



**SOSYAL MEDYADAN ELDE EDİLEN BÜYÜK VERİ BAZ ALINARAK
SOSYAL OLAY ANALİZİ GERÇEKLEME**

Muhammed ŞEN

**YÜKSEK LİSANS TEZİ
BİLGİSAYAR BİLİMLERİ ANABİLİM DALI**

**GAZİ ÜNİVERSİTESİ
BİLİŞİM ENSTİTÜSÜ**

ŞUBAT 2016

Muhammed ŞEN tarafından hazırlanan “SOSYAL MEDYADAN ELDE EDİLEN BÜYÜK VERİ BAZ ALINARAK SOSYAL OLAY ANALİZİ GERÇEKLEME” adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ / OY ÇOKLUĞU ile Gazi Üniversitesi Bilişim Enstitüsü Bilgisayar Bilimleri Anabilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Doç. Dr. Hacer KARACAN

Bilgisayar Mühendisliği, Gazi Üniversitesi

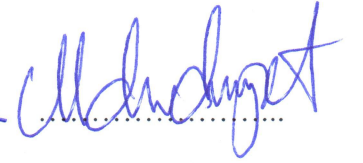
Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum/~~onaylamıyorum~~



Başkan : Prof. Dr. M. Ali AKCAYOL

Bilgisayar Mühendisliği, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum/~~onaylamıyorum~~



Üye : Doç. Dr. Osman ABUL

Bilgisayar Mühendisliği, TOBB ETÜ

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum/~~onaylamıyorum~~



Tez Savunma Tarihi: 11/02/2016

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

.....
Doç. Dr. Nurettin TOPALOĞLU

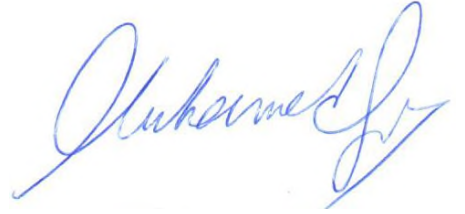
Bilişim Enstitüsü Müdür

ETİK BEYAN

Gazi Üniversitesi Bilişim Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.



Muhammed

ŞEN11/02/2016

SOSYAL MEDYADAN ELDE EDİLEN BÜYÜK VERİ BAZ ALINARAK SOSYAL OLAY ANALİZİ GERÇEKLEME

(Yüksek Lisans Tezi)

Muhammed ŞEN

GAZİ ÜNİVERSİTESİ

BİLİŞİM ENSTİTÜSÜ

Şubat 2016

ÖZET

Sosyal Ağların günlük hayatımızda etkileri gün geçtikçe artmaktadır. Birçok organizasyon, olay artık sosyal ağlar üzerinden planlanıp, gerçekleşmektedir. Twitter ilk kurulduğu 2006 yılından itibaren bu konuda çığır açan bir öneme sahip olmuştur. Son dönemlerde, sosyal ağlar, insanların birbiri ile iletişim ve bilgi edinmesinden çok, özellikle sosyal olayların ateşlendiği, planlandığı ve grupsal hareketlenmelerin yaşandığı ortam haline gelmiştir. Sosyal ağların etkin hale gelmesi ile bir çok akademik çalışma gerçekleştirilmiştir. Doğal dil işleme (NLP), graf teorisi, veri ve metin madenciliği, makine öğrenme teknikleri gibi akademik konular, Sosyal Medya Platformu'ndan anlık veri toplama veya toplanmış verilerden analize kadar Sosyal Medya Sistemi'nin farklı adımlarında uygulanmıştır. Bu yüksek lisans tez çalışmasında, literatür araştırması ile birlikte, toplanmış belli bir döneme ait Türkçe Sosyal Medya verisinden, sosyal olay analizinin gerçekleşmesi, veri etki analizinin değerlendirilmesi amaçlanmıştır. Büyük veride gerçekleştirileceği için yeni geliştirilen metin tabanlı sorgulama yöntemleri denenmiştir.

Bilim Kodu : 902.1.014
Anahtar Kelimeler : Veri madenciliği, sosyal ağlar, etki analizi
Sayfa Adedi : 45
Danışman : Doç. Dr. Hacer KARACAN

PERFORMING SOCIAL EVENTS ANALYSIS BASED ON BIG DATA DERIVED
FROM SOCIAL MEDIA

(M. Sc. Thesis)

Muhammed ŞEN

GAZİ UNIVERSITY

INSTITUTE OF INFORMATICS

February 2016

ABSTRACT

The influence of Social Networks to our daily life has been increasing day by day. Most of organizations or events are planned and carried out via Social Networks. Twitter has a significant place in social networks since the year of 2006 when Twitter was established. Recently, social networks have become a platform that social events are started on and a platform that group movements lives on. After Social Networks became effective, a lot of academic research has been carried out. In these academic studies many methodologies and concepts were used such as data/text mining, machine learning techniques, natural language processing (NLP), graph theory, rule based systems and visualization on Social Media platform. In this master's thesis, social event analysis and data impact analysis from Turkish social media data of determined period are aimed with deep research of literature. Because of using big data, text based query methods which are developed newly are experimented.

Science Code : 902.1.014
Key Words : Data Mining, social networks, influence analysis
Page Number : 45
Supervisor : Assoc. Prof. Dr. Hacer KARACAN

TEŞEKKÜR

Çalışmalarım sırasında değerli zamanını ve yardımlarını esirgemeyen, pozitif yaklaşımları ile çalışma azmimi ayakta tutan saygıdeğer hocam Doç. Dr. Hacer KARACAN'a, bir çok konuda destek olan kadim dostum Tahir YILMAZ'a, hep desteklerini gördüğüm, her işi kolaylaştıran canım aileme, desteği, sevgisi ve ilgisi ile güç bulduğum, zorlukların üstesinden gelmemde en büyük katkı sahibi olan sevgili eşim Elif Şen'e, teşekkürü borç bilir ve teşekkür etmekten onur duyarım.

İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT.....	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	viii
RESİMLERİN LİSTESİ	ix
ŞEKİLLERİN LİSTESİ.....	x
SİMGELER VE KISALTMALAR.....	xi
1. GİRİŞ.....	1
2. TEMEL BİLGİLER	5
2.1. Twitter.....	5
2.2. Twitter API	7
2.3. Analitik Hiyerarşi Süreci.....	8
3. LİTERATÜR ARAŞTIRMASI	17
3.1. Duygu Durum Analizi.....	17
3.2. Özetleme	18
3.3. Olay Tespiti.....	18
3.4. Öneri Sistemleri ve Benzer Grupların Tespiti.....	19
3.5. Etki Ölçümü	19
4. UYGULANAN YÖNTEM	21
4.1. Twitter Veri Toplama.....	21
4.2. Veri Hazırlama	25
4.3. Etki Değeri Hesaplama	34
5. SONUÇ VE DEĞERLENDİRMELER	37
KAYNAKLAR	41
ÖZGEÇMİŞ	45

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 2.1. Önem Skalası	9
Çizelge 2.2. RI Değerleri	12
Çizelge 2.3. Örnek Faktör Karşılaştırma Matrisi	14
Çizelge 2.4. Örnek Faktör Karşılaştırma Matrisi Hesaplama	14
Çizelge 2.5. Örnek Faktör Karşılaştırma Matrisi Ağırlık	14
Çizelge 2.6. Örnek Faktör Ağırlık Değerleri	15
Çizelge 2.7. EF'ye Göre Alternatif Karşılaştırma Matrisi	15
Çizelge 2.8. TY'ye Göre Alternatif Karşılaştırma Matrisi	15
Çizelge 2.9. İR'ye Göre Alternatif Karşılaştırma Matrisi	15
Çizelge 2.10. İM'ye Göre Alternatif Karşılaştırma Matrisi	15
Çizelge 2.11. Ağırlıklandırılmış Alternatif Karşılaştırma Matrisi	16
Çizelge 2.12. Alternatif ve Faktör Ağırlık Matrisleri	16
Çizelge 2.13. Alternatif-Değer Çizelgesi	16
Çizelge 4.1. Örnek Kullanıcı-Grup Özet Tablosu	28
Çizelge 4.2. Tweet CRC32 Tablosu	29
Çizelge 4.3. UID-TUR Tablosu	30
Çizelge 4.4. Tweet-Retweet-Reply Özet Tablosu	33
Çizelge 4.5. AHP Metrik Önem Matrisi	34
Çizelge 4.6. Etki Faktör Ağırlık Değerleri	34

RESİMLERİN LİSTESİ

Resim	Sayfa
Resim 2.1. Resimli Twitter Paylaşımı	6

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 2.1. Twitter Streaming API çalışma modeli ..	8
Şekil 2.2. Tipik AHP Modeli	9
Şekil 4.1. Sistem Mimarisi.....	22
Şekil 4.2. Kullanıcı Bulma Algoritması.....	23
Şekil 4.3. Liste Kodlamaları	23
Şekil 4.4. Kullanıcı Doğrulama	24
Şekil 4.5. Kullanıcı Pasif Etme	24
Şekil 4.6. Kullanıcı Dinleme.....	25
Şekil 4.7. Grup 1 için Toplam Kullanılan Hashtag ve Kullanıcı ilişkisi	26
Şekil 4.8. Grup 2 için Toplam Kullanılan Hashtag ve Kullanıcı ilişkisi	26
Şekil 4.9. Grup 3 için Toplam Kullanılan Hashtag ve Kullanıcı ilişkisi	27
Şekil 4.10. CRC32 Çevirme.....	29
Şekil 4.11. Toplam Retweet-Reply Hesaplama Algoritması	32

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklama
AHP	Analitik Hiyerarşi Süreci-Prosesi
API	Uygulama Programlama Arayüzü (Application Programming Interface)
VPS	Sanal Özel Sunucu (Virtual Private Server)

1. GİRİŞ

Günümüzde sosyal ağların gelişimi ile insanların kendilerini ifade etme yöntemleri, birbirleri ile iletişime geçme şekilleri, haber yapma ve yayma metotları gibi daha bir çok konuda değişim gerçekleşmiştir. Özellikle Blog sitelerin yaygınlaşması ve insanların bu yolla hem kendi günlüklerini tutması hem de diğer insanlar ile paylaşımlara geçmesi, daha hızlı ve takibi kolay olan ve yayılımı daha fazla olan microblog siteleri ile iletişim çağının şuana kadar ki doruğuna ulaşmıştır.

Microblog sitelerini blog sitelerinden ayıran en önemli özellik anlık olaylara odaklanmasıdır. Kullanıcının o an ifade etmek istediğine odaklanır ve bu sebeple paylaşılan mesajda içerik ve dosya boyutu küçüktür. İlk açılan microblog tumblelogs'dur. Microblogta en önemli gelişme 2006 yılında Twitter'ın ortaya çıkması ile olmuştur [1].

Twitter kurulduğu 2006 yılından günümüze birçok sosyal olayın paylaşıldığı, yorum yapıldığı hatta eyleme geçilmesinde başlangıç noktası olduğu bir platform halini almıştır. Örnelemek gerekirse, "Arap Baharı" olarak adlandırılan, Tunus'ta başlayan ve ilerleyen süreçte Orta Doğu ve Kuzey Afrika'yı saran olaylar dizisinde, sosyal medyanın etkisi çok önemli bir yer tutar. Çünkü burada, sosyal medya iletişim ve örgütlenme aracı olarak kullanılmıştır. Bu süreçteki toplumların hareketi bazı kesimlerce "sosyal medya devrimi" olarak bile adlandırılmıştır [2]. Türkiye'de özellikle "Gezi Parkı" olayları sosyal medyanın toplum hareketlerine etkisi konusunda iyi bir örnek teşkil eder. 31 Mayıs 2013 tarihinde gezi parkı ile ilgili 24 saat içinde yaklaşık 2 Milyon tweet atılmıştır. Dönemin ilgili hashtagleri #direngeziparkı 950 bin tweet, #occupygezi 170 bin tweet, #geziparki: 50 bin tweet, içerisinde geçmiştir [3].

Twitter 06.12.2015 verilerine göre kayıtlı kullanıcı sayısı 645 Milyon 750 bin, aktif kullanıcı sayısı 289 Milyon'dur. Hergün yaklaşık 135 bin yeni kullanıcı üye olmaktadır. Twitter'ı aylık olarak ziyaret eden tekil ziyaretçi sayısı 190 Milyon'dur. Günlük atılan tekil tweet sayısı 58 Milyon'dur [4]. İstatisklerin de gösterdiği gibi dünya üzerinde twitter kullanımı ve etkisi çok gelişmiş ve gelişmektedir.

Twitter üzerine bir çok konuda akademik çalışma gerçekleştirilmiştir. Twitter verisi üzerinde NLP, graf teorisi, veri ve metin madenciliği, makine öğrenme teknikleri gibi konularda akademik çalışmalar yapılmıştır. Yapılan çalışmalar konu başlıkları halinde bu tez çalışmanın ilerleyen bölümlerinde mevcuttur. Özellikle kullanıcı etki analizine dair çalışmalar mevcut olup ancak çalışmaların çoğu büyük veri üzerinde gerçekleştirilmemiştir.

Bu tez çalışmasında, ülkemizde de oldukça fazla kullanılan Twitter'dan elde edilen büyük veri üzerinde tweet etki analizi yapılması amaçlanmıştır. Burada problemin tanımı yapıp, gerçek verilere ve limitlemelere dayanan bir çalışma gerçekleştirilmiştir.

Etki analizi çalışması için, Türkiye'de gerçekleşen bir sosyal olay ele alınmıştır. Bu çalışmada, 10 Temmuz – 11 Ağustos 2014 tarihleri arasında atılan tüm Türkçe tweetler toplanmıştır. Tüm Türkçe tweetlerin toplanması .Twitter'ın sağladığı JSON API'si ile yapılmıştır. Tüm tweetleri elde etme işleminde API'nin sınırlamaları mevcuttur. Örneğin, tweet atan kullanıcının toplam takipçi (follower) sayısını alabilirsiniz ancak takipçi listesini canlı olarak almak, Twitter'ın API limitlerinden dolayı mümkün olmamaktadır. Türkiye'de o dönem aktif kullanıcı sayısı yaklaşık 4 Milyon'dur. Günlük atılan toplam tweet (retweet dahil) 12 milyon civarındadır. Bu tez çalışmasına konu olan toplam tweet sayısı 355 milyon 360 bin 510'dur. Tüm Türkçe tweetler toplandıktan sonra, spam tweetler temizlenmiştir. Temizleme işlemi için, birkaç yöntem önerilmiş olup, özellikle aynı tweeti atanların tespit işlemi uygulanmıştır.İlgili sosyal olayın farklı grupları için kullanılan hashtagler belirlenmiştir. Hashtagleri kullanan kullanıcılar, destek grupları olarak kümelendirilmiştir. Twitter API'sinin izin verdiği ölçüde elde edilen verilerden, tweet etki analizi için kullanılabilecek metrikler oluşturulmuştur. Elde edilen metrikler ile etki değeri için bu tez çalışmasında AHP yöntemi ile hesaplama gerçekleştirilmiştir.

Önerilen yeni yöntem ile gerçek ve büyük veride başarılı sonuçlar elde edilmiştir. Canlı büyük verinin elde edilme yöntemi ve elde edilen veride gerekli hesaplamaları yapmak için ElasticSearch gibi yeni teknolojilerin kullanımını, önerilen hesaplama şeklinin diğer literatür çalışmalarına göre gerçek veride uygulanması ve grupsal olarak diğer literatür çalışmalarında olmaması, bu tez çalışmasının katkılarıdır.

Bölümlerin devamı şu şekilde organize edilmiştir.

Bölüm 2, Genel bilgiler bölümüdür. Twitter hakkında genel bilgileri sunar. Çalışmamızda gerçekleştirdiğimiz yaklaşımın Twitter'ın kullanımı ile bağlantısını verir. Ayrıca, Analitik Hiyerarşi Sürecinin tanımlamasını yapar.

Bölüm 3, Literatür taramasıdır. Twitter üzerine gerçekleştirilen akademik çalışmaların geniş bir özetini sunar.

Bölüm 4, çalışmanın ön hazırlığını ve uygulanan yöntemi belirtir. Verinin elde edilmesi, temizlenmesi ve etki değerinin hesaplama yöntemi anlatılmıştır.

Bölüm 5, Sonuç ve gelecek çalışmalar hakkında bilgi verir.

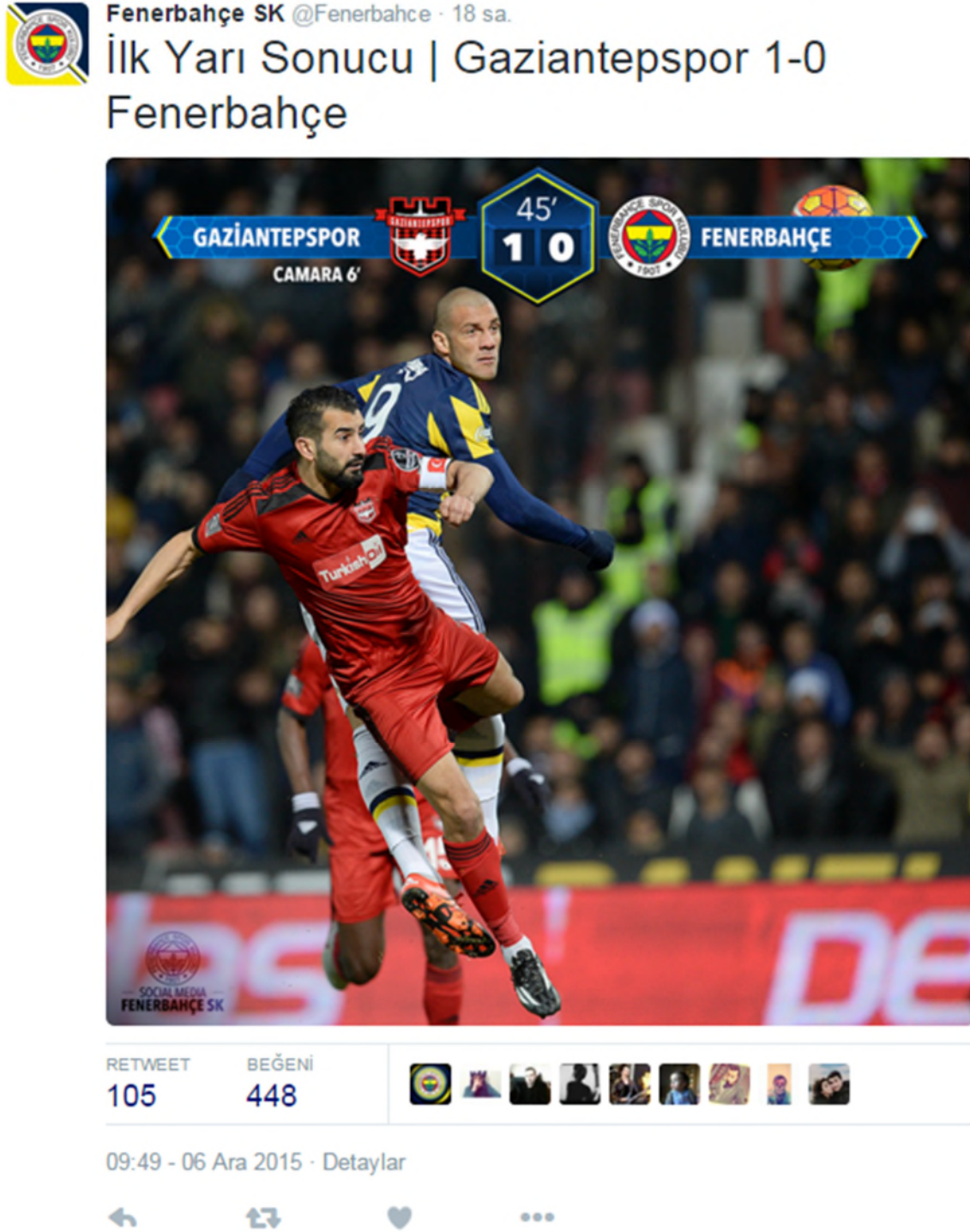
2. TEMEL BİLGİLER

2.1. Twitter

Twitter, Mart 2006’da Jack Dorsey, Evan Williams, Biz Stone and Noah Glass tarafından bulunan ve Temmuz 2006’da yayına giren, online sosyal ağ servisi [5]. Twitter 06.12.2015 verilerine göre kayıtlı kullanıcı sayısı 645 Milyon 750 bin, aktif kullanıcı sayısı 289 Milyon’dur. Hergün yaklaşık 135 bin yeni kullanıcı üye olmaktadır. Twitter’ı aylık olarak ziyaret eden tekil ziyaretçi sayısı 190 Milyon’dur. Günlük atılan tekil tweet sayısı 58 Milyon’dur [4]. Twitter, kişisel bilgilerin paylaşıldığı mikroblog servisi olarak kullanıldığı gibi, birçok marka için pazarlama aracı olarak da kullanılmaktadır. Dünyada şuan özellikle sosyal hareketlenmeler konusunda dikkat çekmektedir.

Twitter, 140 karakterlik mesajlar ile takip ettiklerinize veya hesabınız açık ise online tarama sonucu sizin hesabınızı görenlere, fikir, düşünce ve tepkilerinizi belirttiğiniz bir platform sunar. 140 karakterlik bu mesajlara “tweet” denir. Bu demektir ki ifade etmek istediğiniz tek mesajda 140 karakteri geçemez. Ancak tweet gönderim sayısında herhangi bir sınırlama yoktur. Kullanıcılar genellikle attıkları ilk tweetten sonra ifade etmek istediklerine devam için tweetin başına “+” veya “1,2,3” gibi numaralandırma yolunu seçerler. Ayrıca, tweet içerisinde link, resim veya video paylaşımı yapabilirler [6]. Resim 2.1.’de örnek bir resimli tweet gösterilmiştir.

Twitter’da kullanıcıların, takipçileri (followers) ve takip ettikleri (following-frineds) vardır. Kullanıcının paylaştığı tweetler, takipçilerin “Zaman Akışı” adı verilen anasayfalarında güncelden geçmişe doğru sıralanır [6].



Resim 2.1. Resimli Twitter Paylaşımı [7].

Bir tweet paylaşıldıktan sonra, tweet başkaları tarafından retweet edilebilir. Retweet, atılan orijinal tweeti gören kişilerin, kendi takipçileri (followers) tarafından da görülebilmesi için tekrar göndermesidir. Bongwon'un çalışmasına göre, bilgi yayılımının temelini retweet mekanizması oluşturur [8]. Kullanıcılar, beğendikleri veya hatırlamak isteyecekleri

tweetleri, favorilerine ekleyebilirler. Aynı zamanda kullanıcılar, atılan tweetin altına kendi yorumlarını da ekleyebilirler.

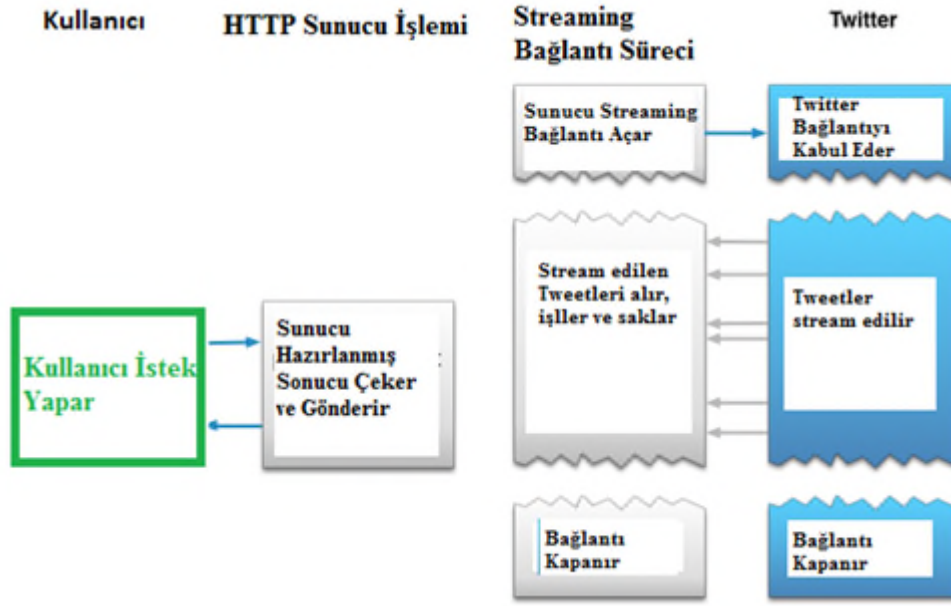
Kullanıcılar attıkları tweetlerde başka kullanıcılardan bahsedebilir veya atıf yapabilirler. Bu kullanım şekline “mention” denir [6]. Tweetin herhangi bir yerinde, bahsedilen kişinin kullanıcı adı “@username” şeklinde görülür.

Twitter, popülerite ve etki oluşturmak için “hashtag” diye adlandırılan bir yapı sunmuştur. Hashtag, belirli bir olayı veya içeriği belirten, bir kelimedenden veya kelime kombinasyonlarından oluşan ve “#” önekini alan belirteçlerdir. Genellikle, tweetin konusunu ve insanlar arasında paylaşımını arttırarak ilgili konuya desteği ve dağıtık tartışmaları organize etmek için kullanılır [6]. Böylelikle bağlantıları olmayan kullanıcılar da aynı konu hakkında görüşlerini belirtebilir.

“TrendingTopic” kavramı da genel olarak hashtaglerden ortaya çıkar. Twitter, kullanıcılar arasında geçen en çok terimi veya hashtagı alır ve belli bir aralık için kullanıcıların en çok bahsettiği 10 konuyu gösterir. Bu konulara “TrendingTopic” denir.

2.2. Twitter API

Twitter API, geliştiricilere Twitter’ın sunucularından veri alımına olanak sağlayan servisedir. İki farklı şekilde hizmet sunmaktadır. REST API ve Streaming API. Twitter REST API istek tabanlı çalışmaktadır [9]. HTTP isteği ile Twitter sunucularına bağlantı kurulur, istek yapılır ve cevap hazır olana kadar beklenir ve cevap JSON formatında alınır [9]. REST API daha çok önceden tanımlı bilgileri dinlemek için kullanılır. Ancak, ihtiyaç canlı dinleme ise, bu durumda REST API kullanılmaz Streaming API kullanılır [9]. Streaming API, tekil sorgulamalar yerine, Twitter sunucularına bağlantı kurar, bağlantı açık kaldığı sürece gelen twitler, API aracılığı ile direk istekte bulunan sunucuya aktarılır. Bu model için filtreleme ve ayırıştırma gereklidir [10]. Şekil 2.1’de sistemin nasıl çalıştığı gösterilmiştir.



Şekil 2.1. Twitter Streaming API çalışma modeli [10].

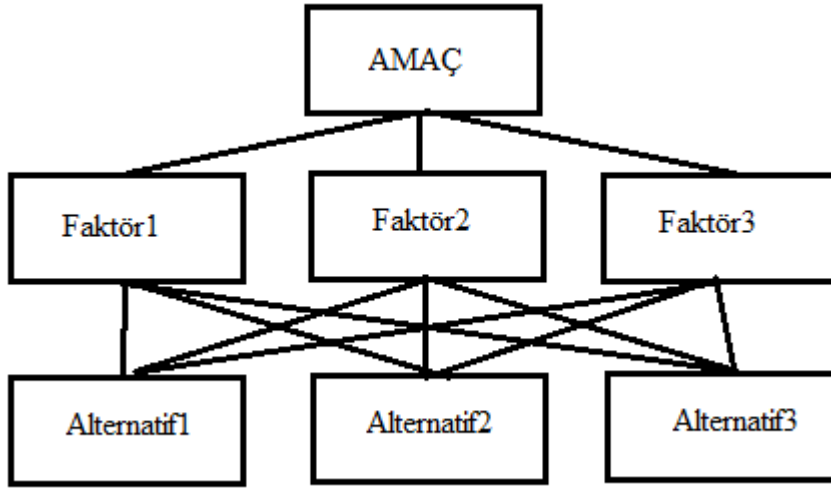
Sistemin belli limitlemeleri mevcuttur. Özellikle canlı dinleme yapıldığında tüm veriyi çekmek imkânsızlaşmaktadır. Sunucu başına 5000 kullanıcının tweet takibine izin vermektedir. Kullanıcıların takipçileri ve takip ettikleri kullanıcıları çekmek için her sorguya 30 kullanıcı adı döner ve sonraki sorgulama için 15 dakika bekleme gerektirir [10]. Bu sebeple, canlı dinlemede kullanıcıların takip ettikleri ve takipçilerin listesi alınamamaktadır.

2.3. Analitik Hiyerarşi Süreci

Analitik Hiyerarşi Süreci, Saaty tarafından geliştirilen, çoklu nitelik modelleme metodolojisidir [11].

AHP, çoklu kriterlerin bulunduğu, karmaşık problemleri çözmek üzere dizayn edilmiştir. AHP, ihtiyaç sahiplerinin özelliklerini belirlediği, nitelik ve sayısal açıdan en iyi kararı vermek için kullanılan, güçlü ve esnek karar verme sürecidir. AHP, sadece karar vermeyi sağlamaz, ayrıca neden en iyi olduğuna dair sayısal sonuçlar ile sebep sunar [11].

Tipik bir AHP modelinde, Şekil 2.1’de gösterildiği gibi Amaç, Faktör ve Alternatifler bulunur.



Şekil 2.2. Tipik AHP Modeli

AHP, hesaplamasında öncelik olarak faktörler arasında birebir karşılaştırma matrisi oluşturulur. Faktörler arası karşılaştırma matrisi, $n \times n$ boyutlu bir kare matristir. Bu matrisin köşegeni üzerindeki matris bileşenleri 1 değerini alır.

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \quad (2.1)$$

Faktörlerin birebir karşılıklı karşılaştırılmasında Çizelge 2.1.'deki önem skalası kullanılır.

Çizelge 2.1. Önem Skalası

Önem Değerleri	Değer Tanımları
1	Her iki faktörün eşit öneme sahip olması durumu
3	1. Faktörün 2. faktörden daha önemli olması durumu
5	1. Faktörün 2. faktörden çok önemli olması durumu
7	1. Faktörün 2. faktöre nazaran çok güçlü bir öneme sahip olması durumu
9	1. Faktörün 2. faktöre nazaran mutlak üstün bir öneme sahip olması durumu
2,4,6,8	Ara değerler

Örneğin birinci faktör üçüncü faktöre göre karşılaştırmayı yapan tarafından daha önemli görünüyorsa, bu durumda karşılaştırma matrisinin birinci satır üçüncü sütun bileşeni ($i = 1, j = 3$), 3 değerini alacaktır. Aksi durumda yani birinci faktörün üçüncü faktörle karşılaştırılmasında, daha önemli tercihi üçüncü faktörden yana kullanılacaksa bu durumda karşılaştırma matrisinin birinci satır üçüncü sütun bileşeni $1/3$ değerini alacaktır. Aynı karşılaştırmada birinci faktörle üçüncü faktörün karşılaştırılmasında faktörler eşit öneme sahip oldukları yönünde tercih kullanılıyorsa bu durumda bileşen 1 değerini alacaktır.

Karşılaştırmalar, karşılaştırma matrisinin tüm değerleri 1 olan köşegeninin üstünde kalan değerler için yapılır. Köşegenin altında kalan bileşenler için ise doğal olarak (2.2) formülünü kullanmak yeterli olacaktır.

$$a_{ji} = \frac{1}{a_{ij}} \quad (2.2)$$

Yukarıda verilen örnek dikkate alınırsa karşılaştırma matrisinin birinci satır üçüncü sütun bileşeni ($i = 1, j = 3$) 3 değerini alıyorsa, karşılaştırma matrisinin üçüncü satır birinci sütun bileşeni ($i=3, j=1$), (2.13) formülünden $1/3$ değerini alacaktır.

Karşılaştırma matrisi, faktörlerin birbirlerine göre önem seviyelerini belirli bir mantık içerisinde gösterir. Ancak bu faktörlerin bütün içerisindeki ağırlıklarını, karşılaştırma matrisini oluşturan sütun vektörlerinden yararlanılır ve n adet ve n bileşenli B sütun vektörü oluşturulur.

$$B_i = \begin{bmatrix} b_{11} \\ b_{21} \\ \cdot \\ \cdot \\ \cdot \\ b_{n1} \end{bmatrix} \quad (2.3)$$

B sütun vektörlerinin hesaplanmasında (2.4) formülü kullanılır.

$$b_{ij} = \frac{a_{ij}}{\sum_{i=1}^n a_{ij}} \quad (2.4)$$

Tüm değerlendirme kriterleri için anlatılan adımlar tekrarlandığında kriter sayısı kadar B sütun vektörü elde edilecektir. n adet B sütun vektörü, bir matris formatında bir araya getirildiğinde ise aşağıda gösterilen C matrisi oluşturulacaktır.

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{bmatrix} \quad (2.5)$$

C matrisinden yararlanarak, kriterlerin birbirlerine göre önem değerlerini gösteren yüzde önem dağılımları elde edilebilir. Bunun için (2.6) formülünde gösterildiği gibi C matrisini oluşturan satır bileşenlerinin aritmetik ortalaması alınır ve Öncelik Vektörü olarak adlandırılan W sütun vektörü elde edilir.

$$w_i = \frac{\sum_{j=1}^n c_{ij}}{n} \quad (2.6)$$

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \end{bmatrix} \quad (2.7)$$

AHP karşılaştırmalardaki tutarlılığın ölçülebilmesi için bir süreç önermektedir. Sonuçta elde edilen Tutarlılık Oranı (CR) ile, bulunan öncelik vektörünün ve dolayısıyla faktörler arasında yapılan birebir karşılaştırmaların tutarlılığın test edilebilmesi imkanını sağlamaktadır. AHP, CR hesaplamasının özünü, faktör sayısı ile Temel Değer adı verilen (λ) bir katsayının karşılaştırılmasına dayandırmaktadır. λ 'nın hesaplanması için öncelikle A karşılaştırma matrisi ile W öncelik vektörünün matris çarpımından D sütun vektörü elde edilir.

$$D = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \times \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \end{bmatrix} \quad (2.8)$$

(2.9) formülünde tanımlandığı gibi, bulunan D sütun vektörü ile W sütun vektörünün karşılıklı elemanlarının bölümünden her bir değerlendirme faktörüne ilişkin temel değer (E)

elde edilir. Bu değerlerin aritmetik ortalaması ((2.10) formülü) ise karşılaştırmaya ilişkin temel değeri (λ) verir.

$$E_i = \frac{d_i}{w_i} \quad (i = 1, 2, \dots, n) \quad (2.9)$$

$$\lambda = \frac{\sum_{i=1}^n E_i}{n} \quad (2.10)$$

λ hesaplandıktan sonra Tutarlılık Göstergesi (CI), (2.11) formülünden yararlanarak hesaplanabilir.

$$CI = \frac{\lambda - n}{n - 1} \quad (2.11)$$

Son aşamada ise CI, *Random Gösterge (RI)* olarak adlandırılan ve Çizelge 2.2’ de gösterilen standart düzeltme değerine bölünerek ((2.12) formülü) CR elde edilir. Çizelge 2.2’ den faktör sayısına karşılık gelen değer seçilir.

Çizelge 2.2. RI Değerleri

N	RI	N	RI
1	0	8	1,41
2	0	9	1,45
3	0,58	10	1,49
4	0,90	11	1,51
5	1,12	12	1,48
6	1,24	13	1,56

$$CR = \frac{CI}{RI} \quad (2.12)$$

Hesaplanan CR değerinin 0.10 dan küçük olması karar vericinin yaptığı karşılaştırmaların tutarlı olduğunu gösterir. CR değerinin 0.10’ dan büyük olması ya AHP’ deki bir hesaplama hatasını ya da karar vericinin karşılaştırmalarındaki tutarsızlığını gösterir.

Daha sonraki aşamada, her bir faktör açısından karar noktalarının yüzde önem dağılımları belirlenir. Diğer bir deyişle birebir karşılaştırmalar ve matris işlemleri faktör sayısı kadar (n kez) tekrarlanır. Ancak bu kez her bir faktör için karar noktalarında kullanılacak G karşılaştırma matrislerinin boyutu mxm olacaktır. Her bir karşılaştırma işleminden sonra

$m \times 1$ boyutlu ve değerlendirilen faktörün karar noktalarına göre yüzde dağılımlarını gösteren S sütun vektörleri elde edilir.

$$S_i = \begin{bmatrix} s_{11} \\ s_{21} \\ \cdot \\ \cdot \\ \cdot \\ s_{m1} \end{bmatrix} \quad (2.13)$$

Bu aşamada öncelikle, yukarıda anlatılan n tane $m \times 1$ boyutlu S sütun vektöründen meydana gelen ve $m \times n$ boyutlu K karar matrisi oluşturulur. Karar matrisi aşağıda tanımlanmıştır:

$$K = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ s_{m1} & s_{m2} & \dots & s_{mn} \end{bmatrix} \quad (2.14)$$

Sonuçta karar matrisi W sütun vektörü (öncelik vektörü) ile aşağıdaki gibi çarpıldığında ise m elemanlı L sütun vektörü elde edilir. L sütun vektörü karar noktalarının yüzde dağılımını verir. Bu dağılım aynı zamanda karar noktalarının önem sırasını da gösterir.

$$L = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ s_{m1} & s_{m2} & \dots & s_{mn} \end{bmatrix} \times \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \end{bmatrix} = \begin{bmatrix} l_{11} \\ l_{21} \\ \cdot \\ \cdot \\ \cdot \\ l_{m1} \end{bmatrix} \quad (2.15)$$

Örnek için;

EF, TY, İR ve İM olmak üzere 4 faktörü olan bir karar sisteminde A,B ve C olan 3 alternatif olduğunu varsayalım. Öncelikle faktörler arasında, karşılaştırma matrisi oluşturalım;

Çizelge 2.3. Örnek Faktör Karşılaştırma Matrisi

FAKTÖRLER	EF	TY	İR	İM
EF	1	1/5	3	4
TY	5	1	9	7
İR	1/3	1/9	1	1
İM	1/4	1/7	1	1

Matris oluşturulduktan sonra, sütun bazlı her sütunun toplamını bulup, her hücre değerine bölelim.

Çizelge 2.4. Örnek Faktör Karşılaştırma Matrisi Hesaplama

FAKTÖRLER	EF	TY	İR	İM
EF	1/(79/12)	1/5/(458/315)	3/(14)	4/(14)
TY	5/(79/12)	1/(458/315)	9/(14)	7/(14)
İR	1/3/(79/12)	1/9/(458/315)	1/(14)	1/(14)
İM	1/4/(79/12)	1/7/(458/315)	1/(14)	1/(14)

Bu işlemten sonra her satırı toplayıp aritmetik ortalaması alıp ağırlık vektörümüzü oluşturalım.

Çizelge 2.5. Örnek Faktör Karşılaştırma Matrisi Ağırlık

FAKTÖRLER	EF	TY	İR	İM	Satır Ort.
EF	0,1519	0,1376	0,2143	0,2029	0,2029
TY	0,7595	0,6878	0,6429	0,5385	0,6571
İR	0,0506	0,0764	0,0714	0,0769	0,0689
İM	0,0380	0,0983	0,0714	0,0769	0,0711

Çizelge 2.6. Örnek Faktör Ağırlık Değerleri

FAKTÖRLER	AĞIRLIK
EF	0,2029
TY	0,6571
İR	0,0689
İM	0,0711

karar verilmesi gereken 3 alternatifi, her faktöre göre karşılaştırma matrisi oluşturulur.

Çizelge 2.7. EF'ye Göre Alternatif Karşılaştırma Matrisi

EF	A	B	C
A	1	3	1
B	1/3	1	1/5
C	1	5	1

Çizelge 2.8. TY'ye Göre Alternatif Karşılaştırma Matrisi

TY	A	B	C
A	1	6	1/3
B	1/6	1	1/9
C	3	9	1

Çizelge 2.9. İR'ye Göre Alternatif Karşılaştırma Matrisi

İR	A	B	C
A	1	1/3	1
B	3	1	5
C	1	1/5	1

Çizelge 2.10. İM'ye Göre Alternatif Karşılaştırma Matrisi

İM	A	B	C
A	1	1/3	1/2
B	3	1	3
C	2	1/3	1

Oluşturulan alternatiflere göre karşılaştırma matrislerinden, yine sütun toplam ve satır ortalama işlemleri yaparak ayrı ayrı matrislerden elde edilen ağırlık matrisi oluşturulur. En son karşılaştırma matrisi elde edilir.

Çizelge 2.11. Ağırlıklandırılmış Alternatif Karşılaştırma Matrisi

	EF	TY	İR	İM
A	0,4055	0,2819	0,1867	0,1594
B	0,115	0,0598	0,6555	0,5889
C	0,4796	0,6583	0,1578	0,2519

Elde edilen karşılaştırma matrisi ile ağırlık matrisi çarpılınca, alternatiflerin aldığı matematiksel değerler elde edilmiş olur.

Çizelge 2.12. Alternatif ve Faktör Ağırlık Matrisleri

	EF	TY	İR	İM
A	0,4055	0,2819	0,1867	0,1594
B	0,115	0,0598	0,6555	0,5889
C	0,4796	0,6583	0,1578	0,2519

X

FAKTÖRLER	AĞIRLIK
EF	0,2029
TY	0,6571
İR	0,0689
İM	0,0711

Çizelge 2.13. Alternatif-Değer Çizelgesi

ALTERNATİFLER	PUAN
A	0,2917
B	0,1497
C	0,5586

Bu sonuç ile seçilmesi gereken en yüksek değeri alan C alternatifi olduğu görülmektedir [12].

3. LİTERATÜR ARAŞTIRMASI

Mikroblog servisleri üzerine birçok akademik çalışma yapılmıştır. Bu çalışmaların bazıları konu başlıklarına ayırarak belirtilmiştir.

3.1. Duygu Durum Analizi

Mikroblog servisleri üzerine en çok çalışma yapılan konulardandır.

Claster, Cooper ve Sallis, 80 milyon Tweetten Tayland'ın turizm sektörü üzerine etkisine dair Duygu Durum Analizi yapmışlardır. Özetlemeler ve kümeleme teknikleri kullanılmıştır. Duygu durum için “binary-choice keyword” algoritması kullanılmıştır. Özdüzenleyici haritalar kullanılarak duygu kutupları belirlenmiş. Önce bir duygu durum kavramı kelime kütüphanesi oluşturulmuş ve oluşturulan bu kütüphane kullanılmıştır [13]. Claster, Dinh Quoc, Shanmuganathan, filmler hakkında atılan tweetler üzerinde duygu durum analizi yapmışlardır. Denetimsiz yapay sinir ağları, çoklu bilgi tabanı yaklaşımlar ve Özdüzenleyici haritalar kullanmışlardır [14]. Claster, Hung ve Dinh Cooper, Meksika turizm sektörü ile ilgili 70 milyon Tweet üzerinde duygu durum analizi yapmışlardır. Naive Bayes ve yapay sinir ağları kullanılmıştır [15]. Celikyilmaz, Hakkani-Tur ve Junlan, makine öğrenme teknikleri kullanarak Twitter üzerinde duygu durum analizi yapmışlardır. Polar ve nötr olacak şekilde iki gruba ayrılmış. Olumlu ya da olumsuz duygu içeren tweetlere polar denilmiş, kalanlara nötr denilmiştir. Gürültülü tweetler metin normalizasyonundan geçirildikten sonra tasnif edilmiştir. Manuel etiketlenmiş tweetler ile sistemin başarısı ölçülmüştür [16]. Hao Ming, Stream twitter verisi üzerinde görsel duygu durum analizi yapmıştır. Başlıklara göre, tweetin ilginçliklerine göre ve lokasyon bazlı görselleştirmeler, yoğunluk analizlerini mümkün kılmıştır [17]. Quercia, Ellis, Capra, ve Crowcroft, etkili kullanıcıların dil kullanımlarını incelemiştir. Sonuçlarda etkili kullanıcıların genelde negatif duygu durumu tercih ettiklerini görmüşlerdir. Bu nedenle bir Twitter kullanıcısının etkisinin sadece ilgili graf ile değil ayrıca kullanıcının kişiliği ve duyguları ile de ölçülmesi gerektiğini belirtilmişlerdir [18]. Gruz, Doiron ve Mai, j. 2010 kış olimpiyatları esnasında atılan tweetleri incelenmişler. Duygu durum analizi yapmışlardır. İnsanlar genel olarak negatif yerine pozitif tweetler attığını ve pozitif tweetlerin negatiflere göre daha çok retweet edildiğini görmüşlerdir. Duygu içeren mesajların kullanıcının ağdaki durumu (centrality) ile ilişkisini tespit etmişlerdir. Negatif tweet atanların, daha üretken ve daha çok tweet attığını belirlemişlerdir [19].

3.2. Özetleme

Inouye ve Kalita, Twitter üzerinde özetleme algoritmalarını karşılaştıran bir çalışma gerçekleştirmişlerdir. Yöntemlerin başarısını manuel yapılmış özetlemeler ile karşılaştırarak yapmışlardır [20]. Zhenhua ve Mehrotra, Sumblr adını verdikleri twitter zaman akışını özetleyen algoritma geliştirmişlerdir. Klasik doküman özetleme algoritmalarından farklı olarak sadece kısa metinlere odaklanmışlardır [21]. Sanghoon, Shakya, Sunderraman, ve Belkasim, sosyal medya platformlarında atılan mesajları konuya göre özetleyerek kullanıcıya sunulması üzerine çalışılmışlardır. Çalışmalarında, Hadoop ve Hbase altyapısı kullanılmışlardır [22].

3.3. Olay Tespiti

Xinyue, Tokarchuk, Cuadrado ve Poslad, Twitter üzerinde olayların ve alt olayların tespitine dair çalışma yapmışlardır. Veri seti olarak 2012 Londra Olimpiyat Oyunları seçmişlerdir [23]. Chung-Hong, Hsin-Chang, Tzan-Feng ve Wei-Shiang, Twitter üzerinden “spatio-temporal” olay tespit sistemi tasarlamışlardır [24]. Chun-Shuo, Meng-Fen, Wen-Chih, ve Chien-Cheng, kullanıcıların ilgililerine uygun olayları bir Olay Grafi şeklinde sunan bir çalışma gerçekleştirmişlerdir. Olayları haber metinlerinden elde edip, o olay ile ilgili tweetleri toplamışlardır [25]. Pohl, Bouchachia ve Hellwagner, Twitter kullanarak kriz yönetimlerinde krizle alakalı alt olayları bulmaya çalışmışlardır. Lokasyon bilgisinden olayın yeri, paylaşılan resim, video veya metinden de ne ile ilgili olduğu anlamaya çalışmışlardır [26]. Ayrıca, Pohl ve arkadaşları, sosyal medyanın acil durum yönetimi için kullanılması anlattıkları bir makale daha yayınlamışlardır. Özellikle alt olayların tespitine çalışmışlardır. Sandy kasırgası esnasındaki alt olayların tespiti makaleye konu olmuştur [27]. Yi-Shiang, Jyun-Yu ve Pu-Jen, Twitter üzerinden bir olayın ne zaman başladığı ve ne zaman bittiğini belirlemeye yönelik bir çalışma gerçekleştirmişlerdir [28]. Giacobe, Hyun-Woo ve Faraz, DARPA’nın 10 tane kırmızı balonu Amerika’nın değişik yerleri üzerinde uçurup bu yerlerin tespiti için yaptığı yarışmada önerdikleri yaklaşım için bir makale yazmışlardır. Yarışmayı MIT üniversitesi kazanmıştır. Kendileri 10. Olmuşlardır. Ancak twitter verisinden yola çıktıkları için önem arz etmektedir [29]. Liangzhe ve Prakash, twitter verisi ile grip salgının belirlenmesine yönelik bir çalışma gerçekleştirmişlerdir [30].

3.4. Öneri Sistemleri ve Benzer Grupların Tespiti

Huajing Yuan, Wang-Chien, Giles ve Meng-Chang Twitter kullanıcısının ilgi alanlarına daha uygun olan sosyal içerik öneri sistemi tasarlamışlardır. Bunun için özel bir indeks manüpilasyon algoritması gerçekleştirmişlerdir. Geri alma performansının ve ilginin arttırılmasını çalışmalarında hedeflemişlerdir [31]. Jonnalagedda ve Gauch, kullanıcıların ilgileri ile açık Twitter trendleri arasındaki girişimi ölçerek kullanıcıya özel haber tavsiye sistemi geliştirmişlerdir. Sonuç olarak tavsiye sistemlerde kişisel ilgilerin açık trend toplara göre daha etkili olduğu anlaşılmıştır. İkisi birleşince daha kaliteli öneri sistemleri gerçekleştirilebileceği belirtmişlerdir [32].

3.5. Etki Ölçümü

Twitter kullanıcı etki analizi yapan ilk çalışmalardan bir tanesini H. Kwak ve arkadaşları gerçekleştirmiştir [33]. Bu çalışmada, 3 etki değerini hesaplamak için, takipçi sayısı, “PageRank” ve retweet sayılarını faktör olarak kullanmışlardır [33]. J. Weng, E.-P. Lim, J. Jiang, ve Q. He, hashtaglerden kullanıcıları konu bazlı etki değerlerini ölçmüşlerdir. Bu çalışmayı yaparken, takipçi sayısının birincil öncelik olmadığını göstermiş ve “PageRank” algoritmasına benzer bir algoritmayı önermişlerdir [34]. M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi ,kullanıcı etki değerini hesaplamak için takipçi sayısı, retweet sayısı ve mention sayısını kullanmışlardır [35].

Rashid Hussain, Hameed ve Hegde, Twitter verilerinin Bulut Hesaplama yöntemi kullanarak analiz edilmesi üzerinde çalışmışlardır. Büyük sosyal network graflarının etki ölçümü için kullanılabileceğini göstermişlerdir. Bunun için, Python’un ücretsiz paketi olan NetworkX kullanmışlardır. Sadece kullanıcı ve retweet bağlantıları üzerinden çalışmalarını gerçekleştirmişlerdir [36].

Subbian ve Melville, Sosyal ağdaki en önemli aktörleri bulmaya çalışmışlardır. İlk bakışta takipçi sayısının önemli olduğunu düşünmüşler ancak retweet ve mentionın daha önemli olduğuna karar vermişlerdir. 11.11.2009 ile 26.11.2009 tarihleri arasında 40 milyon kullanıcıyı kapsayan veri toplamışlardır. Buradan sadece 9625 kullanıcının retweet grafini incelemişlerdir. “Supervised Kemeny Ranking (SKR)” kullanmışlardır [37].

Yong, Zhang Jiaqin ve Tingting, Sina-Weibo’da kullanıcıların etki ölçülmesine dair bir çalışma gerçekleştirmişlerdir. Kullanıcıların takipçi sayıları ve diğer kullanıcılar ile etkileşimleri analiz edilmiştir. Çalışmada takipçi, retweet ve reply bilgileri kullanılmıştır. Bu veriler kullanılarak ağırlıklandırılmış bir network oluşturulmuştur. Daha sonra “Weighted

PageRank” ve “Hypertext Induced Topic Selection” algoritmaları kullanılarak kullanıcının etkisi ölçülmüştür. Kendileri, Spearman’ın etki korelasyon katsayısını kullanarak, etki ölçümü yapmışlardır ve diğer iki algoritma ile karşılaştırmışlardır. Çalışmalarında, 1142 kullanıcı, 645 bin 631 tweet, 129 bin 88 retweet ve 225 bin 213 reply üzerinden değerlendirme yapmışlardır [38].

Yuhong, Li ve Tie, Analitik Hiyerarşi Süreci’ni kullanarak, Çin’in micro-blog platformu olan T.Sina üzerinden tweet etki değeri hesaplamışlardır. Çalışmalarını sadece 93 tweet üzerinde gerçekleştirmişlerdir. Etki değeri ölçüm parametresi olarak; yorum sayılarını, iletme sayılarını, yayımlandığı tarihi, takipçi sayılarını, eşitlenen etiketleri, T.Sina’ya özgü ilan panosu yayınlanması ve kimlik yönetimi olarak belirlemişlerdir [39]. Ancak bu çalışma geniş çapta bir tweet etki değerini hesaplamadan uzak yapılmıştır. Sadece 93 tweet üzerinden yapılması ve farklı grupları göz önünde bulundurulmaması yanıltıcı olabileceği düşünülmektedir.

4. UYGULANAN YÖNTEM

Bu tez çalışmasında, Türkçe Twitter verisinden, sosyal olay analizinin gerçekleştirilmesi, tweet etki analizinin değerlendirilmesi amaçlanmıştır. Amacımız, Twitter’da, farklı gruplara ayrılmış kullanıcıların, gruplar arası etkisi değerlendirilmiştir. Bir tweetin, kendi grubundaki bir kullanıcı tarafından retweet veya reply edilmesinden önce, diğer gruplar tarafından retweet veya reply edilmesinin daha önemli olduğu ve diğer gruplara etki edilip edilmediği, bu öncelik göz önünde bulundurularak her tweet için bir etki değeri hesaplaması gerçekleştirilmiştir.

Çalışma 3 aşamaya ayrılmıştır. Öncelikle verinin toplanması aşaması gerçekleştirilmiştir. Burada sadece Türkçe tweetler toplanmıştır. Bunun için Türkçe’de en çok kullanılan kelimeler belirlenmiş ve bu kelimeleri geçiren kullanıcıların takibi sağlanmıştır. Bölüm 4.1 bu aşamaya detaylı bir açıklama getirmektedir.

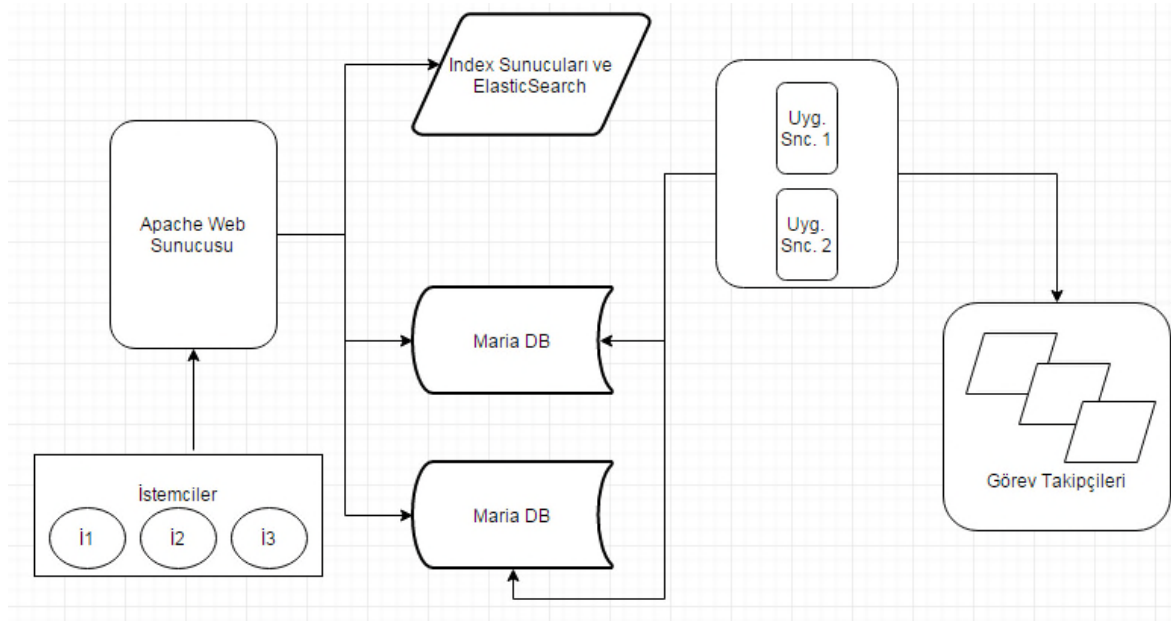
İkinci aşama olarak verinin hazırlanması süreci gerçekleştirilmiştir. Çalışmada sosyal olayı değerlendirmesi yapıldığından dolayı, olayın gerçekleştiği tarihten 1 ay öncesine kadar olan tüm tweetler, çalışmanın hızlanması için ayrı bir sunucuya aktarılmıştır. Burada öncelikle kullanıcıların kullandığı hashtag’lere göre gruplara ayrılmıştır. Spam kullanıcı tespiti gibi çalışmalar yapılarak veri temizlenmesi sağlanmıştır. Bölüm 4.2 bu aşamaya detaylı bir açıklama getirmektedir.

Son olarak, temizlenmiş veri üstünden, AHS (Analitik Hiyerarşi Süreci) yöntemi kullanılarak her tweet için etki değeri hesaplanmıştır. Bölüm 4.3 bu aşamaya detaylı bir açıklama getirmektedir.

4.1. Twitter Veri Toplama

Twitter verisi toplamak için Twitter sağlamış olduğu API kullanıldı. API’nin kullanım limitleri olduğu için ölçeklendirme çok önemli bir kavram haline geldi. Sistem mimarisi Şekil 4.1’de gösterildiği şekilde tasarlandı. Örneğin, Twitter tek API key için sadece 5000 kullanıcı takip izni vermektedir. Bu sebeple, ‘TaskTrackers’ olarak adlandırdığımız VPS’ler ile veri toplama işlemi gerçekleştirilmiştir. İhtiyaç oldukça sadece tasktracker sayısı artırılmıştır. Toplamda yaklaşık 4 Milyon aktif kullanıcı takip edildi ve günlük yaklaşık 12 Milyon tweet toplanması sağlanmıştır. Aynı zamanda, tasktracker sunucuları takip ettikleri kullanıcıların hareketlerini, paylaşımlarını (tweet, retweet vs) eşzamanlı olarak veri tabanına kaydetmekle sorumlulardır. Toplanan ham verilerin işlenebilir hale gelmesi ayrı bir süreç olarak ele alınmıştır. Burada veri tabanından çekilen veriler, index sunucularında işlenerek,

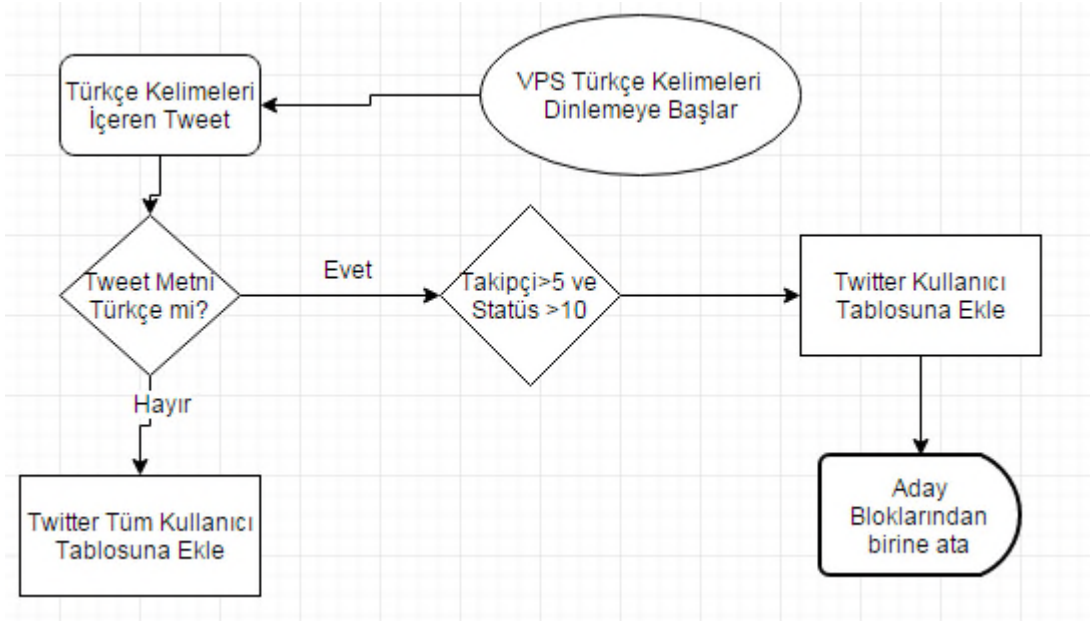
sorgulanabilir formatta tutulmuştur. Bu iş için elasticsearch teknolojisinin gücü kullanılmıştır. Text bazlı sorgulamalarda, No-Sql yaklaşımıyla gayet iyi performanslar sağlayan Elasticsearch bu şekilde ki büyük verilerde çok iyi sonuçlar üretmiştir. Indexlenen verilere ayrıca Kibana web arayüzü tarafından erişim sağlanmıştır.



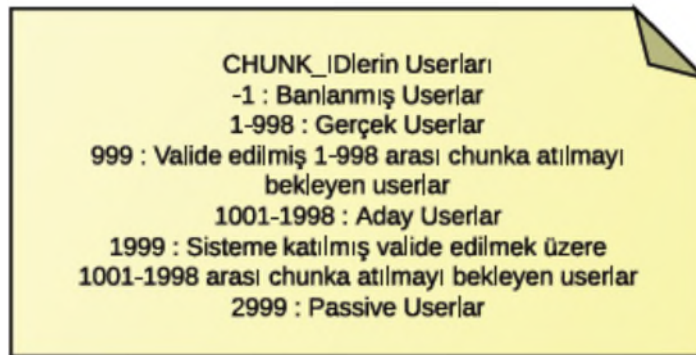
Şekil 4.1. Sistem Mimarisi

Bazı temel işlerin işleyişini detaylandıran aşağıdaki şemalarda, algoritmaların tasarımı görülmektedir. Yukarıda bahsedilen rollerin bu işleyişlerdeki görevleri de anlatılmaktadır.

Twitter Kullanıcı Bulma: Tasktracker'lar Türkçe kelime paylaşımı yapan hesapları bulmaya çalışır. VPS'ler önce oluşturmuş olduğumuz Türkçe kelime listesine göre Türkçe kelimeleri dinlemeye başlar. Tweet metni Türkçe ise, eğer takipçi sayısı 5'den büyük ve 10 taneden fazla tweeti var ise aday kullanıcı listesine konur ve daha sonra doğrulama işlemi gerçekleştirilir. Şekil 4.2'de algoritmanın işleyişi gösterilmiştir. Şekil 4.3'de liste kodlamalarının açıklaması verilmiştir.

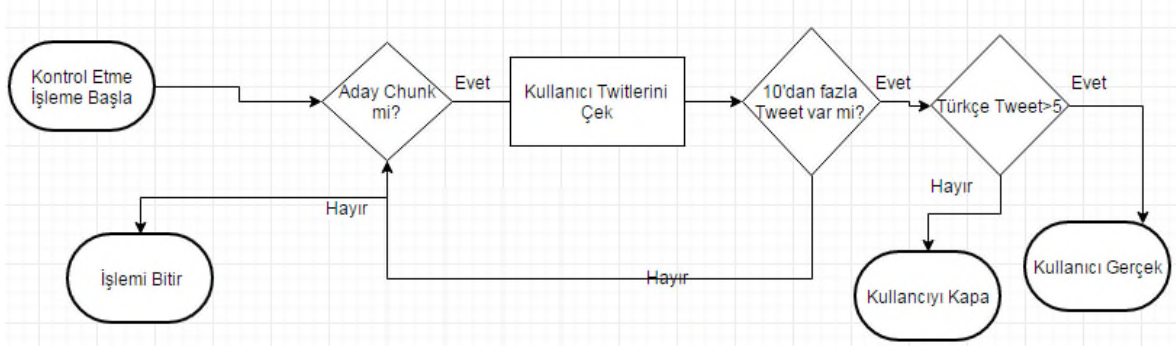


Şekil 4.2. Kullanıcı Bulma Algoritması



Şekil 4.3. Liste Kodlamaları

Twitter Kullanıcı Doğrulama: Tasktracker’lar havuzda bulunan kullanıcıların herbiri için 10 taneden fazla tweeti varsa ve bu tweetlerde kullanılan kelimelerin 5 veya daha fazlası Türkçe ise kullanıcıyı doğrulayarak aktif dinlemeye başlar. Şekil 4.4’de algoritmanın işleyişi gösterilmiştir.



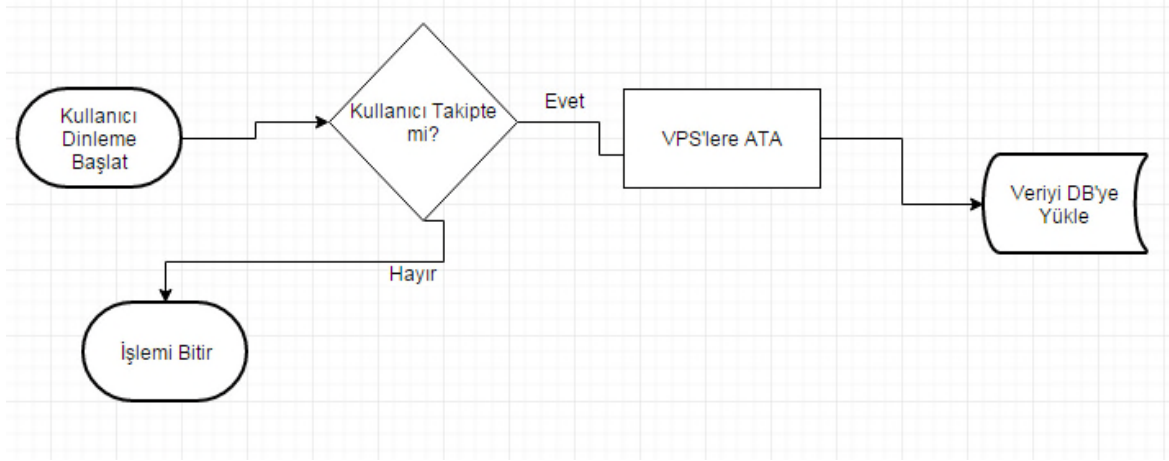
Şekil 4.4. Kullanıcı Doğrulama

Pasif Kullanıcı Keşfi: Tasktracker’lar kendi dinlediği kullanıcı setinde son tweeti 1 ay veya daha eski olan kullanıcıları pasif dinlemeye çeker. Bu çalışma performans açısından gereksiz işleri yapmamak adına gerçekleştirilir. Eğer kullanıcı tekrar tweet atmaya başlarsa, statüsünü tekrar aktif hale getirilir. Şekil 4.5’de algoritmanın işleyişi gösterilmiştir.



Şekil 4.5. Kullanıcı Pasif Etme

Kullanıcı Dinleme: Tasktracker’lar kendi kullanıcı setindeki kullanıcıların tüm paylaşımlarını anlık olarak dinleyerek sisteme kaydederler. Kayıt işlemi Cassandra DB üstünde gerçekleşir. Şekil 4.6’de algoritmanın işleyişi gösterilmiştir.

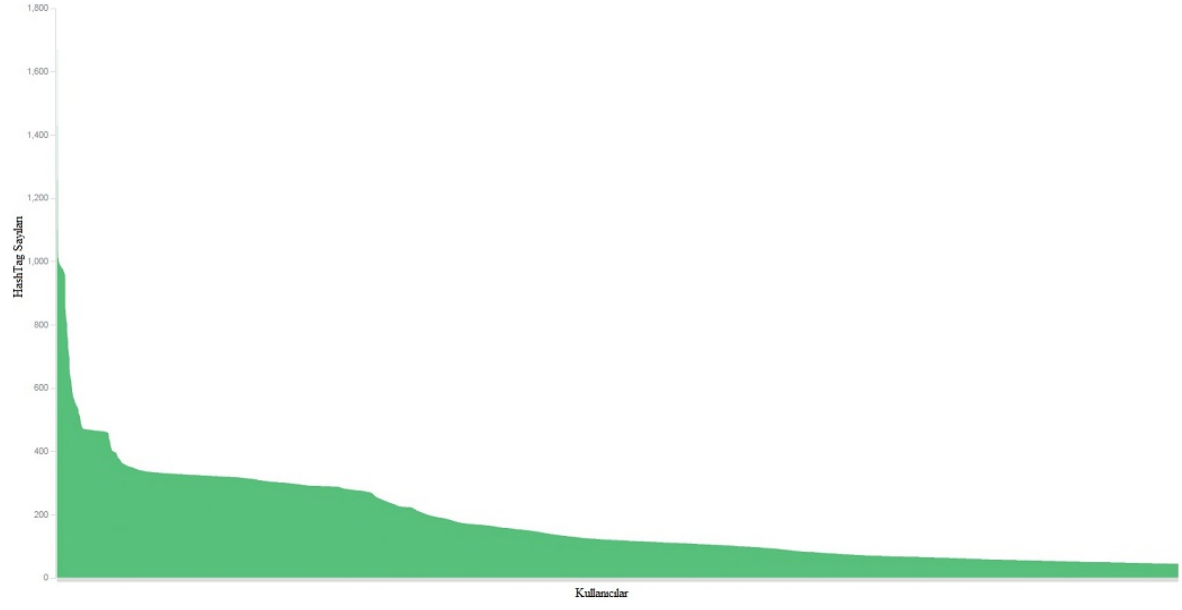


Şekil 4.6. Kullanıcı Dinleme

4.2. Veri Hazırlama

Veri toplama işlemlerinden sonra, Türkiye’de gerçekleşen sosyal olayın, olayın başlama tarihinden 1 ay öncesine kadar atılan tüm Türkçe Tweetler baz alınmıştır. Buna göre, 10 Temmuz 2014 00.00 ile 11 Ağustos 2014 00.00 arasında atılan toplam 355 Milyon 360 bin 510 tweet değerlendirmeye alınmıştır. Sadece bu veri üzerinde çalışma yapılması için tüm bu veriler 192GB RAM, 32Core CPU 2,60GHz işlemcili ayrı bir sunucuya aktarılmıştır. Sunucuda, tüm veriler Maria DB üzerine aktarılmıştır. Daha sonra Elasticsearch’e Maria DB ham data olarak gösterilip indexlenmesi sağlanmıştır. Tüm çalışma boyunca hibrid yapı kullanılmıştır. Maria DB üzerinde öncelikle ham veri bulundurulmuştur. Daha sonra özet tablolar Maria DB üzerinde tutulmuştur. Elasticsearch, verilere hızlı erişim ve özet tablo oluşturulacak verilerin çıkarımı için kullanılmıştır. Yapılan test çalışmalarında, çoklu veride Elasticsearch 20 kata kadar daha hızlı veriye erişim sağladığı görülmüştür. Bu sebeple, Maria DB ilişkisel veriler için kullanılmıştır.

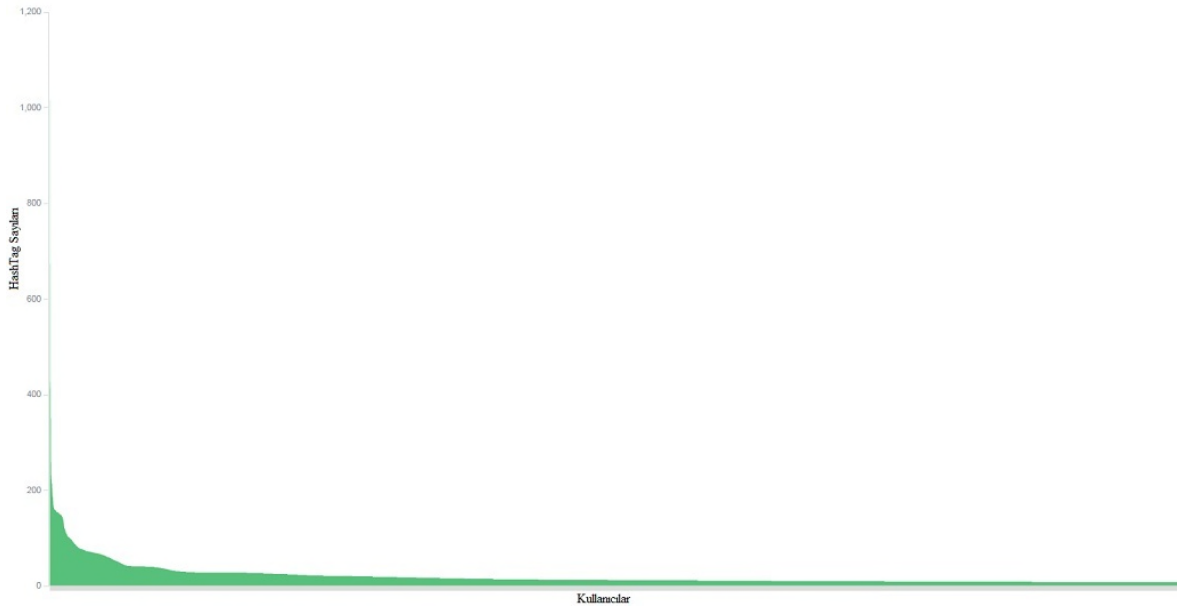
Kullanıcıları gruplara ayırma işlemi, ilgili sosyal olayı desteklemek için kullandıkları hashtagler ile yapılmıştır. 3 Farklı grup için, 3 farklı hashtag grubu oluşturulmuştur. Grup 1 için oluşturulan hashtag grubunda 18 tane hashtag, Grup 2 için oluşturulan hashtag grubunda 5 tane hashtag, Grup 3 için oluşturulan hashtag grubunda 7 tane hashtag bulunmaktadır. Seçilen hashtagler, tüm veri üzerinde kullanılan hashtagler analiz edilerek ilgili gruplara destek hashtagleri ayıklanarak oluşturulmuştur. Oluşturulan hashtag gruplarındaki hashtaglerden en az bir tanesini ilgili tarihte geçiren kullanıcı ilgili gruba dahil edilmiştir. Hashtag gruplarının toplam kullanım sayıları ile kullanıcı ilişkisi, Grup 1 için Şekil 4.7, Grup 2 için Şekil 4.8, Grup 3 için Şekil 4.9’de gösterilmiştir.



Şekil 4.7. Grup 1 için Toplam Kullanılan Hashtag ve Kullanıcı ilişkisi



Şekil 4.8. Grup 2 için Toplam Kullanılan Hashtag ve Kullanıcı ilişkisi



Şekil 4.9. Grup 3 için Toplam Kullanılan Hashtag ve Kullanıcı ilişkisi

Grafiklerden de görülebileceği gibi , hashtaglerin toplam kullanım sayıları birden düşüyor. Bunun sebebi Twitter ortamında birçok spam Tweet hesabının olmasıdır. Tezin ilerleyen bölümlerinde spam tweet kullanıcı temizleme işlemleri detaylı bir şekilde ele alınmıştır.

Oluşturulan gruplara göre kullanıcıların özet bilgileri, Maria DB üzerinde özet tabloya kaydedilmiştir. Bu tabloda, kullanıcı ID'si, hastagleri kaç defa kullandığı, o dönemde toplam kaç tane tweet attığı ve kullandığı hashtaglerin hangi gruba ait olduğu bilgisi tutulmuştur. Çizelge 4.2.1 örnek olarak göstermektedir. Bu tabloda UID, kullanıcının sistemdeki ID numarasını, CNT ilgili kullanıcının, ilgili hashtag grubuna ait [16] toplam hashtag kullanım sayısını, TUR değerlerin hangi gruba ait olduğunu ve son olarak TOTALTWEETCNT kullanıcının belirlenen tarih aralığında attığı toplam Tweet sayısını gösterir.

Bu tabloda, toplam 238 bin 720 hashtag kullanan kullanıcı kaydı bulunmaktadır. Toplam 185 bin 612 tekil kullanıcı bulunmaktadır. Tekil kullanıcı ile hashtag gruplarını kullanan kullanıcı sayısının farklı olması, farklı hashtag gruplarının, ilgili süre içerisinde aynı kullanıcı tarafından kullanıldığını göstermektedir. Tezin ilerleyen bölümlerinde, bu özet tabloda, veri temizleme ve düzeltme çalışmaları ile nasıl tekilleştirildiği anlatılmıştır.

Grup 1 için toplam hashtag kullanımı 2 milyon 660 bin 671, Grup 2 için 281 bin 972, Grup 3 için 316 bin 344 olarak belirlenmiştir.

Çizelge 4.1. Örnek Kullanıcı-Grup Özet Tablosu

UID	CNT	TUR	TOTALTWEETCNT
2345300976	1758	1	20586
2539302773	1510	1	14769
968431578	1436	1	14791
96541644	1369	1	65800

Oluşturulan özet tabloda ikinci aşama olarak kullanıcıları sadece bir gruba dahil etme işlemi gerçekleştirilmiştir. Eğer bir kullanıcı farklı iki grubun hashtag gruplarını aynı sayıda kullanmış ise bu kullanıcı tablodan çıkarılarak değerlendirmeye alınmamıştır. Bu şekilde 2853 kullanıcı değerlendirme dışına alınmıştır. Tabloda, birden çok hashtag grubunu kullanan kullanıcılar, en çok kullandığı hashtag grubunun üyesi sayılmıştır.

Spam Twitter hesaplarını belirlemek için algoritma gerçekleştirilmiştir. Gerçekleştirilen algoritmada, özet tablomuzdaki her kullanıcı için, ayrı ayrı olarak ilk 10 ve son 10 twittin CRC32 değerleri hesaplanmıştır.



Şekil 4.10 CRC32 Çevirme

CRC32'ye çevrildikten sonra Maria DB'de Çizelge 4.2'deki gibi tablo ilk 10 tweet için ayrı son 10 tweet için ayrı tablo oluşturuldu.

Çizelge 4.2. Tweet CRC32 Tablosu

UID	SHASH
2345300976	3824932648
2539302773	30056733
968431578	2888926727
96541644	264689236
2215696116	430329092

Oluşturulan tablolar incelendiğinde, son 10 tweet için oluşturulan CRC32 tablosunda, 1907 farklı CRC32 değerinin toplamda 13415 tekil kullanıcı tarafından oluşturulduğu, ilk 10 tweet için 2020 farklı CRC32 değerinin toplamda 19962 tekil kullanıcı tarafından oluşturulduğu görülmüştür. İki tablodaki ortak kullanıcılar değerlendirildikten sonra, oluşturmuş olduğumuz özet tablodan toplamda 30573 kullanıcı çıkarılmıştır. Son olarak spam kullanıcıları da çıkarıldıktan sonra oluşturmuş olduğumuz özet tablodan, sadece kullanıcı ID ve o kullanıcının bağlı olduğu grubu gösteren yeni bir tablo türetilmiştir. Tablonun içeriği Çizelge 4.3’de gösterilmiştir.

Çizelge 4.3. UID-TUR Tablosu

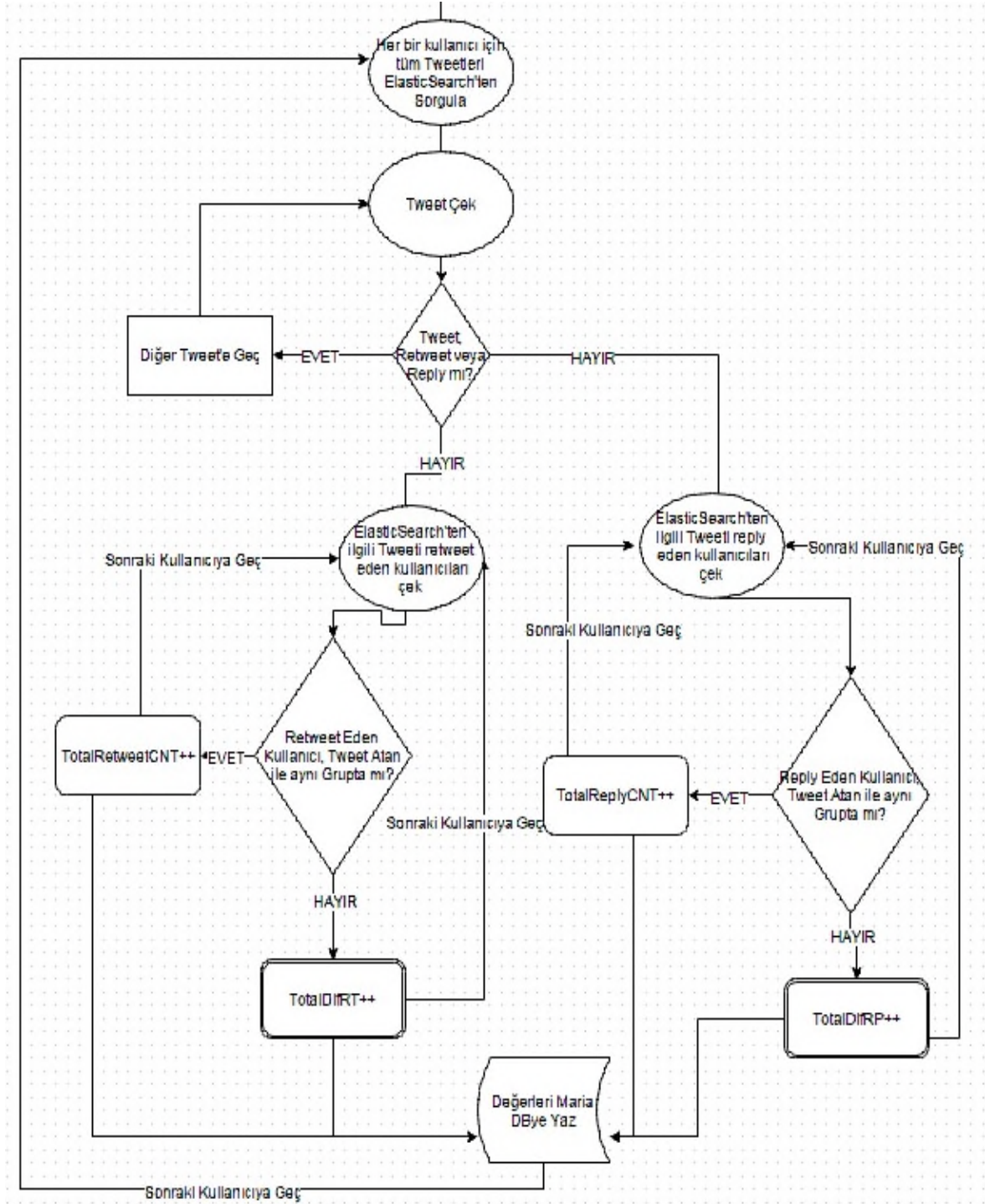
UID	TUR
2853	2
53973	1
611093	2
774519	2
1016321	2

Oluşturulan tabloda toplamda 152 bin 186 kullanıcı bulunmaktadır. Grup 1 için kullanıcı sayısı 72 bin 610, Grup 2 için kullanıcı sayısı 43 bin 875, Grup 3 için kullanıcı sayısı 35 bin 701 olarak bulunmuştur. Bu kullanıcılar, belirlenen tarih aralığında toplam 78 milyon 926 bin 232 tweet atmıştır.

Bir sonraki aşama olarak, türettiğimiz yeni tablodaki kullanıcıların, ilgili dönemde attığı her bir tweet için kullanıcının kendi grubu tarafından yapılan toplam retweet sayısı, toplam reply sayısı, kullanıcının kendi grubu dışındaki kullanıcılar tarafından yapılan toplam retweet sayısı, toplam reply sayısı, ilgili tweet atıldığında kullanıcının toplam takipçi sayısı ve toplam arkadaş sayısını gösteren yeni bir özet tablo Maria DB’de oluşturulmuştur.

Programda öncelikle Maria DB’de oluşturulmuş olan gruplarına göre kullanıcılar, bir array’e atılmıştır. Böylelikle, tweetlerin hangi grup kullanıcılar tarafından retweet veya reply edildiği array’den kontrol edilip tekrar veritabanı sorgulama yapılmayarak daha hızlı çalışması amaçlanmıştır. Kullanıcılar Array’e alınırken, kullanıcı ID’li array indexi bize direk o kullanıcının grubunu verecek şekilde eşitleme yapılarak, Array içinde tekrar bir

arama algoritması çalıştırmanın önüne geçilmiştir. `Array[UID]=UID_TUR` şeklinde eşitleme yapılarak bu işlem sağlanmıştır. Daha sonra, her kullanıcı ID için ElasticSearch'ten kullanıcıların tweetleri çekilmiştir. Her bir tweet tek tek değerlendirilerek, o tweet kaç defa aynı gruptan kullanıcılar tarafından retweet veya reply ve kaç defa farklı gruptaki kullanıcılar tarafından retweet veya reply edildiği hesaplanmıştır. Retweet veya reply edip ancak bizim tablomuzdaki hiçbir gruba dahil olmayan kullanıcılar, tweeti atan kullanıcı ile aynı grupta olduğu varsayılmıştır. Toplam retweet ve reply sayıları bu varsayım üzerine hesaplanmıştır. Elde edilen sonuçlar, her tweet için yeni bir satır olacak şekilde Maria DB'ye işlenmiştir. Şekil 4.2'de algoritmanın işleyişi gösterilmiştir.



Şekil 4.11. Toplam Retweet-Reply Hesaplama Algoritması

Bu algoritma, 16 ayrı process ile çalışan program yazıldı. Tüm tablo, 12 saatte 192GB RAM, 32Core CPU 2,60GHz işlemcili makine tarafından oluşturuldu. Tabloya sadece retweet ve reply'ı olan tweetler yazıldı. Oluşan tabloda toplam tekil kayıt sayısı 3 milyon 502 bin 771'dir.

Çizelge 4.4. Tweet-Retweet-Reply Özet Tablosu

Tweet ID	UID	Follower	Retweet	Reply	difCntRT	difCntRP
496355731428614144	68034431	4453618	10981	699	60	50
496991839787155456	296104214	24445	8359	13	2375	2
496054460331098113	1022502697	19208	9227	0	1066	0
496027652009246720	1022502697	16700	9081	0	1077	0
496338023861481472	296104214	24470	7539	25	2388	3
490495295370051584	296104214	24097	8007	42	1650	5
497653429192646656	296104214	24464	6956	12	2376	3
488639266118172672	224609329	317494	6778	507	516	93
495209157310816257	68034431	4431271	6932	670	251	62
493626278562775040	68034431	4407880	7045	670	97	41

Burada Tweet ID, tweetin ham verideki ID'sidir. UID, tweeti atan kullanıcının ham verideki ID'sidir. Follower, tweet atıldığında, tweeti atan kullanıcının sahip olduğu takipçi sayısıdır. Retweet, tweetin, tweeti atan kullanıcı ile aynı gruptaki kullanıcılar tarafından kaç defa retweet edildiğini gösterir. Reply, tweetin, tweeti atan kullanıcı ile aynı gruptaki kullanıcılar tarafından kaç defa reply edildiğini gösterir. DifCntRT, tweetin, tweeti atan kullanıcıdan farklı gruptaki kullanıcılar tarafından kaç defa retweet edildiğini gösterir. DifCntRP, tweetin, tweeti atan kullanıcıdan farklı gruptaki kullanıcılar tarafından kaç defa reply edildiğini gösterir.

Tweet-Retweet-Reply özet tablosu ile birlikte, tweetlerin etki değerleri hesaplamak için gerekli metrikler elde edilmiş olmaktadır.

4.3. Etki Değeri Hesaplama

Her bir tweetin etki değerini hesaplamak için gerekli bütün metrikler önceki bölümde açıklandığı gibi elde edilmiştir. Bu çalışmamızda, tweetin, tweeti atan kullanıcının bulunduğu gruptan farklı kullanıcılar tarafından retweet veya reply edilmesinin diğer metriklerden daha değerli olduğu düşüncesi işlenmiştir. Bu sebeple, etki değeri belirlemek için AHP'nin tek boyutlu yöntemi hesaplama metodu olarak seçilmiştir. Burada, etki değerini hesaplama için kullanılacak metrikler önceki bölümde anlatılan tweet-retweet-reply özet tablosundaki, Follower, Retweet, Reply, difCntRT, difCntRP metrikleridir.

AHP yöntemi için önce metriklerin önem skalası belirlenir. Bu tez çalışmasında önem skalasına ait matris Çizelge 4.3.1'de gösterilmiştir.

Çizelge 4.5. AHP Metrik Önem Matrisi

	Follower	Retweet	Reply	difRetweet	difReply
Follower	1	1/5	1/3	1/9	1/7
Retweet	5	1	5/3	5/9	5/7
Reply	3	3/5	1	1/3	3/7
difCntRT	9	9/5	3	1	9/7
difCntRP	7	7/5	7/3	7/9	1

Çizelge 4.1'de belirtilen matristen, etki faktör ağırlık değerleri Çizelge 4.3.2'deki gibi elde edilmiştir.

Çizelge 4.6. Etki Faktör Ağırlık Değerleri

Follower	0,04	EF
Retweet	0,2	ER
Reply	0,12	EP
difCntRT	0,36	EDR
difCntRP	0,28	EDP

Bir tweetin etki değeri hesaplamak için 4.1'deki eşitlik uygulandı.

$$TR_i = \left(\frac{F_i}{\Sigma F_j} \times EF \right) + \left(\frac{RT_i}{\Sigma RT_j} \times ER \right) + \left(\frac{RP_i}{\Sigma RP_j} \times EP \right) + \left(\frac{DRT_i}{\Sigma DRT_j} \times EDR \right) + \left(\frac{DRP_i}{\Sigma DRP_j} \times EDP \right) \quad (4.1)$$

TR_i , Tweetin etki değerini gösterir. F_i , ilgili tweetin toplam takipçi sayısı, ΣF_j karşılaştırma yapılan diğer tweetlerin takipçi sayılarının toplamı, EF , AHP'den elde edilen Follower için etki faktör değeridir. RT_i , tweeti atan kullanıcı ile aynı grupta olan kullanıcıların toplam yaptığı retweet sayısı, ΣRT_j , karşılaştırma yapılan diğer tweetlerin aynı gruptaki kullanıcılar tarafından yapılan toplam retweet sayısı, ER , AHP'den elde edilen Retweet için etki faktör değeridir. RP_i , tweeti atan kullanıcı ile aynı grupta olan kullanıcıların toplam yaptığı reply sayısı, ΣRP_j , karşılaştırma yapılan diğer tweetlerin aynı gruptaki kullanıcılar tarafından yapılan toplam reply sayısı, EP , AHP'den elde edilen Reply için etki faktör değeridir. DRT_i , tweeti atan kullanıcıdan farklı grupta olan kullanıcıların toplam yaptığı retweet sayısı, ΣDRT_j , karşılaştırma yapılan diğer tweetlerin farklı gruptaki kullanıcılar tarafından yapılan toplam retweet sayısı, EDR , AHP'den elde edilen difCntRT için etki faktör değeridir. DRP_i , tweeti atan kullanıcıdan farklı grupta olan kullanıcıların toplam yaptığı reply sayısı, ΣDRP_j , karşılaştırma yapılan diğer tweetlerin farklı gruptaki kullanıcılar tarafından yapılan toplam reply sayısı, EDP , AHP'den elde edilen difCntRP için etki faktör değeridir.

5. SONUÇ VE DEĞERLENDİRMELER

Bu tez çalışmasında, ülkemizde gerçekleşen sosyal bir olayın twitterda farklı gruplara yönelik etki analizi yapılmıştır. Bu amaçla, büyük veri seti üzerinde, değişik veri temizleme ve düzeltme işlemleri uygulanmış ve sonrasında Analitik Hiyerarşi Süreci yöntemi uygulanarak etki değeri hesaplaması gerçekleştirilmiştir.

Twitter üzerine yapılan birçok çalışma genel olarak örneklem verileri üzerinden veya belli kriterlerle küçültülmüş veri seti üzerinden yapıldığı görülmüştür. Çalışmanın doğruluğu ve yapılacak etki analizinin iyi sonuç getirmesi için gerçek twitter verisi üzerinde bu çalışma gerçekleştirilmiştir. Bunun için, Türkiye’de gerçekleşen bir sosyal olay temel alınmıştır. Bu sosyal olayda insanların gruplara ayrıldığı görülmüştür. Etkili bir tweetin, belli fikirlerde zaten bir araya gelmiş grup içinde yayılmasından öte, farklı görüşteki gruplara nüfus etmesinin gerçek gösterge olacağı düşüncesi ile bu çalışma gerçekleştirilmiştir. Bu olayın gerçekleştiği ve çalışmalarının yapıldığı 1 aylık döneme ait tüm tweetler toplanmıştır. Toplamda toplam 355 Milyon 360 bin 510 tweet değerlendirmeye alınmıştır. Büyük verilerde çalışmanın zorluğu, bu tez çalışmasında ortaya konmuştur. Öncelikle büyük veride, özellikle internet verilerinde, spam veya gereksiz diyebileceğimiz çok veri bulunmaktadır. Bu verilerin temizleme ve düzeltme işlemi başlı başına bir çalışma konusudur. Bölüm 4’te detaylıca anlatıldığı gibi gerekli temizleme ve düzeltme işlemleri gerçekleştirilmiştir. Ayrıca, bu tez çalışmasında veriye hızlı erişim ve hesaplamalar için yeni teknoloji olan ElasticSearch kullanılmıştır. Daha hızlı çözüm için ElasticSearch’ün yanında hibrid şekilde Maria DB kullanılmıştır.

Çalışmada, tüm tweetlerden, kullanılan hashtaglere göre twitter kullanıcıları üç gruba ayrılmıştır. Bu şekilde toplamda, 152 bin 186 Twitter kullanıcısı değerlendirmeye alınmıştır. Grup 1 için kullanıcı sayısı 72 bin 610, Grup 2 için kullanıcı sayısı 43 bin 875, Grup 3 için kullanıcı sayısı 35 bin 701 olarak bulunmuştur.

Veriler, temizlenip düzenlendikten sonra özet çalışmamızda etki değerini oluşturacak metriklerin hesaplanması işlemi gerçekleştirilmiştir. Bunun için atılan bir tweetin atıldığı andaki takipçi sayısı, aynı grup içerisinde yapılan toplam retweet sayısı, aynı grup içerisinde yapılan toplam reply, farklı gruplar tarafından yapılan toplam retweet sayısı ve farklı gruplar tarafından yapılan toplam reply sayıları hesaplanmıştır.

Tweet etki değeri hesaplamak için Analitik Hiyerarşi Süreci kullanılmıştır. Bu süreçte bir tweet diğer tweetlere göre değerlendirilmiştir. 3 milyon 502 bin 771 tweet için etki değerleri hesaplanmıştır.

H. Kwak ve arkadaşlarının ilk olarak gerçekleştirdiği kullanıcı etki analizinde takipçi sayıları, page-rank ve retweet sayılarını kullanmışlardır [33]. Bu çalışmada, bunlara ek olarak reply sayıları etki değerine faktör olarak eklenmiştir, ayrıca, J. Weng, E.-P. Lim, J. Jiang, ve Q. He yaptığı çalışmada gösterdiği gibi, takipçi sayısının en önemli faktör olmadığını gerçeği kullanılmıştır [34]. Weng ve arkadaşlarının yaptığı, hashtag bazlı konu ayırımından farklı olarak, tez çalışmasında aynı konuda ancak farklı düşüncelere sahip kullanıcılar göz önünde bulundurulmuştur. Bu yaklaşım ile gerçek etki değerinin, farklı düşünce gruplarına erişim ile olabileceği belirlenmiştir.

Subbian ve Melville'in yaptığı çalışmada 40 milyon kullanıcıyı kapsadıkları değerlendirilse, network grafını 9625 kullanıcı üzerinden oluşturmuşlardır [37]. Yine aynı şekilde Yuhong, Li ve Tie, Analitik Hiyerarşi Süreci'ni kullanarak, Çin'in micro-blog platformu olan T.Sina üzerinden tweet etki değerini sadece 93 tweet üzerinden gerçekleştirmişlerdir [39]. Ancak bu çalışmalar geniş çapta bir tweet etki değerini hesaplamadan uzak yapılmıştır. Gerçek veriler üzerinden bir yöntem denenmemiştir. Bu tez çalışmasında, özellikle büyük veri üzerinde, geleneksel veri tabanı yöntemi ve yeni teknoloji olan yapısal olmayan veri sorgulama yöntemi ile verilere hızlı erişim sağlanarak, AHP yöntemi gerçek veriler üzerinde denenmiştir. Sonuçlar gerçeğe uygun bir şekilde elde edilmiştir.

Sonuçların oldukça başarılı olduğu görüldü. 494438997163515905 tweet id ile 441956711 id'li kullanıcının attığı tweet 0,05164 etki değeri ile birinci sırada yer almıştır. Tweeti atan kullanıcı Grup 2 içerisinde olduğu görülmüştür. İlgili döneme ait sosyal medyada paylaşılan videolar değerlendirildiğinde en çok paylaşılan video ve resim, çalışmamızda ortaya çıkan etki değeri en yüksek tweet ile örtüştüğü görülmüştür. İçeriği ile örtüşen resim yaklaşık 183 bin 100 ve video ise 4454 defa paylaşılarak dönemin en çok paylaşımları olmuştur. Yine çalışmamızda gruplara göre etki değerlendirmesinin toplam dağılımına bakıldığında, ilgili dönemde gerçekleşen sosyal olay ile örtüştüğü görülmüştür. Gruplara ait tüm tweetlerin etki değerleri toplandığında ve yüzde dağılımı yapıldığında Grup1 %41, Grup2 %35, Grup3 %24 olarak belirlenmiştir.

Çalışmada ayrıca herhangi veri temizleme işlemi yapılmadan tüm tweetler üzerinde etki değeri hesaplaması yapılmıştır. Spammerların da dahil olduğu hesaplamada en etkili

495886100343169024 id numaralı tweet 0,06213 etki değeri almıştır. Veri temizleme işlemi sonrası en etkili tweet olan 494438997163515905 id'li tweet 0,04502 etki değeri olarak ikinci sırada yer almıştır. Gruplara ait tüm tweetlerin etki değerleri toplandığında ve yüzde dağılımı yapıldığında Grup1 %45, Grup2 %33, Grup3 %22 olarak belirlenmiştir. Etki değeri yüksek olanın değişimi, spam kullanıcıların otomatik bir şekilde retweet etmesine bağlanmaktadır.

Sosyal medya, günümüzde oldukça gelişen, insanların davranışlarını sergiledikleri, birbirlerini etkilemeye çalıştıkları büyük bir platform haline gelmiştir. Bu tarz büyük verilerde çalışma yapmak akademik camianın ilgisine girmiş ve buralardan birçok yeni veri elde edilmiş ve edilmektedir. Etki analizleri, bu konuda önemli bir yer tutmaktadır. Çalışmamız, bugüne kadar yapılan etki analizi çalışmalarına farklı bir açıdan katkı sağlaması yönünden önem arz etmektedir.

Yapılan çalışmada tweetin etki değeri hesaplanırken, , farklı konu grupları için etkili kullanıcıların tespiti önemli bir çalışma konusu olarak karşımıza çıkacaktır.

KAYNAKLAR

1. İnternet: Microblog (2015) Web: <https://tr.wikipedia.org/wiki/Mikroblog> 06 Aralık 2015'de alınmıştır.
2. İnternet: KORKMAZ, A. (Mayıs, 2013). Arap Baharı Sürecinde İnternet ve Sosyal Medyanın Rolü 2013. Web: www.inlcs.org/online/Book14.pdf adresinden 14 Eylül 2015'de alınmıştır.
3. İnternet: Gezi Parkı Verileri (Haziran, 2013). Web: <http://www.ntv.com.tr/turkiye/gezi-parki-icin-24-saatte-2-milyon-tweet,EgVS7DclOUWNf0NZks6zUA> 18 Eylül 2015'de alınmıştır.
4. İnternet: Twitter Statistics (2015, December). Web: <http://www.statisticbrain.com/twitter-statistics/> 06 Aralık 2015'de alınmıştır.
5. İnternet: Twitter (2015). Web: <https://en.wikipedia.org/wiki/Twitter> 06 Aralık 2015'de alınmıştır.
6. İnternet: Twitter Help (2015). Web: <https://support.twitter.com/articles/20169578> 18 Eylül 2015'de alınmıştır.
7. İnternet: Örnek Twitter (Aralık, 2015). Web: <https://twitter.com/Fenerbahce> 06 Aralık 2015'de alınmıştır.
8. Suh, B., Lichan, H., Pirolli, P. and Chi, EH. (20-22 Aug. 2010). *Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network*. Paper presented at Social Computing (SocialCom), 2010 IEEE Second International Conference on, Minneapolis, MN.
9. İnternet: Twitter REST API (2015). Web: <https://dev.twitter.com/rest/public> 06 Aralık 2015'de alınmıştır.
10. İnternet: Twitter Streaming API (2015). Web: <https://dev.twitter.com/streaming/overview> 06 Aralık 2015'de alınmıştır.
11. Saaty, T.L., (1980). *The Analytic Hierarchy Process*. New York: McGrawHill.
12. İnternet: Analitik Hiyerarşi Süreci (Temmuz, 2008). Web: <http://www.hasanbaltalar.com/index.php?id=43> 06 Aralık 2015'de alınmıştır.
13. Claster, WB., Cooper, M., Sallis, P. (28-30 Sept. 2010). *Thailand -- Tourism and Conflict: Modeling Sentiment from Twitter Tweets Using Naïve Bayes and Unsupervised Artificial Neural Nets*. Paper presented at Computational Intelligence, Modelling and Simulation (CIMSIM), 2010 Second International Conference on, Bali.
14. Claster, WB., Dinh Quoc, H., Shanmuganathan, S. (28-30 July 2010). *Unsupervised Artificial Neural Nets for Modeling Movie Sentiment*. Paper presented at Computational Intelligence, Communication Systems and Networks (CICSyN), 2010 Second International Conference on, Liverpool.

15. Claster, WB., Hung, D., Cooper, M. (15-17 Dec. 2010). *Bayes and unsupervised artificial neural nets for Cancun tourism social media data analysis*. Paper presented at Nature and Biologically Inspired Computing (NaBIC), 2010 Second World Congress on, Fukuoka.
16. Celikyilmaz, A., Hakkani-Tur, D. and Junlan, F. (12-15 Dec. 2010). *Probabilistic model-based sentiment analysis of twitter messages*. Paper presented at Spoken Language Technology Workshop (SLT), 2010 IEEE, Berkeley, CA.
17. Ming, H., Rohrdantz, C., Janetzko, H. and Dayal, U., Keim, DA., Haug, L., et al. (23-28 Oct. 2011). *Visual sentiment analysis on twitter data streams*. Paper presented at Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on, Providence, RI.
18. Quercia, D., Ellis, J., Capra, L. and Crowcroft, J. (9-11 Oct. 2011). *In the Mood for Being Influential on Twitter*. Paper presented at Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, Boston, MA.
19. Gruz, A., Doiron, S., Mai, P. (4-7 Jan. 2011). *Is Happiness Contagious Online? A Case of Twitter and the 2010 Winter Olympics*. Paper presented at System Sciences (HICSS), 2011 44th Hawaii International Conference on, Kauai, HI.
20. Inouye, D., Kalita, JK. (9-11 Oct. 2011). *Comparing Twitter Summarization Algorithms for Multiple Post Summaries*. Paper presented at Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, Boston, MA.
21. Zhenhua, W., Lidan, S., Ke, C. and Gang, C., Mehrotra, S. (2015). On Summarization and Timeline Generation for Evolutionary Tweet Streams. *Knowledge and Data Engineering, IEEE Transactions on*.27(5):1301-1315.
22. Sanghoon, L., Shaky, S., Sunderraman, R. and Belkasim, S. (17-20 Nov. 2013). *Real Time Micro-blog Summarization Based on Hadoop/HBase*. Paper presented at Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on, Atlanta, GA.
23. Xinyue, W., Tokarchuk, L., Cuadrado, F. and Poslad, S. (25-28 Aug. 2013). *Exploiting hashtags for adaptive microblog crawling*. Paper presented at Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on, Niagara Falls, ON.
24. Chung-Hong, L., Hsin-Chang, Y., Tzan-Feng, C. and Wei-Shiang, W. (25-27 July 2011). *A Novel Approach for Event Detection by Mining Spatio-temporal Information on Microblogs*. Paper presented at Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on, Kaohsiung.
25. Chun-Shuo, L., Meng-Fen, C., Wen-Chih, P. and Chien-Cheng, C. (21-23 Sept. 2011). *An event-based POI service from microblogs*. Paper presented at Network Operations and Management Symposium (APNOMS), 2011 13th Asia-Pacific, Taipei.

26. Pohl, D., Bouchachia, A., Hellwagner, H. (12-15 Dec. 2012). *Automatic Identification of Crisis-Related Sub-events Using Clustering*. Paper presented at Machine Learning and Applications (ICMLA), 2012 11th International Conference on, Boca Raton, FL.
27. Pohl, D., Bouchachia, A., Hellwagner, H. (4-7 Dec. 2013). *Online Processing of Social Media Data for Emergency Management*. Paper presented at Machine Learning and Applications (ICMLA), 2013 12th International Conference on, Miami, FL.
28. Yi-Shiang, T., Jyun-Yu, J., Pu-Jen, C. (4-7 Dec. 2012). *Event Duration Detection on Microblogging*. Paper presented at Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on, Macau.
29. Giacobe, NA., Hyun-Woo, K., Faraz, A. (8-10 Nov. 2010). *Mining social media in extreme events : Lessons learned from the DARPA network challenge*. Paper presented at Technologies for Homeland Security (HST), 2010 IEEE International Conference on, Waltham, MA.
30. Liangzhe, C., Tozammel Hossain, KSM., Butler, P. and Ramakrishnan, N., Prakash BA. (14-17 Dec. 2014). *Flu Gone Viral: Syndromic Surveillance of Flu on Twitter Using Temporal Topic Models*. Paper presented at Data Mining (ICDM), 2014 IEEE International Conference on, Shenzhen.
31. Huajing, L., Yuan, T., Wang-Chien, L. and Giles, CL., Meng-Chang, C. (20-22 Aug. 2010). *Personalized Feed Recommendation Service for Social Networks*. Paper presented at Social Computing (SocialCom), 2010 IEEE Second International Conference on, Minneapolis, MN.
32. Jonnalagedda, N., Gauch, S. (17-20 Nov. 2013). *Personalized News Recommendation Using Twitter*. Paper presented at Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on, Atlanta, GA.
33. H. Kwak, C. Lee, H. Park, and S. Moon. (2010). *What is twitter, a social network or a news media?*. Paper presented at the 19th International Conference on World Wide Web, ser. WWW '10 on, New York, NY, USA.
34. J. Weng, E.-P. Lim, J. Jiang, and Q. He. (2010). *Twitterrank: Finding topic-sensitive influential twitterers*. Paper presented at the Third ACM International Conference on Web Search and Data Mining, ser. WSDM '10 on, New York, NY, USA.
35. M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. (23-26 May. 2010). *Measuring user influence in twitter: The million follower fallacy*. Paper presented at ICWSM 10: Proceedings of international AAAI Conference on Weblogs and Social , Washington.
36. Rashid Hussain, A., Hameed, MA., Hegde, NP. (11-14 Dec. 2011). *Mining Twitter using cloud computing*. Paper presented at Information and Communication Technologies (WICT), 2011 World Congress on, Mumbai.

37. Subbian, K., Melville, P. (9-11 Oct. 2011). *Supervised Rank Aggregation for Predicting Influencers in Twitter*. Paper presented at Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, Boston, MA.
38. Yong, Z., Jiaqin, M., Tingting, H. (Oct. 30 2012-Nov. 1 2012). *User influence analysis on micro blog*. Paper presented at Cloud Computing and Intelligent Systems (CCIS), 2012 IEEE 2nd International Conference on, Hangzhou.
39. Yuhong, G., Li, K., Tie, S. (4-7 Dec. 2012). *Personalized Tweet Ranking Based on AHP: A Case Study of Micro-blogging Message Ranking in T.Sina*. Paper presented at Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on, Macau.

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, Adı : ŞEN, Muhammed
Uyruğu : T.C.
Doğum Yeri : Gaziantep
Medeni Hali : Evli
İletişim : muhammedsen@gmail.com

Eğitim Derecesi	Okul / Program	Mezuniyet tarihi
Yüksek Lisans	Gazi Üniversitesi / Bilgisayar Bilimleri	Devam ediyor
Lisans	İstanbul Teknik Üni / Bilgisayar Müh.	2005

İş Deneyimi, Yıl	Yer	Görev
2012 - Devam	Arinkom Bilgi Teknolojileri	CTO

Yabancı Dil

İngilizce



GAZİ GELECEKTİR..