedir. Bu değerde oluşan hatanın nedeni Bölüm 4 'de kısıtlarda bahsedildiği üzere tanımlanan örüntülerin QA olan ve olmayan iki sınıfı da modelleyebiliyor olmasından kaynaklanmaktadır.

Bulma değeri gerçekte QA sınıfına ait olan tvitlerin yüzde kaç doğrulukla tespit edilebildiğini göstermektedir. Bu değerdeki hata ise hiçbir şekilde modellenemeyen

tvitlerden kaynaklanmaktadır. Tvitler son derece kuralsız yazılabilmektedir. Bu durum tviti modelleyebilmeyi zorlaştırmaktadır.

Uygulamanın, yedi farklı soru türünden QA sınıfındaki tvitleri algılamaya yönelik oluşturulmasının nedeni bu soru tipinin modellenmeye en yatkın grubu oluşturmasıdır. Soru türleri içerisinde en yaygın bulunan FK ve RQ sınıflarını birbirinden ayırabilmenin anlamsal bir analiz katmadan mümkün olmayacağı düşünülmektedir. Çalışmanın kapsamı anlamsal analiz olmadığı için ilgili iki soru türü ayırt edilmeye çalışılmamış olup, bu işlem kendi başına bir çalışma konusu olacaktır. Ayrıca, veri seti içerisinde sayıca baskın olan bir grubu tespit etmek yerine sayıca az olan bir grubu tespit edebilmek amaçlanmıştır.

QA algılama için oluşturulan veri setinde toplam 73.060 tvit bulunmaktadır. Bunlardan 4.511 tanesi ise QA sınıfına aittir. QA sınıfındaki tvit sayısı toplam tvit sayısına oranlandığında veri setinin %6,17'sinin QA sınıfındaki tvitlerden oluştuğu görülür. Bu oran göz önüne alındığında elde edilen F1-Puanı değerinin başarılı olduğu anlaşılmaktadır. Ayrıca, Çizelge 5.5'de de görüleceği gibi, QA sınıfına ait olmayan 68.549 tvitten sadece 2.372 tanesinin uygulama tarafından QA olarak etiketlenmiş olması geliştirilen uygulanmanın doğru modellendiğini de göstermektedir.

BÖLÜM 6

SONUÇ VE ÖNERİLER

Web 2. 0 teknolojisi ile Twitter en yaygın kullanılan mikro blog servislerinden biri haline gelmiştir. Her geçen gün kullanıcı ve paylaşılan içerik sayısı artmakta olan Twitter ile kullanıcılar sadece içerik okuyucu olmaktan çıkmış, içerik sağlayıcı konumuna gelmiştirler. Twitter aracılığıyla kullanıcılar görüşlerini, düşüncelerini, günlük aktivitelerini, ilgi alanlarını birbirleriyle paylaşabilmektedir. Kullanıcılar mobil erişim imkanı ve kullanım kolaylığı neticesinde Twitter gibi sosyal ağları bilgi paylaşımı yapmak ve sorularına cevap almak için de sıklıkla kullanmaktadır.

Soru algılama, Doğal Dil İşleme alanında bilgi çıkarımı ve bilgiye erişim amacıyla, kurallı veya kuralsız bir derlemden soru anlamı taşıyan cümlelerin tespit edilmesidir. İlgili alandaki çalışmalar, insanların bilgi ihtiyaçlarını Internet üzerinden gidermeye başlaması ile birlikte forum siteleri ve mikro blog servislerindeki paylaşımlar gibi kuralsız metinlere yönelmiş olup, Türkçe için pek fazla çalışma bulunmamaktadır. Bu çalışmada, Türkçe tvitlerden oluşan bir veri setinden, Şartlı Rastgele Alanlar metodu ile soru içeren tvitlerin tespit edilmesi işlemi gerçekleştirilmiştir.

Geliştirilen çalışmada kullanılacak olan veri setindeki tvitlerden, sistem için anlamsız olan veri temizlendikten sonra, kural tabanlı bir metot ile soru olmaya aday tvitler belirlenmiştir. Aday tvitler; soru içerip içermediklerine ve içerdikleri sorunun türüne göre etiketlendirilmiştir. Veri setinde yedi adet soru türü bulunmaktadır. Ardından Şartlı Rastgele Alanlar metodu ile aday tvitlerden soru içerenler %87 başarı ile tespit edilmiştir. Ayrıca, soru içeren tvitlerden, mevcut soru türlerinden biri olan soruyu ve cevabını içeren tvitler %68 başarı ile tespit edilmiştir.

Soru içeren tvitlerin tespit edilmesi aşamasında, bu alanda çok yaygın kullanılan, soru işareti, soru kelimeleri gibi dile bağımlı, soru cümlesi oluşturma kurallarından yararlanılmıştır. Kuralların küçük boyutlu sözlükler ile örüntü şeklinde tanımlanması başarıyı artırmıştır. Soruyu ve cevabı içeren tvitleri tespit edebilmek için ise, tvitler analiz edilmiş ve bu türdeki soruların genellikle karşılıklı konuşma şeklinde olduğu görülmüştür. Bu nedenle özellikler, karşılıklı konuşmaları modelleyebilecek şekilde örüntüler kullanılarak oluşturulmuştur. Aynı şekilde örüntüler küçük boyutlu sözlükler ile beslenmiş ve böylece modellenebilen tvit sayısı artırılarak başarı yükseltilmiştir.

Tvitler, son derece kuralsız bir yapıya sahip olup, günlük konuşma dilinde yazılmaktadır ve kendine özgü bir jargonu bulunmaktadır. Verinin kuralsız doğası modellenebilmesini zorlaştırmaktadır. Soru içeren tvitlerin tespitinde, tanımlanan bir takım özellikler; Türkçe' de, soru cümlesi oluşturmanın dışında farklı görevlerde de kullanılabilmesi, tvitlerin yazım kurallarına dikkat edilmeden yazılması nedeniyle, soru içeren tvitleri modelleyebildiği gibi soru içermeyen tvitleri de modelleyebilmektedir. Bu durum çakışık sınıf problemini ortaya çıkarmakta ve başarıyı etkilemektedir. Veri setindeki bir soru türü olan hem soruyu hem de cevabı içeren tvitlerin tespit edilmesi aşamasında ise özelliklerin genellikle karşılıklı konuşmayı modelleyecek şekilde tanımlanması, karşılıklı konuşmaların soru-cevap çiftini içerebileceği gibi, soru içermeyen iki cümleden de oluşabiliyor olması başarıyı etkilemiştir. Her iki aşamada da özelliklerin hassaslaştırılması ve sayısının artırılması, doğru bulunabilen özelliklerin sayısının artmasına aynı zamanda yanlış bulunan negatif örneklerinde sayısını arttırmaktadır. Bu nedenle özellikler her iki hata türünün oranını dengede tutabilecek şekilde tanımlanmıştır. Ayrıca, bazı tvitler hiçbir şekilde modellenemediğinden tespit edilememektedir. Çalışmaya anlamsal özellikler eklenerek her iki sınıf için de çakışık sınıf probleminden kaynaklanan hatalar azaltılabilir. Ayrıca, sisteme tvitlerdeki kelimelerin biçimbirimsel analizini yapacak bir sistem entegre edilerek Türkçe' deki eş sesli kelimelerden kaynaklanan hatalar azaltılabilir.

KAYNAKLAR

[1] Efron, M. and Winget, M., (2010) "Questions are content: a taxonomy of questions in a microblogging environment." In Proc. of ASIST '10.

[2] Paul, S.A., Chi, E., (2011a). "Is Twitter a Good Place for Asking Questions?". A Characterization Study. AAAI Conference on Weblogs and Social Media (ICWSM '11).

[3] Morris, M. R., Teevan, J., and Panovich, K., (2010). "What do people ask their social networks, and why?: a survey study of status message q&a behavior." In Proceedings of the 28th international conference on Human factors in computing systems, CHI '10: 1739–1748. New York, NY, USA: ACM.

[4] Paul, S.A., Chi, E., (2011b). "What is a Question? Crowdsourcing Tweet Categorization." Position paper at the workshop on Crowdsourcing and Human Computation at the ACM Conference on Human Factors in Computing Systems (CHI '11).

[5] Li ,B., Si, X, Lyu, Michael R., King, I., Chang, Edward Y., (2011). "Question Identification on Twitter". CIKM'11, Glasgow, Scotland, UK.

[6] Büyükşener, E., (2009). "Türkiye'de Sosyal Ağların Yeri ve Sosyal Medyaya Bakış".XIV. Türkiye'de İnternet Konferansı.

[7] Twitter, https://twitter.com/, 12.10.2013.

[8] Twitter, tr.wikipedia.org/wiki/Twitter, 12.10.2013.

[9] About Twitter, Inc., about.twitter.com/company, 12.10.2013

[10] Cingiz, M. Ö, Diri, B., (2012). "Content Mining of Microblogs", IEEE International Symposium on Foundation of Open Source Intelligence and Security Informatics, FOSINT- SI 2012, İstanbul.

[11] Go, A., Bhayani, R., Huang, L., (2009). "CS224N Project Report". Stanford.

[12] Bravo-Marquez, F., Mendoza, M., Poblete, B.,(2013). "Combining strengths, emotions and polarities for boosting Twitter sentiment analysis", WISDOM-13,New York.

[13] Popescu, A. M., Pennacchiotti, Marco., (2010). “Detecting Controversial Events from Tweitter”. CIKM-10, Toronto, Ontario, Canada.

[14] Klein, B., Castanedo, F., Elejalde, I., Lopez-de-Ipina, D., Nespral, A., (2013). “Emergency Event Detection in Twitter Streams Based on Natural Language Processing”. Ubiquitous Computing and Ambient Intelligence.

[15] Michelson, M., Macskassy, S. A., (2010). “Discovering Users' Topics of Interest On Twitter”. AND'10, Toronto, Ontario, Canada.

[16] Agarwal, P.(2013). Prediction of Trends in Online Social Netwok, Yüksek lisans Tezi, Indian Institute of Technology, New Delphi.

[17] Asur, S., Huberman, B. A., (2010). “Predicting The Future With Social Media”. Web Intelligence and Intelligent Agent Technology (WI-IAT)

[18] Wang, X., Gerber, M. S., Brown, D.E., (2012). “Automatic Crime Prediction Using Events from Twitter Posts”. Social Computing, Behavioral-Cultural Modeling and Prediction.

[19] Shamma, D.A., Kennedy, L., ve Churchill, E.F., (2009). “Tweet the debates: understanding community annotation of uncollected sources”, WSM’09: Proceedings of first SIGMM Workshop on Social media, 3-10, ACM, New York, NY, USA

[20] Değirmencioğlu, E.A., (2010). Exploring Area-Specific Mikroblogger Social Networks, Master Tezi, Boğaziçi Üniversitesi.

[21] Balahur, A., Boldrini, E., (2010). “Going Beyond QA Systems: Challenges and Keys in Opinion Question Answering”, Coling 2010,

[22] Dent, K., Paul, S., (2011). “Through the Twitter Glass: Detecting Questions in Micro-Text”. Analyzing Microtext: Papers from the 2011 AAAI Workshop.

[23] Cong, G., Wang, L., Lin, C.Y., Song, S. I., Sun, Y., (2008). “Finding Question-Answer Pairs from Online Forums”, SIGIR’08, Singapore.

[24] Wang, K., Chua, T.S., (2010) “Exploiting salient patterns for question detection and question retrieval in community-based question answering”. In COLING ’10.

[25] Hong, L., Davison, B. D., (2009). “A Classification-based Approach to Question Answering in Discussion Boards”, SIGIR’09, Boston, USA.

[26] Ding, S., Cong, G., Lin, C.Y., Zhu, X., (2008). “Using conditional random fields to extract contexts and answers of questions from online forums.” In ACL.

[27] Shrestha, L., McKeown, K., (2004). “Detection of question-answer pairs in email conversations”. In Proc. of COLING.

[28] Lafferty, J., McCallum, A., Pereira, F. C. N., (2001) “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”. ICML’01,

[29] Rabiner, L., Juang, B. H., (1993). “Fundamentals of Speech Recognition. Prentice Hall Signal Processing Series.” Prentice-Hall, Inc.

[30] Tür, G., Hakkani-Tür, D. Ve Oflazer, K., (2003). "A Statistical Information Extraction System for Turkish", Natural Language Engineering, 9, 2: 181-210.

[31] Rabiner, L.R., (1989) . "A tutorial on Hidden Markov models and selected applications in speech recognition". Proceeding of the IEEE, 2:257-285,.

[32] Anderson, J. R., Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (1985). Machine Learning: An Artificial Intelligence Approach,Cilt 1.

[33] Manning, C., Schutze, H. (1999). "Foundations of Statistical Natural Language Processing", MIT Press, Massachussetts.

[34] Wallach, H. M., (2004). "Conditional Random Fields: An Introduction", University of Pennsylvania Technical Report MS-CIS-04-21.

[35] MacQueen, J. B., (1967). "Some Methods for classification and Analysis of Mutivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, 1:281-297.

[36] MacCallum, A., Freitag, D., Pereira, F., (2000). "Maximum Entropy Markov models for information extraction and segmentation", In international Conference on Machine Learning.

[37] Clifford, P., (1990). Markov random fields in statistics. In Geoffrey Grimmett and Dominic Welsh, editors, Disorder in Physical Systems: A Volume in Honour of John M. Hammersley, 19–32. Oxford University Press.

[38] Clifford, P., (1990). Markov random fields in statistics. In Geoffrey Grimmett and Dominic Welsh, editors, Disorder in Physical Systems: A Volume in Honour of John M. Hammersley, 19–32. Oxford University Press.

[39] Shannon, C. E., (1948). A mathematical theory of communication. Bell System Tech. Journal, 27: 379–423 and 623–656.

[40] Jaynes, E. T., (1957). Information theory and statistical mechanics. The Physical Review, 106(4):620–630.

[41] Mallet, http://mallet.cs.umass.edu/ , 27.12.2013.

[42] Machine Learning with MALLET, http://mallet.cs.umass.edu/mallet-tutorial.pdf, 27.12.2013.

[43] Fan, J. J., Su, K. Y., (1993). "An Efficient Algorithm for Matching Multiple Patterns".

[44] McNaughton. R., Yamada, H., (1960). "Regular Expressions and State Graphs for Automata".

[45] Thompson, K., (1968). "Regular Expresion Search Algorithm" .

[46] K-Fold Cross Validation, http://statweb.stanford.edu/~tibs/sta306b/cvwrong.pdf, 03.01.2014.

[47] Confusion Matrix, http://en.wikipedia.org/wiki/Confusion_matrix, 03.01.2014.

[48] Sign in with your Twitter Acoount, https://dev.twitter.com/apps, 04.01.2014.

[49] Twitter with Java, http://www.java-tutorial.ch/framework/twitter-with-java-tutorial, 04.01.2014.

[50] Grishman, R., Sundheim, B., (1996). 'Message Understanding Conference-6: A Brief History'.

[51] Message Understanding Conference Proceedings MUC-7, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html, 10.01.2014.

[52] Mallet, http://mallet.cs.umass.edu/, 10.01.2014.

[53] Soru Cümlesi, http://tr.wikipedia.org/wiki/Soru_c%C3%BCmlesi, 10.01.2014.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Zeynep Banu ÖZGER

Doğum Tarihi ve Yeri : 03.01.1985, Kahramanmaraş

Yabancı Dili : İngilizce

E-posta : zeynep@ce.yildiz.edu.tr

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Lisans	Bilgisayar Mühendisliği	Yıldız Teknik Üniversitesi	2011
Ön lisans	Bilg. Tek. ve Prog.	Sütçü İmam Üniversitesi	2007

İŞ TECRÜBESİ

Yıl	Firma/Kurum	Görevi
2013-	Yıldız Teknik Üniversitesi	Araştırma Görevlisi
2011	Akedaş	Bilgisayar Mühendisi