T.C.
DÜZCE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TWITTER TABANLI DUYGU ANALİZİ

ABDULLAH TALHA KABAKUŞ

DOKTORA TEZİ
ELEKTRİK-ELEKTRONİK VE BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI

DANIŞMAN
DOÇ. DR. RESUL KARA

DÜZCE, 2017

T.C.
DÜZCE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TWITTER TABANLI DUYGU ANALİZİ

Abdullah Talha KABAKUŞ tarafından hazırlanan tez çalışması aşağıdaki jüri tarafından Düzce Üniversitesi Fen Bilimleri Enstitüsü Elektrik-Elektronik ve Bilgisayar Mühendisliği Anabilim Dalı'nda **DOKTORA TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Doç. Dr. Resul KARA

Düzce Üniversitesi

Jüri Üyeleri

Doç. Dr. Resul KARA
Düzce Üniversitesi _____

Doç. Dr. Pakize ERDOĞMUŞ
Düzce Üniversitesi _____

Yrd. Doç. Dr. İbrahim Alper DOĞRU
Gazi Üniversitesi _____

Yrd. Doç. Dr. Mehmet Şimşek
Düzce Üniversitesi _____

Yrd. Doç. Dr. Okan Erkaymaz
Bülent Ecevit Üniversitesi _____

Tez Savunma Tarihi: 25/05/2017

BEYAN

Bu tez çalışmasının kendi çalışmam olduğunu, tezin planlanmasından yazımına kadar bütün aşamalarda etik dışı davranışımın olmadığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara kaynak gösterdiğimi ve bu kaynakları da kaynaklar listesine aldığımı, yine bu tezin çalışılması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını beyan ederim.

25 Mayıs 2017

Abdullah Talha KABAKUŞ

TEŞEKKÜR

Doktora öğrenimimde ve bu tezin hazırlanmasında gösterdiği her türlü destek ve yardımdan dolayı çok değerli hocam Doç. Dr. Resul KARA'ya en içten dileklerimle teşekkür ederim.

Eğitim hayatım boyunca bana her zaman ilmin önemini öğütleyen, benden her türlü fedakarlığı, yardımı ve desteğini esirgemeyen sevgili aileme, rahmetli anneanneme ve çalışma arkadaşlarıma sonsuz teşekkürlerimi sunarım.

25 Mayıs 2017 **Abdullah Talha KABAKUŞ**

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ VII

ÇİZELGE LİSTESİ IX

KISALTMALAR XI

SİMGELER XII

ÖZET XIII

ABSTRACT XIV

EXTENDED ABSTRACT XV

1. GİRİŞ 1

2. İLGİLİ ÇALIŞMALAR 7

3. MATERYAL VE METOT 16

3.1. TWEET YAPISI 16

3.2. N-GRAM VE POLARİTE TANIMLARI 17

3.2.1. n-gram Tanımı 17

3.2.2. Polarite Tanımı 18

3.3. DUYGU SÖZLÜKLERİ 18

3.3.1. WordNet 18

3.3.2. SentiWordNet 19

3.3.3. SenticNet 20

3.3.4. AFINN 22

3.3.5. SentiStrength 23

3.3.6. NoSlang.com 23

4. TWITTER TABANLI DUYGU ANALİZİ 27

4.1. TWITTER UYGULAMA GELİŞTİRME ARAYÜZLERİ (API) 30

4.2. TWITTER API LİMİT KONTROLÜ 32

4.3. HIZLI DUYGU ANALİZİ 32

4.3.1. Emoji Analizi 36

4.4. METİN ÖNİŞLEME 38

4.4.1. Noktalama İşaretlerinin Temizlenmesi 39

4.4.2. Yeni Satır İfadelerinin Temizlenmesi 39
4.4.3. Twitter'a Özgü Karakter ve İfadelerin Temizlenmesi 39
4.4.4. Web Adreslerinin (URL) Temizlenmesi 40
4.4.5. Harf Tekrarlarının Düzenlenmesi 40
4.4.6. Kısaltmaların Düzenlenmesi 41
4.4.7. Yazım Hatası Düzeltme 43
4.4.7.1. *Levenshtein Uzaklık Algoritması 43*
4.4.7.2. *Hatalı Kelime Tespiti ve Düzeltmesi 45*
4.5. DOĞAL DİL İŞLEME VE DİLBİLİMSEL METOT TABANLI DUYGU ANALİZİ 48
4.5.1. n-gram Polarite Hesabının Yapılması 49
4.5.2. unigram'ların Polarite Hesabının Yapılması 49
4.5.3. Nihai Duygu Tespiti 51
4.6. HARF TEKRARLARININ DEĞERLENDİRİLMESİ 52
4.7. HASHTAG'LERİN DEĞERLENDİRİLMESİ 52
4.8. OLUMSUZLUK DEĞERLENDİRMESİ 54
5. DENEYSEL SONUÇLAR VE TARTIŞMA 57
5.1. TWEET SEVİYESİNDE BAĞLAM DUYARLILIĞI 67
5.2. CANLI OLAY ANALİZLERİ 72
5.2.1. Apple iPhone 7 Özel Etkinliği 72
5.2.2. Apple MacBook Pro Özel Etkinliği 74
5.3. BİLİNEN KISITLAR 76
5.4. DUYGU ANALİZİ SÜRESİNİN KISALTILMASI 78
5.5. TWITTER SPAM TESPİT MODÜLÜ 80
6. SONUÇLAR VE ÖNERİLER 93
7. KAYNAKLAR 96
ÖZGEÇMİŞ 108

ŞEKİL LİSTESİ

Sayfa No

Şekil 1.1. Wemogee uygulamasının (a) afazi olmayan modu ve (b) afazi modu için örnek kullanıcı görsel arayüzü. [36] 3
Şekil 3.1. Twitter kullanıcı takip ilişkisi. 16
Şekil 3.2. SenticNet üzerinde RDF/XML formatında örnek bir kaynak tanımı. 21
Şekil 3.3. NoSlang.com kısaltma arama web arayüzü. 24
Şekil 3.4. NoSlang.com üzerinde örnek bir kısaltmanın ve açık formunun kodlanması. 25
Şekil 3.5. NoSlang.com üzerinde aranan kısaltmaya yönelik sonuç bulunamaması durumunda görüntülenen web arayüzü. 26
Şekil 4.1. TwitterSentiDetector sisteminin duygu analizi için kullandığı niteliklerin örneklerle gösterimi. 28
Şekil 4.2. TwitterSentiDetector iş akışı önizlemesi. 29
Şekil 4.3. TwitterSentiDetector'de kullanılan kalan Twitter sorgu limitini sorgulayan Java diliyle yazılmış metot. 32
Şekil 4.4. Tweet içerisindeki her bir kelimeyi UTF-8 kodlamasına dönüştürerek emoji kod karşılıklarını tespit eden Java diliyle yazılmış metot. 38
Şekil 4.5. Örnek bir tweet üzerinden tweet'i oluşturan kelimelerin UTF-8 karşılıkları.. 38
Şekil 4.6. “has” kelimesi için WordNet üzerinden kelimelere yönelik temel halinin bulunması. 41
Şekil 4.7. Levenshtein Uzaklık Algoritması'nın sözde kodu. 43
Şekil 4.8. “a happy day” ve “beautiful day” örneklerinin graf veri yapısında depolanma yöntemi. 44
Şekil 4.9. Cümlelerin graf veritabanında depolanması için kullanılan algoritmanın sözde kodu. 45
Şekil 4.10. Alternatif kelime ile hatalı yazılan kelimenin en büyük ortak kısmındaki harf sayısını hesaplayan Java diliyde yazılmış metot. 46
Şekil 4.11. Hashtag ayrıştırma algoritması sözde kodu 54
Şekil 4.12. “#epicfail” hashtag'inin TwitterSentiDetector tarafından ayrıştırılması. 54
Şekil 5.1. TwitterSentiDetector veri setleri üzerinde duygu analizi yapılmasını sağlayan kullanıcı görsel arayüzü. 60
Şekil 5.2. TwitterSentiDetector detaylı tweet analiz sonuçlarının görüntülendiği kullanıcı görsel arayüzü. 61
Şekil 5.3. Anlık duygu analizi servisi ile Twitter üzerinden girilen anahtar kelimeye yönelik tweet toplanmasını için geliştirilmiş kullanıcı görsel arayüzü. 61
Şekil 5.4. Anlık duygu analizi servisi sonucunda aranan anahtar kelimeye yönelik sonuçları grafiksel olarak özetleyen kullanıcı görsel arayüzü. 62
Şekil 5.5. Girilen kaynaklara yönelik polarite yükünün duygu sözlüklerinde aranmasını sağlayan kullanıcı görsel arayüzü. 62
Şekil 5.6. Veri setleri ve çalışma moduna göre test sonuçlarına yönelik karışıklık matrisinin ve gösterge (gauge) grafikleriyle sistem hassasiyetinin izlenebilmesi için geliştirilen istatistik kullanıcı görsel arayüzü. 63
Şekil 5.7. Sanders Analytics veri seti kullanılarak sırasıyla üçlü ve ikili mod için elde edilen ROC eğrileri. 65
Şekil 5.8. STS veri seti kullanılarak sırasıyla üçlü ve ikili mod için elde edilen ROC eğrileri. 66
Şekil 5.9. Twitter üzerinden anlık tweet'lerin çocuk-atası hiyerarşisinde ilişkisel

veritabanında depolanması için öne sürülen algoritma. 68
Şekil 5.10. Bağlam duyarlı olarak depolanan tweet'lerin duygu analizi için öne sürülen algoritma. 68
Şekil 5.11. Apple iPhone 7 Özel Etkinliği süresince “#iPhone7” hashtag'i ile gönderilen tweet'lerin duygu analiz sonucunu gösteren kullanıcı görsel arayüzü. 73
Şekil 5.12. Apple MacBook Pro özel etkinliği süresince “#MacBookPro” hashtag'i ile gönderilen tweet'lerin duygu analiz sonucunu gösteren kullanıcı görsel arayüzü. 75
Şekil 5.13. URLVoid web kullanıcı görsel arayüzü. 84
Şekil 5.14. URLVoid API üzerinden http://bevrich.com zararlı URL'inin sorgulanması sonucunda elde edilen XML formatındaki cevap. 86

ÇİZELGE LİSTESİ

Sayfa No

Çizelge 3.1. Twitter'a özel semboller ve tweet içindeki kullanım amaçları. 17
Çizelge 3.2. Türleriyle beraber n-gram örnekleri. 18
Çizelge 3.3. WordNet kelime eş küme çeşitleri. 19
Çizelge 3.4. NoSlang.com sözlüğü kullanılarak yapılan örnek kısaltma dönüşümleri. . 24
Çizelge 3.5. Tez çalışmasında kullanılan duygu sözlüklerinin karşılaştırılması. 26
Çizelge 4.1. Twitter uygulama geliştirme arayüzü işlem operatörleri. 31
Çizelge 4.2. Wikipedia'da tabanlı pozitif ve negatif olarak sınıflandırılan yüz ifadeleri. 33
Çizelge 4.3. TwitterSentiDetector tarafından değerlendirilen emojiler, UTF-8 karşılıkları ve yansıttıkları duyguya yönelik gülücük eşleşme listesi. 34
Çizelge 4.4. Raez ve arkadaşları tarafından öne sürülen gülmeye karşılık gelen metinsel ifadeler [41]. 36
Çizelge 4.5. “happyyy” kelimesi için soldan ve sağdan başlayarak tekrar eden harf düzenlemesi sonuçları. 41
Çizelge 4.6. NoSlang.com üzerinden açık formları elde edilen örnek kısaltmalar ve açık formları. 42
Çizelge 4.7. Anlamsal olarak Levenshtein Uzaklık Algoritması ile yanlış, TwitterSentiDetector ile doğru olarak düzeltilen örnek yazım hatalı ifadeler. 48
Çizelge 4.8. Apache OpenNLP POS etiketleri. 51
Çizelge 4.9. Councill ve arkadaşları [137] tarafından öne sürülen olumsuzluk anlamı katan ek ve kelimeler. 55
Çizelge 4.10. TwitterSentiDetector tarafından olumsuzluk değerlendirmesi için kullanılan destekleyici ve çevirici bağlaçlar. 56
Çizelge 5.1. Tez çalışmasında sistemin testi için kullanılan veri setlerine ait polarite dağılımı. 57
Çizelge 5.2. ROC analizi etiketlerinin TwitterSentiDetector üçlü mod duygu analizi için karşılıkları. 58
Çizelge 5.3. TwitterSentiDetector'ün Sanders Analytics veri seti kullanılarak elde edilen üçlü mod testlerinin karışıklık matrisi. 59
Çizelge 5.4. TwitterSentiDetector'ün STS veri seti kullanılarak elde edilen üçlü mod testlerinin karışıklık matrisi. 60
Çizelge 5.5. TwitterSentiDetector Sanders Analytics veri seti kullanılarak ikili mod testlerinin karışıklık matrisi. 64
Çizelge 5.6. TwitterSentiDetector STS veri seti kullanılarak ikili mod testlerinin karışıklık matrisi. 65
Çizelge 5.7. TwitterSentiDetector UMICH SI650 veri seti kullanılarak ikili mod testlerinin karışıklık matrisi. 65
Çizelge 5.8. TwitterSentiDetector yüz ifadeleri ile metinsel gülmeler ve bağlaç analizi kullanılmaması durumunda üçlü mod için veri setlerine göre duygu tespit hassasiyeti. 67
Çizelge 5.9. TwitterSentiDetector yüz ifadeleri ile metinsel gülmelerden ve bağlaç analizinden faydalanılmasına göre ikili mod için veri setlerine göre duygu tespit hassasiyeti. 67
Çizelge 5.10. NSFAS veri seti için bağlam analizi kullanılması durumunda elde edilen

sonuçlar. 69
Çizelge 5.11. NSFAS veri seti için bağlam analizi kullanılmaması durumunda elde edilen sonuçlar. 70
Çizelge 5.12. TwitterSentiDetector ile ilgili diğer çalışmaların duygu analizi yöntemlerinin karşılaştırılması. 70
Çizelge 5.13. Apple iPhone 7 özel etkinliği önemli anlarının zaman bilgisi ve TwitterSentiDetector tarafından anlık olarak tespit edilen pozitif tweet oranları. 74
Çizelge 5.14. Apple MacBook Pro Özel Etkinliği önemli anlarının zaman bilgisi ve TwitterSentiDetector tarafından tespit edilen pozitif tweet oranları. 75
Çizelge 5.15. Olumsuzluk eki veya bağlaç olmadan zıt duyguları birarada bulunduran tweet örnekleri. 76
Çizelge 5.16. Veri setlerinde yanlış etiketlenmiş örnek tweet'ler ve beklenen duygu etiketleri. 77
Çizelge 5.17. Test veri setindeki kaynaklara yönelik ilişkisel veritabanı (MySQL) ve bellek içi veritabanı (Redis) kullanılması durumunda elde edilen erişim süreleri. 80
Çizelge 5.18. Kısaltılmış URL'ler, kullanılan kısaltma servisi ve elde edilen normal formları. 82
Çizelge 5.19. URLVoid'in üzerinden servis sağladığı güvenlik servisleri. 83
Çizelge 5.20. Kullanıcı bazlı spam tespit nitelikleri. 87
Çizelge 5.21. Tweet bazlı spam tespit nitelikleri. 89
Çizelge 5.22. TwitterSentiDetector spam tespit modülü tarafından göz ardı edilen nitelikler ve gerekçeleri. 90
Çizelge 5.23. Spam tespit modülü veri seti istatistikleri. 91
Çizelge 5.24. Test sonuçlarının karmaşıklık matrisi. 92
Çizelge 5.25. Twitter spam tespiti için göz ardı edilen niteliklerin sisteme dahil edilmesinde durumunda sistemin doğruluğu ve hassasiyeti. 92

KISALTMALAR

ANEW	Affective Norms for English Words
API	Application Programming Interface
CART	Classification and Regression Tree
CSS	Cascading Style Sheets
CRUD	Create, Read, Update, Delete
DALY	Disability-Adjusted Life Year
FN	False Negative
FP	False Positive
GUI	Graphical User Interface
HTML	Hypertext Transfer Language
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
MB	Megabyte
NoSQL	Not Only SQL
NSFAS	National Student Financial Aid Scheme
OWL	Web Ontology Language
PDT	Pacific Daylight Time
POS	Parts of Speech
RDF	Resource Description Language
RDFS	Resource Description Language Schema
RDMS	Relational Database Management System
REST	REpresentational State Transfer
ROC	Receiver Operating Characteristic
SA	Sentiment Analysis
STS	Stanford Twitter Sentiment
SVM	Simple Vector Machines
TN	True Negative
TP	True Positive
TSA	Twitter Sentiment Analysis
TT	Trending Topic
URL	Uniform Resource Locator
XML	Extensible Markup Language
WHO	World Health Organization

SİMGELER

@	Twitter kullanıcı adı öneki
#	Hashtag

ÖZET

TWİTTER TABANLI DUYGU ANALİZİ

Abdullah Talha KABAKUŞ
Düzce Üniversitesi
Fen Bilimleri Enstitüsü, Elektrik-Elektronik ve Bilgisayar Anabilim Dalı
Doktora Tezi
Danışman: Doç. Dr. Resul KARA
Mayıs 2017, 107 sayfa

Duygu analizi sosyal medyanın özellikle de yapısal ve erişime açık veri sağladığı için Twitter'ın doğuşuyla beraber çok daha önemli hale gelmiştir. TwitterSentiDetector, tez çalışması kapsamında geliştirilen yazım hataları, harf tekrarları, hashtag, gülücük, emoji ve metinsel gülme kullanımı gibi Twitter'da kullanılan resmi olmayan dilden kaynaklı farklılıkları hedef alan, alandan bağımsız ve eğiticisiz Twitter duygu analizi aracıdır. TwitterSentiDetector, polarite değerleri geniş kullanımı olan duygu sözlüklerinden elde edilen polarite değerleri üzerinden doğal dil işleme tekniklerinin yanı sıra öne sürülen dilbilimsel metotları kullanarak tweet'lerin duygularını pozitif, negatif ve nötr olarak sınıflandırmaktadır. Geniş kullanımı olan ve elle duygu etiket tespiti yapılmış Twitter verisetleri üzerinde yapılan testlere göre TwitterSentiDetector %69'a kadar doğru duygu tespiti gerçekleştirebilmektedir. Hedef duygu kümesi pozitif ve negatife indirgendiği zaman ise, doğru duygu tespit oranı %87'ye kadar yükselmektedir. Sonuçlar, aynı veri seti bağlam duyarlı duygu analizi modülü ile oldukça benzer hesaplanmıştır ki bu durum iki yöntemin de geçerliliğini ispatlamaktadır. Bir Twitter duygu analizi servisinin Twitter'da hala yaygın olan ve amaçlarından birisi de Twitter tabanlı servislerin geçerliliğini gasp etmek olan spam tweet'lerden haberdar olması gerekmektedir. Deneysel sonuçlara göre, TwitterSentiDetector'e entegre edilen spam tespit modülünün doğruluğu 0,943 olarak hesaplanmıştır. Yine TwitterSentiDetector'le entegre çalışan graf tabanlı cümle seviyesinde yazım hatası kontrol uygulama çatısı üç farklı listeden elde edilen İngilizce'de en çok yazım yanlışı yapılan kelimeler kullanılarak test edildiğinde doğruluğu 0,84 olarak hesaplanmıştır.

Anahtar sözcükler: Twitter duygu analizi, Duygu analizi, Doğal dil işleme, Sosyal medya madenciliği, Duygu tespiti

ABSTRACT

SENTIMENT ANALYSIS BASED ON TWITTER

Abdullah Talha KABAKUS
Duzce University
Graduate School of Natural and Applied Sciences, Department of Electrical-Electronics and Computer Engineering
Doctoral Thesis
Supervisor: Assoc. Prof. Dr. Resul KARA
May 2017, 107 pages

Sentiment analysis has become more crucial after the rise of social media, especially for the Twitter since it provides structured and publicly available data. TwitterSentiDetector is a domain-dependent and unsupervised Twitter sentiment analyzer that focuses on the differences occurred by the informal language used in Twitter such as spelling mistakes, letter repetitions, usage of hashtags, emoticons, emojis, and laughs. TwitterSentiDetector uses natural language processing techniques alongside the proposed linguistic methods to classify sentiments of tweets into positive, negative, and neutral through the polarity scores obtained from widely used sentiment lexicons. According to tests on the widely-used Twitter datasets that contain manually detected sentiment labels alongside tweets, TwitterSentiDetector's sentiment detection ratio is calculated as up to 69%. When the target sentiment classes are decreased to positive and negative, the detection ratio is increased up to 87%. The results are calculated very similarly when the same dataset is evaluated by the proposed tweet-level context aware sentiment analysis module which confirms the validity of each approach. A Twitter sentiment analyzer services should be aware of spam since it is still widespread in Twitter and one of its aim is to hijack the validity of the services based on Twitter. According to the experimental results, the integrated spam detection framework's accuracy is calculated as 0.943. Similarly, when the developed graph based sentence level spell checking framework which is integrated into TwitterSentiDetector is evaluated using the most commonly misspelled words in English which are based on three lists, the accuracy is calculated as 0.84.

Keywords: Twitter sentiment analysis, Sentiment analysis, Natural language processing, Social media mining, Sentiment detection

EXTENDED ABSTRACT

SENTIMENT ANALYSIS BASED ON TWITTER

Abdullah Talha KABAKUS
Duzce University
Graduate School of Natural and Applied Sciences, Department of Electrical-Electronics and Computer Engineering
Doctoral Thesis
Supervisor: Assoc. Prof. Dr. Resul KARA
May 2017, 107 pages

1. INTRODUCTION

Sentiment analysis has become more crucial after the rise of social media, especially the Twitter since it provides structured and publicly available data. The proposed work namely TwitterSentiDetector is a domain-dependent and unsupervised Twitter sentiment analyzer that focuses on the differences occurred by the informal language used in Twitter such as intentional or unintentional spelling mistakes, usage of emoticons, emojis, and laughs. Spam which is defined as the unsolicited messages is still widespread on Twitter despite the actions taken by Twitter. In order to provide more accurate result in terms of sentiment detection in Twitter, the spam tweets are needed to be filtered before starting the analysis. Thus, a Twitter sentiment detection approach should contain spam detection module. Similarly, in order to correct the spelling mistakes which are widespread on Twitter, TwitterSentiDetector checks tweets for possible spelling mistakes thanks to the developed graph based sentence level spell checking framework.

2. MATERIAL AND METHODS

TwitterSentiDetector uses natural language processing techniques alongside the proposed linguistic methods to classify sentiments of tweets into positive, negative, and neutral through the polarity scores obtained from the widely-used sentiment lexicons. In order to overcome the differences as the result of the language used in Twitter, TwitterSentiDetector uses more comprehensive techniques compared to traditional sentiment analysis approaches such as slang translation, emoticon, emoji, and laugh detection, and hashtag processing. Thanks to the developed sentence level spell checking framework which is based on a graph model, TwitterSentiDetector is able to correct

spelling mistakes which cannot be detected by spell checking approaches based on edit distance method. TwitterSentiDetector provides a service to collect and make a conclusion in terms of detected sentiment for the entered keywords through Twitter. In order to eliminate spam tweets, a spam detection framework is developed and embedded into TwitterSentiDetector.

3. RESULTS AND DISCUSSIONS

According to tests on widely used Twitter datasets that contain manually detected sentiments labels alongside tweets, TwitterSentiDetector's sentiment true detection ratio is calculated as up to 69% for the trinary mode. When the target sentiment classes are decreased to positive and negative which is known as binary mode, the true detection ratio is increased up to 87%. The results are calculated very similarly when the same dataset is evaluated by the proposed tweet-level context aware sentiment analysis module which confirms the validity of each approach. According to the experimental results, the developed spam detection framework's accuracy is calculated as 0.943. The developed graph based sentence level spell checking framework is evaluated using the most commonly misspelled words in English which are based on three lists, the accuracy is calculated as 0.84.

4. CONCLUSION AND OUTLOOK

TwitterSentiDetector is a domain-independent and unsupervised sentiment analyzer framework which is specifically designed for Twitter. The services which are developed for Twitter should be aware of the differences occur by the architecture of Twitter and the usage practices of its users. Therefore, a Twitter sentiment analyzer service is needed to overcome the differences between the traditional texts which are the result of the language used in Twitter such as spelling mistakes, letter repetitions, usage of hashtags, emoticons, emojis, and laughs. Another aspect that Twitter sentiment analyzer services should be aware of is spam.

1. GİRİŞ

İnsanların fikirlerinin diğerleri üzerinde ciddi etkileri vardır; bu fikirler bir ürün, ünlü bir kişi, bir firma, bir şehir veya ülke veya bir politikacı ile ilgili olabilir. İnternet kullanıcılarının çok büyük bir çoğunluğunu oluşturan sosyal medya kullanıcıları, sosyal medya platformlarında günlük yaşamdaki hemen hemen her konuyla ilgili duygu ve düşüncüleri paylaşmak için kullanmaktadır [1]. Bu konular, insanların gündelik olaylara verdiği tepkileri, politikacıları, ünlüleri, gazetecileri, sporcuları, ürünleri, firmaları ve dünyadaki olayları kapsayacak kadar çeşitlidir [2]–[11]. Twitter, kullanıcılarının "*tweet*" adı verilen durum mesajlarının paylaştığı ve Nisan 2015'de yayınlanan rapora göre 5 yılda 10 katına çıkan aylık 302 milyon aktif kullanıcısıyla en popüler mikroblog servisidir [12], [13]. Bu popülerliğin sonucu olarak Twitter'da dakikada ortalama 350 bin ve günlük olarak 500 milyon tweet atıldığı raporlanmıştır [14]–[16]. Çevrimiçi servisleri kullanan erişkinlerin %19'u Twitter'ı kullanmaktadır [17]. Rakamların açıkça belirttiği üzere, Twitter üzerinde oldukça yoğun bir veri trafiği bulunmaktadır ve gelecek yıllarda da bu yoğunluğun artacağı ön görülmektedir. Bu popülerlik ve sonucunda oluşan büyük veri, Twitter'ı iyi bir umumi veri kaynağı haline getirmekte ve endüstrinin ve akademinin ilgisini üzerine çekmektedir. Twitter'da günlük 1 milyar tweet atılmasına rağmen, Twitter 15 milyar uygulama geliştirme arayüzü (Application Programming Interface – API) çağrısı almaktadır. Bu sayı, Facebook ve Instagram gibi popüler sosyal medya platformlarına ve Google, Yandex ve Bing gibi arama motorlarına gelen çağrı sayısının yaklaşık olarak üç katıdır [18]. Bunun en önemli sebebi, Twitter verisinin sıklıkla kurum ve kuruluşlar tarafından ürünleri hakkında kullanıcılarının ne düşündüğünü veya çok çeşitli konularda kullanıcıların genel eğilimlerini daha iyi öğrenmek için kullanılmasıdır [19]–[21]. Bunun yanı sıra Twitter'ı akademik araştırmalarda diğer sosyal medya platformlarından daha popüler olmasını sağlayan başlıca sebep, Twitter'ın veri erişimini herkese (sisteme üye olan veya olmayan tüm kullanıcılara) açık tutmasıdır [17], [22]. Ayrıca Twitter, diğer sosyal medya platformlarına göre daha yapısal bir veri formatı sunmaktadır. Bu sebeplerden ötürü, Twitter Duygu Analizi (Twitter Sentiment Analysis – TSA) doğal dil işleme (Natural Language Processing - NLP) ve duygu analizi (Sentiment Analysis - SA) konusunda çalışan araştırmacılar için oldukça popüler bir

araştırma konusu haline gelmiştir. Sosyal medyadaki verinin çeşitliliği ve büyüklüğü, duygu analizinin insanlar tarafından yapılmasını imkânsız hale getirmektedir. Bu durum otomatize edilmiş bir duygu analizi sisteminin gerekliliğini doğurmuştur [19].

Görsel yüz ifadeleri (emoji) Twitter'da hatta genel olarak sosyal medya ve anlık mesajlaşma servisleri tarafından oldukça sıklıkla kullanılmaktadır. Emoji'lerin bir sunum yönteminin yanı sıra yazılı iletişimde yazarın niyetini belirtme, sosyokültürel farklılıklar ve yazarın kimliği yansıtmak gibi birçok rol oynadığı belirtilmektedir [23]–[25]. Instagram tarafından yayınlanan rapora göre Instagram'da paylaşılan metinlerin neredeyse yarısı emoji içermektedir [26]. Yapılan araştırmada Twitter'da noktalama ve harflerden oluşturulan gülücüklerin (emoticon) kullanımının azaldığı, emoji'lerin kullanımının arttığı ve emoji'lerin emoticon'ların yerini alacağı belirtilmektedir [27], [28]. Emoji'ler duygu ifade etmeyi kolaylaştırırken aynı zamanda görsel ifadeler ve el-kol işaretleri (gesture) barındırdığından beynin belirli bölgelerinde meydana gelen hasardan ötürü konuşma dillerini kavrayamayan ve kullanamayan afazi (aphasia) hastaları [29] için bir iletişim kaynağı olabileceği düşünülmektedir [30]–[35]. Bu doğrultuda yakın zamanda (Nisan 2017) Samsung "*Wemogee*"[1] isimli afazi hastalarına yönelik duygu ve düşünceleri konuşma diline çeviren bir uygulamayı kullanıma sürmüştür [36]. Şekil 1.1'de *Wemogee* uygulamasının örnek kullanıcı görsel arayüzü (graphical user interface – GUI) sunulmuştur. Bu uygulama ile afazi hastaları gerek kendi aralarında gerekse diğer kullanıcılar ile iletişim kurma imkânı sağlamaktadır. Benzer olarak Twitter'da da afazi hastaları emoji'leri kullanarak duygu ve düşüncelerini attıkları tweet'lerle paylaşabilecektir. Bu bağlamda, tez çalışmasında öne sürülen sistem, duygu analizi esnasında emoji'lerden faydalanmaktadır.

[1] http://wemogee.com



(a) Afazi olmayan mod (b) Afazi modu

Şekil 1.1. Wemogee uygulamasının (a) afazi olmayan modu ve (b) afazi modu için örnek kullanıcı görsel arayüzü. [36]

Whitelaw ve arkadaşları [37], duygu analizini "*belgeleri hedefe göre pozitif veya negatif olarak etiketlemek*" olarak tanımlamaktadır. Duygu analizi yaklaşımları, duyguları genellikle pozitif, negatif veya nötr olarak sınıflandırmaktadır [2], [38], [39]. Duygu analizi yeni bir araştırma alanı değildir, birçok araştırmacı duygu analizi konusunda yıllardır birçok çalışma öne sürmüşlerdir. Ancak bu çalışmalar film incelemeleri ve kitap bölümleri gibi geniş metinlerin analizine yoğunlaşmıştır [40]. Bu klasik yaklaşımlar, tweet'lerle düz metinler arasında yapısal ve dilbilimsel olarak birçok farklılığı sahip olmasından ötürü Twitter duygu analizinde kullanılamamaktadır. Bu farklılıkların temel sebepleri, tweet'lerin 140 karakterle sınırlandırılmış olan boyutu ve sosyal medyanın dilbilimsel olarak esnek olan kendine özgü bir dili olmasıdır [41]–[46]. Tweet'ler ile klasik metinler arasındaki farklar aşağıdaki gibi listelenebilir:

- Klasik metinlerde (1) kısaltmalar, (2) duyguları ifade etmek için harfler ve

noktalama işaretlerinin çeşitli kombinasyonlarıyla elde edilen yüz ifadeleri (emoticon) ve (3) görsel yüz ifadeleri (emoji) kullanılmazken, Twitter'da var olan 140 karakter sınırı sebebiyle kullanıcıların duygu ve düşünceleri ifade etmeleri birkaç kelimeye sıkıştırıldığından Twitter'da kısaltmalar ve yüz ifadeleri sıklıkla kullanılmaktadır [41], [46], [47].

- Tweet'lerde kullanıcılar ve konular arasındaki iletişimi sağlamak için Twitter'a özel sembol ve ifadeler (`@`, `#`, `RT`) kullanılmaktayken, klasik metinlerde böyle bir durum söz konusu değildir.
- Twitter (genel anlamda sosyal medya) kullanıcıları, duygularını vurgulamak için kelime içinde harf tekrarlarına ("*verryy*", "*happpy*", gibi) ve fazladan noktalama işaretlerine ("???", "!!!", "!??", gibi) sıklıkla yer vermektedirler [39], [48], [49].
- Klasik metinlerde duygu ve düşünceler dilbilimsel olarak kurallı bir şekilde ifade edilirken, Twitter'da maksimum 140 karakter olan mesaj limiti sebebiyle [48] "sosyal medya İngilizcesi" diye nitelendirilen, dilbilimsel ve sözlüksel kurallardan uzak hatalı kelimelerin, kısaltmaların, kasti kelime hatalarının ve yüz ifadelerinin sıklıkla kullanıldığı bir ifade biçimi mevcuttur [43], [50], [51].
- Bazı tweet'ler hem pozitif hem de negatif duyguları beraber barındırabilmektedir [52] ve 140 karakter sınırı sebebiyle kısa metinlerde nihai duyguyu tespit etmek, karakter sınırı olmayan klasik metinlere göre kıyaslandığında oldukça zor olmaktadır [53].
- Klasik metinler, yazarlarının bir konudaki duygu ve düşüncelerini özetlemekteyken, tweet'ler daha çok gündelik olarak yazıldığından çok çeşitli konularla ilgili olabilmektedir [54].
- Amaçlarından birisi de Twitter üzerinden servis sağlayan araçları manipüle etmek olan spam tweet'lerin Twitter'da oldukça yaygın olmasından ötürü duygu analizi öncesinde spam tweet'lerin filtrelenmesi gerekmektedir [55]–[60].

Twitter üzerinden kullanıcıların duygu ve düşüncelerinin analiz edilmesi, Twitter'ın efor sarf etmeden, düşük maliyetli ve daha geniş kitleye ulaşabilen bir anket sistemi haline dönüşmesini sağlamaktadır [61]–[63]. Bu sayede, tüketiciler herhangi bir ürünü satın almadan önce ürün hakkında araştırma yapabilecek ve ürün kullanıcılarının

memnuniyetini öğrenebilmektedirler [48], [64], [65]. Pazarlamacılar ise kendi firma ve ürünleriyle ilgili müşteri memnuniyetini, taleplerini tespit edebilecek [66] ve bir ürün çıkarmadan önce müşterinin o ürüne olan ilgisini gözlemleyebilmektedirler [67]. Twitter'ı oldukça etkin ve yoğun şekilde kullanan politikacılar [4], [5], [68]–[70] da öne sürdükleri politikalarının toplumda nasıl karşılık bulduğunu gözlemleyebilmektedirler. Literatürdeki diğer çalışmalarda Twitter duygu analizinin seçim [4], [44], [70]–[74] ve film gişe [37], [40], [75] tahminlerinde de sıklıkla kullanılarak başarılı sonuçlara erişilmektedir.

Dünya Sağlık Örgütü (World Health Organization - WHO)[2], 2020 yılında dünya çapında kalp rahatsızlıklarından sonra "kayıp sağlıklı yılların" (Disability-Adjusted Life Year - DALY) en önemli ikinci gerekçesinin depresyon olacağını öngörmektedir [76]. Günümüzde ise 15-44 yaş aralığı için şimdiden "kayıp sağlıklı yılların" en önemli ikinci gerekçesidir [77]. Depresyonun bir alt türü olan "Maskeli Depresyon" hastaları, duygularını günlük hayatta saklama eğilimindedir [78]–[80]. Uzmanlar, psikoanaliz açısından insanların duyguları ifade ederken kullanılan kelimelerin seçimi, kişinin ruh durumu hakkında çok değerli bilgiler barındırdığını belirtmektedir [81]–[83]. Bu durumda kişilerin attığı tweet'ler psikoanaliz açısından zengin bir veri kaynağı olacağı düşünülmektedir. Depresyon hastalarının izleminde, hastaların Twitter'da attıkları tweet'lerin periyodik olarak değerlendirilmesinin hastalık sürecini değerlendirme açısından önemli sonuçlar barındıracağı düşünülmektedir.

Klasik metinlerle tweet'ler arasındaki farklar ele alındığında Twitter üzerinden duygu analizi yapmak ve nihai duyguyu tespit etmek için klasik metin tabanlı duygu analizi yöntemlerinden daha kapsamlı bir yaklaşıma ihtiyaç duyulduğu görülmektedir. Tez çalışması ile öne sürülen TwitterSentiDetector isimli sistem, Twitter'da kullanılan dilbilimsel kurallardan uzak, gayri resmi dil kullanımından kaynaklı tüm bu çeşitlilikleri doğru bir şekilde ele alıp, İngilizce dilinde yazılmış tweet'leri dilbilimsel ve sözlüksel yöntemlerden faydalanarak yapılan duygu analizleri sonucunda pozitif, negatif ve nötr olmak üzere üç grupta sınıflandırmaktadır. TwitterSentiDetector, tweet'lerdeki konu çeşitliliği göz önünde bulundurularak tanım kümesinden ve bağlamdan bağımsız,

[2] http://www.who.int

öğreticisiz bir sistem olarak geliştirilmiştir. TwitterSentiDetector, tweet içerisindeki kelime ve kelime gruplarının duygu yükünü, literatürde kabul görmüş duygu sözlüklerini kullanarak belirlemektedir.

Tez çalışması şu şekilde yapılandırılmıştır: 2. bölümde ilgili çalışmalar sunulmuştur. 3. bölümde materyal ve metotlara yer verilmiştir. 4. bölümde tez çalışması ile öne sürülen sistem gerçekleştirim detaylarıyla birlikte sunulmuştur. 5. bölümde deneysel sonuçlar ve tartışmaya yer verilmiştir. Son olarak, 6. bölümde ise sonuçlar ve önerilere yer verilmiştir.

2. İLGİLİ ÇALIŞMALAR

Raez ve arkadaşları [41], WordNet aracılığıyla oluşturulan ağırlıklandırılmış düğümleri SentiWordNet ve Rastlantısal Yürüme Algoritması kullanarak nihai duygu analizi tespitinde bulunan eğiticisiz ve tanım kümesinden bağımsız Twitter tabanlı duygu analizi yöntemi öne sürmüşlerdir. Raez ve arkadaşları, tez çalışmasında öne sürülen yaklaşımda olduğu gibi duygu analizine başlamadan önce tweet'leri metin önişleme sürecine tabi tutmuşlardır. Raez ve arkadaşları, TwitterSentiDetector'daki metin ön işlemede kullanılan ve tweet'e anlam kattığı görülen hashtag'leri, metin önişleme aşamasında temizleyerek duygu analizinde kullanmamaktadır. Bu yaklaşımın, duygunun polarite tespitini olumsuz etkileyeceği düşünüldüğünden TwitterSentiDetector, hashtag'leri 4. bölümde tartışıldığı üzere duygu analizi açısından faydalı bir nitelik olarak kullanmaktadır. Raez ve arkadaşlarının öne sürdükleri sistem ile TwitterSentiDetector arasındaki bir diğer fark ise polarite tespiti için kullanılan duygu sözlük sayısıdır. Raez ve arkadaşlarının öne sürdükleri yaklaşım, polarite tespitini sadece SentiWordNet üzerinden yapmaktayken, TwitterSentiDetector birden çok duygu sözlüğünden faydalanarak kelime ve kelime gruplarına yönelik daha tutarlı bir polarite değeri elde etmektedir. TwitterSentiDetector ile Raez ve arkadaşlarının öne sürdüğü yaklaşımdaki metin önişleme sürecindeki bir diğer fark ise tekrarlı karakterlerin düzenlenmesinde meydana gelmektedir. Raez ve arkadaşları tekrarlı harfleri ikiye indirgemekte iken TwitterSentiDetector karakter sayısına sabit bir sınırlama getirmemekte, kelimenin WordNet üzerinde tanımlı anlamlı en küçük haline dönüştürmektedir. İki yaklaşımın karakter düzenlemesi bir örnek üzerinden incelenecek olursa, "*verrryy*" kelimesini Raez ve arkadaşlarının öne sürdüğü yaklaşım "*verryy*" haline dönüştürecekken TwitterSentiDetector anlamlı hali olan "*very*" şekline çevirecektir. İlgili çalışmada Twitter'dan tweet çekmek için kullandıkları web servisin negatif yüz ifadesi içeren tweet'lere yönelik arama sonuçlarında ":P" ve ":PP" ifadelerini içerdiği raporlanmıştır. TwitterSentiDetector'da Twitter API için geliştirilmiş Java tabanlı ve açık kaynak kodlu Twitter4J[3] kütüphanesi kullanılmakta ve böyle bir sorunla karşılaşılmamaktadır. Raez ve arkadaşları, rastlantısal yürüme algoritması yerine destek vektör makinası (Support Vector Machines - SVM) kullanıldığında duygu analizi başarımının düştüğünü raporlamışlardır. Çalışmada zıt yüz ifadelerini bir arada barındıran tweet'lerin anlam karmaşası içerdiği düşünüldüğünden bu tip tweet'ler değerlendirme dışı

[3] http://twitter4j.org

bırakılmıştır.

Sentimentor [39], Naive Bayes sınıflandırıcısından faydalanarak tweet'leri pozitif, negatif ve objektif olmak üzere üç sınıfta gruplandırmaktadır. Tez çalışmasındaki yaklaşımdan farklı olarak Sentimentor, sadece unigram ve bigramlardan faydalanırken tez çalışmasında öne sürülen yaklaşımda unigram, bigram'ların yanı sıra trigram ve four-gram'lardan faydalanılarak tweet'lerde gerek kısaltmalarla gerekse düz metin halinde kullanılan tüm kelime öbeklerinin polarite hesabı hedeflenmiştir. Raez ve arkadaşlarının öne sürdüğü yaklaşımdaki gibi, Sentimentor da kelimelerde karakter tekrarlarını ikiye düşürmektedir. TwitterSentiDetector ise tekrarlı (recursive) bir yaklaşımdan faydalanarak anlamlı en sade halini WordNet üzerinden kontrol ederek harf indirgemesi yapmaktadır. Sentimentor, "*a*", "*an*", "*the*", "*and*" gibi etkisiz kelimeler (stop words) diye isimlendirilen ve tek başına anlamı olmayan kelimeleri tweet içerisinden çıkartmaktadır. Ancak bu kelimeler "*laugh a lot*", "*a little*", "*hustle and bustle*" gibi kelime grupları içinde kullanılması durumunda anlamlı olduğundan duygu analizi açısından önem arz edecek şekilde tweet'lerde sıklıkla kullanılmaktadır. Bundan dolayı tez çalışmasında öne sürülen yaklaşımda etkisiz kelimeler tweet'lerden çıkartılmamaktadır. Ayrıca bu kelimeler, kelime öbekleri (n-gram) içerisinde kullanılmaması durumunda WordNet üzerinden polarite değeri hesaplandığında objektif değerlere sahip olmasından ötürü tweet'lerden çıkartılmaması, duygu analizini olumsuz etkilememektedir. Ayrıca Sentimentor'ün TwitterSentiDetector'e göre diğer bir kısıtı ise duygu yükü bakımından oldukça önemli olan yüz ifadelerinin [41] metin önişleme sürecinde tweet'den temizlenmesiyle duygu analizinde kullanılmamasıdır.

Barbosa ve Feng [2], tweet içerisindeki kelimelerle ilgili çeşitli tanımlayıcı veri (meta-data) ve tweet'in nasıl yazıldığı ile ilgili söz dizimi (syntax) niteliklerinden faydalanarak tweet'lerde duyguları tespit etmeyi sağlayan bir yaklaşım öne sürmüşlerdir. Öne sürülen yaklaşım duygu analizini nesnellik tespiti ve öznellik tespiti olmak üzere iki aşamada yapmaktadır. Nesnellik tespiti sürecinde tweet'in duygu yükü açısından nesnel ya da öznel olup olmadığı belirlenmektedir. Öznellik tespitinde ise duygu yükü içeren tweet'lerin polaritesi tespit edilerek negatif ve pozitif olarak sınıflandırılmaktadır. Çalışmada kısaltmaların açık halleri elle çevrimiçi çeşitli kaynaklardan temin edilirken, tez

çalışmasıyla öne sürülen yaklaşımda çevrimiçi kısaltma sözlüğü kullanıldığından kısaltma veri tabanı sürekli büyümektedir. Ayrıca kullanılan çevrimiçin kısaltma sözlüğü kontrol edilerek sunulduğundan yanlış kısaltma açıklamalarının önüne geçilmektedir. İlgili çalışmada öznellik tespiti için 8.000 kaynak tanımı içeren bir veri setinden faydalanılırken, tez çalışmasında kullanılan sözlükler yaklaşık 135.000 kaynak tanımı bulundurarak çok daha kapsamlı bir bilgi tabanı sunmaktadır. Çalışmada duygu analizi için kullanılan "tweet içerisinde web adresi (link), hashtag içeriyor mu?", "tweet içerisinde büyük harfle başlayan kelime sayısı", "tweet bir kullanıcıya cevap veriyor mu?", "tweet alıntı (retweet) durumu" niteliklerinin duygu analizi açısından önem arz etmediği düşünülmektedir.

O'Connor ve arkadaşları [44], Twitter'daki verileri kullanarak fikir analizi ve anketler yapılmasını hedeflemişlerdir. İlgili çalışmada kullanılan dilbilimsel tabanlı doğal dil işleme yöntemlerinin tweet'lerde kullanılan ve dilbilim kurallarına aykırı olan, içerisinde sıklıkla kısaltmalar, harf ve noktalama işareti tekrarları, Twitter'e özel karakterler ve yüz ifadeleri barındıran "sosyal medya İngilizcesi"'ni analiz etmekte yetersiz kalacağı düşünülmektedir. Ayrıca çalışmada sınıflandırma yapmak için kullanılan 1.600 pozitif ve 1.200 negatif kelime içeren veri setinin duygu analizi açısından TwitterSentiDetector'ün kullandığı SenticNet, SentiWordNet, WordNet, AFINN gibi sözlüklerle karşılaştırıldığında oldukça kısıtlı kaldığı düşünülmektedir. TwitterSentiDetector, sosyal medyada kullanılan dilbilimsel kurallardan uzak dil kullanımından doğan çeşitliliklerin üstesinden gelecek şekilde tasarlanmıştır.

Agarwal ve arkadaşları [49], ağaç çekirdek (tree kernel) tabanlı bir yaklaşımla Twitter tabanlı duygu analizi yaklaşımı sunmaktadır. Öne sürülen yaklaşımda, tez çalışmasında olduğu gibi kısaltmaların açık anlamlarının elde edilmesi için çevrimiçi kısaltma sözlüğü olan NoSlang.com'dan faydalanılmıştır. Ancak Agarwal ve arkadaşları, çalışma esnasında NoSlang.com'da tanımlı olan kısaltmaları derleyerek 5.184 kısaltma içeren bir liste oluşturmuştur. Tez çalışmasında öne sürülen yaklaşımda ise bu çevrimiçi sözlük anlık olarak sorgulanmakta ve sabit bir kısaltma listesi kullanılmamaktadır. Böylelikle sürekli büyüyen, dinamik bir kısaltma bilgi tabanının elde edilmesi hedeflenmektedir. Agarwal ve arkadaşları, tweet'ler içerisinde tekrar eden karakter barındıran kelimelerdeki

tekrar sayısını sabit olarak üçe düşürmüşlerdir. Tez çalışmasında öne sürülen yaklaşımda ise tekrar düzenlemesi sabit bir sayıya değil; kelime anlamlı en sade hale gelene kadar yapılmaktadır. Öne sürülen çalışmanın bir kısıtı da kullanılan duygu sözlüğünün (Dictionary of Affect in Language (DAL)) 1989 yılında geliştirilmesinden ötürü mikrobloglarla beraber ortaya çıkan kelimelere yönelik duygu kuvveti (polarite) tanımı sunmamasıdır. Bu kelimeler, mikrobloglara özgü olduklarından ve mikrobloglarda sıklıkla kullanıldığından, tweet'lere yönelik duygu analizi yapılması için mutlaka değerlendirilmesi gerekmektedir. Bu sebepten ötürü TwitterSentiDetector, AFINN gibi mikrobloglarda kullanılan ve mikrobloglara özgü olan kelimeleri hedef alan duygu sözlüklerini kullanmaktadır.

Go ve arkadaşları [48], Naive Bayes, Maximum Entropy (MaxEnt), SVM gibi makine öğrenmesi algoritmalarını kullanarak Twitter üzerinde duygu analizi yöntemini öne sürmüşlerdir. TwitterSentiDetector, Go ve arkadaşlarının öne sürdüğü yaklaşımın aksine eğitime ihtiyaç duymamaktadır. Go ve arkadaşları, tweet'leri pozitif ve negatif olmak üzere iki sınıfta değerlendirmektedir. Nötr (objektif) tweet'lerin göz ardı edilmesi çalışmanın bir kısıtı olmaktadır. Tez çalışmasında öne sürülen yaklaşımda sınıflandırma yapılırken nötr duygu grubu da dikkate alınacaktır. Go ve arkadaşlarının sistem başarımını test ederken duygu yükü açısından oldukça anlamlı olan ve tweet'e ait nihai duygu yükü ile ilgili değerli bilgi barındıran yüz ifadelerini dikkate almamalarının, çalışmanın bir diğer kısıtı olduğu düşünülmektedir. Tez çalışmasında öne sürülen yaklaşımda yüz ifadeleri duygu analizi yapılırken kullanılmaktadır. Go ve arkadaşlarının öne sürdüğü yaklaşım, 174 pozitif ve 185 negatif kelime içeren sabit bir veri seti kullanılarak tweet'in içindeki kelimelerin negatif ve pozitif olarak sayılmasını temel almaktadır. Tez çalışmasında yaklaşık 120.000 kaynak tanımı barındıran ve dinamik bir bilgi bankası sunan duygu sözlüklerinden faydalanılarak daha kapsamlı bir bilgi bankasından duygu analizi yapılması hedeflenmiştir. Go ve arkadaşları, tweet'ler içerisindeki karakter tekrarlarını ikiye düşürmektedir. Tez çalışmasında öne sürülen yaklaşımda ise kelimeler WordNet üzerinden kontrol edilerek anlamlı en sade hale getirilmektedir. Raez ve arkadaşlarının [41] raporladığı gibi Go ve arkadaşları da Twitter API üzerinden veri çekerken “:(” ifadesini içeren aramalarda “:P” içeren tweet'lerin de sonuç olarak döndürüldüğünü raporlamıştır. İlgili çalışmada sonuç olarak SVM algoritmasının diğer makine öğrenmesi algoritmalarına (Naive Bayes, Maximum

Entropy) göre duygu analizinde daha başarılı sonuç verdiği raporlanmıştır. Ayrıca çalışmada, yüz ifadeleri ile sistemin eğitilmesi durumunda makine öğrenmesi algoritmalarının %80'den daha yüksek doğruluk oranına sahip olduğu raporlanmıştır.

Bravo-Marquez ve arkadaşları [84], çeşitli dilbilimsel sözlükleri ve duygu analiz metotlarından türetilen fikir kuvvetliliği, duygu ve polarite göstergelerini, analiz nitelikleri olarak değerlendiren eğiticili bir yaklaşım öne sürmüşlerdir. Çalışmada SVM, CART (Classification and Regression Tree), J48 ve Naive Bayes gibi makine öğrenmesi algoritmaları kullanılarak öne sürülen yaklaşım test edilmiş ve SVM'nin diğer algoritmalara göre daha başarılı sonuç verdiği çalışmada raporlanmıştır. Tez çalışmasında öne sürülen yaklaşım, eğitime ihtiyaç duymamaktadır. Böylelikle yetersiz ve homojen olmayan eğitimden kaynaklanabilecek hatalara maruz kalmaması hedeflenmiştir. Çalışmada, pozitif ve negatif ifadeleri beraber barındıran tweet'lerden nihai duygu tespiti yapılmasıyla ilgili net bir çözüm yolu olmadığı raporlanmıştır. Çalışmada, çok çeşitli duygu sözlüklerinden faydalanılmasına rağmen n-gram'larla ilgili duygu yükü tanımı sağlayan en kapsamlı duygu sözlüğü olan SenticNet'den faydalanılmamasının bir eksiklik olduğu düşünülmektedir.

SentiStrength [85], dilbilimsel (sözlüksel olmayan) etkenlerden de faydalanarak duygu analizinde bulunan sözlük tabanlı bir sınıflandırıcıdır. SentiStrength hem eğiticili hem de eğiticisiz olarak duygu analizi yapabilmektedir. TwitterSentiDetector gibi SentiStrength hem kelime hem de cümle seviyesinde duygu analizi yapmaktadır. SentiStrength, hem ikili (pozitif, negatif) hem de üçlü (pozitif, negatif, nötr) hedef duygu kümelerine yönelik duygu yükleri üretebilmektedir. SentiStrength, pozitif duygu kuvvetini [+1, +5] aralığında, negatif duygu kuvvetini ise [-1, -5] aralığında tanımlamaktadır. SentiStrength, versiyon 2 ile ön tanımlı güçlendirici, ters çevirici kelime listelerinden, duygu ifadelerinden ve noktalama işaretlerinden faydalanarak duygu polaritesi hesaplamaktadır. SentiStrength'in duygu analizi esnasında faydalandığı başlıca kriterler şu şekilde sıralanabilir: (1) İnsan kaynaklı duygu ve duygu kuvveti etiketlenmiş duygu kelime listesi, (2) duyguyu kuvvetlendiren ve azaltan destekleyici kelime listesi, (3) bazı yaygın kullanılan kelime veya kelime gruplarına yönelik duygu tespiti için deyim listesi, (4) duygu polaritesini ters çevirmek için kullanılan çevirici kelime listesi, (5) polarite tanımlı

yüz ifade listesi, (6) noktalama işaretleri ve (7) harf tekrarları.

Jiang ve arkadaşları [86], hedef (target) ve bağlam duyarlı (context-aware) bir yaklaşım öne sürerek tweet'lere yönelik duygu analizi tespiti yapılırken ilgili diğer tweet'ler de değerlendirmekte ve nihai duygu hedefe göre belirlenmektedir. Bu bağlam, incelenen tweet'i atan kullanıcının yakın zamanda aynı hedefle ilgili attığı tweet'leri, retweet yaptığı tweet'leri ve cevap verilen tweet'leri kapsamaktadır. Bu yaklaşımın özellikle kısa ve cevap niteliğinde olan tweet'leri analiz ederken olumlu sonuç vereceği düşünülmektedir; ancak tweet'lerin gündelik fikirleri yansıttığı ve insanların bir konuyla ilgili düşüncelerinin kısa zaman aralığında da değişebileceği göz önünde bulundurularak, "yakın zaman" tanımının doğru yapılması önem arz etmektedir. Aksi takdirde bağlam duyarlılığın, sınıflandırmaya olumsuz etki yapabileceği düşünülmektedir.

Saif ve arkadaşları [20] öne sürdükleri yaklaşımda duygu analizi yaparken kelimelere yönelik (1) benzer bağlamsal semantik ve duygu içeren ve (2) tweet'lerde birlikte kullanılan kavramları da değerlendirmektedir. Öne sürülen yaklaşım, geliştirilen SentiCircle isimli kelimelere yönelik kavramsal birliktelik taşıyan ifadeleri geometrik çember üzerinde trigonometrik kurallar kullanarak tanımlayan bir model tabanlıdır. Y ekseni kelimelere yönelik duyguları, X ekseni ise bu duygunun kuvvetini belirtmektedir. Öne sürülen yaklaşımın duygu analizi yaparken ön tanımlı varlık listesini kullanmasının tweet'lerdeki konu çeşitliliği göz önüne alındığında çalışmanın bir kısıtı olduğu düşünülmektedir.

Burnap ve arkadaşları [22] gerçek bir olaya yönelik Twitter'daki gerilimi tespit etmeyi sağlayan bir yaklaşım öne sürmüşlerdir. Öne sürülen yaklaşım, bir olayla ilgili üyelik kategorileri (membership categories) ve kategoriye bağlı aktiviteler (category bound activities) tanımlarına dayandığından, konuya ve konuyla ilgili kişilere bağımlı bir çözüm içermektedir. TwitterSentiDetector ise Twitter'daki içerik çeşitliliği göz önüne alınarak özellikle konudan bağımsız bir sistem olarak tasarlanmıştır. Burnap ve arkadaşlarının öne sürdüğü yaklaşım, "ırkçılığa" özgü ve yaşanmış bir olaydaki kişilere bağlı bir çözüm sunduğundan, kullanım alanı TwitterSentiDetector'e göre oldukça kısıtlıdır. Öne sürülen algoritmanın çalışması için olayla ilgili en az iki aktörün bulunması gerekmektedir. Bu

da gerçek hayattaki örnekler göz önüne alındığında çalışmanın kullanım alanını kısıtladığı düşünülmektedir. Bunun yanı sıra öne sürülen çalışmada sosyal medyada oldukça sık kullanılan ve duygu polaritesi açısından önemli değer taşıyan kısaltmaların değerlendirmeye alınmaması öne sürülen çalışmanın önemli bir başka kısıtıdır.

Vilares ve arkadaşları [5], İspanyolca tweet'lere yönelik sözlüksel ve sözdizimsel yöntemlerden faydalanarak polarite sınıflandırılması yapmayı sağlayan bir yaklaşım öne sürmüşlerdir. Çalışmada sözlüksel yaklaşımların yanı sıra sözdizimsel yaklaşımların kullanılmasının tweet'lere yönelik duygu analizini olumlu şekilde etkilediği gösterilmiştir. Öne sürülen yaklaşım temelde İspanyolca tweet'leri hedef alsa da İngilizce'ye de uyarlanmıştır. Çalışmada metin ön işleme aşamasında başta ve sonda yer alan hashtag'ler, TwitterSentiDetector'ün aksine tweet içerisinden temizlenmiştir. Özellikle sonda yer alan hashtag'lerin, tweet içerisindeki nihai duygu aktarmakta sıklıkla kullanıldığı gözlemlendiğinden bu yaklaşımın polarite analizinde olumsuz etkisinin olacağı düşünülmektedir. Ortada yer alan hashtag'ler, birden çok kelimeden oluşuyorsa ayrıştırılmadan doğrudan analiz yapılmaktadır. Bu durumda da ilgili hashtag'e yönelik duygu sözlüklerinden polarite karşılığı bulunamayacaktır. TwitterSentiDetector, birden çok kelime içeren hashtag'leri geliştirilen algoritma ile ilgili kelimelere ayrıştırmaktadır.

Liu ve arkadaşları [54], belirli alanlarla ilgili veri setleri kullanılarak eğitilen sınıflandırıcıların, diğer alanlarda iyi performans gösteremediğini ve duygu kelimelerinin polarite karşılıklarının konulara doğrudan bağlı olduğunu düşünerek Topic-Adaptive Sentiment Classification (TASC) isimli yarı eğitimli ve konu uyumlu duygu sınıflandırma modelini öne sürmüşlerdir. Öne sürülen model, duygu sınıflandırmaya ortak nitelikler ve çeşitli konularla ilgili duygu etiketli verilerinden başlamakta ve duygu kelimeleri, yazar duyguları ve tweet'de bahsedilen hesaplara (@-network) ağına bağlı olarak konu uyumlu bir niteliklerden faydalanarak nihai sınıflandırma yapmaktadır. Zhang ve arkadaşları [87], kullanıcıların duygularını, kullanıcıların geçmişteki duygu kayıtları ve sosyal ağlarını modelleyen dinamik sürekli faktör grafik modeli kullanarak kullanıcıların duygularını tahmin etmektedirler. Tan ve arkadaşları [88] ise yaptıkları çalışmada sosyal ilişkilerin kullanıcı seviyesindeki duygu analizini ilerletmek için kullanılabileceğini öne sürmüşlerdir. Liu ve arkadaşları, tweet seviyesindeki niteliklere yoğunlaşarak duygu

sınıflandırması yaparken, Tan ve arkadaşları kullanıcı seviyesindeki niteliklere yoğunlaşmışlardır.

Zhang ve arkadaşları [66], sözlük ve öğrenmeye dayalı yöntemlere dayalı Twitter duygu analizi yaklaşımı öne sürmüşlerdir. Bu yaklaşım, kaynaklara yönelik polarite tanımlarını duygu sözlüklerinin yanı sıra bağlamsal analizler kullanarak beraberinde kullanılan her bir kaynak, SVM kullanılarak otomatik olarak eğitime uğratılmış yüksek sayıdaki tweet içerisinde istatistiki olarak değerlendirilmekte ve buna göre duygu polarite tanımı yapılmaktadır. TwitterSentiDetector, eğitime ihtiyaç duymadan duygu analizi yapmak üzere tasarlanmıştır. Böylelikle, eğitim veri setinin kalitesizliğinden doğabilecek yanlış analizlerin önüne geçilmesi ve sosyal medyanın doğasına uygun olarak tanım kümesinden bağımsız duygu analiz yöntemi öne sürülmüştür. İlgili çalışmada, duygu yükü belirten hashtag'ler elle sisteme dahil edilirken, TwitterSentiDetector, duygu sözlükleri ve dilbilimsel metotlar tabanlı hashtag analizi yapmaktadır. Duygu belirten hashtag'lerin el ile sisteme dahil edilmesi Twitter'da kullanılan dil çeşitliliği göz önüne alındığında eksik kaynak tanımlama yapılmasına sebep vereceği düşünülmektedir. Bu sebepten ötürü ilgili çalışmanın aksine TwitterSentiDetector, sözlük tabanlı çözüm öne sürmektedir.

Ikram ve arkadaşları [53], açık kaynak kodlu ürünlere yönelik Twitter tabanlı duygu analizi sistemi öne sürmüşlerdir. Öne sürülen yaklaşım, tanım kümesi bağımlı olduğundan TwitterSentiDetector'e göre kullanım yeri oldukça dardır. Öne sürülen model, açık kaynak kodlu ürünlere yönelik duygu analizi öne sürdüğünden tanım kümesine özgü altı adet tanıma etkeni üzerine inşa edilmiştir. Bu durumun öne sürülen modelin Twitter'ın kullanıcı çeşitliliği ve yapısına uygun olan dinamik bir model değil, statik bir model olmasına sebep olduğu düşünülmektedir. Ayrıca çalışmada, sadece isimler ve sıfatlar duygu polaritesini belirleyen cümlenin öğeleri olarak kullanıldığından TwitterSentiDetector'de olduğu gibi bağlaçların duygu analizine olan etkisinden faydalanılmamaktadır. Bu durumun, anlam çevirici bağlaçlar/zarflar barındıran tweet'lerde, ilgili tweet'in polaritesi değerlendirilirken önemli bir eksikliğe sebep olacağı düşünülmektedir.

Kanavos ve arkadaşları [89], makine öğrenmesi metotlarının yanı sıra doğal dil işleme

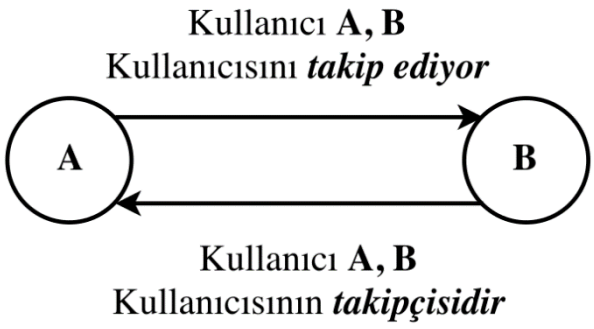
teknikleri ve Apache Spark projesinin makine öğrenmesi kütüphanesi olan MLlib'i kullanarak Twitter'da duygu sınıflandırması çalışması öne sürmüşlerdir. İlgili çalışmada vektör nitelikleri olarak (1) unigram, bigram ve trigram'ları, (2) tweet'in içerisinde kullanıcı adı, hashtag, URL olup olmamasını ve (3) tweet içeriğinin POS etiketleriyle analiz edilmesi sonucunda tweet'de bulunan isim, sıfat, zarf, fiil, ünlem sayıları kullanılmıştır. TwitterSentiDetector, ilgili çalışmadan farklı olarak duygu analizi açısından oldukça önemli olduğu düşünülen (1) four-gram'ları, (2) son zamanlarda Twitter'da akıllı telefonların popülerliği ve web istemcilerinde olan kullanım imkânı sebebiyle emoji'leri, (3) metinsel gülmeleri, (4) tweet içerisindeki nihai duyguyu değiştirebilen bağlaçları ve (5) kısaltmaları da duygu analizi esnasında dikkate almaktadır.

3. MATERYAL VE METOT

Bu bölümde, tez çalışmasının temellerini içeren bileşenlerden bahsedilecektir.

3.1. TWEET YAPISI

Tweet, Twitter kullanıcılarının paylaştığı durum mesajlarıdır. Tweet'lerin karakter boyutu 140 karakterle sınırlandırılmıştır. Tweet'ler metinlerin yanı sıra fotoğrafları, video'ları, web adreslerini (link) ve çeşitli sembollerin kombinasyonlarıyla elde edilen ve duygu belirten yüz ifadelerini içerebilmektedir. Ayrıca tweet'lerde, kullanıcılar ve konularla ilgili iletişimi ve bağlantıyı sağlamak üzere Twitter'a özel bazı semboller kullanılmaktadır. Çizelge 3.1'de bu semboller ve bu sembollerin tweet'lerdeki kullanım amaçları listelenmiştir. Twitter'da kullanıcılar, kendisiyle ilgili olduğunu düşündüğü diğer kullanıcı hesaplarını takip edebilmektedir. Twitter'da kullanıcı "*arkadaşlık, takip*" ilişkisi, diğer birçok sosyal medya platformunda olduğu gibi tek yönlü değil, çift yönlüdür. Kullanıcı, zaman tünelinde (timeline) listelenmiş olan tweet'leri (1) beğenebilmekte (like), (2) retweet fonksiyonunu kullanarak kendi zaman tünelinde görüntüleyebilmekte, (3) ilgili tweet'lere cevap (reply) verebilmektedir. Tüm bu etkileşimler, ilgili tweet'in kullanıcısına anında bildirilmektedir. Twitter kullanıcı takip ilişkisi, Şekil 3.1'de sunulmuştur. Ayrıca Twitter kullanıcıları, kendilerine özgü olan kullanıcı adlarının başına "@" karakteri ekleyerek diğer kişilerden bahsedebilmekte (mention) ve bu bahis diğer etkileşimlerde olduğu gibi ilgili kullanıcılara anında bildirilmektedir.



Şekil 3.1. Twitter kullanıcı takip ilişkisi.

Çizelge 3.1. Twitter'a özel semboller ve tweet içindeki kullanım amaçları.

Sembol	Kullanım Amacı
@	Twitter kullanıcılarından, kullanıcı adları vasıtasıyla bahsetmek ve onları Twitter tarafından sunulan bilgilendirme mekanizmasıyla haberdar etmek için kullanılır.
# (hashtag)	(1) Tweet'in hangi konuyla ilişkili olduğunu belirtmek ve (2) tweet içerisindeki nihai duyguyu veya düşünceyi özetlemek için kullanılır.
RT (retweet)	Tweet'lerin en başında kullanılan bu ifade, tweet'in başka bir kullanıcıdan alınarak kullanıcının kendi zaman tünelinde yayınlandığını belirtmek için kullanılır.

3.2. N-GRAM VE POLARİTE TANIMLARI

3.2.1. n-gram Tanımı

n-gram, bilişimsel dilbilimde (*computational linguistics*) bir cümleye ait ardışıl kelime ve kelime gruplarını tanımlar. *n,* kelime grup boyutunu tanımlar. *n* değeri 1 için unigram, *n* değeri 2 için bigram, *n* değeri 3 için trigram özel isimlendirmeler bulunmaktadır. *n*, 3'den büyük olduğunda ise kelime grupları, boyutla beraber adlandırılmaktadır. Örneğin *n* değeri 4 için oluşacak grup four-gram, *n* değeri 5 için five-gram olarak adlandırılmaktadır. n-gram örnekleri türleriyle beraber Çizelge 3.2'de listelenmiştir.

Çizelge 3.2. Türleriyle beraber n-gram örnekleri.

n-gram Türü	n-gram Örnekleri
four-gram	poor as a church mouse, live long healthy life
trigram	follow rule blindly, lack of confidence
bigram	run out, take care, take medicine, heart attack, trojan horse, living dead
unigram	good, lucky, ill, relieved, cancer

3.2.2. Polarite Tanımı

Metin tabanlı duygu analizi, metni oluşturan kelime ya da kelime gruplarının polaritesinin hesaplanması temeline dayanmaktadır. Erişime açık sözlükler, kelime ve kelime gruplarının polaritesi [-1, +1] aralığında ondalık bir değer olarak tanımlamaktadır. -1'e olan yakınlık kelimenin negatifliğini, +1'e olan yakınlık ise kelimenin pozitifliğini ifade etmektedir. Pozitif bir cümleye örnek "*Bugün hava çok güzel, çok mutluyum*" verilebilir. Negatif bir cümleye örnek ise "*iOS5'i hiç beğenmedim*" verilebilir. Nötr'lüğü ise 0 ve 0'a olan yakınlık belirtmektedir.

3.3. DUYGU SÖZLÜKLERİ

Duygu sözlükleri, duygu analizi yaklaşımları tarafından tweet'i oluşturan kelime ve kelime gruplarına yönelik polarite değeri hesaplamak için kullanılan sözlüklerdir. Duygu sözlükleri kelime ve kelime gruplarına yönelik duyguyu pozitif ve negatif kutuplarda derecelendirmek için polarite değerleri tanımlamaktadırlar. İlerleyen başlıklarda kelime ve kelime gruplarına pozitif, negatif ve nötr polarite değerleri tanımlayan erişime açık çeşitli sözlükler sunulmuştur.

3.3.1. WordNet

WordNet, temeli 1985 yılına dayanan, psikoloji profesörü George Miller tarafından

başlatılmış olan bünyesinde 155,287 İngilizce kelime barındıran elektronik İngilizce sözlüktür [90]. WordNet, her bir kelimeyi isim, fiil, sıfat ve zarf olarak dört kategoriye ayırmaktadır ve kelimeler arasındaki eş kümeleri (synsets) sunmaktadır. Bu eş kümeler Çizelge 3.3'de sunulan sözlüksel çeşitlere sahiptir.

Çizelge 3.3. WordNet kelime eş küme çeşitleri.

Eş Küme Adı	Açıklama	Örnek
hypernym	Kelimeler arası özel-genel ilişkisini tanımlar.	kangal-köpek
hyponym	Kelimeler arası genel-özel ilişkisini tanımlar.	köpek-kangal
meronym	Bütüne ait parça tanımı sağlar.	pencere-bina
holonym	Parça-bütün ilişkisini tanımlar.	bina-pencere
troponym	Kelimeler arası özelleştirme tanımı yapar.	kahkaha atmak – gülmek
entailment	Ön koşul tanımını sağlar.	uyumak - horlamak

3.3.2. SentiWordNet

SentiWordNet 3.0 [91], WordNet'den türetilmiş, kelimelerin eş küme grupları halinde toplandığı [92] alandan bağımsız erişime açık İngilizce sözlüktür. SentiWordNet 3.0'ın kaynağı, Esuli ve Sebastiani [93], [94] tarafından öne sürülen SentiWordNet sözlüğüdür. SentiWordNet, WordNet'de tanımlı olan tüm eş kümelere [0, 1] aralığında pozitif, negatif ve objektif polarite değerleri tanımlar. SentiWordNet, 117,659 kaynağa duygu tanımı yapmaktadır [42]. SentiWordNet, SenticNet'den farklı olarak, üzerinde tanımladığı kaynaklara yönelik polarite değerini [0, 1] aralığında yapmaktadır [54]. Tez çalışmasıyla öne sürülen sistem, tweet içindeki n-gramları -1 negatifi, 0 nötr'lüğü ve +1 pozitifliği belirtmek üzere [-1, +1] aralığında tanımlandığından SentiWordNet'den elde edilen

polarite değerlerinin [-1, +1] aralığına uygun şekilde genişletilmesi gerekmektedir. Bunun için her bir kaynağa ait tanımlı olan eş kümelerin polarite değerleri ağırlıklı olarak toplanarak [-1, +1] aralığında nihai bir polarite değeri elde eden bir Java sınıfından[4] faydalanılmıştır. SentiWordNet, yeni sürümü olan SentiWordNet 3.0'da sunulan rastlantısal yürüyüş (random walk) yaklaşımı ile veri setindeki polarite değerlerinin geliştirildiği öne sürülmektedir [91]. SentiWordNet, duygu analizi araştırmacıları tarafından tanım kümesinden bağımsız olarak kavramların duygu boyutları hakkında bilgi sağlaması nedeniyle kullanılmaktadır [95], [96]. SentiWordNet, duygu analizini sentetik olarak ele aldığından sadece unigram'lara polarite tanımı yapmakta, kelime gruplarını (n-gram'lar) bir bütün olarak değerlendirememektedir. Böylece "*a lot*", "*a little*", "*take care*", "*32 teeth*" gibi tweet'lerde oldukça sık kullanılan ve bütün anlamıyla kelime kelime işlenmesiyle çıkacak anlam farklılıklarının ayırt edilebilmesini sağlayamamaktadır. Bu sebepten ötürü tek başına kullanılarak cümlelere yönelik duygu analizi yapılmasının yeterli olmayacağı, yüksek yanlış tespit oranına sebebiyet vereceği düşünülmektedir.

3.3.3. SenticNet

SenticNet [97], [98], sentik hesaplama yaklaşımları kullanarak kavramlara polarite değerleri atayan erişime açık bir duygu sözlüğüdür [99], [100]. SenticNet 2, üzerinde anlamsal ve duygu tabanlı bilgilerle ilişkilendirilmiş 14,244 kaynak tanımı barındırmaktadır [98]. SenticNet, SentiWordNet'den farklı olarak, unigram'ların yanısıra bigram, trigram ve four-gram'lara da polarite tanımı yapmaktadır. Kaynaklar, üzerindeki tanımlama ve ilişkilerin makinalar tarafından okunabilir olması için SenticNet üzerindeki veriler anlamsal kaynak tanımlama biçimi olan RDF/XML formatında tanımlanmıştır. Java tabanlı açık kaynak kodlu bir yazılım kütüphanesi olan Apache Jena[5], RDF (Resource Description Framework), RDFS (Resource Description Framework Schema), OWL (Web Ontology Language), bellek, ilişkisel veritabanları ve Triple Store gibi çeşitli semantik web veri kaynakları üzerinde oluşturma, okuma, güncelleme ve oluşturma (Create, Read, Update, Delete - CRUD) işlemleri yapmayı sağlamaktadır [101]. SenticNet, bir RDF notasyonu olan RDF/XML tabanlı bir sözlük olduğundan n-gram'ların polarite değerleri Apache Jena kullanılarak SenticNet'den elde

[4] http://sentiwordnet.isti.cnr.it/code/SentiWordNetDemoCode.java
[5] https://jena.apache.org

edilebilmektedir. SenticNet tanımladığı kaynakların polarite değerini [-1, 1] arasında yapmaktadır. SenticNet, polarite değerinin yanı sıra hoşnutluk (pleasantness), ihtimam (attention), hassaslık (sensitivity) ve eğilim (aptitude) tanımları da yapmaktadır. XML (Extensible Markup Language) formatında veri tanımlamanın JSON (JavaScript Object Notation) gibi diğer veri sunum formatlarına göre en büyük dezavantajı, Şekil 3.2'de sunulduğu üzere veri etiketlerinin tanımlanmak istenen veriden çok daha fazla olması sebebiyle veri boyutunun oldukça büyümesidir. Bu sebepten ötürü, RDF/XML formatında tanımlı olan kaynaklar, Apache Jena kütüphanesi aracılığıyla çıkartılıp, polarite değerleriyle birlikte ilişkisel veritabanı yönetim sistemi (relational database management system - RDMS) üzerinde depolanmıştır. Böylelikle kaynaklara yönelik polarite sorgu süresi ortalama 120 saniyeden ortalama 10 saniyeye düşürülerek sorgu performansı yaklaşık 12 kat hızlandırılmıştır. Ayrıca bu işlemle SenticNet veri boyutu RDF/XML formatında 39.2 MB (megabyte) iken ilişkisel veritabanı sorgu dili (Structured Query Language – SQL) formatında 1.2 MB'a düşerek yaklaşık 33 kat boyuttan kazanç elde edilmiştir.

```
<rdf:Description rdf:about="http://sentic.net/api/en/concept/32_teeth">
    <rdf:type rdf:resource="http://sentic.net/api/concept"/>
    <text xmlns="http://sentic.net/api">32 teeth</text>
    <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/
    snot"/>
    <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/
    bad_taste"/>
    <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/
    mouth_part"/>
    <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/
    gum"/>
    <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/
    near_mouth"/>
    <pleasantness xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/
    XMLSchema#float">0.965</pleasantness>
    <attention xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/
    XMLSchema#float">0.775</attention>
    <sensitivity xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/
    XMLSchema#float">0</sensitivity>
    <aptitude xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/
    XMLSchema#float">0.968</aptitude>
    <polarity xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/
    XMLSchema#float">0.903</polarity>
</rdf:Description>
```

Şekil 3.2. SenticNet üzerinde RDF/XML formatında örnek bir kaynak tanımı.

3.3.4. AFINN

Affective Norms for English Words (ANEW), Bradley ve Lang [102] tarafından geliştirilen ve insanların kelimelere yönelik psikolojik tepkilerine göre [84] 1,034 İngilizce kelimeye duygu yükü yükleyen bir sözlüktür. ANEW, mikrobloglar popüler olmadan önce geliştirilmesinden ötürü, sosyal medyada sıklıkla kullanılan kısaltmalara yönelik kaynak tanımı yapmamaktadır. AFINN, Nielsen [103] tarafından ANEW sözlüğünün, mikrobloglarda sıklıkla kullanılan kelimeler ve kısaltmaları da dahil edecek şekilde genişletilmesiyle elde edilen ve özellikle mikrobloglarda kullanılan dili hedef alan duygu sözlüğüdür. AFINN, bünyesinde 2,477 İngilizce kelimeye yönelik [-5, +5] aralığında polarite tanımı barındırmaktadır. Bravo-Marquez ve arkadaşları, AFINN sözlüğünün boyut olarak SentiWordNet gibi duygu sözlüklere göre küçük olmasına rağmen diğer sözlüklerde olmayan mikroblog kısaltmalarının AFINN'de tanımlı olduğunu raporlamışlardır [84]. Pozitif kelimeleri duygu yüküne göre [+1, +5] aralığında, negatif kelimeleri ise duygu yüküne göre [-5, -1] aralığında tanımlamaktadır. Negatiflik duygu yükü -1'den -5'e doğru, pozitiflik duygu yükü ise +1'den +5'e doğru kuvvetlendirilmektedir. Tez çalışmasında pozitif ve negatif duygu kutupları için kaynak tanımının [-1, +1] aralığında yapılması hedeflendiğinden ötürü bu sözlüğün kaynak tanımında kullanılabilmesi için değer kümesinin [-1, +1] aralığına göre ölçeklendirilmesi gerekmektedir. Bunun için *P(Scaled-AFINN)* ölçeklendirilmiş yeni polarite değerini ve *P(AFINN)* sözlükte tanımlı polarite değerini göstermek üzere, Denklem (3.1)'deki gibi bir ölçeklendirmeden faydalanılmıştır:

$$P(Scaled - AFINN) = \frac{P(AFINN)}{5} \tag{3.1}$$

AFINN veri seti, düz metin dosyasında anahtar-değer ikilileri şeklinde erişime sunulmuştur. Sorgu performansının artırılması için, veri setindeki anahtar ve değerler ilişkisel veritabanı sistemi üzerinde bir tabloda depolanmıştır.

3.3.5. SentiStrength

SentiStrength, ek dilbilimsel bilgi ve kurallardan faydalanarak kısa ve dilbilimsel olarak kuralsız İngilizce metinlere yönelik duygu tespiti yapan sözlük tabanlı duygu analiz sistemidir [85]. SentiStrength, çevirici kelime listesi, deyim listesi, tekrarlı harf ve noktalama işaretleri, deyim listesi ve polarite tanımı yapılmış yüz ifade listesinden gibi dilbilimsel yaklaşımları da değerlendirerek 2.489 kaynağa yönelik duygu tanımı barındırmaktadır [85]. Sözlük, girdi olarak tanımlanan her bir kelime ya da cümleye karşılık pozitif, negatif ve nötr duygu yükü tanımlamaktadır. Pozitif polarite değerlendirmesini [+1, +5], negatif değerlendirmesini [-1, -5], nötr değerlendirmesini -1, 0, +1 olarak yapmaktadır. TwitterSentiDetector, tweet'lere yönelik duygu yükü hesaplanırken, n-gram polarite değerlerini [-1, +1] aralığında alınmasını öne sürdüğü için, hesaplanan polarite sonucunun Denklem (3.1) 'de olduğu gibi sadeleştirilmesi gerekmektedir. Ayrıca SentiStrength her bir kaynağa karşılık pozitif, negatif ve nötr polarite değerleri hesapladığından dolayı TwitterSentiDetector, SentiStrength'den faydalanarak yapılan kelime ve kelime gruplarına yönelik polarite hesabında en kuvvetli polarite değerini dikkate almaktadır. Thelwall ve arkadaşları, SentiStrength üzerinden kelime ya da cümle bazlı duygu tespiti yapmak için uygulama geliştirme arayüzü sunmaktadırlar.

3.3.6. NoSlang.com

lol, tyvm gibi kısaltmalar Twitter kullanıcıları tarafından, Twitter'da tanımlı olan 140 karakter limiti olması sebebiyle tweet'lerde yer kazanmak amacıyla sıklıkla kullanılmaktadır [41], [48]. Kısaltmaların barındırdıkları anlamlar sebebiyle tweet'lerin duygu analizi yapılırken göz ardı edilmemesi, doğru duygu tespiti açısından önem arz etmektedir. Yaptığımız araştırmalara göre, kısaltmalara yönelik kelime ve kelime gruplarına polarite değeri sunan erişime açık sözlük bulunmamaktadır. Ancak çeşitli web siteleri kısaltmaların uzun formlarını paylaşmaktadır. NoSlang.com[6], Şekil 3.3'de sunulan web tabanlı arama arayüzü ile kısaltmaların açık hallerini sunan çevrimiçi kısaltma sözlüğüdür.

[6] http://www.noslang.com



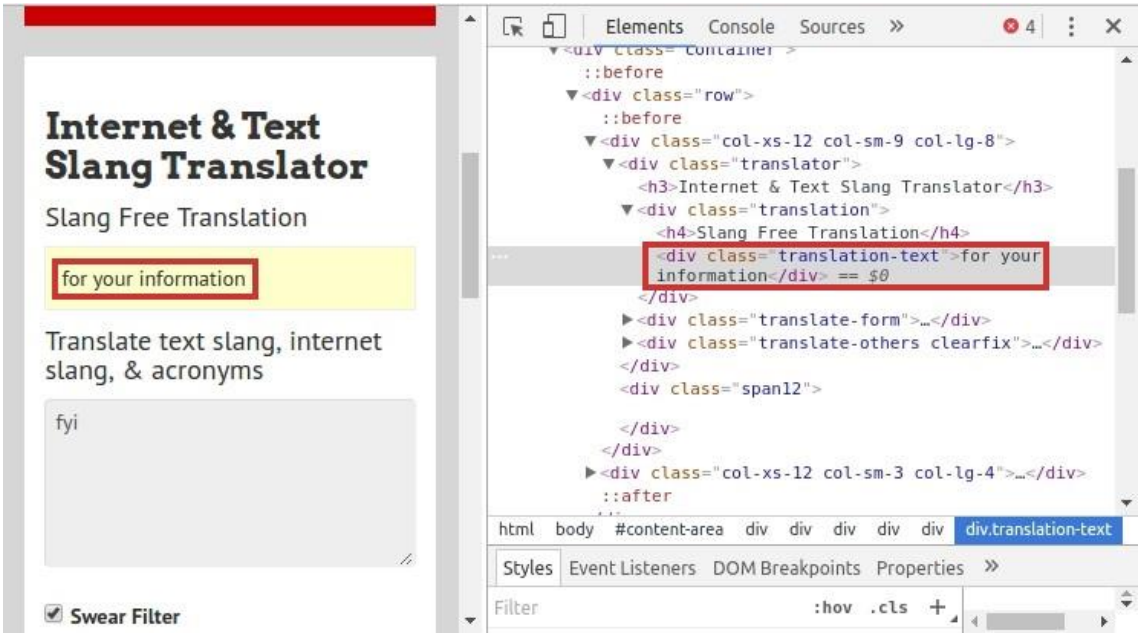
Şekil 3.3. NoSlang.com kısaltma arama web arayüzü.

NoSlang.com, kullanıcılarına kısaltma ekleme imkânı sağlaması ve bu kısaltmaları doğruladıktan sonra yayınlaması sebebiyle diğer çevrimiçi kısaltma sözlüklerinden ayırt edilmektedir. NoSlang.com, kontrollü ve sürekli büyüyen bir veritabanına sahip olmasından dolayı tez çalışmasında tercih edilmiştir. Çizelge 3.4'de NoSlang.com'dan faydalanılarak yapılan örnek kısaltma dönüşümleri sunulmuştur.

Çizelge 3.4. NoSlang.com sözlüğü kullanılarak yapılan örnek kısaltma dönüşümleri.

Kısaltma	Açık Hali
lol	laughing out loud
gr8	great
bff	best friend forever
thx	thanks, thank you
fyi	for your information
rite	right

NoSlang.com, üzerinde tanımlı olan kısaltma veritabanını erişime açık bir şekilde sunmadığından ve sorgulama için bir uygulama geliştirme arayüzü veya web servis sağlamadığından dolayı tez çalışması kapsamında NoSlang.com üzerinde arama yapıp, sonuç sayfasındaki HTML (Hypertext Markup Language) kodunu ayrıştıran bir servis geliştirilmiştir. Bu servis aracılığıyla parametre olarak verilen kısaltmaların açık formları NoSlang.com üzerinden elde edilmektedir. Bu servis, "`<abbr>`" etiketi içerisinde açık hali verilen kısaltmaları sonuç sayfasından ayrıştırmaktadır. NoSlang.com üzerinde kısaltmanın ve açık formunun HTML kodlarında nasıl kodlandığı Şekil 3.4'de sunulmuştur.



Şekil 3.4. NoSlang.com üzerinde örnek bir kısaltmanın ve açık formunun kodlanması.

Aranan kısaltmanın NoSlang.com üzerinde bulunamaması durumunda ise Şekil 3.5'de gösterilen ve "`translation-text`" türünde özel bir CSS (Cascading Style Sheets) sınıfıyla etiketlenerek üzerinde "*None of the words you entered are in our database. Found a word we're missing? Add it to our dictionary.*" ifadesi bulunan sonuç sayfası görüntülemektedir. Tez çalışması kapsamında, NoSlang.com üzerinden kısaltma sorgusu için geliştirilen servis ayrıştırma sonucunda yukarıdaki ifadeyle karşılaşıldığında, aranan kısaltmaya yönelik sonuç bulunamadığı kabul edilmiştir.



Şekil 3.5. NoSlang.com üzerinde aranan kısaltmaya yönelik sonuç bulunamaması durumunda görüntülenen web arayüzü.

Tez çalışmasında kullanılan duygu sözlüklerinin kaynak sayısı, polarite değer aralığı ve kaynak türü açısından karşılaştırılması Çizelge 3.5'de sunulmuştur.

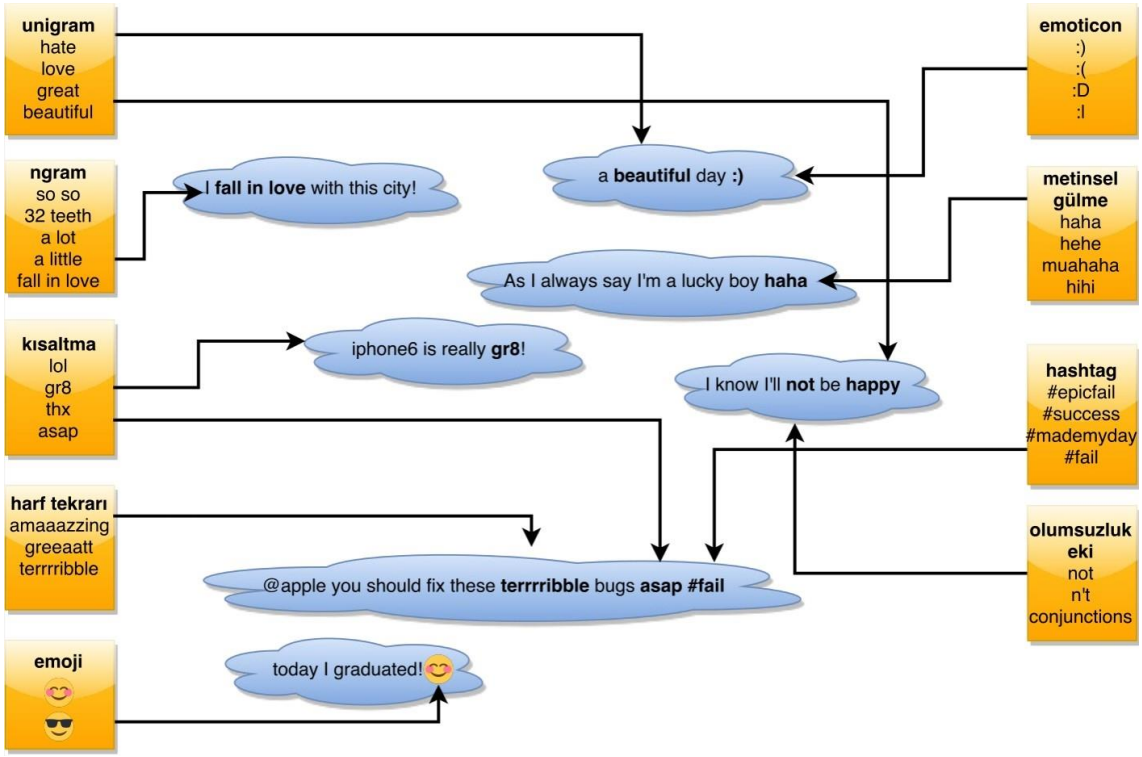
Çizelge 3.5. Tez çalışmasında kullanılan duygu sözlüklerinin karşılaştırılması.

Duygu Sözlüğü	Kaynak Sayısı	Polarite Değer Aralığı	Kaynak Türü
SentiWordNet	117,659	[0, +1]	unigram
SenticNet	14,244	[-1, +1]	unigram, n-gram
AFINN	2,477	[-5, +5]	unigram
SentiStrength	2,489	[-5, +5]	unigram

4. TWITTER TABANLI DUYGU ANALİZİ

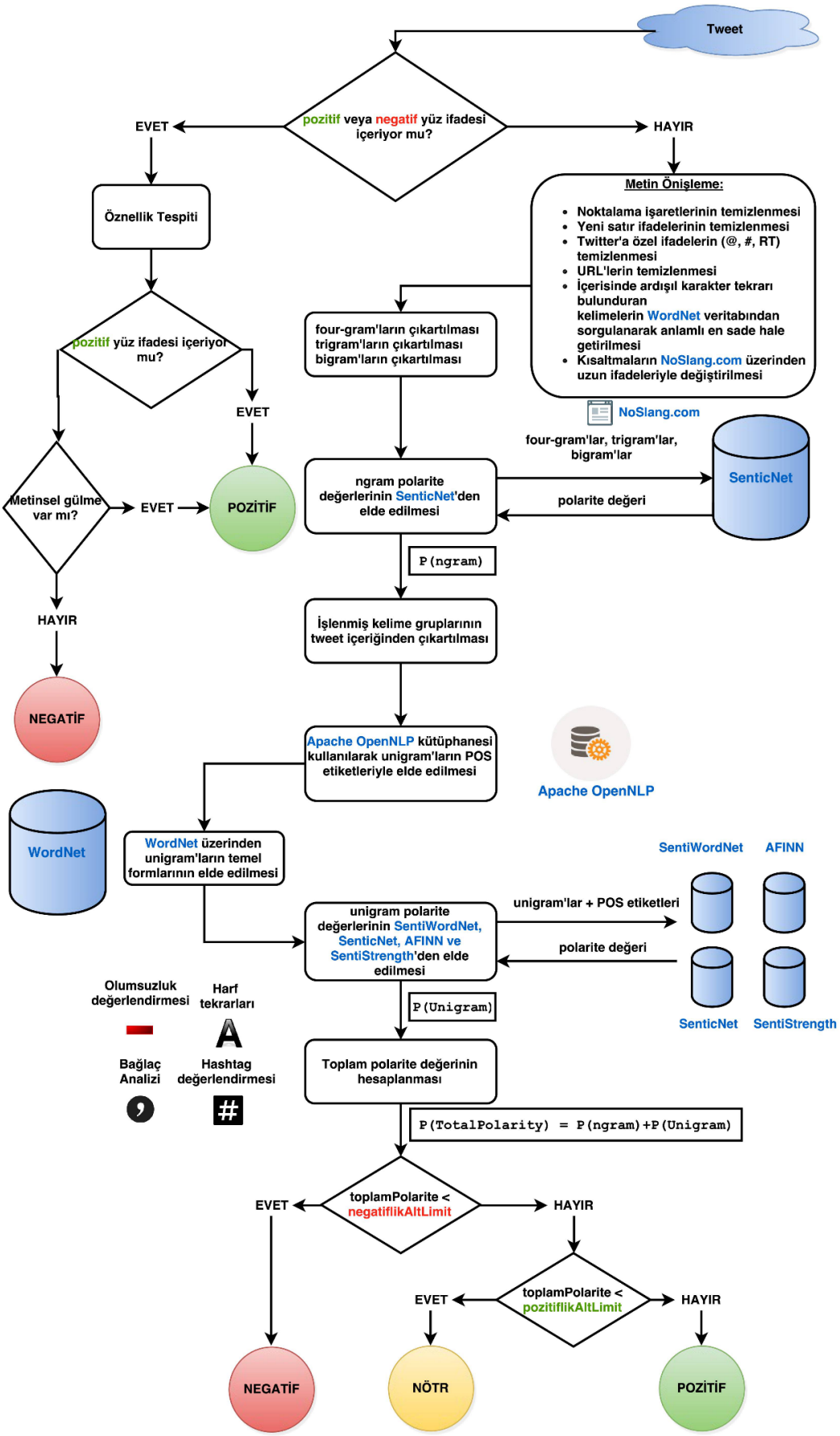
TwitterSentiDetector, tez çalışması kapsamında geliştirilen, hedeften (target) ve bağlamdan (context) bağımsız, doğal dil işleme ve dilbilimsel kural tabanlı eğiticisiz bir Twitter tabanlı duygu analizi sistemidir. Türkçe'de Zemberek[7] gibi doğal dil işleme kütüphaneleri olmasına rağmen henüz yeterli kararlılığa ulaşmaması ve literatürde geçerliliği olan duygu sözlüklerinin İngilizce olması sebebiyle, tez çalışmasında geliştirilen sistemin kaynak dili İngilizce olarak belirlenmiştir. Bu sebeple çalışmanın ismi "Twitter Sentiment Detector (Twitter Duygu Algılayıcısı)" kısaltmasını belirtmek üzere TwitterSentiDetector olarak belirlenmiştir. TwitterSentiDetector, geleneksel duygu analizi yöntemleriyle analiz edilemeyecek kısaltmalar, harf ve noktalama işareti tekrarları, yüz ifadeleri, hashtag'ler, dilbilimsel kurallardan uzak kelime ve cümleler gibi mikrobloglara özel kullanım dilindeki farklılıkları göz önünde bulundurularak geliştirilmiştir. TwitterSentiDetector, bir tweet'e yönelik duygu analizi yaparken unigram'lardan, n-gram'lardan (bigram, trigram ve four-gram), görsel yüz ifadelerinden, noktalama işaretleriyle oluşturulan gülücük sembollerinden, metinsel gülmelerden, hashtag'lerden, kısaltmalardan, karakter tekrarlı (vurgulu) kelimelerden ve duygu çevirici ek ve kelimelerden (olumsuz ekleri ve bağlaçlar) faydalanmaktadır. Twitter'da kullanılan dilbilimsel kurallardan uzak dil dikkate alınarak daha duyarlı duygu analizi gerçekleştirebilmek için söz dizimi ve bağlam duyarlı yazım hatası düzeltme modülü sisteme entegre edilmiştir. Twitter'da oldukça yaygın olan ve amaçlarından birisi de Twitter'da çalışan servislerin çalışmasını manipüle etmek olan ilgisiz tweet'ler, duygu analizinin daha doğru bir şekilde yapılabilmesi için geliştirilen spam tespit modülüyle duygu analizi süreci öncesinde filtrelenmiştir. Şekil 4.1'de tez çalışması kapsamında geliştirilen TwitterSentiDetector sisteminin duygu analizinde kullandığı nitelikler örneklerle görselleştirilmiştir.

[7] https://code.google.com/p/zemberek/



Şekil 4.1. TwitterSentiDetector sisteminin duygu analizi için kullandığı niteliklerin örneklerle gösterimi.

Duygu analizi sistemleri tarafından kelime ve kelime gruplarını duygu anlamı (polarite) yüklemek için duygu sözlükleri kullanılmaktadır. Duygu sözlüklerinin üzerinde tanımlı olan kaynaklara yönelik polarite tanım aralığı değişkenlik göstermekle beraber genel eğilim [-1, +1] aralığıdır. Bu sebepten ötürü TwitterSentiDetector de, -1 negatifliği, 0 nötr'lüğü ve 1 pozitifliği belirtmek üzere tweet'lere yönelik polarite değeri [-1, +1] aralığında hesaplayarak nihai duygu değerlendirmesini pozitif, negatif ve nötr olmak üzere üç sınıfta yapmaktadır. Bir tweet'in işlenerek duygu tespitinin yapılması için TwitterSentiDetector Şekil 4.2'de sunulan ardışıl süreçlere (boru hattı – pipeline) sahiptir. TwitterSentiDetector tweet'lere yönelik duygu analizini yaparken üç temel sürece sahiptir: (1) Hızlı duygu analizi, (2) tweet ön işleme ve (3) doğal dil işleme ve dilbilimsel yöntemler kullanılarak duygu sözlükleri tabanlı duygu analizi. Her bir süreç ilerleyen alt başlıklarda sunulmuştur.



Şekil 4.2. TwitterSentiDetector iş akışı önizlemesi.

4.1. Twitter Uygulama Geliştirme Arayüzleri (API)

Twitter, sunduğu REST (REpresentational State Transfer) tabanlı arama uygulama geliştirme arayüzü aracılığıyla erişime açık (public) olan tweet'lerin sorgulanmasını sağlamaktadır. TwitterSentiDetector, Twitter REST web servisleriyle haberleşmek için geliştirilmiş olan Twitter4J[8] isimli Java kütüphanesini kullanmaktadır. Twitter arama servisi, son bir hafta içerisinde yayınlanmış tweet'ler arasından örneklem veri döndürmektedir [104]. Twitter, yayınlandığı dokümantasyonda arama uygulama geliştirme arayüzü ile yapılan aramalarda içerik açısından yakınlığa odaklanıldığını, konuyla ilgili tüm tweet'lere ulaşmak için duraksız iletim (streaming) uygulama geliştirme arayüzünün kullanılması gerektiğini belirtmektedir [104]. Twitter uygulama geliştirme arayüzleri, Çizelge 4.1'de listelendiği üzere gelişkin arama (query) işlem operatörlerine sunmaktadır. Twitter üzerinde arama, sunulan çok çeşitli arama işlem operatörleri ile detaylı bir şekilde yapılabilmektedir. Twitter sunduğu uygulama geliştirme arayüzleri ile Twitter üzerinden veri çekilmesine imkân sağlamaktadır. Bu amaçla Twitter'da arama (Search API) ve duraksız iletim uygulama geliştirme arayüzleri (Streaming API) sunulmaktadır. Arama uygulama geliştirme arayüzü Twitter'da indekslenen son bir haftada atılmış tweet'ler üzerinde arama yapılmasını sağlarken, duraksız iletim uygulama geliştirme arayüzü ise gerçek zamanlı (real-time) olarak Twitter'da atılan tweet'lerin çekilmesini sağlamaktadır. Bu uygulama geliştirme arayüzleri, RESTful web servisler ile hizmet verdiğinden programlama dillerinden bağımsız bir şekilde Twitter'la iletişim kurulmasını sağlamaktadır. TwitterSentiDetector, hem arama uygulama geliştirme arayüzü kullanarak son bir haftalık veri üzerinde hem de duraksız iletim uygulama geliştirme arayüzünü kullanarak gerçek zamanlı veriler üzerinde duygu analizi yapılması için servisler barındırmaktadır.

[8] http://twitter4j.org

Çizelge 4.1. Twitter uygulama geliştirme arayüzü işlem operatörleri.

İşlem Operatörü	Aranacak Tweet'ler
Windows 10	İçerisinde hem “Windows” hem de “10” ifadesi bulunan tweet'ler
“Windows 10”	İçerisinde tam olarak “Windows 10” ifadesi bulunan tweet'ler
Windows OR 10	İçerisinde “Windows” ve/veya “10” ifadesi bulunan tweet'ler
Windows -10	İçerisinde “Windows” bulunup “10” ifadesi bulunmayan tweet'ler
from:xyz	xyz kullanıcısından atılan tweet'ler
to:xyz	xyz kullanıcısına cevap olarak atılan tweet'ler
@xyz	İçerisinde xyz kullanıcısı geçen tweet'ler
Android since:2016-01-26	26 Ocak 2016 tarihinden sonra atılmış olan ve içerisinde “Android” ifadesi geçen tweet'ler
Android until:2016-01-26	26 Ocak 2016 tarihine kadar atılmış olan ve içerisinde “Android” ifadesi geçen tweet'ler
Android lang:tr	İçerisinde “Android” ifadesi bulunan ve Türkçe olarak yazılmış tweet'ler
Android result_type=recent	İçerisinde “Android” ifadesi bulunan yakın zamanda yazılmış tweet'ler
Android result_type=popular	İçerisinde “Android” ifadesi bulunan popüler tweet'ler
Android result_type=mixed	İçerisinde “Android” ifadesi bulunan popüler veya yakın zamanda atılmış tweet'ler

4.2. Twitter API Limit Kontrolü

Duraksız iletim uygulama geliştirme arayüzü, aranan sorguyla ilgili o andan itibaren kayıpsız olarak tüm tweet'lerin sorgulanmasını sağlamaktadır. Sorgu limiti, kullanıcı erişim jetonu (access token) kullanılırsa 15 dakikada 180 istek, uygulama bazlı erişim jetonu kullanılması durumunda 15 dakikada 450 istektir. 15 dakikalık süre zarfında tanımlanan limitlerin aşılması durumunda uygulama geliştirme arayüzü isteklere cevap vermemektedir. Bu limitin sıklıkla aşılması durumunda, yetkilendirilen kullanıcı veya uygulama kara listeye alınabilmektedir [105]. TwitterSentiDetector, Twitter uygulama geliştirme arayüzlerini kullanarak tweet çekerken, Twitter'ın da tavsiye ettiği şekilde her işlem öncesinde Şekil 4.3'de sunulan metodu kullanarak limitlerin aşılıp aşılmadığını sorgulamaktadır. Limit sorgu servisi, kalan sorgu adetini ve sorgu periyodunun yenilenmesi için kalan süre (saniye) bilgisini döndürmektedir. Limitlerin aşılması durumunda yeni bir sorgu yapılmadan önce tespit edilen kalan süre kadar beklenmektedir.

```
private Integer getRemainingAppRateLimitStatus(Twitter twitter) {
    Integer remainingRateLimit = 0;
    try {
        Map<String, RateLimitStatus> rateLimitStatus = twitter.getRateLimitStatus();
        for (String endpoint : rateLimitStatus.keySet()) {
            if (endpoint.equals("/application/rate_limit_status")) {
                RateLimitStatus appRateLimitStatus = rateLimitStatus.get(endpoint);
                remainingRateLimit = appRateLimitStatus.getRemaining();
                logger.debug("Remaining rate limit: " + remainingRateLimit);
                logger.debug("Seconds until reset: " +
                        appRateLimitStatus.getSecondsUntilReset());
            }
        }
    } catch (TwitterException twitterException) {
        twitterException.printStackTrace();
        logger.error("Unable to get rate limit status!!");
        logger.error("Exception Details: " + twitterException.getMessage());
    }
    return remainingRateLimit;
}
```

Şekil 4.3. TwitterSentiDetector'de kullanılan kalan Twitter sorgu limitini sorgulayan Java diliyle yazılmış metot.

4.3. Hızlı Duygu Analizi

Duygu analizinin ilk aşaması olan bu aşamada, doğal dil işleme ve dilbilimsel metotlar kullanılarak detaylı duygu analizi yapılmadan, tweet içerisinde kullanılan ve nihai duyguyu belirttiği düşünülen, duygu analizi açısından anlamlı olan harf ve noktalama

işaretleri tabanlı yüz ifadeleri (emoticon), emoji ve metinsel gülmelerin tespiti yapılmaktadır [48], [66]. Yapılan araştırmalar, insanların yüz yüze iletişim kurarken sözsel olmayan ifadelerden faydalandığını belirtmektedir [106], [107]. Yüz ifadeleri bu doğrultuda metinsel olarak ince ruh hallerini ifade etmeyi, ironi, iğneleme anlamı yüklemeyi, duygu kuvvetlendirmeyi mümkün kılmaktadır [107]. Bunun için TwitterSentiDetector aşağıdaki iş akışını uygulamaktadır:

- Nesnellik Tespiti (Objectivity Detection): Tweet içinde yüz ifadesi, metinsel gülme veya emoji olup-olmadığının tespiti yapılır.

Öznellik Tespiti (Subjectivity Detection): Çizelge 4.2'deki Wikipedia tabanlı [108] yüz ifadesi listesinden faydalanılarak ilgili tweet'deki yüz ifadesinin hangi duyguyu yansıttığı tespit edilir. Yüz ifadelerinin yanı sıra,

- Çizelge 4.3'de sunulan emoji-yüz ifadesi eşleşme tablosu kullanılarak emojiler de TwitterSentiDetector tarafından duygu analizinde kullanılmaktadır. Çizelge 4.4'de Raez ve arkadaşları [41] tarafından öne sürülen metinsel gülme kelimeleri ile gülme tespiti yapılmakta ve ilgili tweet pozitif olarak sınıflandırılmaktadır.

Çizelge 4.2. Wikipedia'da tabanlı pozitif ve negatif olarak sınıflandırılan yüz ifadeleri.

Sınıflandırma Grubu	Yüz İfadeleri
Pozitif olarak değerlendirilen yüz ve metin ifadeleri	:) :-) =) =-) :D =D :-D =-D ;) ;-) ^-^ ^^ =d :d =-d :-d (= (: (-= (-: ♥ <3 :} :-} ;} ;-} :] :-] :o) :o) :3 :c) :> =] :^) :つ) 8-D 8D x-D X-D XD =-D =-3 =3 B^D :'-) :')
Negatif olarak değerlendirilen yüz ifadeleri	:(:-(=(=-(:'(='(;(;-(:{ :-{ :| :-| :[:-[>:[:-c :c :-< :つC :< :-|| :@ >:(:'-(:'(

Çizelge 4.3. TwitterSentiDetector tarafından değerlendirilen emojiler, UTF-8 karşılıkları ve yansıttıkları duyguya yönelik gülücük eşleşme listesi.

Emoji	UTF-8 Karşılığı	Gülücük Eşleşmesi	Emoji	UTF-8 Karşılığı	Gülücük Eşleşmesi
😉	f09f9889	;)	😁	f09f9881	:)
😊	f09f988a	:)	😂	f09f9882	:D
😜	f09f989c	:P	😃	f09f9883	:)
😄	f09f9884	:D	😅	f09f9885	:D
😆	f09f9886	XD	😉	f09f9889	;)
😊	f09f988a	:)	😋	f09f988b	:P
😌	f09f988c	^^	😍	f09f988d	<3
😏	f09f988f	:)	😒	f09f9892	:|
😓	f09f9893	:(	😔	f09f9894	:(
😖	f09f9896	:(	😘	f09f9898	<3
😚	f09f989a	<3	😜	f09f989c	:P
😝	f09f989d	:P	😞	f09f989e	:(

Emoji	UTF-8 Karşılığı	Gülücük Eşleşmesi	Emoji	UTF-8 Karşılığı	Gülücük Eşleşmesi
	f09f98a0	:@		f09f98a1	:@
	f09f98a2	:'(		f09f98a3	:(
	f09f98a4	:@		f09f98a5	:'(
	f09f98a8	:(		f09f98a9	:(
	f09f98aa	:(		f09f98ab	:(
	f09f98ad	:'(		f09f98b0	:'(
	f09f98b1	:(		f09f98b2	:(
	f09f98b3	:(		f09f98b5	:(
	f09f998b	:)		f09f998c	:)
	e29da4	♥		f09f918d	;)
	f09f918c	;)		f09f9293	:)
	f09f9294	:(		f09f9295	:)
	f09f9296	:)		f09f9297	:)
	f09f929d	:)		f09f929e	:)

Emoji	UTF-8 Karşılığı	Gülücük Eşleşmesi	Emoji	UTF-8 Karşılığı	Gülücük Eşleşmesi
😀	f09f9880	:)	😇	f09f9887	:)
😈	f09f9888	:)	😎	f09f988e	:)
😕	f09f9895	:(	😗	f09f9897	<3
😚	f09f9899	<3	😛	f09f989b	:P
😟	f09f989f	:(	😦	f09f98a6	:(
😧	f09f98a7	:(	😬	F09f98ac	:@

Çizelge 4.4. Raez ve arkadaşları tarafından öne sürülen gülmeye karşılık gelen metinsel ifadeler [41].

haha, hehe, hihi, huhu, hoho, muaha, buaha

4.3.1. Emoji Analizi

Emojiler, insanların duygularını daha kolay ve otomatik bir şekilde ifade etmesini sağlayan görsel sembollerdir [109]. Emoji kelimesi, Japonca'da "*resim*" ve "*karakter*" kelimelerinin birleşimi manasına gelmektedir. Emojiler, akıllı cihazlardaki uygulamaların emoji klavye desteğiyle beraber çevrimiçi anlık mesajlaşma ve -posta servislerinin yanı sıra sosyal medyada da gün geçtikçe daha da yaygınlaşmaktadır [110]–[113]. Emojilerin dilden bağımsız olarak insanlar arasında iletişim kurulmasında önemli bir işlevselliği bulunmaktadır. Emogi, duygusal pazarlama platformu tarafından yayınlanan rapora göre, çevrimiçi kullanıcılarının %92'si, duygularını ifade ederken

emojilerden faydalanmaktadır [110]. 212 ülkeden elde edilen 3,88 milyon veride yapılan analiz sonucunda, 2015 yılında dünyanın en çok kullanılan emojisi, %15.2 kullanım oranıyla "gülmekten gözlerinden yaş gelmek" emojisi olarak raporlanmıştır [114]. Aynı zamanda bu emoji, bu popülerliğin sonucu olarak Oxford Sözlüğü'nde "*2015 yılının kelimesi*" ilan edilmiş ve ilk defa metinsel bir kelime yerine görsel bir ifade "*Oxford Sözlüğü yılın kelimesi*" olarak seçilmiştir [115]. "*Gülmekten gözlerinden yaş gelmek*" emojisi, Twitter tarafından 14,5 milyar kullanım sayısıyla "Twitter tarihinin en çok kullanılan emojisi" olarak ilan edilmiştir [116].

Emojiler, tweet içerisinde Unicode formatında kodlanmaktadırlar ve bu emoji kodları 2009 yılında standartlaşmıştır [117], [118]. UTF-8 karakter kodlaması sayesinde Unicode (ya da genel anlamdaki karşılığı ile UTF-16) karakter kodlaması ile kodlanamayacak bazı karakterin de kodlanabiliyor olması sebebiyle emoji-emoticon dönüşümü esnasında UTF-8 karakter kodlaması temel alınmıştır [119]. UTF-16 veya UTF-32 kodlamanın, UTF-8 kodlamaya göre bir diğer dezavantajı ise baytlarının sırasının oluşturulan makinanın mimarisine olan bağımlılığıdır [120]. Emoji yüz ifadeleri, UTF-8 kodları ve yansıttıkları duyguya yönelik elle belirlenen gülücük eşleşmeleriyle beraber duygu analizi esnasında karakter kodundan tespit edilmek üzere veritabanında depolanmıştır.

Çizelge 4.3'de sunulan emojiler, TwitterSentiDetector tarafından ayrıştırılarak UTF-8 karşılıkları ve yansıttıkları duyguya yönelik gülücük eşleşmesi gerçekleşmektedir. Java programlama dili kullanılarak geliştirilen ve Şekil 4.4'de sunulan metot, tweet içerisindeki her bir kelimeyi UTF-8 kodlamasına dönüştürmekte ve ilgili kelimenin yüz ifadesi emoji olduğunu UTF-8 karşılığının "`f0`" veya "`e2`" bayt dizisi ile başlamasından tespit etmektedir. Şekil 10'da örnek bir tweet üzerinden tweet'i oluşturan kelimelerin UTF-8 olarak nasıl kodlandığı sunulmuştur. Emojilerin yan yana kullanılması ihtimali düşünülerek bu UTF-8 karşılıkları, her bir emojinin UTF-8 kodlama uzunluğu olan sekizerli bloklara bölünmüştür. Her bir blok, veritabanından sorgulanarak emojilere karşılık gelen yüz ifadesi ile değiştirilmiştir. Böylelikle TwitterSentiDetector duygu analizine tabi tutulacak tweet'i, metinsel gülme, gülücük ve emoji tabanlı analiz ederek pozitiflik ifade eden nitelik bulunamaması durumunda hızlı duygu analizi sürecine başlamaya hazır hale getirmiştir.

```
public static List<String> getEmojiCode(String tweet) throws UnsupportedEncodingException {
    List<String> emojiCodes = new ArrayList<String>();

    String[] parts = tweet.split(" ");

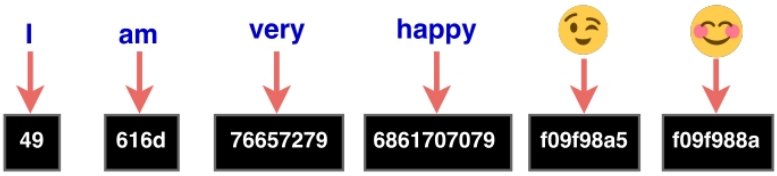
    for(String part : parts) {
        byte[] bytes = part.getBytes();

        String tmp = "";
        for(byte b : bytes) {
            tmp += Integer.toHexString(b & 0xFF);
        }

        System.out.println(part + " --> " + tmp);
        if(tmp.startsWith("f0")) {
            for(int i = 0; i < tmp.length(); i+= 8) {
                System.out.println("Emoji Code: " + tmp.substring(i, i+8));
                emojiCodes.add(tmp.substring(i, i+8));
            }
        }
    }

    return emojiCodes;
}
```

Şekil 4.4. Tweet içerisindeki her bir kelimeyi UTF-8 kodlamasına dönüştürerek emoji kod karşılıklarını tespit eden Java diliyle yazılmış metot.



Şekil 4.5. Örnek bir tweet üzerinden tweet'i oluşturan kelimelerin UTF-8 karşılıkları.

4.4. Metin Önişleme

Metin önişleme süreci, doğal dil işleme ve duygu analizi sürecine başlamadan önce tweet'lerde bulunan ve dilbilimsel olarak anlam ifade etmeyen karakterlerin, noktalama işaretlerinin ve Twitter'a özel sembollerin tweet içeriğinden temizlenmesini içerir. Tweet'leri kitap, dergi, makale gibi diğer metinlerden ayıran en önemli özellik, içerisinde dilbilimsel kurallara aykırı kelimelerin ve dilbilimsel olarak anlamsız olan sembollerin kullanılmasıdır. Dilbilimsel olarak anlamı olmayan ancak duygu analizi için anlamlı olan yüz ifadeleri, kısaltmalar ve hashtag'lerin algılanması doğru duygu analizi için önem arz etmektedir.

4.4.1. Noktalama İşaretlerinin Temizlenmesi

Tweet'lerde soru anlamı ifade etmek için “?”, bir heyecan, haykırış, aşırılık belirtmek için “!”, tweet sonunu belirtmek için “.” ve tweet içinde anlamsal bütünlüğü sağlamak için “,”, “;” noktalama işaretleri sıklıkla kullanılmaktadır. Bu kullanım amaçlarının yanı sıra tweet'lerde duyguyu kuvvetlendirmek (vurgulamak) için noktalama işareti tekrarlarına yer verilmektedir. Noktalama işaretleri, dilbilimsel açıdan anlam ifade etmediğinden ve duygu sözlüklerinde karşılık bulmadıklarından dolayı duygu analizi süreci başlamadan önce noktalama işaretlerinin tweet içerisinden temizlenmesi gerekmektedir.

4.4.2. Yeni Satır İfadelerinin Temizlenmesi

Twitter sunduğu uygulama geliştirme arayüzü (API)[9] ile Twitter'dan arama terimi ve kullanıcı bazlı veri çekilmesini sağlamaktadır. Bazı Twitter kullanıcıları, attıkları tweet içerisindeki içeriği listelemek veya kullanıcıların dikkatini çekmek için ilgili tweet’i

[9] https://dev.twitter.com/overview/documentation

birden fazla satıra yazmaktadırlar. Bu durum, Twitter web arayüzünde ilgili HTML etiketleri (“<br>”) kullanılarak görüntülendiğinden, birden çok satır şeklinde son kullanıcılara gösterilmektedir. Ancak Twitter'ın sunduğu API ile çekilen tweet'lerde birden fazla satır olması durumunda dönen cevaplarda yeni satırı ifade etmek için “\n” sembolü kullanılmaktadır. Dilbilimsel olarak anlamsız olan bu karakterlerin duygu analizi süreci başlamadan önce tweet içeriğinden temizlenmesi gerekmektedir.

4.4.3. Twitter’a Özgü Karakter ve İfadelerin Temizlenmesi

Twitter, kullanıcıların birbirleriyle olan etkileşimini ve tweet'lerin konularla olan bağlantısını sağlamak amacıyla tweet içerisinde çeşitli ön tanımlı karakter ve ifadelerden faydalanmaktadır. Çizelge 3.1'de kullanım amaçlarıyla beraber listelenen Twitter’a özgü karakter ve ifadeler dilbilimsel açıdan bir anlam ifade etmediğinden duygu analizi süreci başlamadan önce temizlenmesi gerekmektedir.

4.4.4. Web Adreslerinin (URL) Temizlenmesi

Kullanıcılar, attıkları tweet'lerde metin içeriğinin yanı sıra fotoğraf, video ve web adresleri (link) paylaşabilmektedirler. Metin dışındaki tüm bu içerikler, link olarak ilgili tweet içeriğine eklenmektedir. Web adresleri, dilbilimsel açıdan bir anlam ifade etmediğinden duygu analizi süreci başlamadan önce temizlenmesi gerekmektedir.

4.4.5. Harf Tekrarlarının Düzenlenmesi

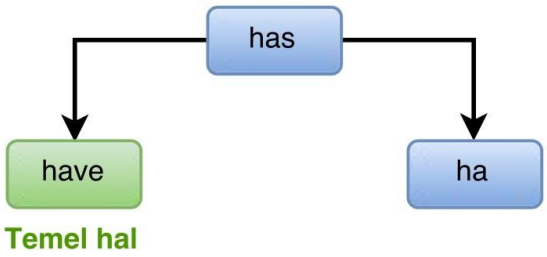
Tweet'lerde kullanıcılar duygularını ifade ederken kuvvetlendirmek için kelime içerisindeki harfleri tekrarlı olarak kullanabilmektedir. Ardışıl olarak tekrar eden harfler düzenlenmeden doğal dil işleme ve duygu analizi sürecine başlanırsa, dilbilimsel sözlüklerde kelime ya da kelime gruplarının temel halleri tanımlı olduğundan karşılıklar bulunamayacaktır. Örnek olarak “*I am verrryyy happyy!!*” gibi bir tweet ele alındığı zaman “*verrryyy*” kelimesi “*very*” kelimesine, “*happyy*” kelimesi “*happy*” kelimesine dönüştürüldüğü zaman dilbilimsel sözlüklerde karşılıkları bulunabilecektir. Literatürde, tekrar eden harf sayılarının teke [121], ikiye [39], [48] veya üçe [49] düşürülmesi yaklaşımları mevcuttur; ancak bu yaklaşımlar genelleme içerdiğinden yanlış tespitlere

sebebiyet vereceği düşünülmektedir. Yukarıdaki örnek cümlede tekrar eden harfler teke düşürüldüğü zaman "*very*" kelimesi doğru düzeltilecek; ancak "*happy*" kelimesi "*hapy*" olarak anlamsız bir hale dönüştürülecektir. Benzer şekilde tekrar eden harflerin ikiye düşürülmesi ise "*very*" kelimesini dilbilimsel olarak anlamsız olan "*verryy*" kelimesine dönüştürecektir. TwitterSentiDetector, tekrar eden harfleri WordNet üzerinden kontrol ederek anlamlı en sade hale getirmektedir. Bu esnada literatürdeki diğer çalışmalarda belirtilmemiş olan bir diğer kritik nokta ise harf düzenlemesinin kelimenin neresinden başlanacağıdır. Harf düzenlemesinin doğru bir düzeltmenin uygulanabilmesi için İngilizce'deki vurgu yapısı göz önüne alındığında kelimenin sonundan başlanması gerekmektedir. Sondaki ikiden fazla olan harf tekrarları düzenlendikten sonra (w_1) TwitterSentiDetector önce sağdan sola doğru geriye kalan harf tekrarlarını düzenlemektedir (w_2). Bu düzenleme sonrasında anlamlı bir kelime bulunamaması durumunda ilgili kelimenin sondaki harf tekrarları düzenlenmiş hali (w_1) üzerinden soldan sağa doğru harf düzenlemesi yapılarak (w_3) kelime anlamlı hale getirilmeye çalışılmaktadır. Her bir düzenleme adımında elde edilen yeni kaynak, WordNet üzerinden kontrol edilmekte ve anlamlı bir kaynağın bulunması durumunda süreç sonlandırılmaktadır. Çizelge 4.5'de soldan ve sağdan başlayarak "*happyy*" kelimesinin üzerindeki tekrar eden harf düzenlemesi sonuçları sunulmuştur.

Çizelge 4.5. "*happyyy*" kelimesi için soldan ve sağdan başlayarak tekrar eden harf düzenlemesi sonuçları.

Adım	Soldan Sağa Doğru Düzenleme	Sağdan Sola Doğru Düzenleme
1. Adım	happyy	happyy
2. Adım	hapyy	happy
3. Adım	hapy	

WordNet üzerinden kelimelerin temel halleri (kökleri) bulunurken, tek bir değer sunmak yerine morfolojide bağlantılı olarak tanımlı olan tüm bağlantılı kelimeleri sunmaktadır. Aranan kelimeye yönelik birden çok kök elde edilmesi durumunda doğrudan son sunulan değerin alınması yerine temel bağlantılı değerin kök ya da köklerine inilmesi gerekmektedir. Şekil 4.6'da bu durum "*has*" kelimesi üzerinden görselleştirilmiştir.



Şekil 4.6. "*has*" kelimesi için WordNet üzerinden kelimelere yönelik temel halinin bulunması.

4.4.6. Kısaltmaların Düzenlenmesi

Kısaltmalar, 140 karakter limiti olan tweet'lerde karakterden yer kazanmak için sıklıkla kullanılmaktadır. Kısaltmalar, mevcut halleriyle dilbilimsel açıdan bir anlam ifade etmemektedir. Ancak duygu anlamı taşıyan kısaltmaların düzenlenerek polarite değerlerinin hesaplanması, sistemin doğru duygu analizi yapılabilmesi için gerekmektedir. Yapılan literatür taramalarından elde edilen bilgilere göre, kısaltmalara yönelik erişime açık bir sözlük bulunmamaktadır. Ancak çeşitli web siteleri kısaltmaları sundukları web kullanıcı arayüzleri aracılığıyla anlamlandırmaktadır. Tez çalışması kapsamında geliştirilen servis ile bir çevrimiçi kısaltma sözlüğü olan NoSlang.com üzerinden kısaltmaların açık formaları elde edilerek duygu sözlükleri üzerinden polarite değerleri elde edilmiştir. Çizelge 4.6'da NoSlang.com üzerinden açık formları elde edilen örnek kısaltmalar açık formlarıyla beraber listelenmiştir.

Çizelge 4.6. NoSlang.com üzerinden açık formları elde edilen örnek kısaltmalar ve açık formları.

Slang	Meaningful Form
lol	laughing out loud
gr8	great
lmfao	laughing my f**king a** off

Slang	Meaningful Form
thx/tx	thank you
wtf	what the f**k
icymi	in case you missed it

Yukarıdaki tablodan da görülebileceği üzere, NoSlang.com içerisinde müstehcen ifade barındıran kelimelere yönelik açık form dönüşümü yaparken bazı harfleri sansürlemektedir. Buna çözüm olarak Wiktionary[10] isimli açık kaynaklı sözlük tarafından sunulan ve 285 adet kısaltmaya yönelik sansürsüz dönüşüm barındıran sözlük [122] TwitterSentiDetector veritabanına yüklenmiştir. NoSlang.com üzerinden çevrimiçi olarak sorgulanan kısaltmalara yönelik elde edilen açılım, sansür içerdiğinde TwitterSentiDetector, Wiktionary sözlüğünü kullanarak sansürsüz açılımı elde etmektedir. Tez çalışması kapsamında, kısaltmaların hem açılımını hem de TwitterSentiDetector tarafından hesaplanan polarite değerini sunan erişime açık bir web servisin geliştirilip, sunulması hedeflenmektedir.

4.4.7. Yazım Hatası Düzeltme

Tweet'lerde gerek resmi olmayan dilin kullanılması gerekse 140 karakter limitinden ötürü yazım hatalı kelimelere sıklıkla rastlanmaktadır. Bir tweet içerisindeki duygu belirten kelimeden meydana gelebilecek tek bir yazım hatası kelimenin hatta tüm tweet'in duygu değerinin kaybedilmesine sebebiyet verebilmektedir. Bu sebeple, doğru bir duygu analizinin yapılabilmesi için yazım hatalı kelimelerin düzenlenmesi gerekmektedir. Geleneksel yöntemler, yazım hatalı bir kelimeyi düzenlerken içerisinde bulunduğu cümleye bakmadan kendisine en benzer alternatif kelime ile düzeltmektedir. TwitterSentiDetector, geleneksel yazım tespit ve düzeltme yaklaşımları yerine özgün bir yaklaşım öne sürerek, yazım hatalı kelimeyi bağlam duyarlı bir şekilde kendisinden önceki ve sonraki kelime/kelimelere göre doğrusuyla değiştirmektedir.

[10] https://www.wiktionary.org

4.4.7.1. *Levenshtein Uzaklık Algoritması*

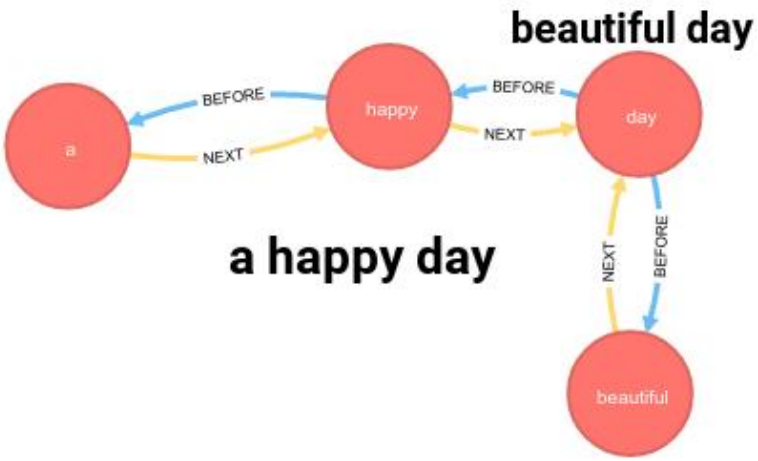
Levenshtein Uzaklık Algoritması, iki kelimenin benzerliğini, bir kelimenin diğerinin aynısı olması için gerekli tekil işlem (ekleme, silme veya yer değiştirme) adım sayısı olarak hesaplayan algoritmadır. s_1 ve s_2 kelimeleri (string) için algoritmanın sözde kodu Şekil 4.7'de verilmiştir.

```
int m[i, j] = 0
for i ← 1 to |s1|
do m[i, 0] = i
for j ← 1 to |s2|
do m[i, 0] = j
for i ← 1 to |s1|
do for j ← 1 to |s2|
     do m[i, j] = min{m[i-1, j-1] + if(s1[i] = s2[j]) then
0 else 1 fi, m[i-1, j] + 1, m[i, j-1] + 1}
return m[|s1|, |s2|]
```

Şekil 4.7. Levenshtein Uzaklık Algoritması'nın sözde kodu.

Hatalı kelimelerin bağlamsal olarak kendisinden önceki ve sonraki kelimelere bakılarak düzeltilirken kullanılacak bir bilgi bankası oluşturmak için Open American National Corpus (OANC) [123] veri setinden 46.808 ve Leipzig Corpora [124] veri setinden ise 100.000 cümle graf formatında neo4j[11] isimli yüksek ölçeklenebilir doğal graf veritabanında depolanmıştır. Veri setlerindeki her bir cümleyi oluşturan kelimeler, grafın düğümlerini oluşturmaktadır. Cümledeki her bir kelime, kendisinden önceki kelimeye `BEFORE`, kendisinden sonraki kelimeye ise `NEXT` ilişkisi ile bağlanmıştır. Böylelikle toplamda 146.808 cümle, iki yönlü graf içerisinde tanımlanmıştır. "*a happy day*" ve "*beautiful day*" örneklerinin graf veri yapısında depolanma yöntemi Şekil 4.8'de sunulmuştur.

[11] https://neo4j.com



Şekil 4.8. "*a happy day*" ve "*beautiful day*" örneklerinin graf veri yapısında depolanma yöntemi.

Cümleyi oluşturan her bir kelimenin graf veritabanında temsil ettiği bir düğüm olup olmadığına bakılarak, olmaması durumunda kelimeyi temsil eden yeni bir düğüm oluşturulmaktadır. Böylelikle her bir kelimenin diğer kelimelerle olan yakınlığı, kurduğu ilişki sayısından elde edilebilmektedir. Sonrasında ilgili kelime ile kendisinden önceki ve sonraki kelime arasında bağlantı (ilişki) oluşturulmaktadır. Şekil 4.9'da cümlelerin graf veritabanında depolanması için kullanılan algoritmanın sözde kodu sunulmuştur.

```
veri setindeki her cümle için dön
   cümleyi kelimelere ayır (arr_w)
   dizideki (arr_w) her bir kelime(w_i)için dön
      w_i'nin öncesindeki (w_p) ve sonrasındaki (w_n) kelimeyi
bul
      eğer w_i graf içinde düğüm olarak yer alıyorsa
         w_p, w_i ve w_n arasında ilişki oluştur
      eğer sonu
      değilse
         w_i'yi temsil edecek düğümü oluştur
         w_p, w_i ve w_n arasında ilişki oluştur
      değilse sonu
```

```
   döngü sonu
döngü sonu
```

Şekil 4.9. Cümlelerin graf veritabanında depolanması için kullanılan algoritmanın sözde kodu.

4.4.7.2. *Hatalı Kelime Tespiti ve Düzeltmesi*

Cümleyi oluşturan her bir kelimenin geçerliliği, WordNet sözlüğü üzerinden kontrol edilmektedir. Kelimenin WordNet'de karşılığı bulunamaması durumunda, kelimenin yazım hatalı bir kelime olduğu varsayılmaktadır. Bu durumda, yazım hatalı kelime, graf veritabanında aranarak bulunmakta ve kendisinden önceki ve sonraki kelimeler de `BEFORE` ve `NEXT` bağlantıları kullanılarak elde edilmektedir. Graf veritabanı, SQL benzeri bildirimsel (declarative) bir sorgu dili Cypher sorgu dili kullanılarak sorgulanmıştır. Cypher sorgu dili kullanılarak graf üzerinde çift yönlü ilişkilerin yönlerine göre sorgulama yapılabildiğinden, ters sorgular için ayrı bir sorgunun yazılmasına ihtiyaç duyulmamakta, sadece sorgudaki ilişkinin yönü değiştirilerek önceki ve sonraki kelimeler kolaylıkla sorgulanabilmektedir [125]. Yazım hatalı kelimenin yerine alternatif kelimeler, (1) yazım hatalı kelimeden önce gelen kelimeye bakılarak graf üzerinden bu önceki kelimeyi takip eden kelime listesi, (2) yazım hatalı kelimeden sonra gelen kelimenin graf üzerinde öncesinde gelen kelime listesinden oluşmaktadır. Sonuç olarak bulunan alternatif kelimelerin, hatalı kelimeye olan yakınlığı, (1) kelimenin kullanım sıklığı, (2) Levenshtein Uzaklık Algoritması kullanılarak elde edilen, alternatif kelimenin hatalı yazılan kelimeye olan metinsel benzerliği, (3) alternatif kelime ile hatalı yazılan kelimenin en büyük ortak kısmındaki harf sayısına bağlı olarak hesaplanmaktadır. neo4j graf veritabanının geliştirilmesinde de kullanılan Java programlama dili kullanılarak geliştirilen ve alternatif kelime ile hatalı yazılan kelimenin en büyük ortak kısmındaki harf sayısını hesaplayan Java diliyle yazılmış metot Şekil 4.10'da sunulmuştur.

```
private static Integer calculateCommonChars(String s1, String s2) {
    Integer commonLetters = 0;

    String bigWord = null, smallWord = null;

    if(s1.length() >= s2.length()) {
        bigWord = s1; smallWord = s2;
    } else {
        bigWord = s2; smallWord = s1;
    }

    for(int i = 0; i < smallWord.length(); i++) {
        int seqLength = smallWord.length() - i;
        for(int j = seqLength; j > 0; j--) {
            String tmpWord = smallWord.substring(i, i + j);
            if(bigWord.contains(tmpWord) && tmpWord.length() > commonLetters) {
                commonLetters = tmpWord.length();
            }
        }
    }

    return commonLetters;
}
```

Şekil 4.10. Alternatif kelime ile hatalı yazılan kelimenin en büyük ortak kısmındaki harf sayısını hesaplayan Java diliyde yazılmış metot.

s, yazım hatalı kelime (w_i) ile bir alternatif kelime arasındaki (w_n) benzerliği, $lev(w_i, w_n)$ bu iki kelimenin Levenshtein Uzaklık değerini, $lev(w_i, w_n)$ alternatif kelimenin w_i'nin önceki ve sonrasındaki kelimelere göre kullanım sıklığını, $com(w_i, w_n)$ yazım hatalı kelime ile alternatif kelime arasındaki en büyük ortak kısmın harf sayısını belirtmek üzere, TwitterSentiDetector iki kelime arasındaki benzerliği Denklem (4.1)'de sunulan formülle hesaplamaktadır. Formülden de açıkça görülebileceği üzere, hesaplanan benzerlik (s), alternatif kelimenin kullanım sıklığı ve ortak harf sayısı ile doğru, Levenshtein Uzaklık değeri ile ters orantılı olarak değişmektedir.

$$s(w_i, w_n) = \left(\frac{freq(w_i, w_n) \times com(w_i, w_n)^3}{lev(w_i, w_n)^3} \right) \quad (4.1)$$

Geleneksel yazım hatası düzeltme yöntemleri, yazım hatalı kelimeyi beraber kullanılan kelimelerle birlikte değerlendirmeyerek sadece kendisine metinsel açıdan en benzer

kelimeyi bulmaktadır. Bu yaklaşım, bağlamsal analiz yöntemleri kullanılmadan, doğrudan metinsel benzerliğe göre yazım hatası düzeltmesi yaptığından

Çizelge 4.7'de sunulan anlamsal olarak tutarsız düzeltmelere sebebiyet vermektedir. TwitterSentiDetector, hatalı bir kelimeyi "kendisine metinsel olarak en benzer kelime" yaklaşımının yerine "kelimenin bağlamında en çok kullanılan benzer kelime" yaklaşımı ile düzeltmektedir.

Öne sürülen yaklaşım, Wikipedia[12]'da sunulan ve üç farklı liste [126] tabanlı sıklıkla yanlış yazılan 188 İngilizce kelime barındıran liste [127]–[129] kullanılarak test edilmiş ve öne sürülen yaklaşımın doğruluğu %84 olarak bulunmuştur. Öne sürülen yaklaşım, cümle tabanlı yazım hatası düzeltme yaklaşımı öne sürdüğünden, liste içerisindeki her bir kelime başka bir kelime ile birlikte kullanılarak test edilmiştir. Listedeki kelimelerden 14 tanesi ("*hors d'oeuvres*", "*memento*", gibi) bağlamsal analiz için oluşturulan bilgi bankasında bulunamamıştır. Bu kelimeler çıkartıldığında, öne sürülen yaklaşımın doğruluğu %90,8'e çıkmaktadır. Çalışmada kullanılan bilgi bankasının, anlamsal ve dilbilimsel olarak doğru cümleleri içeren veri setleri dahil edilerek kolaylıkla genişletilebilmesi mümkündür. Yanlış düzeltmeler incelendiğinde (1) bilgi bankasında tanımlı olmayan kelimelerin varlığından, (2) öne sürülen düzeltmenin, doğru düzeltmeye daha az benzemesine rağmen bağlamsal olarak yanlış yazılan kelimenin öncesindeki ve sonrasındaki kelimelerle daha sık kullanılmasından kaynaklandığı görülmüştür.

[12] https://www.wikipedia.org

Çizelge 4.7. Anlamsal olarak Levenshtein Uzaklık Algoritması ile yanlış, TwitterSentiDetector ile doğru olarak düzeltilen örnek yazım hatalı ifadeler.

Yazım Hatalı İfade	Levenshtein Uzaklık Algoritması		TwitterSentiDetector	
	Düzeltme	Levenshtein Uzaklık Değeri	Düzeltme	Levenshtein Uzaklık Değeri
development *fase*	*faze*	1	*phase*	2
crew *custome*	*custom*	1	*costume*	2
radar *sencor*	*censor, sensor*	1, 1	*sensor*	1
wheter forecast	*whether*	1	*weather*	3
academic *celender*	*calender*	1	*calendar*	2
flaot warning	*float, flood*	2, 2	*flood*	2
psychiatric *petienhe*	*patience*	2	*patient*	3
Christmas *presenthe*	*present, presence*	2, 2	*present*	2
first-quarter *prophit*	*prophet*	1	*profit*	2
quate village	*quite*	1	*quiet*	3

4.5. Doğal Dil İşleme ve Dilbilimsel Metot Tabanlı Duygu Analizi

Metin önişleme süreci sonrasında, dilbilimsel açıdan anlamsız karakter ve ifadelerden, yazım hatalarından temizlenen tweet, doğal dil işleme ve duygu sözlükleri kullanılarak duygu analizine tabi tutulmaya hazır hale getirilmiştir. Bu aşamada, doğal dil işleme yöntemleri ile işlenen kelime (unigram) ve kelime grupları (n-gram), duygu sözlükleri aracılığıyla duygu yükü bakımından anlamlandırılacak ve ilgili tweet, dilbilimsel kurallar uygulanarak pozitif, negatif veya nötr olmak üzere sınıflandırılacaktır.

4.5.1. n-gram Polarite Hesabının Yapılması

SenticNet, n-gram'lara yönelik polarite tanımı barındıran tek duygu sözlüğü olduğundan tweet'den çıkartılan four-gram'lara, trigram'lara ve bigram'lara yönelik polarite değerleri sırasıyla SenticNet'den elde edilmektedir. SenticNet'de polarite karşılığı bulunan n-gram'lar tweet içeriğinden çıkartılarak tekrar polarite hesabına dahil olması engellenmektedir. Bu esnada n-gram'lara yönelik olumsuzluk eki varsa bu olumsuzluk ekleri de polarite hesabı yapılırken tekrar değerlendirilmemesi için n-gram'larla beraber tweet metninden çıkartılmaktadır. Denklem (4.2)'de formülize edildiği üzere toplam n-gram polarite değeri *(P(n-gram))*, four-gram *(P(four-gram))*, trigram *(P(trigram))* ve bigram *(P(bigram))* polarite değerlerini toplanmasıyla elde edilmektedir.

$$P(ngram) = P(four - gram) + P(trigram) + P(bigram) \quad (4.2)$$

4.5.2. unigram'ların Polarite Hesabının Yapılması

SenticNet'de karşılığı bulunan four-gram, trigram ve bigram'ların tweet'den çıkartılması sonucunda kalan tweet, unigram'lara ayrıştırılır. unigram'lara ait polarite hesabı yapılırken SentiWordNet, SenticNet, AFINN ve SentiStrength sözlüklerinden faydalanmıştır. SentiWordNet, kaynaklara yönelik polarite tanımını cümlenin öğesine göre yapmaktadır. Java tabanlı ve açık kaynak kodlu Apache OpenNLP[13] kütüphanesi, sunduğu büyük eğitim seti ile girdi olarak verilen cümleyi özne, yüklem, nesne, edat, zarf gibi cümlenin öğelerine ayırmakla beraber nesnelere yönelik insan, zaman, para, lokasyon, yüzde ve organizasyon gibi tür tespitinde de bulunabilmektedir [130]–[132]. Ayrıca barındırdığı uygulama geliştirme arayüzleri ile cümle tespitinde ve cümleyi belirteçlere (token) ayırmayı sağlar [130]–[133]. Unigram'lar, Apache OpenNLP kütüphanesi ile cümlenin öğelerine (Parts Of Speech - POS) dönüştürülerek

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

Çizelge 4.8'de listelenen POS etiketleri ile birlikte bir dizide tutulmaktadır. Duygu sözlüklerinde kaynakların temel halleri (fiiller için temel halleri, çoğul isimler için tekil halleri) tanımlı olduğundan unigram'ların SentiWordNet'de aranmaya başlamadan önce temel hallerine dönüştürülmesi gerekmektedir. WordNet, kendisine girdi olarak verilen kelimelerin türlerine göre kendi veritabanı üzerinden sorgulayarak temel hallerini sunmaktadır. WordNet kullanılarak unigram'lar temel hallerine dönüştürülmüştür. İngilizce'de "*homograph*" diye adlandırılan Türkçe'de "*eşsesli*" diye nitelendirilen kelimeler, cümlenin öğesine göre fiil, isim ya da sıfat/zarf olarak kullanılabildiğinden ("*fine*", "*mine*", "*like*", "*fear*", gibi) ve kullanım yerine göre cümleye kattıkları anlamlar değişebildiğinden temel hallerine dönüştürülen unigram'lar tespit edilen POS etiketlerine göre duygu sözlüklerinde aranmaktadır. SentiWordNet, diğer duygu sözlüklerinden farklı olarak POS etiketlerine göre polarite tanımı yapmaktadır. Çizelge 3.5'de sunulduğu üzere, AFINN ve SentiStrength duygu sözlükleri kaynaklara yönelik polarite değerlerini [-5, +5] aralığında tanımladığından TwitterSentiDetector'ün tanım aralığı olan [-1, +1] aralığına ölçeklendirilmiştir. SentiWordNet, sağladığı Java servisi[14] ile [0, 1] aralığında olan kaynak tanımını [-1, +1] aralığına ölçeklendirilmesini sağlamaktadır. Duygu sözlüklerinin polarite tanım aralıkları düzenlendikten sonra, Denklem (4.3)'da sunulduğu üzere unigram'lara yönelik elde edilen SentiWordNet polarite değeri *(P(SWN))*, SenticNet polarite değeri *(P(SenticNet))*, AFINN polarite değeri *(P(AFINN))* ve SentiStrength polarite değeri *(P(SentiStrength))* toplanarak, unigram'a ait polarite değeri barındıran sözlük sayısına *(n(lexicon))* bölünerek unigram'a ait nihai unigram polarite değeri *(P(unigram))* hesaplanmaktadır.

$$P(unigram) = \sum \frac{P(SWN) + P(SenticNet) + P(AFINN) + P(SentiStrength)}{n(lexicon)} \quad (4.3)$$

[14] http://sentiwordnet.isti.cnr.it/code/SentiWordNetDemoCode.java

Çizelge 4.8. Apache OpenNLP POS etiketleri.

Etiket	Etiket Türü	Etiket	Etiket Türü	Etiket	Etiket Türü
CC	Coordinating Conjuction	*NNS*	Noun, plural	*TO*	to
CD	Cardinal Number	*NNP*	Proper noun, singular	*UH*	Interjection
DT	Determiner	*NNPS*	Proper noun, plural	*VB*	Verb, base form
EX	Existential there	*PDT*	Predeterminer	*VBD*	Verb, past tense
FW	Foreign word	*POS*	Possessive ending	*VBG*	Verb, gerund
IN	Preposition	*PRP*	Personal pronoun	*VBN*	Verb, past participle
JJ	Adjective	*PRP$*	Possessive pronoun	*VBP*	Verb, non-3rd person singular present
JJR	Adjective, comparative	*RB*	Adverb	*VBZ*	Verb, 3rd person singular present
JJS	Adjective, superlative	*RBR*	Adverb, comparative	*WDT*	Wh-determiner
LS	List item marker	*RBS*	Adverb, superlative	*WP*	Wh-pronoun
MD	Modal	*RP*	Particle	*WP$*	Possessive wh-pronoun
NN	Noun, singular	*SYM*	Symbol	*WRB*	Wh-adverb

4.5.3. Nihai Duygu Tespiti

Tweet'e ait nihai polarite değeri *(P(TotalPolarity))*, Denklem (4.5)'de formülize edildiği üzere n-gram'lara ait polarite değeri (P(n-gram)) ile unigram'lara ait polarite değeri *(P(unigram))* toplanarak elde edilmektedir. Pozitif polarite değerleri olumlu duyguları, negatif polarite değerleri ise olumsuz duyguları yansıtmaktadır. Teorik olarak geriye kalan tek değer (0) nötr duyguyu tanımlamaktadır. Ancak elle duygu etiketi yapılmış örnek veri setleri incelendiğinde, nötr olarak etiketlenen tweet'lerin hiçbir zaman 0'a eşit olmadığı görülmüştür. Bunun için nötr duygu kümesi tek değer tanımı yerine [-0,1, +0,1]

aralığında değerlendirilerek esneklik kazandırılmıştır. Bu aralığın negatif başlangıç noktası olan -0,1 negatiflik alt limitini, pozitif başlangıç noktası olan +0,1 ise pozitiflik alt limitini tanımlamaktadır. Sonuç olarak elde edilen nihai polarite değeri *(P(TotalPolarity))*, Denklem (4.5)'de sunulduğu üzere negatiflik alt limitin altında ise sonuç negatif; değilse ve pozitiflik alt limitten küçükse sonuç nötr, pozitiflik alt limitten büyükse sonuç pozitif olarak bulunmaktadır. İki şarta da uymuyorsa tweet nötr olarak nitelendirilmektedir.

$$P(TotalPolarity) = P(unigram) + P(n-gram) \tag{4.4}$$

$$Sonuç = \begin{cases} negatif, & P(TotalPolarity) < -0{,}1 \\ nötr, & -0{,}1 \leq P(TotalPolarity) \leq +0{,}1 \\ pozitif, & +0{,}1 < (PTotalPolarity) \end{cases} \tag{4.5}$$

4.6. Harf Tekrarlarının Değerlendirilmesi

Twitter kullanıcıları, paylaştıkları tweet içerisinde belirtmek istedikleri duyguyu kuvvetlendirmek için sıklıkla yoğun duygu yükü taşıyan kelimelerde harf tekrarları yapmaktadırlar [39], [48], [134]. Metin önişleme aşamasında TwitterSentiDetector, kelimelerdeki harf tekrarlarını kelimelere yönelik duygu sözlüklerinde karşılık bulunabilmesi için düzenlemekte, ilgili kelimeyi WordNet üzerinden yinelemeli olarak kontrol ederek anlamlı en sade hale getirmektedir. Polarite tanımı elde edilmesi için gerekli olan bu aşamada, bir kısıt olarak kullanıcının duygu vurgusu kaybedilmektedir. Bu sebepten ötürü tekrarlı harf düzenlemesine maruz kalan kelimelerin polaritesi, yapılan testler sonucunda bulunan en optimum değer olan 1,5 polarite kuvvetlendiricisi ile kuvvetlendirilmiştir.

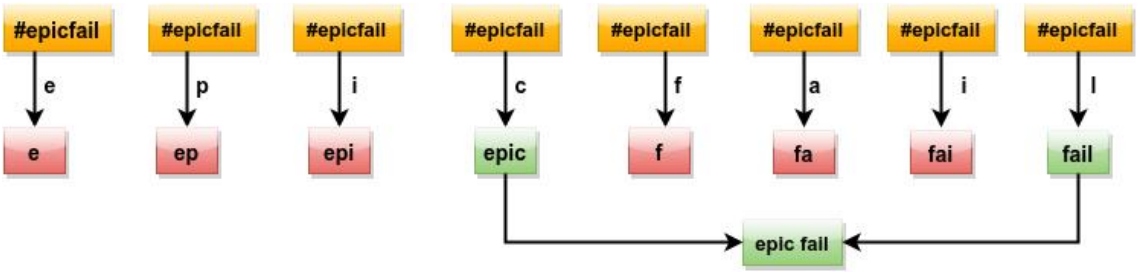
4.7. Hashtag'lerin Değerlendirilmesi

Literatürdeki birçok çalışma hashtag'leri duygu analizi için kullanmamakta ve basitçe tweet önişleme aşamasında tweet içeriğinden temizlemektedir. Bu yaklaşımın aksine, hashtag'lerin tweet'e ait konuyu, tweet'in bağlamını (context) belirtmenin [135], [136]

yanı sıra tweet'in nihai duygu ve düşüncesini belirtmek için de kullanılmasından [49] ötürü TwitterSentiDetector hashtag'leri duygu analizi açısından değerli kabul etmekte [47] ve hashtag'lerden duygu analizi esnasında faydalanılmaktadır. Hashtag'ler "*happiness*" gibi bir kelime veya "*epicfail*" gibi birleşik kelimeler içerebilmektedir. Hashtag'ler boşluk karakteri barındırmadığından, birleşik kelimeler aralarında boşluk olmadan yanyana yazılmaktadırlar. Böyle durumlarda hashtag'i oluşturan kelimelerin ayrıştırılıp, anlamlı kelimeler haline dönüştürülmesi gerekmektedir. Bunun için TwitterSentiDetector'de kullanılan hashtag ayrıştırma algoritmasının sözde kodu Şekil Şekil 4.11'de sunulduğu üzere şu şekilde çalışmaktadır: Başlangıçta TwitterSentiDetector tüm hashtag'i tek bir kelime (unigram) gibi düşünerek duygu sözlükleri üzerinden anlamlandırmaya çalışmaktadır. Sonuç bulunması durumunda hashtag'in tek bir kelimeden oluştuğu anlaşılmakta ve diğer unigram'lara yönelik polarite hesabı sürecinden geçirilmektedir. Sonuç bulunamama durumunda ise hashtag ayrıştırma süreci başlamaktadır. Başlangıçta *belirteç* hashtag'in ilk harfine ve *son indeks belirtecin* bir sağına yerleştirilmektedir. Adım adım *belirtecin* ile *son indeks* arasında kalan harf gruplarının WordNet'de tanımlı olup olmadığı kontrol edilmekte ve anlamlı olmaması durumunda *son indeks* bir sağa kaydırılmaktadır. Anlamlı bir kelime bulunduğu zaman bulunan kelime hashtag'den ayrıştırılmakta ve *belirteç son indeksin* bir sağını gösterecek şekilde kaydırılmaktadır. Bu döngü, *son indeks* hashtag'in sonuna gelene kadar devam etmektedir. *Son indeks* hashtag'in sonuna geldiği zaman ise ayrıştırılan kelimeler aralarında boşluk bırakılarak birleştirilmekte ve ilgili hashtag çıkartılarak yerine anlamlandırılmış kelime ya da kelime grupları tweet içeriğine eklenmektedir. "*#epicfail*" örnek hashtag'i için ayrıştırma süreci Şekil 4.12'de görselleştirilmiştir.

```
belirteci 0, son indeksi 1 olarak ata
ayrıştırılan hashtag'i boş string olarak belirle
son indeks < hashtag uzunluğu olduğu sürece dön
    eğer hashtag'in belirteç ve son indeks arasındaki kısım WordNet'de tanımlıysa
        ilgili kısmı ayrıştırılan hashtag'le birleştir, belirtece anlamlı son indeks+1'i ata, son indekse belirteç+1'i ata
    eğer sonu
    değilse
        son indeksi 1 artır
    değilse sonu
dön sonu
```

Şekil 4.11. Hashtag ayrıştırma algoritması sözde kodu



Şekil 4.12. "*#epicfail*" hashtag'inin TwitterSentiDetector tarafından ayrıştırılması.

4.8. Olumsuzluk Değerlendirmesi

Olumsuzluk anlamı taşıyan kelimelerin tweet'lerde kullanılması, tweet'deki duyguyu tam tersine çevirdiğinden olumsuz anlamı katan ek ("*not*") ve kelimelerin tespit edilmesi ve duygu analizinde önemli bir kriter olarak değerlendirilmesi gerekmektedir. Apache

OpenNLP kütüphanesi bünyesindeki Tokenizer[15] sınıfı, sağladığı eğitim setini kullanarak cümleyi olumsuzluk ekleriyle beraber belirteçlere (tokens) ayırmaktadır. TwitterSentiDetector bu kütüphaneyi kullanarak İngilizce'de sıklıkla kullanılan yardımcı fiille birleşik yazılmış olumsuzluk eklerini tespit edilebilmektedir. Bu sayede TwitterSentiDetector, Tablo 13'de listelendiği üzere Councill ve arkadaşlarının [137] yaptığı gibi yardımcı fiillerle birleşik yazılmış olan olumsuzluk eklerini metin bazlı karşılaştırma yaparak tweet içerisinde aramamakta; doğrudan olumsuzluk eklerini tespit edebilmektedir. Olumsuz anlamı katan kelime ve ekler, sadece duygunun polaritesini ters çevirmektedir; duygu kuvveti ise aynı kalmaktadır. Sayısal olarak bu anlam polarite değerinin -1 ile çarpılması, yani işaret değiştirmesi anlamına geldiğinden TwitterSentiDetector olumsuz anlamı katan kelime ve ekleri tespit ettiği zaman hesaplanan polarite değerini -1 ile çarpmaktadır. Tweet içerisinde olumsuz eki tespit edildiği zaman ekten sonraki duygu yükü taşıyan fiilin polaritesi -1 ile çarpılarak tam tersine çevrilmektedir. Buradaki kritik nokta, olumsuzluk ekinden sonraki ilk kelimenin fiil yerine zarf olabileceği düşünülerek, ardışıl olarak fiilin aranmasıdır. Örneğin "*I don't really like ...*" cümlesinde anlam bildiren "*really*" değil, "*like*" kelimesidir. Bundan dolayı "*not*" olumsuzluk ekinden sonraki fiil aranmaktadır. Örnek olarak "*I am not happy*" cümlesinde "*happy*" kelimesinin SentiWordNet'deki karşılığı 0,695'dır. Kendisinden önce olumsuzluk anlamı katan "not" kelimesi bulunduğundan bu cümlede "*not happy*" kelimelerinin polaritesi -0,695 olarak hesaplanacaktır.

Çizelge 4.9. Councill ve arkadaşları [137] tarafından öne sürülen olumsuzluk anlamı katan ek ve kelimeler.

hardly	*lacks*	*never*	*none*	*not*	*cant*
lack	*neither*	*no*	*nothing*	*n't*	*cannot*
lacking	*nor*	*nobody*	*nowhere*	*aint*	*darent*
dont	*doesnt*	*didnt*	*hadnt*	*hasnt*	*havnt*
havent	*isnt*	*mighnt*	*mustnt*	*neednt*	*oughtnt*
shant	*shouldnt*	*wasnt*	*wouldnt*	*without*	

[15] https://opennlp.apache.org/documentation/1.5.2-incubating/apidocs/opennlp-tools/opennlp/tools/tokenize/package-summary.html

TwitterSentiDetector'ü diğer ilgili çalışmalardan ayıran bir diğer nokta ise olumsuzluk anlamını, sadece olumsuzluk eklerine göre değil; bağlaçlardan da faydalanarak değerlendirmesidir. Literatürdeki bazı çalışmalar [49], zıt duyguları bir arada bulunduran tweet'ler belirsiz olarak değerlendirmekte ve analiz dışı bırakmaktadır. Veri setleri incelendiğinde bu tip tweet'lerin genellikle "*but, despite that, however*" gibi çevirici bağlaçlar ile zıt duyguları birbirine bağlayarak tweet'in nihai duygu polaritesini ters çevirdiği görülmüştür. TwitterSentiDetector, çevirici bağlaçların kullanıldığı tweet'lerde, ilgili tweet'e yönelik duygu polarite hesabını, son duyguyu belirttiği için çevirici bağlaçtan sonraki kısıma bakarak yapmaktadır. Ayrıca veri setlerindeki örneklerden "*additionally, similarly, also*" gibi destekleyici bağlaçların aynı duyguları birbirine bağladığı ve duygunun kuvvetlendirilmesi amacıyla kullanıldığı görülmüştür. Duygu analizi için TwitterSentiDetector tarafından kullanılan destekleyici ve çevirici olarak nitelendirilen bağlaçlar Çizelge 4.10'da listelenmiştir.

Çizelge 4.10. TwitterSentiDetector tarafından olumsuzluk değerlendirmesi için kullanılan destekleyici ve çevirici bağlaçlar.

Bağlaç Türü	Bağlaçlar
destekleyici	*additionally, similarly, also, furthermore, besides, likewise, moreover*
çevirici	*but, despite that, however, even though, nevertheless, nonetheless, in spite of, although, notwithstanding, on the contrary*

5. DENEYSEL SONUÇLAR VE TARTIŞMA

Twitter duygu analizi sistemleri, öne sürdükleri yaklaşımların başarımını Sanders Analytics[16] ve Stanford Twitter Sentiment (STS)[17] gibi bilinen, erişime açık veri setleri üzerinden değerlendirmektedir. Bu veri setleri, tweet içeriğinin yanı sıra elle etiketlenmiş duygu etiketlerini de barındırmaktadır. Sanders Analytics veri seti, pozitif, negatif, nötr ve ilgisiz olmak üzere dört duygu etiketiyle tanımlanmış 5.512 tweet barındırmaktadır. İlgisiz olarak etiketlenen tweet'ler, bu duygu grubunun TwitterSentiDetector'ün hedef kümesi dahilinde olmadığından veri setinden çıkartılmıştır. Bunun yanı sıra anlamsız veya İngilizce olmayan tweet'ler de veri setlerinden çıkartılmıştır. Tez çalışmasında sistemin testi için kullanılan veri setlerine ait polarite dağılımı Çizelge 5.1'de sunulmuştur.

Çizelge 5.1. Tez çalışmasında sistemin testi için kullanılan veri setlerine ait polarite dağılımı.

Veri Seti	Pozitif	Negatif	Nötr	Toplam
Sanders Analytics	1.139	668	1.228	3.035
STS	217	168	111	496

Tez çalışması kapsamında geliştirilen sistem, TwitterSentiDetector, literatürde sıklıkla kullanılan bu iki veri seti kullanılarak test edilmiştir. Sistemin amacı tweet'leri duygu kümelerine sınıflandırmak olduğundan, performans değerlendirmesi, sınıflandırıcıları görselleştirmek, düzenlemek ve performanslarına göre seçmek için sıklıkla kullanılan Alıcı İşlem Karakteristiği (Receiver Operating Characteristic – ROC) analizi yöntemi [66], [95] kullanılarak yapılmıştır. Pozitif (P) üzerinde pozitif veya negatif duygu yükü bulunan öznel tweetleri temsil etmekteyken, negatif (N) üzerinde pozitif veya negatif duygu yükü bulunmayan nötr tweet'leri temsil etmektedir. ROC analizi etiketlerinin TwitterSentiDetector üçlü mod duygu analizi için karşılıkları Çizelge 5.2'de sunulmuştur. Hassaslık (sensitivity), özgüllük (specificity) ve doğruluk (accuracy) ROC analizi

[16] http://www.sananalytics.com/lab/twitter-sentiment/
[17] http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip

sonuçlarının performansının nihai değerlendirmesi için kullanılan terimlerdir [138]. Hassaslık, sistemin pozitif örnekleri tespit etmedeki başarısını göstermekte ve doğru sınıflandırılan pozitif örneklerin *(True Positive - TP)* tüm pozitif örneklere *(Positive - P)* oranı olarak hesaplanmaktadır. Özgüllük, sistemin negatif örnekleri tespit etmedeki başarısını göstermekte ve yanlış sınıflandırılan negatif örneklerin *(False Negaive - FN)* tüm negatif örneklere *(Negative - N)* oranı olarak hesaplanmaktadır. Son olarak doğruluk, sistemin doğru tespit başarısını göstermekte ve doğru tespitlerin *(TP+TN)* tüm örnek sayısına *(P+N)* oranı olarak hesaplanmaktadır. Bu kavramlar, Denklem (5.1)'de formülize edilmiştir.

Çizelge 5.2. ROC analizi etiketlerinin TwitterSentiDetector üçlü mod duygu analizi için karşılıkları.

Etiket	Anlamı
True Positive *(TP)*	Doğru sınıflandırılan öznel tweetler
True Negative *(TN)*	Doğru sınıflandırılan nesnel tweetler
False Positive *(FP)*	Yanlış sınıflandırılan öznel tweetler
False Negative *(FN)*	Yanlış sınıflandırılan nesnel tweetler

$$\begin{aligned} Hassaslık &= TP \div (TP + FN) \\ Özgüllük &= TN \div (TN + FP) \\ Doğruluk &= (TP + TN) \div (TP + TN + FP + FN) \end{aligned} \tag{5.1}$$

Yapılan testler sonucunda TwitterSentiDetector'ün doğruluk oranı, üçlü mod için sırasıyla Tablo 17 ve Tablo 18'de ROC analizleri sunulduğu üzere Sanders Analytics veri seti için 0,66 ve STS veri seti için ise 0,6935 olarak hesaplanmıştır. Twitter, günlük yaşamla ilgili bütün konularda kullanıcı duygu ve düşüncelerine yönelik tweet'ler içerdiğinden, TwitterSentiDetector özellikle alandan bağımsız bir sistem olarak öne sürülmüştür. Bu sebepten ötürü test verisindeki tweet'ler spor, teknoloji, kişisel duygular gibi çeşitli alanlardan seçilmiştir. TwitterSentiDetector sistemi, erişime açık ve elle duygu etiketi yapılmış veri setlerine göre test etmek ve sonuçlara yönelik karmaşıklık matrisini

(confusion matrix) görüntülemek için Şekil 5.1'de sunulan kullanıcı görsel arayüzüne sahiptir. Ayrıca Şekil 5.2'de sunulan arayüz ile test edilen her bir tweet'e yönelik tekrarlı harf düzenlemesi, kısaltma dönüşümü, olumsuzluk değerlendirmesi, bağlaç değerlendirmesi, hashtag işlemesi, emoji, gülücük ve metinsel gülme tespiti gibi yapılan işlemlerin sıralı kayıtları (log) görüntülenebilmekte ve tweet'i oluşturan unigram ve n-gram'lara ait polarite değerleri ayrı ayrı listelenebilmektedir. Bu arayüzler, test sonuçlarındaki yanlış tespitlerin analiz edilerek sistemin başarımının artırılmasını kolaylaştırmaktadır. Ayrıca testlere yönelik duygu tespiti yapılırken hangi dönüşüm ve düzenlemelerin yapıldığının izlenmesini sağlamaktadır. Bu arayüzlerin yanı sıra TwitterSentiDetector'ün sunduğu arayüzler şu şekilde sıralanabilir: (1) Tekil tweet analizcisi ile girilen tek bir tweet'in sistem üzerinde analiz edilmesini sağlayan kullanıcı görsel arayüzü, (2) anlık duygu analizcisi ile anahtar kelime bazlı olarak aranan Twitter'daki anlık tweet'lerin toplanması için Şekil 5.3'de sunulan kullanıcı görsel arayüzü, (3) anlık olarak Twitter'dan elde edilen tweet'lere yönelik duygu analizinin yapılmasıyla elde edilen sonuçların Şekil 5.4'de sunulduğu üzere grafiksel olarak özetlenmesi için geliştirilen kullanıcı görsel arayüzü, (4) Şekil 5.5'de sunulan kaynak sözlüğü ile girilen kavramların duygu sözlüklerinde aranması için geliştirilen kullanıcı görsel arayüzü, ve (5) Şekil 5.6'da sunulan istatistik arayüzü ile veri setleri ve çalışma moduna göre test sonuçlarına yönelik karışıklık matrisinin ve gösterge (gauge) grafikleriyle sistem hassasiyetinin izlenebilmesi için geliştirilen kullanıcı görsel arayüzü. Duygu analizi sonuçlarının görselleştirilmesi, kullanıcıların sonuçları kolay bir şekilde anlamasını sağlayacaktır [137]. Bu sebeple analizler sonucunda duygu sınıflandırması yapan birçok çalışma, duygu sınıflarına yönelik sonuçları grafiklerle, tablolarla sunmaktadır [137]. Aynı sebepten dolayı TwitterSentiDetector kullanılarak duygu analizi tespiti için Twitter üzerinden yapılan tüm aramalar grafiklerle görselleştirilerek sunulmuştur. Sonuçlar, duygu sınıflarına göre yeşil pozitif duyguları, kırmızı negatif duyguları ve sarı nötr duyguları belirtecek şekilde renklendirilmiştir.

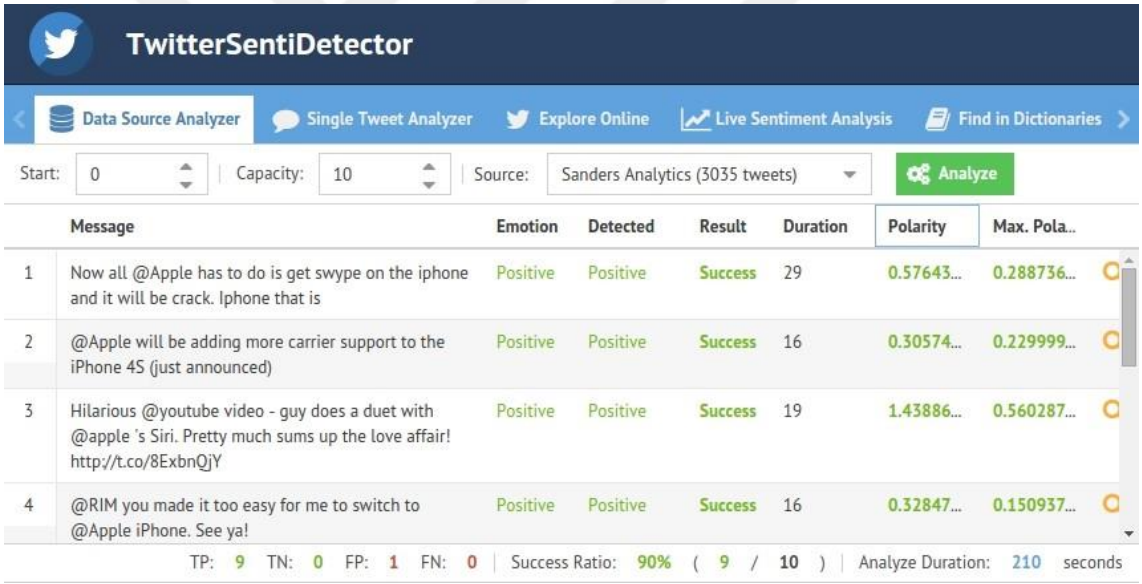
Çizelge 5.3. TwitterSentiDetector'ün Sanders Analytics veri seti kullanılarak elde edilen üçlü mod testlerinin karışıklık matrisi.

	Pozitif (P)	Negatif (N)	Toplam
Doğru (T)	1.527	476	2.003

	Pozitif (P)	Negatif (N)	Toplam
Yanlış (F)	280	752	1.032
Toplam	1.807	1,228	3.035

Çizelge 5.4. TwitterSentiDetector'ün STS veri seti kullanılarak elde edilen üçlü mod testlerinin karışıklık matrisi.

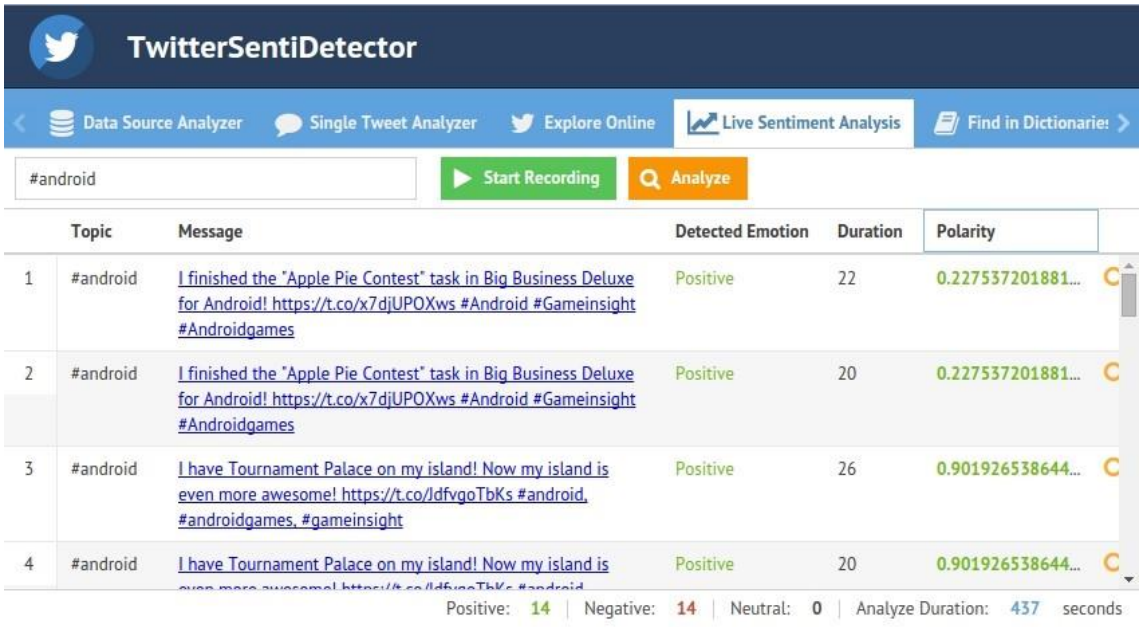
	Pozitif (P)	Negatif (N)	Toplam
Doğru (T)	309	35	344
Yanlış (F)	76	76	152
Toplam	385	111	496



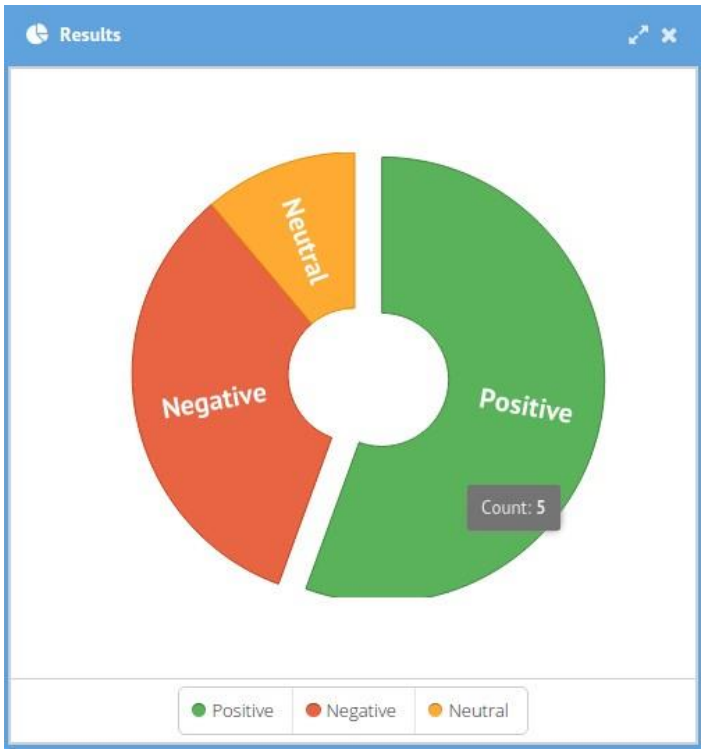
Şekil 5.1. TwitterSentiDetector veri setleri üzerinde duygu analizi yapılmasını sağlayan kullanıcı görsel arayüzü.



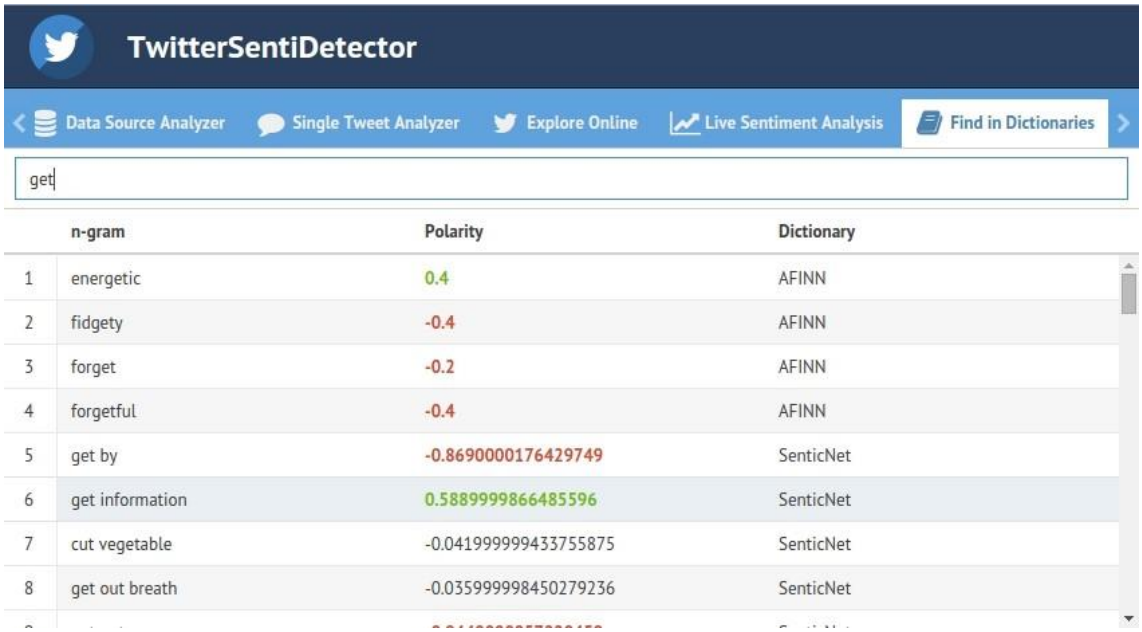
Şekil 5.2. TwitterSentiDetector detaylı tweet analiz sonuçlarının görüntülendiği kullanıcı görsel arayüzü.



Şekil 5.3. Anlık duygu analizi servisi ile Twitter üzerinden girilen anahtar kelimeye yönelik tweet toplanmasını için geliştirilmiş kullanıcı görsel arayüzü.

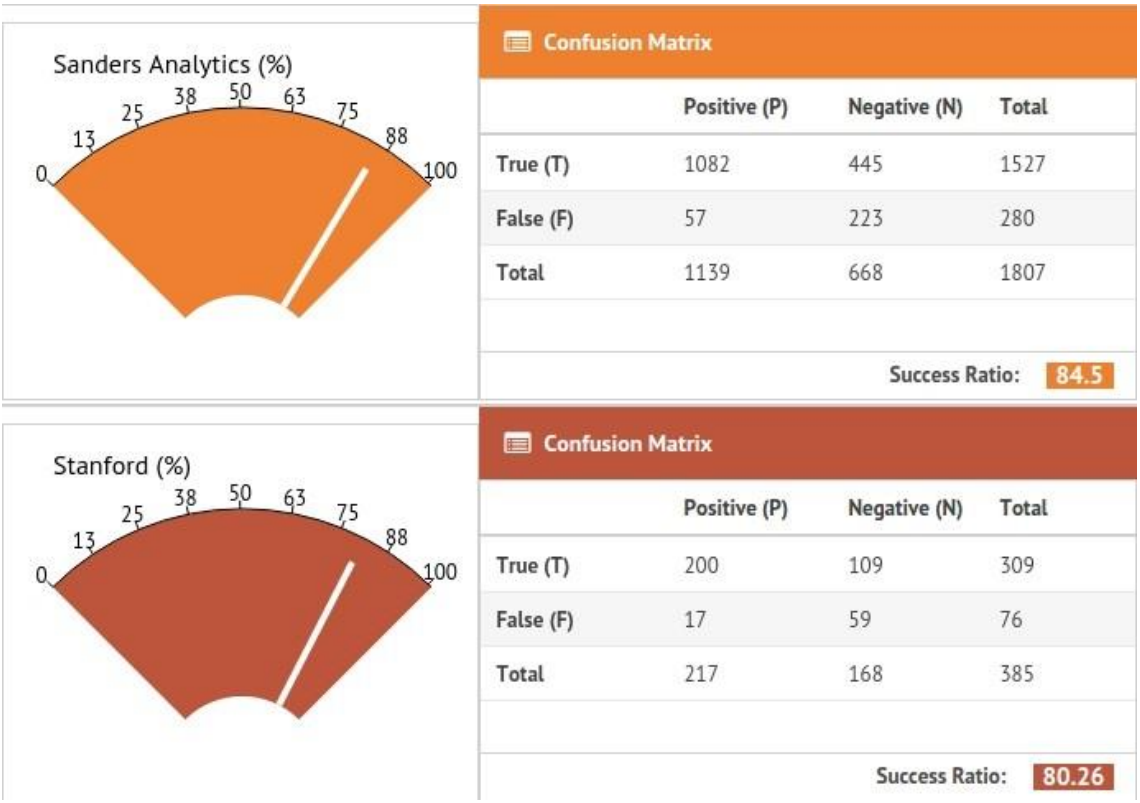


Şekil 5.4. Anlık duygu analizi servisi sonucunda aranan anahtar kelimeye yönelik sonuçları grafiksel olarak özetleyen kullanıcı görsel arayüzü.



	n-gram	Polarity	Dictionary
1	energetic	0.4	AFINN
2	fidgety	-0.4	AFINN
3	forget	-0.2	AFINN
4	forgetful	-0.4	AFINN
5	get by	-0.8690000176429749	SenticNet
6	get information	0.5889999866485596	SenticNet
7	cut vegetable	-0.041999999433755875	SenticNet
8	get out breath	-0.035999998450279236	SenticNet

Şekil 5.5. Girilen kaynaklara yönelik polarite yükünün duygu sözlüklerinde aranmasını sağlayan kullanıcı görsel arayüzü.



Şekil 5.6. Veri setleri ve çalışma moduna göre test sonuçlarına yönelik karışıklık matrisinin ve gösterge (gauge) grafikleriyle sistem hassasiyetinin izlenebilmesi için geliştirilen istatistik kullanıcı görsel arayüzü.

Twitter tabanlı duygu analizi sistemleri, sistemin başarımını test ederken literatürde kabul görmüş ve araştırmacıları tarafından erişime açık olan Sanders Analytics, STS veri setlerinden faydalanılmaktadır. Bu veri setleri tweet içeriğinin yanı sıra tweet'lere yönelik elle sınıflandırma yapılmış duygu etiketleri de barındırmaktadır. Bu duygu etiketlerinden nötr (objektif) duygu kümesi incelendiğinde çeşitli çalışmalarda raporlandığı gibi pozitif ve negatif duygu yüklerini beraber içerdiği gözlemlenmiştir [44], [48]. Bu sebepten ötürü ikili duygu değerlendirmesi (pozitif-negatif) yapılması halinde sistemin başarım testinin daha doğru sonuç vereceği düşünülmektedir. Aynı sebeple, tez çalışmasında geliştirilen sistem, literatürdeki diğer duygu analizi sistemleri gibi parametrik olarak hem üçlü (trinary) hem de ikili (binary) modda duygu tespiti yapabilecek şekilde tasarlanmıştır. İkili mod başarımında üçlü modda kullanılan veri setleri kullanılarak yalnızca mod performans ölçümü yapılmıştır. Yapılan testler sonucunda veri setlerine yönelik ikili mod doğruluğu, sırasıyla ROC analiz sonuçları Çizelge 5.5 ve

Çizelge 5.6'da sunulduğu üzere Sanders Analytics veri seti için 0,845, STS veri seti için 0,8026 bulunmuştur. Sonuç olarak, beklenildiği gibi TwitterSentiDetector ikili modda çalıştırılması durumunda sistemin hassasiyeti 0,11 ile 0,18 arasında artmıştır. TwitterSentiDetector ikili mod duygu analizi mekanizması, Michigan Üniversitesi tarafından paylaşılan UMICH SI650[18] isimli veri seti de kullanılarak test edilmiş, sistemin doğruluğu Çizelge 5.7'de karmaşıklık matrisi sunulduğu üzere 0,871 olarak hesaplanmıştır. UMICH SI650 veri seti 3.995 pozitif, 3.091 negatif olmak üzere toplam 7.086 adet sosyal medyadan elde edilmiş örneklem barındırmaktadır. Sunulan veri setinde, duygular elle pozitif ve negatif olarak etiketlendirilmiştir. Şekil 5.7'de Sanders Analytics ve Şekil 5.8'de STS veri seti kullanılması durumunda sırasıyla üçlü ve ikili mod için elde edilen ROC eğrileri sunulmuştur. İkili mod için pozitiflik *(P)* pozitif olarak etiketlenen tweet'leri, negatiflik *(N)* ise negatif olarak etiketlenen tweet'leri belirtmektedir. İkili modda sistemin hassasiyetinin yükselmesindeki başlıca sebepler şöyle sıralanabilir: (1) Veri setlerinde nötr olarak etiketlenmiş tweet'ler öznel (pozitif, negatif) duygular barındırma, (2) nötr duygu grubu için ayrılmış polarite değer aralığı pozitif ve negatif duygu grubuna göre oldukça kısıtlı olması, (3) SentiStrength hariç diğer duygu sözlüklerinin, nötr'lük polaritesi tanımlamamaları, pozitif ve negatif eksende kelime ve kelime gruplarına polarite tanımlamaları, ve (4) duygu etiketleme öznel bir iş olması nedeniyle [84], [139] ortak bir yaklaşımın bulunmaması.

Çizelge 5.5. TwitterSentiDetector Sanders Analytics veri seti kullanılarak ikili mod testlerinin karışıklık matrisi.

	Pozitif (P)	Negatif (N)	Toplam
Doğru (T)	1.082	445	1.527
Yanlış (F)	57	223	280
Toplam	1.139	668	1.807

[18] https://inclass.kaggle.com/c/si650winter11

Çizelge 5.6. TwitterSentiDetector STS veri seti kullanılarak ikili mod testlerinin karışıklık matrisi.

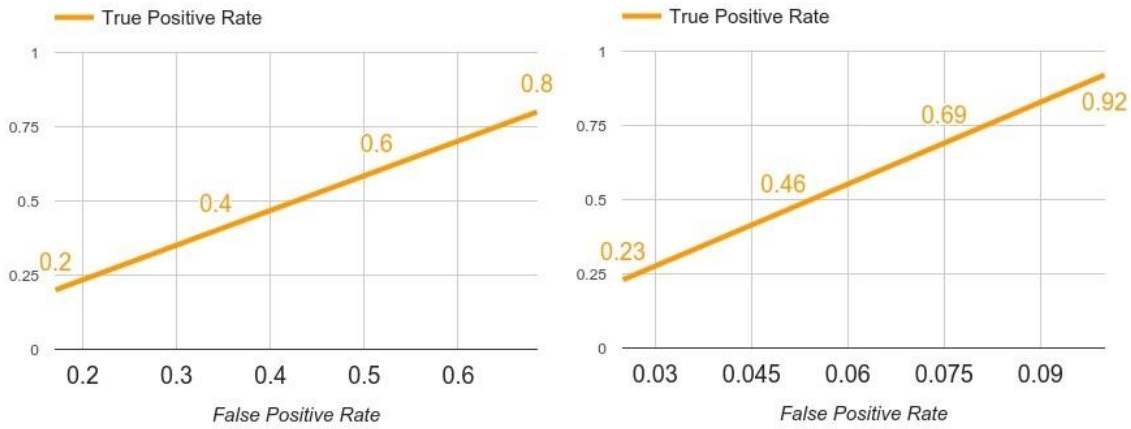
	Pozitif (P)	Negatif (N)	Toplam
Doğru (T)	200	109	309
Yanlış (F)	17	59	76
Toplam	217	168	385

Çizelge 5.7. TwitterSentiDetector UMICH SI650 veri seti kullanılarak ikili mod testlerinin karışıklık matrisi.

	Pozitif (P)	Negatif (N)	Toplam
Doğru (T)	3.308	2.866	6.174
Yanlış (F)	687	225	912
Toplam	3.995	3.091	7.086



Şekil 5.7. Sanders Analytics veri seti kullanılarak sırasıyla üçlü ve ikili mod için elde edilen ROC eğrileri.



Şekil 5.8. STS veri seti kullanılarak sırasıyla üçlü ve ikili mod için elde edilen ROC eğrileri.

Geliştirilmiş olan, kendine özgü bağlaç analizi mekanizması ile TwitterSentiDetector, diğer birçok Twitter tabanlı duygu analizi sisteminden ayrılmaktadır. Sırasıyla Çizelge 5.8 ve Çizelge 5.9'da aynı veri setleri için yüz ifadeleri ile metinsel gülmelerin ve bağlaç analizinin kullanılmaması durumunda üçlü ve ikili mod için TwitterSentiDetector'ün başarımı sunulmuştur. Test sonuçlarına göre bağlaç analizinin devre dışı bırakılması durumunda TwitterSentiDetector'ün başarımı 1,04 ile 1,6 arasında veri setine ve çalışma moduna göre gerilemiştir. Bu durumun başlıca sebebinin, Twitter'da kullanıcıların zıt duyguları bir arada kullanması olduğu düşünülmektedir.

Yüz ifadeleri ve metinsel gülmeler mikrobloglarda sıklıkla kullanıldığından TwitterSentiDetector, duygu analizi sürecinde yüz ifadeleri ve metinsel gülmelerden faydalanmaktadır. Aynı veri setleri üzerinde yapılan testler sonucunda yüz ifadeleri ve metinsel gülmelerden faydalanılmaması durumunda TwitterSentiDetector'ün başarımı 3,66 ile 5,14 arasında veri setine ve çalışma moduna göre gerilediği görülmüştür. Bu durumun başlıca sebebinin, Twitter'da içerisinde yüz ifadeleri ve metinsel gülmelerin de sıklıkla kullanıldığı resmi olmayan bir ifade dilinin kullanılması olduğu düşünülmektedir. Ayrıca emojiler, Twitter'da son yıllarda yaygınlaştığından dolayı literatürde kullanılan veri setlerinde emojiler kullanılmamaktadır. Duygu analizi için oldukça değerli olduğu düşünülen emojilerin duygu analizi başarımını gerçekte daha çok artıracağı düşünülmektedir.

Çizelge 5.8. TwitterSentiDetector yüz ifadeleri ile metinsel gülmeler ve bağlaç analizi kullanılmaması durumunda üçlü mod için veri setlerine göre duygu tespit hassasiyeti.

Veri seti	Yüz ifadesi ve metinsel gülme analizi devre dışı	Bağlaç analizi devre dışı
Sanders Analytics	62,34	64,42
STS	65,32	68,15

Çizelge 5.9. TwitterSentiDetector yüz ifadeleri ile metinsel gülmelerden ve bağlaç analizinden faydalanılmasına göre ikili mod için veri setlerine göre duygu tespit hassasiyeti.

Veri seti	Yüz ifadesi ve metinsel gülme analizi devre dışı	Bağlaç analizi devre dışı
Sanders Analytics	79,36	82,9
STS	75,58	79,22

5.1. Tweet Seviyesinde Bağlam Duyarlılığı

Bazı tweet’lere (özellikle diğer tweet'lere cevap niteliğinde olan kısa tweet'lere) yönelik duygu tespiti, bağlamları değerlendirilmeden mümkün olmamaktadır. Bu tip tweet'lere yönelik duygu analizinde bulunabilmek için tweet bağlamının da değerlendirilmesi gerekmektedir. Literatürde kabul görmüş olan Twitter duygu analizi veri setleri, tweet akışına yönelik bilgi ihtiva etmediğinden bağlam duyarlı duygu analizi, Twitter'ın sunduğu Duraksız İletim Uygulama Geliştirme Arayüzü (Streaming API) kullanılarak elde edilen tweet'ler üzerinde yapılmıştır. Twitter'dan aranan kelimeye (konu, hashtag veya Twitter kullanıcı adı) yönelik anlık tweet'leri çocuk-ata (child-parent) ilişkisini gözeterek ilişkisel veritabanında saklanması için öne sürülen algoritma Şekil 5.9’da sunulmuştur.

```
1: Aranan konuyla ilgili tweet'in (t_child) Twitter Uygulama Geliştirme Arayüzü kullanılarak elde edilmesi
2: t_child'in cevap olarak yazıldığı tweet'in (t_parent) bulunması
3: t_parent boş olmadığı sürece dön
4: t_parent'ı t_child'ın atası olarak işaretle ve ilişkisel veritabanında depola
5: Döngüyü sonlandır
```

Şekil 5.9. Twitter üzerinden anlık tweet'lerin çocuk-atası hiyerarşisinde ilişkisel veritabanında depolanması için öne sürülen algoritma.

Depolanan her bir tweet'e yönelik duygu tespiti yapılırken, veritabanında hiyerarşik olarak depolanan diğer tweet'ler de dikkate alınmaktadır. Bunun için Şekil 5.10'da sözde kodu sunulan algoritmadan faydalanılmaktadır.

```
1: Ele alınan tweet'e (t_base) yönelik duygu tespiti yapılır
2: eğer t_base öznel bir duyguya sahipse
3: eğer sonu
4: değilse
5: t_base'in bağlamsal tweet'lerini bul (t_context)
6: t_context kapsamındaki her bir tweet'in duygu sınıfını tespit et
7: t_context kapsamındaki duygu sınıflarını say
8: t_base'in duygu sınıfını t_context kapsamında en çok tekrar duygu sınıfı olarak belirle
9: değilse sonu
```

Şekil 5.10. Bağlam duyarlı olarak depolanan tweet'lerin duygu analizi için öne sürülen algoritma.

olan Ulusal Öğrenci Finansal Destek Taslağı (National Student Financial Aid Scheme - NSFAS)'nın ödemelerini yapmayan öğrencileri kara listeye almasıyla[19] ilgili tweet'ler toplanmıştır. Kendi veri setimizi toplamamızdaki sebep, konuyla ilgili diğer veri setlerinin yazarlarının Twitter'ın kuralları gereğince veri setinin paylaşımının mümkün olmadığı yanıtının alınmasıdır. Bu amaçla konuyla ilgili bağlamında en az bir cevap içeren 275 tweet toplanmıştır. TwitterSentiDetector, bağlamsal ilişkiler kullanılarak ve kullanılmayarak test edilmiş, Çizelge 5.10 ve

[19] http://www.news24.com/SouthAfrica/News/we-will-blacklist-non-paying-students-nsfas-20160121

Çizelge 5.11'de sunulduğu üzere sonuçlar birbirine oldukça yakın elde edilmiştir. Twitter'ın gündelik ve hızla değişen gündemi göz önüne alındığında bir konuyla ilgili uzun süre veri toplamak her zaman mümkün olmamaktadır. Bu sebeple, aranan konu güncelliğini yitirmeden mevcut tweet'ler kullanılarak duygu analizinin yapılması gerekmektedir. Sonuçlar, toplumları ilgilendiren ve gündem oluşturan konularla ilgili duygu analizi yapılırken bağlamın değerlendirilmeden ilk sonuçların hızlı bir şekilde alınabileceğini göstermektedir. Bağlamın göz önüne alınmaması, sistemin duygu analiz performansını bağlamsal analizde her tweet için en az bir cevap olduğu göz önüne alındığında en az iki katına çıkarmaktadır. Bağlamsal tweet'ler kullanılması ve kullanılmaması durumundaki duygu analizi sonuçlarının birbirine oldukça yakın olması, TwitterSentiDetector'ün bağlamsal analize gerek kalmadan bir konuyla ilgili duygu analizini daha hızlı bir şekilde yapabildiğini göstermektedir. Veri setindeki tweet'lerden sadece %5,2'si (7/134) bağlamsal olarak birden fazla tweet içerdiğinden, bağlam değerlendirmesindeki tweet sayısının duygu analizini nasıl etkilediği değerlendirilememiştir.

Çizelge 5.10. NSFAS veri seti için bağlam analizi kullanılması durumunda elde edilen sonuçlar.

Duygu Etiketi	Tweet Sayısı	Tweet Yüzdesi (%)
Pozitif	150	54,55
Negatif	81	29,45
Nötr	44	16

Çizelge 5.11. NSFAS veri seti için bağlam analizi kullanılmaması durumunda elde edilen sonuçlar.

Duygu Etiketi	Tweet Sayısı	Tweet Yüzdesi (%)
Pozitif	71	53
Negatif	40	29,85
Nötr	23	17,15

Makine öğrenmesi yöntemlerini kullanan yaklaşımlar, eğitim yapılan tanım kümelerine özel olduğundan diğer tanım kümelerinde kullanılması durumunda başarısız olmaktadır [84], [86], [140]. Bu yöntemler, sözlüksel yakınlık ve kelimelerin eş dizimliliği üzerinden sınıflandırma yapmaktadır [48], [84], [85]. Mikrobloglarda kuralsız cümleleri barındıran bir dil kullanıldığından ötürü makine öğrenmesi temelli yöntemlerin uygulanması zorlaşmaktadır [86]. Çizelge 5.12'de TwitterSentiDetector ile ilgili diğer çalışmaların duygu analizi yöntemlerinin karşılaştırılması sunulmuştur.

Çizelge 5.12. TwitterSentiDetector ile ilgili diğer çalışmaların duygu analizi yöntemlerinin karşılaştırılması.

Özellik	Jiang ve arkadaşları [86]	Go ve arkadaşları [48]	SentiStrength [85]	Barbosa & Feng [2]	TwitterSentiDetector
Hashtag işleme	Hayır	Hayır	Hayır	Hayır	Evet
Yüz ifadesi işleme	Evet	Evet	Evet	Hayır	Evet
Bağlaç analizi	Hayır	Hayır	Evet	Hayır	Evet
Duygu sözlüklerinin kullanımı	Evet	Hayır	Evet	Hayır	Evet

Özellik	Jiang ve arkadaşları [86]	Go ve arkadaşları [48]	SentiStrength [85]	Barbosa & Feng [2]	TwitterSentiDetector
Doğal dil işleme yöntemlerinin kullanımı	Evet	Evet	Evet	Evet	Evet
Tanım kümesi bağımsızlık	Evet	Evet	Evet	Evet	Evet
Eğitim ihtiyacı	Hayır	Evet	Hayır	Hayır	Hayır
Dinamik tekrarlı harf düzenlemesi	Hayır	Hayır	Evet	Hayır	Evet
Çevrimiçi kısaltma sözlüğü kullanımı	Hayır	Hayır	Hayır	Hayır	Evet
Kısaltma dönüşümü	Hayır	Hayır	Hayır	Hayır	Evet
Bağlam duyarlılığı	Evet	Hayır	Hayır	Hayır	Evet
Hedef duyarlılığı	Evet	Hayır	Hayır	Hayır	Hayır
İkili sınıflandırma modu	Hayır	Evet	Evet	Hayır	Evet
Üçlü sınıflandırma	Evet	Hayır	Evet	Evet	Evet

Özellik	Jiang ve arkadaşları [86]	Go ve arkadaşları [48]	SentiStrength [85]	Barbosa & Feng [2]	TwitterSentiDetector
modu					

5.2. Canlı Olay Analizleri

TwitterSentiDetector'ün duygu analiz doğruluğu ve hassasiyetini değerlendirmek üzere, canlı olaylar esnasında olayla ilgili tweet'ler, Twitter Duraksız İletim Uygulama Geliştirme Arayüzü kullanılarak bir NoSQL (Not Only SQL) veritabanı olan MongoDB[20] veritabanında depolanmış ve sonrasında tweet'ler TwitterSentiDetector ile analiz edilmiştir.

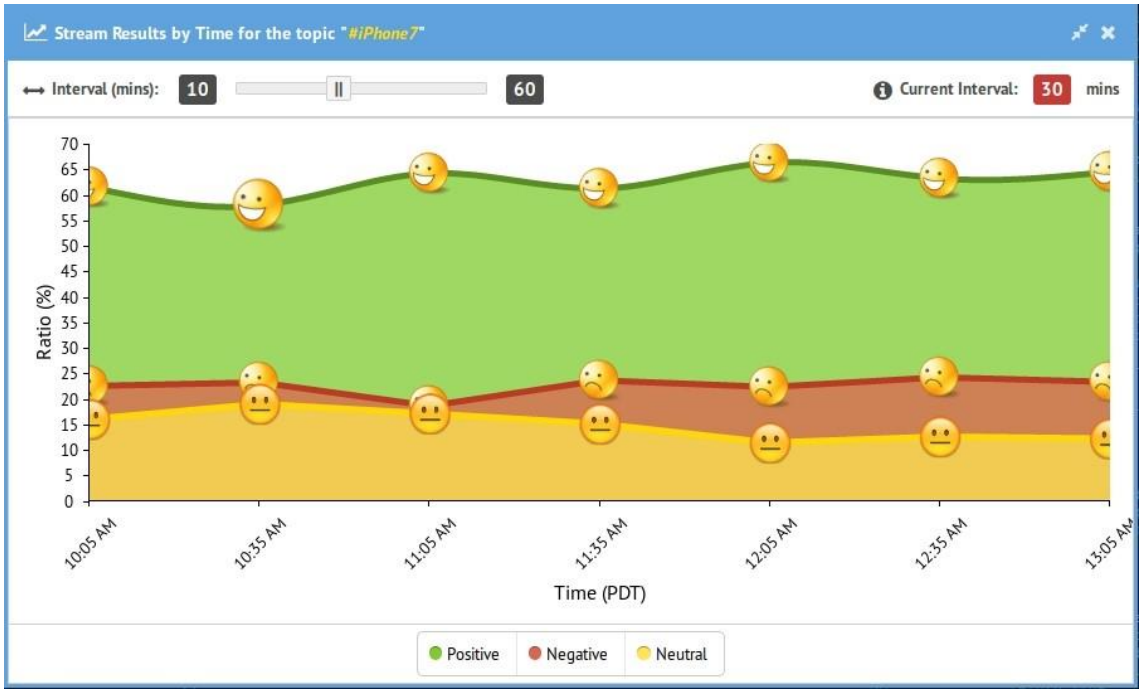
5.2.1. Apple iPhone 7 Özel Etkinliği

Apple'ın iPhone 7'yi tanıttığı iPhone 7 Özel Etkinliği[21], 7 Eylül 2016 tarihi saat 10:00 (Pacific Daylight Time - PDT)'da başlamış olup, iki saat sürmüştür. Etkinlik sırasında "*#iPhone7*" hashtag'ini içeren 23.072 canlı tweet toplanmış ve TwitterSentiDetector'ün veritabanında sistemin duygu analiz hassasiyetini ve doğruluğunu test etmek üzere depolanmıştır. Geliştirilen görsel kullanıcı görsel arayüzü, istenilen zaman aralıklarındaki pozitif, negatif ve nötr tweet oranlarını (%) gösterecek şekilde tasarlanmıştır. Apple iPhone 7 etkinliği esnasındaki toplanan canlı tweet'lerin yarım saatlik periyotlarla analiz sonuçlarının görüntülenmesi geliştirilen kullanıcı görsel arayüzü Şekil 5.11'de sunulmuştur. Geliştirilen kullanıcı görsel arayüzü, zaman aralıklarının istenilen hassasiyete ayarlanabilmesini sağlayarak daha detaylı analiz yapılabilmesini sağlamaktadır. Görselde sunulduğu üzere, Apple CEO'su Tim Cook açılış konuşmasını yapmak üzere sahneye çıktığında, beklenildiği gibi pozitif tweet oranında büyük bir artış görülmüştür. Etkinliğin ilk heyecanı dinmeye başladığında, pozitif tweet oranının biraz düştüğü görülmektedir. Tim Cook, iPhone 7 ile ilgili konuşmaya başladığında pozitif tweet oranının artmaya başladığı ve iPhone 7'ye ait olan su geçirmezlik, geliştirilmiş kamera, daha uzun batarya ömrü, genişletilmiş bellek gibi özellikler duyurulmaya başladığı anda pozitif tweet oranı en yüksek seviyeye ulaşmıştır. TwitterSentiDetector tarafından tespit edilen pozitiflik oranı, iPhone 7'nin özelliklerinin anlatımıyla sonlanan

[20] https://www.mongodb.com
[21] http://www.apple.com/apple-events/september-2016/

etkinliğin sonuna kadar benzer oranda seyir etmiştir. Etkinlik süresince toplanan tweet'ler, bir yeni veya geliştirilen özelliğin ortalama duyurulma süresi olan yarım saatlik periyotlara bölünerek TwitterSentiDetector'ün gerçek zamanlı duygu analiz tespiti ve hassasiyeti test edilmiştir. Etkinliğe ait önemli anlar, ilgili zaman bilgisi ve TwitterSentiDetector tarafından tespit edilen pozitif tweet oranıyla birlikte

Çizelge 5.13'de listelenmiştir.



Şekil 5.11. Apple iPhone 7 Özel Etkinliği süresince "*#iPhone7*" hashtag'i ile gönderilen tweet'lerin duygu analiz sonucunu gösteren kullanıcı görsel arayüzü.

Çizelge 5.13. Apple iPhone 7 özel etkinliği önemli anlarının zaman bilgisi ve TwitterSentiDetector tarafından anlık olarak tespit edilen pozitif tweet oranları.

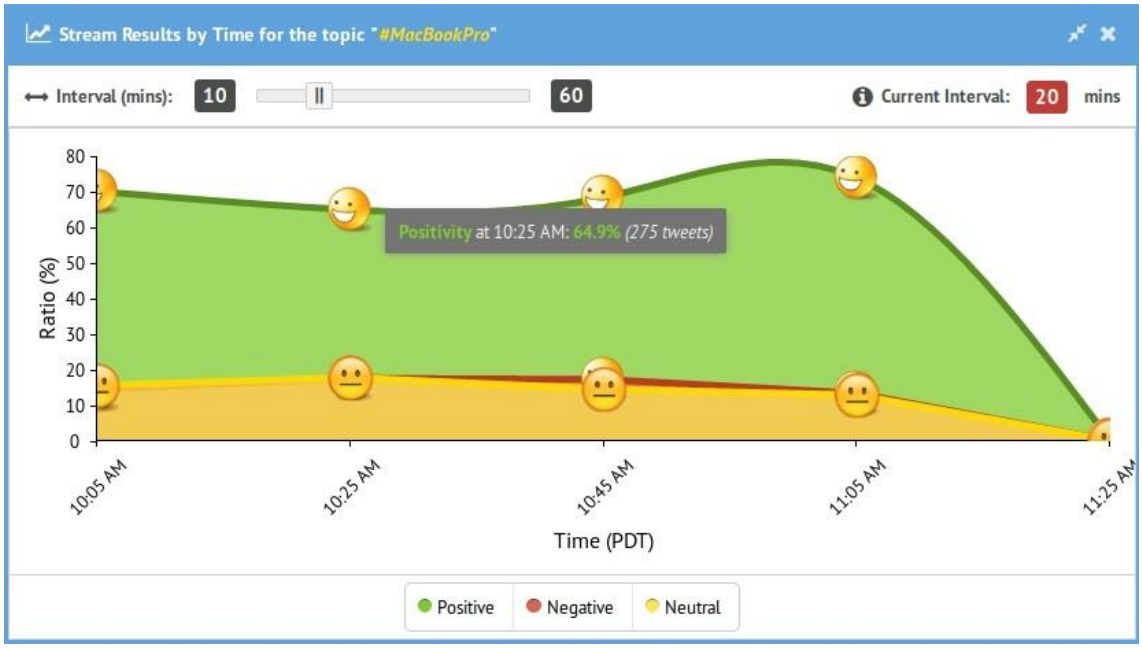
Zaman (PDT)	Olay Açıklaması	Pozitif Tweet Oranı (%)
10:08 AM	Apple CEO'su Tim Cook'un sahneye girişi	75,9
10:12 AM	Tim Cook'un Super Mario'nun Apple App Store'a geleceğini duyuruşu	64,1
10:21 AM	Tim Cook'un Apple Watch'a ait yenilikleri duyurumaya başlaması	61,5
10:56 AM	Tim Cook'un iPhone 7 ile ilgili konuşmaya başlaması	58,6
11:01 AM	iPhone 7'nin ilk gösterimi	65,4
11:04 AM	iphone 7'nin detaylarının sunulmaya başlanması	47

5.2.2. Apple MacBook Pro Özel Etkinliği

Apple MacBook Pro Özel Etkinliği[22], 27 Ekim 2016 tarihi saat 10:00 (PDT)'da başlamış olup, yaklaşık bir buçuk saat sürmüştür. Etkinlik süresince "*#MacBookPro*" hashtag'li 2.366 canlı tweet toplanmış ve TwitterSentiDetector'ün veritabanında sistemin duygu analiz tespit hassasiyetini ve doğruluğunu test etmek üzere depolanmıştır. Şekil 5.12'de etkinlik esnasında gönderilen bu tweet'lerin yirmi dakikalık periyotlarla analizini gösteren kullanıcı görsel arayüzü sunulmuştur. Görselde sunulduğu üzere, pozitif tweet oranı yeni MacBook Pro'nun ilk gösterimi esnasında hızla yükselmiş ve touch bar, touch ID gibi yeni MacBook Pro'nun öne çıkan özellikleri duyurulmaya başladığında en üst seviyeye çıkmıştır. Sunulan yeni özelliklerin yaygın kullanımı olan Final Cut Pro X[23], Adobe Photoshop[24] gibi örnek programlar üzerindeki kullanımı gösterilmeye başladığında pozitiflik oranında düşüş gözlemlenmiştir. Etkinliğin sonuna doğru pozitif tweet oranında gözle görülür bir düşüş görülmüştür. Etkinliğe ait önemli anlar, ilgili zaman bilgisi ve TwitterSentiDetector tarafından tespit edilen pozitif tweet oranıyla birlikte Çizelge

[22] http://www.apple.com/apple-events/october-2016/
[23] http://www.apple.com/final-cut-pro/
[24] http://www.adobe.com/products/photoshop.html

5.14'de listelenmiştir.



Şekil 5.12. Apple MacBook Pro özel etkinliği süresince "*#MacBookPro*" hashtag'i ile gönderilen tweet'lerin duygu analiz sonucunu gösteren kullanıcı görsel arayüzü.

Çizelge 5.14. Apple MacBook Pro Özel Etkinliği önemli anlarının zaman bilgisi ve TwitterSentiDetector tarafından tespit edilen pozitif tweet oranları.

Zaman (PDT)	Olay Açıklaması	Pozitif Tweet Oranı (%)
10:02 AM	Apple CEO'su Tim Cook'un açılış yapmak üzere sahneye girişi	90
10:25 AM	Tim Cook'un yeni MacBook Pro ile ilgili olarak konuşmaya başlaması	66,7
10:30 AM	Yeni MacBook Pro'nun ilk gösterimi	67,8
10:32 AM	MacBook Pro'nun touch bar, touch ID gibi yeni özelliklerinin tanıtılmaya başlanması	67
11:12 AM	Yeni MacBook Pro'nun işletim sistemi olan macOS Sierra'nın tanıtılmaya başlanması	68,9

5.3. Bilinen Kısıtlar

TwitterSentiDetector, olumsuzluk değerlendirmesini hem olumsuzluk eklerine hem de zıt duyguları bir arada kullanmak için genellikle kullanılan bağlaçlara yöre yapmaktadır. Buna rağmen, örnekleri Çizelge 5.15'de listelendiği üzere her ikisini de içermemesine rağmen anlamca negatif olan [65] tweet'ler mevcuttur.

Çizelge 5.15. Olumsuzluk eki veya bağlaç olmadan zıt duyguları birarada bulunduran tweet örnekleri.

Can someone plz explain to me whhyy @apple is distributing 2-3 phones per day to the sprint stores
this situation is perfect opposite of insanely great
any chance @apple will release good headphones
I made a reservation and yet I still have to wait in line
My new car is sweet sync system is #epicfail
Issues couldn't resolve after 1/2 day w/ @ATT & @Apple. Hero of my day!
An apple update has seemed to render my work machine incapable of opening HDV video. Thanks @Apple
Apparently fuzzball crashes on #iOS5. Congratulations @apple on another incompatible upgrade
4 millions in a weekend, 16 #iPhone4S per second. This is madness?! no, this is @Apple !!!
Oh, @apple. Steve obviously had nothing to do with iPhoto, as it's the perfect opposite of 'insanely great'. Get it fixed, please.
@apple, your update server has been timing out for a week. i had to force a restore w/the new #iOS5 software. lost all my stuff. thanks!

Karşılaştırma içeren tweet'lerde zamana bağlı olarak tespit edilen duygunun mevcut durumu mu yoksa önceki durumu mu yansıttığının tespit edilmesi gerekmektedir. Bu tip tweet'lerde duygu tespitinin zaman eklerine göre yapılması gerekmektedir. Örnek olarak

"*iOS4 is better than iOS5*" tweet'i ele alınırsa, zaman eki değerlendirilmezse tweet pozitif olarak etiketlenecektir. Ancak zaman ekiyle beraber değerlendirilmesi durumunda polarite ters çevrilecek ve tweet negatif olarak etiketlenecektir.

Duygu analizi sistemlerinin başarım değerlendirmesi, genellikle elle duygu etiketi yapılmış olan erişime açık veri setleri üzerinden yapılmaktadır. Bu veri setleri incelendiğinde,

Çizelge 5.16'da örnekleri listelendiği üzere yanlış duygularla etiketlenmiş tweet'ler tespit edilmiştir. Bu yanlış etiketler sistemin başarım testlerinin geçerliliğini olumsuz etkilemektedir.

Çizelge 5.16. Veri setlerinde yanlış etiketlenmiş örnek tweet'ler ve beklenen duygu etiketleri.

Tweet	Etiketlenen Duygu	Beklenen Duygu
#DontBeMadAtMeBecause #Android is by far better than @Apple	negatif	pozitif
Dear #Twitter, I missed you today, >> internet was down :/	pozitif	negatif
@loveguru_b i miss my #twitter boo	pozitif	negatif
#Google Apps vs. #Microsoft #Office 365: it is clear that Google Apps should be the top choice http://t.co/KDTD4zRe	negatif	pozitif
@Blackberry & @Facebook U R really about to make me throw this @Blackberry in the trash an get an @Apple iPhone! @Facebook upload issues!	pozitif	negatif
making the switch from @Android to @Apple #iphone #iphone4S #smartphone #stevejobs (@ Apple Store) http://t.co/kj6pJvkH	pozitif	nötr
Telegraph reports that the biggest threat to #facebook is its power users have shifted to #twitter or #google+ http://t.co/JHdTYKCV	pozitif	nötr

Twitter'a özel, sözlüklerden faydalanarak anlam verilemeyecek ancak sosyal medya dilinde anlamlı olan kelimeler de sistemin duygu tespitini zorlaştırmaktadır. Örneğin "*who could be Twitterless?*" gibi bir cümlede "Twitter'sız, Twitter olmadan" anlamı katan "*Twitterless*" kelimesi duygu sözlüklerine göre anlamsızken; sosyal medya dilinde anlamlıdır. Bunun gibi sosyal medyaya özgü, sözlüklerde anlamsal olarak karşılığı olmayan kelimeler sistemin yanlış alarm vermesine sebep olmaktadır. Bunun yanı sıra "*addicted to*" gibi bazı kelimeler tweet içerisinde sözlükteki anlamının tam tersinde kullanılabilmektedir. Örneğin "*I'm addicted to the FIFA 16*" tweet'i, "*addicted to*" ifadesi duygu sözlüklerinde negatif olarak tanımlı olmasına rağmen veri setlerinde kullanım yeri göz önüne alınarak pozitif olarak etiketlenmiştir. Bu tip tweet'lerdeki duygu, tanım kümesine bağlı olduğundan [45], [65], [141], [142] ve TwitterSentiDetector kullanım alanından bağımsız-eğiticisiz bir sistem olduğundan bu tip tweet'lerde yanlış duygu tespiti yapabilmektedir. Aynı kaynaklara yönelik farklı polarite tanımı, ilgili tanım kümesine yönelik etiketlenmiş veri setleri kullanarak alana özgü duygu sözlüklerinin oluşturulmasıyla mümkündür; ancak bu yaklaşım alan bağımlılığa sebep olacaktır [5], [48], [66], [86], [143]. Bu durum da çalışmanın hedef kümesini kısıtlayacak ve diğer alanlarda duygu analizi yapılması durumunda başarısız sonuçlar elde edilmesine sebep olacaktır [85], [143].

5.4. Duygu Analizi Süresinin Kısaltılması

Duygu analizi sistemlerinin başarımı, doğru duygu tespiti ve duygu analiz süresi olmak üzere iki eksende değerlendirilmektedir. Duygu analizi süresi, sosyal medyadaki gündemin gündelik olmasından ötürü önem arz etmektedir. TwitterSentiDetector'ün duygu analiz süresini etkileyen etmenler şu şekilde sıralanabilir: (1) Duygu sözlüklerinden ilgili sözlüklerin sorgulanması, (2) doğal dil işleme yöntemleriyle tweet'lere yönelik cümlenin öğelerine ayrıştırılması ve bağlaç tespiti, (3) tweet'lerin metin ön işleme sürecinden geçirilerek duygu analizi bakımından anlamsız olan kısımlarının çıkartılması, (4) hashtag'lerin işlenerek anlamlı hale getirilmesi, (5) kısaltmaların çevrimiçi sözlükler kullanılarak anlamlı hallerine dönüştürülmesi. Bu etmenlerden duygu analizi süresini kısaltmak için optimize edilebilecek tek etmen, duygu sözlüklerinden kaynak sorgulamasının yapılmasıdır.

Bellek içi veritabanları (in-memory databases) verinin tutarlılığını korunarak yüksek erişilebilirliği sağlama temeline dayanmaktadır [2], [54], [88]. Bu amaçla, ilişkisel veritabanlarına göre verileri disk üzerinde değil bellekte tutan bellek içi veritabanı kullanılarak aranan kaynakların duygu sözlükleri üzerinden daha hızlı bir şekilde sorgulanması sağlanmıştır. Bellek içi veritabanı olarak çeşitli araştırmacılar tarafından yapılan performans testlerine göre Cassandra[25], HBase[26], MongoDB gibi diğer çok kullanılan bellek içi veritabanlarına göre daha performanslı olduğu raporlanan ve Twitter, Instagram, Flickr, GitHub gibi büyük çaplı verileri işleyen şirketler tarafından da kullanılan [144] Redis[27] seçilmiştir [145], [146]. Geliştirilen servis ile ilişkisel veritabanı (MySQL) üzerinde depolanan AFINN ve SenticNet duygu sözlük verileri Redis'e taşınmıştır. Redis ile veri alışverişinde bulunmak için Jedis[28] isimli açık kaynak kodlu Java kütüphanesi kullanılmıştır. Redis, verileri anahtar-değer (key-value) ikilisi şeklinde tuttuğundan ve aynı kaynak tanımları hem AFINN'de hem de SenticNet'de olabileceğinden kaynaklar Redis üzerinde tanımlanırken AFINN için "`afinn-`", SenticNet için "`senticnet-`" ön eki (prefix) kullanılmıştır. Kaynaklar, ilgili duygu sözlüklerinde aranırken bu ön eklerle birleştirilerek aranmıştır.

Sistem testlerinde kullanılan 3.531 adet tweet unigram'lar, bigram'lar, trigram'lar ve four-gram'lara dönüştürülerek elde edilen toplam 186.472 kaynak, hem ilişkisel veritabanında (MySQL) hem de bir kopyası bellek içi veritabanında (Redis) bulunan AFINN ve SenticNet duygu sözlüklerinde aranarak performans kıyaslaması yapılmıştır.

[25] http://cassandra.apache.org
[26] https://hbase.apache.org
[27] https://redis.io
[28] https://github.com/xetorthio/jedis

Çizelge 5.17'de listelenen sonuçlara göre, beklenildiği üzere verilerin ilişkisel veritabanından bellek içi veritabanına taşınması sistemin bu duygu sözlükleri üzerinde veri arama performansını 5,13 kat artırmıştır.

Çizelge 5.17. Test veri setindeki kaynaklara yönelik ilişkisel veritabanı (MySQL) ve bellek içi veritabanı (Redis) kullanılması durumunda elde edilen erişim süreleri.

Kaynak Sayısı	İlişkisel Veritabanı (MySQL) Süresi	Bellek İçi Veritabanı (Redis) Süresi
186.472	91 dakika 23 saniye (5.483 saniye)	17 dakika 48 saniye (1.068 saniye)

TwitterSentiDetector'un kaynaklara yönelik duygu yükü tespitinde kullandığı diğer sözlükler (SentiWordNet ve SentiStrength), kendine özgü veri yapısı ve bu veri yapılarını sorgulamak için geliştirilen kütüphaneler sunduğundan bu kapsamda değerlendirilmemiştir.

5.5. Twitter Spam Tespit Modülü

Twitter, popülerliği ve kolay anlaşılabilir mimarisi, gerçek zamanlı arama ve duygu-düşünce analiz servislerinin tweet'leri analiz ederek dünya üzerindeki gelişmeleri kısa bir gecikmeyle analiz edilebilmesine olanak vermektedir [18], [147]. Bu servislerin doğru sonuç sunabilmesi, "spam" olarak nitelendirilen tweet ve kullanıcıların filtrelenmesine bağlıdır. Twitter spam, istenmeyen, talep edilmemiş ve içerisinde kötücül linkler barındıran tweet'ler olarak tanımlanmaktadır [9], [148]–[150]. Twitter kullanıcıları, spam tweet'lerle başlıca amaçları şu şekilde sıralanabilir: (1) Kullanıcıları kötücül içerik (malware) barındıran harici web sayfaları ve yazılımlara yönlendirme, (2) kimlik hırsızlığı, (3) dolandırıcılık, (4) yönlendirilen linkler üzerinden satış yapma, (5) trend olan başlıkları gasp ederek, Twitter'ın "*cevaplama*" (reply) ve "*bahis*" (mention) özelliklerini kullanarak meşru kullanıcıların dikkatini çekmek [9], [56], [58]–[60], [150]–[156]. Twitter'ın popülaritesi, kolay kullanılabilir yapısı ve sunduğu gelişmiş uygulama geliştirme arayüzü sebebiyle "spam" tweet gönderen Twitter'ı kullanım amacı dışında kullanan kullanıcıların (spammer) da dikkatini çekmiştir. 2014 yılı Nisan ayında Twitter, binlerce tweet gönderen zararlı kullanıcı hesapları tarafından tweet seline uğramıştır [58]. 2014 yılı Ağustos ayında Twitter, aylık aktif kullanıcılarının %8,5'inin (yaklaşık 23 milyon kullanıcı) düzenli güncellemeler için otomatize edilmiş olarak sunucularına bağlandığını açığa vurmuştur [56], [57]. Bu sebeple, herhangi bir konuyla ilgili duygu analizi yapılırken Twitter API ile toplanan tweet'lerden spam tweet'lerin temizlenmesi, çevrimiçi arama ve veri madenciliği araç sonuçlarının doğruluğu için önem arz etmektedir

[9]. Twitter'da spam mesaj dağıtımının büyük bir çoğunluğu web adresleri ile yapıldığından [157], [158] tweet'in web adresi (Uniform Resource Locator - URL) içermesi TwitterSentiDetector spam tespiti için ön kabul olarak belirlenmiştir. Tweet'ler, Twitter'ın mesaj limiti olan 140 karakter sebebiyle az karakter kaplamasının sağlanması ve kötücül URL'leri gizlemek amacıyla genelde web adres kısaltma servisleri kullanılarak elde edilen kısaltılmış URL'ler barındırmaktadır [60], [135], [150], [156], [157], [159]–[163]. Kısaltılmış URL'ler, alan adlarıyla ilgili bilgi içermediğinden kötücül bir web sayfasına veya yazılıma yapılacak yönlendirme kolaylıkla saklanabilmektedir. Ayrıca bu hileden ötürü, geleneksel IP [164], alan adı ve erişim adresi [165] kara listeleme (blacklisting) yöntemleri tabanlı içerik filtreleme yaklaşımları Twitter spam tespiti için yetersiz olmaktadır [150]. Ayrıca kara listeleme yönteminin, kötücül web adreslerinin ilgili bilgi veritabanlarına dâhil edilmesi süreci göz önünde bulundurulduğunda oluşan gecikme süresi sebebiyle kullanıcıları korumak için çok yavaş olduğu belirtilmiştir [156]. Bu amaçla, tez çalışması bünyesinde Twitter'a özgü pratikleri ele alan spam tespit modülü geliştirilmiştir.

Twitter hem kullanıcı destekli hem de otomatize edilmiş servislerle spam ile mücadele etmektedir. Kullanıcılar, rahatsız oldukları hesapları profil sayfaları aracılığıyla gerekçesiyle birlikte Twitter'a bildirebilmektedir. Bu geribildirim mekanizması ile Twitter ilgili hesapları incelemekte ve amaç dışında bir kullanım tespit etmesi durumunda ilgili hesabı engellemektedir. Literatürde oldukça sık bir şekilde bahsedilen ancak Twitter'ın raporuna göre güncelliğini yitiren [151] bir yöntem ise resmi Twitter hesaplarından biri olan "*@spam*" hesabına ilgili kullanıcı adını bildirmektedir [60], [150]. Ayrıca bu yöntemin aldatma ve spam amacıyla kullanılarak suiistimal edildiği raporlanmıştır [60]. Bu yaklaşımların yoğun işi gücü gerektirdiği ve milyarlarca kullanıcısı olan bir platform için yetersiz olduğu düşünülmektedir. Twitter, spam tespiti için şu kriterleri kullanmaktadır [166]: (1) Bir tweet'in bir ya da birden çok hesap üzerinden tekrarlanması, (2) kısa zaman periyodunda çokça kullanıcı takip edilmesi, takipten çıkartılması, (3) hakkında oldukça çok sayıda spam raporunu düzenlenmiş olması, (4) saldırgan bir şekilde tweet'lerin beğenilmesi, retweet edilmesi, (5) kötücül içerik barındıran linklerin paylaşılması, (6) genellikle kişisel durum güncellemeleri yerine link içeren tweet'lerin paylaşılması, (7) trend olan başlıklara alakasız tweet gönderilmesi.

Geliştirilen TwitterSentiDetector Spam Tespit Modülü ile analiz için toplanan her bir tweet'in öncelikle spam olup olmadığı kontrol edilmekte ve spam olarak tespit edilmesi durumunda ilgili tweet'i yok saymaktadır. Bu hileyi önlemek için öncelikle tweet içerisindeki URL'lerin normal formlarına dönüştürülmesi gerekmektedir. TwitterSentiDetector, kısaltılmış URL'leri normal formuna dönüştürmek için ExpandURL[29], CheckShortURL[30] ve Link Expander[31] gibi çevrimiçi servisler yerine yerel bir çözüm olan `curl`[32] komut satırı aracından faydalanmaktadır. Böylelikle URL düzenlemesi için harici bir sisteme bağımlılığa gerek kalmamakta ve analiz süresi ek bir web servis kullanımına gerek kalmadığından büyümemektedir. Ayrıca harici web servisler ve uygulama geliştirme arayüzleri, sistemin kaynaklarının verimli kullanılabilmesi için limitler tanımlamaktadır. Yerel çözüm ile bu tip limitler de ortadan kaldırılmış olmaktadır. Çizelge 5.18'de kısaltılmış URL'ler, kullanılan kısaltma servisi ve `curl` komutu aracılığıyla elde edilen URL'lerin normal formları listelenmiştir.

Çizelge 5.18. Kısaltılmış URL'ler, kullanılan kısaltma servisi ve elde edilen normal formları.

Kısaltılmış URL	Kısaltma Servisi	URL'in Elde Edilen Normal Formu
fb.me/1FuyUcX3F	Facebook	http://javarevisited.blogspot.com.tr/2016/12/difference-between-thread-and-executor.html
goo.gl/nf6dfI	Google	https://madeby.google.com/home/learn/
https://t.co/c5nyOba3SG	Twitter	https://github.com/angular/angular/blob/master/CHANGELOG.md
bit.ly/ZJdItR	bitly	http://www.java67.com/2012/12/difference-between-arraylist-vs-LinkedList-java.html

TwitterSentiDetector, `curl` komutunu yürüterek kendisine parametre olarak girilen kısaltılmış URL'den normal formdaki URL'e dönüşüm gerçekleştirmektedir. Elde edilen normal formdaki URL'lerin spam tespiti için güvenlik servisinden doğrulanması

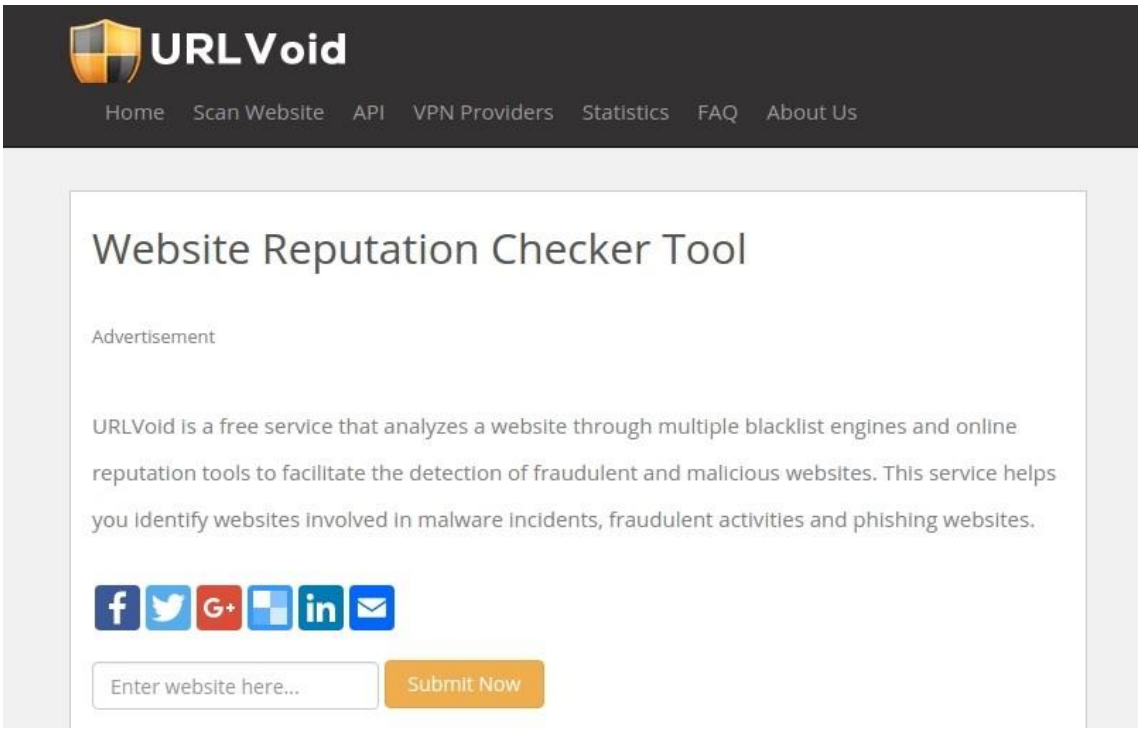
[29] http://www.expandurl.net
[30] http://checkshorturl.com
[31] http://www.linkexpander.com
[32] https://github.com/curl/curl

gerekmektedir. TwitterSentiDetector bu amaçla URLVoid[33] servisini kullanmaktadır. URLVoid, Çizelge 5.19'da sunulduğu üzere içerisinde Fortinet[34], GoogleSafeBrowsing[35], AVGThreatLabs[36], BitDefender[37], Sucuri[38] güvenlik servisleri de dâhil olmak üzere toplam 31 güvenlik servisi üzerinden parametre olarak iletilen URL'i sorgulamakta ve elde ettiği güvenlik itibar (security reputation) sonucunu sunmaktadır. URLVoid'in web kullanıcı görsel arayüzü Şekil 5.13'de sunulmuştur.

Çizelge 5.19. URLVoid'in üzerinden servis sağladığı güvenlik servisleri.

AVGThreatLabs	Avira	Bambenek Consulting	BitDefender
CyberCrime	c_APT_ure	DNS-BH	DrWeb
DShield	Fortinet	GoogleSafeBrowsing	hpHosts
Malc0de	MalwareDomainList	MalwarePatrol	MyWOT
OpenPhish	PhishTank	Quttera	Ransomware Tracker
SCUMWARE	Spam404	SURBL	ThreatCrowd
ThreatLog	urlQuery	URLVir	VXVault
WebSecurityGuard	YandexSafeBrowsing	ZeuS Tracker	

[33] http://www.urlvoid.com
[34] https://www.fortinet.com
[35] https://developers.google.com/safe-browsing/
[36] http://www.avgthreatlabs.com
[37] https://www.bitdefender.com
[38] https://sucuri.net



Şekil 5.13. URLVoid web kullanıcı görsel arayüzü.

URLVoid sunduğu API[39] ile uygulamaların kendisi üzerinden URL bazlı güvenlik sorgulaması yapılabilmesini sağlamaktadır. Sunulan API, HTTP (Hyper Text Transfer Protocol) POST isteği sorgulanmak istenen URL adresi ile sisteme erişilip, XML formatındaki sorgu sonucunun HTTP POST cevabı (response) ile iletilmesini sağlamaktadır. Örnek olarak http://bevrich.com zararlı URL'in URLVoid API ile sorgulanması sonucunda dönen cevap Şekil 5.14'de verilmiştir. URLVoid API'ından cevap olarak XML formatında dönen sonuç, ayrıştırılarak içerisinde tanımlı olan "`<engine>`" etiketleri elde edilmektedir. Elde edilen bu etiketler, hangi güvenlik servislerinde aranan URL'lerin zararlı URL olarak etiketlediğini temsil etmektedir. Ayrıştırma sonucunda elde edilen "`<engine>`" etiket sayısının 0 olması durumunda ilgili URL, TwitterSentiDetector tarafından zararsız olarak değerlendirilmektedir. Aksi takdirde tweet, en az bir güvenlik servisi tarafından zararlı olarak nitelendirilmiş URL barındırdığından spam olarak değerlendirilmekte ve analiz sürecinden çıkartılmaktadır.

[39] http://api.urlvoid.com

```
<?xml version="1.0" encoding="UTF-8"?>
<response>
    <details>
        <host>bevrich.com</host>
        <updated>1487148683</updated>
        <http_response_code>0</http_response_code>
        <domain_age>1301112000</domain_age>
        <google_page_rank>0</google_page_rank>
        <alexa_rank>0</alexa_rank>
        <connect_time>0</connect_time>
        <header_size>0</header_size>
        <download_size>0</download_size>
        <speed_download>0</speed_download>
        <external_url_redirect></external_url_redirect>
        <ip>
            <addr>69.163.192.23</addr>
            <hostname>ps12663.dreamhost.com</hostname>
            <asn>26347</asn>
            <asname>New Dream Network, LLC</asname>
            <country_code>US</country_code>
            <country_name>United States</country_name>
            <region_name>California</region_name>
            <city_name>Brea</city_name>
            <continent_code>NA</continent_code>
            <continent_name>North America</continent_name>
            <latitude>33.9269</latitude>
            <longitude>-117.861</longitude>
        </ip>
    </details>
    <detections>
        <engines>
            <engine>GoogleSafeBrowsing</engine>
            <engine>YandexSafeBrowsing</engine>
            <engine>Fortinet</engine>
```

```
        </engines>
        <count>3</count>
    </detections>
    <page_load>0.01</page_load>
</response>
```

Şekil 5.14. URLVoid API üzerinden http://bevrich.com zararlı URL'inin sorgulanması sonucunda elde edilen XML formatındaki cevap.

Twitter spam tespiti için tez çalışması kapsamında geliştirilen modül hem kullanıcı hem de tweet bazlı nitelikleri kullanmaktadır.

Çizelge 5.20'de sunulan kullanıcı bazlı nitelikler, spam tespiti için tweet sahibi hesapların (kullanıcıların) analizinden faydalanmaktadır. Kullanıcı bazlı niteliklerin birçoğu doğrudan kullanıcının kontrolünde olduğundan manipüle edilmesi mümkün olmaktadır. “Kullanıcının toplam takipçi sayısı”, “kullanıcının tweet'lerinin elde ettiği toplam retweet” sayısı gibi diğer bazı nitelikler ise kurulacak botnet ağı ile kolaylıkla manipüle edilebileceği düşünülmektedir. Bu sebeple, tez çalışmasında doğrudan kullanıcı doğrudan kontrolünde olan ve profil bilgisi kapsamındaki nitelikler spam tespiti için kullanılmamıştır.

Çizelge 5.20. Kullanıcı bazlı spam tespit nitelikleri.

Nitelik	Açıklama	Kullanıcı Kontrollü Mü?
Kullanıcı adı	Kullanıcının tekil kullanıcı adı	Evet
Biyografi	Kullanıcının biyografi bilgisi	Evet
Profil fotoğrafı	Kullanıcının profil fotoğrafı	Evet
Başlık fotoğrafı	Kullanıcının başlık fotoğrafı	Evet
Tema rengi	Kullanıcının tercih ettiği tema rengi	Evet
Doğum tarihi	Kullanıcının doğum tarihi bilgisi	Evet
Anasayfa bilgisi	Kullanıcının anasayfa bilgisi	Evet
Lokasyon	Kullanıcının lokasyon bilgisi	Evet
Hesap oluşturulma tarihi	Hesabın oluşturulma tarihi	Evet
Toplam tweet sayısı	Kullanıcının toplam tweet sayısı	Evet
Takipçi sayısı	Kullanıcının toplam takipçi sayısı	Hayır
Takip edilen kullanıcı sayısı	Kullanıcının takip ettiği kullanıcı sayısı	Evet
Toplam beğeni sayısı	Kullanıcının tweet'lerinin toplam beğeni sayısı	Hayır
Toplam retweet sayısı	Kullanıcının tweet'lerinin elde ettiği toplam retweet sayısı	Hayır
Liste sayısı	Kullanıcının oluşturduğu liste sayısı	Evet
Oluşturulan an sayısı	Kullanıcının oluşturduğu an sayısı	Evet

Spam tweet'ler incelendiğinde kullanıcı bazlı niteliklerle ilgili olarak aşağıdaki sonuçlara varılmıştır:

- Spamcılar (spam tweet atan kullanıcılar) zararsız kullanıcıların dikkatini çekmek

için oldukça fazla kullanıcı takip ettiğinden takip edilen kullanıcı sayısının yüksek olması beklenmektedir.

- Spamcılar zararsız kullanıcılar tarafından takip edilmediğinden takipçi sayısının düşük olması beklenmektedir.
- Spam tweet'ler zararsız kullanıcıların ilgi alanlarına hitap etmediğinden tweet'lere yapılan toplam beğeni ve toplam retweet sayısının düşük olması beklenmektedir.
- Spamcılar zararsız kullanıcıların dikkatini çekmek için çok çeşitli konularda tweet paylaştığından toplam tweet sayısının yüksek olması beklenmektedir.
- Spam tweet'ler zararsız kullanıcıların dikkatini çekmek için bahis (mention) ve hashtag'lerden faydalandığından attıkları tweet'de bahis ve hashtag sayısının yüksek olması beklenmektedir.
- Spam tweet'ler zararsız kullanıcılar tarafından genellikle göz ardı edildiğinden spam tweet'lere yönelik cevap sayısının düşük olması beklenmektedir.
- Spamcılar bir ya da birden çok hesap tarafından otomatize edilmiş benzer ya da aynı içerikli tweet'ler attığından tweet içerik benzerliklerinin yüksek olması beklenmektedir.
- Zararsız kullanıcılar faydalı tweet içeriklerinden dolayı Twitter'da kurulan listelere dahil edildiğinden spamcıların dahil edildiği liste sayısının düşük olması beklenmektedir.

Çizelge 5.21'de sunulan tweet bazlı nitelikler, ilgili tweet'in doğrudan analiz edilmesi sonucunda elde edilen niteliklerdir.

Çizelge 5.21. Tweet bazlı spam tespit nitelikleri.

Nitelik	Açıklama	Kullanıcı Kontrollü Mü?
Gönderici	Tweet'i gönderen kullanıcı bilgisi	Evet
Bahisler	Tweet'de bahis edilen kullanıcılar	Evet
Hashtag'ler	Tweet'deki hashtag'ler	Evet
Link	Tweet içeriğindeki link	Evet
Tweet beğeni sayısı	Tweet'in beğeni sayısı	Hayır
Tweet retweet sayısı	Tweet'in retweet sayısı	Hayır
Tweet cevap sayısı	Tweet'e verilen cevap sayısı	Hayır
Tweet gönderilme tarihi	Tweet'in gönderilme tarihi	Evet
Tweet gönderilme yeri	Tweet'in gönderilme yeri	Evet

Spam tweet'ler incelendiğinde tweet bazlı spam tespit nitelikleriyle ilgili olarak aşağıdaki sonuçlara ulaşılmıştır:

- Spam tweet'ler, zararsız kullanıcıların dikkatini çekmek için linklerden faydalanmaktadır.
- Spam tweet'ler, zararsız kullanıcıların dikkatini çekmek için bahislerden faydalanmaktadır.
- Spam tweet'ler, zararsız kullanıcıların dikkatini çekmek için hashtag'lerden faydalanmaktadır.
- Spam tweet'ler, ilgisiz tweet'ler olduğundan toplam beğeni ve toplam retweet sayısının düşük olması beklenmektedir.

iPhone 7'nin duyurulduğu gün toplanan ve içerisinde "*#iPhone7*" hashtag'i içeren Twitter4J kütüphanesi kullanılarak toplanan İngilizce 77.033 tweet analiz edildiğinde Çizelge 5.22'de gerekçeleri ile belirtilen nitelikler spam tespit modülünde kullanılmamıştır.

Çizelge 5.22. TwitterSentiDetector spam tespit modülü tarafından göz ardı edilen nitelikler ve gerekçeleri.

Nitelik	İhmal Edilme Gerekçesi
Takipçi sayısı	(1) Takipçi sayısı bot'lar kullanılarak kolaylıkla manipüle edilebilir, (2) bazı zararsız kullanıcılar çok fazla tweet göndermediğinden ve Twitter'ı ilgi alanlarını takip etmek için kullandığından çok fazla takipçisi olmamaktadır.
Takip edilen kullanıcı sayısı	(1) Takip edilen kullanıcı sayısı kullanıcı kontrollü olduğundan manipülasyona açıktır, (2) hem zararsız hem de spamcılar kendi amaçları doğrultusunda çok sayıda kullanıcıyı takip edebilmektedir.
Tweet sayısı	(1) Tweet gönderme işlemi kullanıcı kontrollü olduğundan manipülasyona açıktır, (2) spamcıların yanı sıra haber ajansları, sosyal medya uzmanları gibi bazı zararsız kullanıcılar da çok sayıda tweet atabilmektedir, (3) bazı spamcılar, spam tweet'ler popülerliğini kaybettiği zaman ilgili tweet'lerini silmektedirler.
Liste sayısı	Oluşturulan liste sayısı kullanıcı kontrollü olduğundan manipülasyona açıktır.
Hesap Oluşturulma tarihi	Hesabın oluşturulma tarihinin spam tespiti açısından tek başına anlamsızdır.
Minimum tweet gönderme aralığı Ortalama tweet gönderme aralığı	Hem zararsız kullanıcılar hem de spamcılar çeşitli araçlar kullanarak otomatize tweet'ler gönderdiğinden minimum ve ortalama tweet gönderme aralıkları spam tespiti açısından anlamsızdır.

TwittterSentiDetector spam tespit modülü, manipüle edilmesi diğer niteliklere göre oldukça zor olan spam tespiti için Denklem (5.2)'de formülü sunulan ortalama hashtag sayısını ve en yüksek tweet benzerlik niteliklerini kullanmaktadır. Bir kullanıcının tweet'leri içerisindeki en yüksek tweet benzerliği, iki vektör ve arasındaki kosinüs açısından faydalanarak benzerlik ölçümü yapılmasını sağlayan ve metinsel benzerlik tespitinde geçerliliğini ispatlamış olan Cosine benzerlik yöntemi [167]–[169] kullanılarak hesaplanmaktadır. Veri setindeki 77.033 tweet'in 50.490 tekil kullanıcı tarafından atıldığı görülmüştür. Veri seti ile ilgili istatistikler Çizelge 5.23'de sunulmuştur.

Çizelge 5.23. Spam tespit modülü veri seti istatistikleri.

Veri seti	Zararsız Kullanıcı Sayısı	Spam Kullanıcı Sayısı	Toplam
Eğitim	32.392	2.371	35.343
Test	14.111	1.036	15.147
Toplam	47.083	3.407	50.490

Veri seti analiz edildiğinde üç hashtag'in (*#iPhone7*, *#Apple* ve *#iphone7*) tüm tweet'lerin %9'dan daha fazlasında bulunması sebebiyle üçten fazla hashtag barındıran tweet'ler spam olarak kabul edilmiştir. Tez çalışması kapsamında spam tespit modülü için belirlenen nitelikler (1) spamcılar, zararsız kullanıcıların dikkatini çekmek için özellikle trend olan hashtag'lerden ilgili tweet'lerinde faydalanmaktadır, (2) spamcılar bir veya birden fazla hesap tarafından yönetilen özellikle trend olan konuları hedef alan benzer ya da aynı tweet'ler paylaşmaktadır esaslarını temel almaktadır. Sınıflandırıcılar, yeni örneklerin, önceki örneklemlerin gözlemlenmesi sonucu kategorilere ayrılmasını sağlayan makine öğrenmesi yöntemleridir [170]. TwitterSentiDetector spam modülü, geçerliliğini metin ve spam e-posta tespitinde kanıtlamış olan Naive Bayes yöntemi [170]–[173] ile eğitilmiştir. Veri seti %70 eğitim, %30 test olmak üzere paylaştırılmıştır. Bu yaklaşımla öne sürülen spam tespit modülünün doğruluğu ve hassasiyeti sırasıyla 0,943 ve 0,913 olarak hesaplanmıştır. Elde edilen sonuçların karmaşıklık matrisi Çizelge 5.24'de sunulmuştur.

$$OrtalamaHashtagSayısı = \frac{\sum_0^n Toplam\ HashtagSayısı}{n} \quad (5.2)$$

Çizelge 5.24. Test sonuçlarının karmaşıklık matrisi.

Sınıf	Pozitif (P)	Negatif (N)	Toplam
Doğru (T)	786	13.040	13.826
Yanlış (F)	1.071	250	1.321
Toplam	1.857	13.290	15.147

Çizelge 5.22'de gerekçeleriyle beraber sunulduğu üzere bazı nitelikler spam tespit modülü tarafından kullanılmamıştır. Bu niteliklerin sistemin başarımına olan etkisini gözlemlemek için her bir nitelik tek tek sisteme entegre edilmiş ve sistemin başarımı gözlemlenmiştir. Sonuçlar Çizelge 5.25'de sunulduğu üzere bu niteliklerden hiçbiri sistemin doğruluğunu ve hassasiyetini artırmamıştır.

Çizelge 5.25. Twitter spam tespiti için göz ardı edilen niteliklerin sisteme dahil edilmesinde durumunda sistemin doğruluğu ve hassasiyeti.

Dahil Edilen Nitelik	Doğruluk	Hassasiyet
Tweet sayısı	0,878	0,891
Takipçi sayısı	0,938	0,624
Takip edilen kullanıcı sayısı	0,936	0,834
Liste sayısı	0,937	0,836
Minimum tweet gönderme aralığı	0,922	0,939
Ortalama tweet gönderme aralığı	0,943	0,914

6. SONUÇLAR VE ÖNERİLER

TwitterSentiDetector, doğal dil işleme ve dilbilimsel yöntemler kullanarak tweet'leri pozitif, negatif ve nötr olarak üç duygu sınıfıyla etiketleyen alan bağımsız ve eğiticisiz bir Twitter duygu analizi sistemidir. TwitterSentiDetector, elde edilen bilgilere göre İngilizce dışında erişime açık ve literatürde kullanılan duygu sözlüğü olmadığından İngilizce tweet'leri analiz etmektedir. Ancak geliştirilen sistem, modüler olarak tasarlandığından diğer dillere uyarlanabilecek esnekliktedir. TwitterSentiDetector, tanım kümesi bağımsız bir yapıda tasarlandığından makine öğrenmesi temelli diğer Twitter duygu analizi sistemleri gibi eğitime ihtiyaç duymamaktadır. Böylelikle eğitim veri setinin heterojenliği ve dağılımından kaynaklanabilecek olumsuzluklardan uzak bir sistem geliştirilmiştir. TwitterSentiDetector, kaynaklara yönelik polarite tanımını, literatürde kabul görmüş birden çok sözlük üzerinden yaparak daha tutarlı bir yaklaşım önermektedir. Çalışmanın doğruluğu ve hassasiyetinin, makine öğrenmesi yöntemleri kullanılarak ilgili alanda sistemin eğitimden geçirilmesi sonucunda artırılması mümkündür ancak bu durum literatürde de sıklıkla bahsedildiği gibi alan bağımlılığına sebebiyet vereceğinden ve alan bağımlılığın Twitter'ın doğasına aykırı olduğu düşünüldüğünden tez çalışmasında tercih edilmemiştir. Verisetleri incelendiğinde bazı tweet'lerin ihtiva ettiği anlamlar oldukça öznel olduğundan bu tip tweet'lere yönelik doğru duygu tespitinde bulunulabilmesi için tweet sahibinin detaylı analiz edilmesi gerekmektedir. Tez çalışmasının katkıları aşağıdaki gibi listelenebilir:

- TwitterSentiDetector, Twitter'da kullanılan ve dilbilimsel kurallardan uzak kelime ve kısaltmaların, yazım hatalarının, sözlüklerde karşılığı olmayan yeni türetilmiş kelimelerin, yüz ifadelerinin, metinsel gülmelerin, emoji'lerin, harf ve noktalama işareti tekrarlarının, hashtag'lerin sıklıkla kullanıldığı ve geleneksel duygu analizi sistemlerinin uygulanamayacağı durumlarda başarılı sonuçlar veren bir Twitter duygu analizi sistemidir.
- Tweet'ler gündelik hayattaki hemen hemen her olayla ilgili olabildiğinden TwitterSentiDetector özellikle tanım kümesi bağımsız bir duygu analizi sistemi olarak tasarlanmıştır. TwitterSentiDetector böylelikle eğitime ihtiyaç duymamakta, eğitim veri setinin heterojenliği ve dağılımından kaynaklanabilecek

hatalardan etkilenmemektedir.

- TwitterSentiDetector, doğal dil işleme yöntemlerini kullanarak tweet'i oluşturan kelimelerin cümlenin içindeki türünü (POS etiketi) tespit ettiğinden kelime ve kelime gruplarına doğrudan sözlükteki anlamlarına göre değil, tweet içerisindeki asıl anlamlarına göre polarite değeri yüklenmesini sağlamaktadır.
- TwitterSentiDetector, kısaltmaları anlamlandırmak için literatürde sıklıkla kullanılan yöntem olan sabit kısaltma listelerinin kullanımı yerine çevrimiçi ve sürekli büyüyen dinamik bir kısaltma veritabanından faydalanmaktadır.
- TwitterSentiDetector, entegre çalıştığı graf tabanlı cümle seviyesinde yazım hatası düzeltme algoritması ile olası yazım hatalarından kaynaklanabilecek duygu anlamlandıramama durumunun önüne geçilmesini sağlamaktadır.
- TwitterSentiDetector, herhangi bir konuyla ilgili duygu analizi yapmadan önce biriktirdiği tweet'leri spam analizinden geçirerek amaçlarından birisi de otomatize edilmiş sistemlerin analizlerinin geçerliliğini azaltmak olan spam tweet'leri filtreleyerek daha doğru duygu analizinin gerçekleştirilmesini sağlamaktadır.
- TwitterSentiDetector, olumsuzluk tespiti için sadece olumsuzluk eklerinden değil, bağlaçlardan da faydalanmaktadır. Böylelikle zıt duyguları bir arada bulunduran, literatürdeki bazı çalışmalar tarafından belirsiz (ambiguous) olarak nitelendirilen ve değerlendirme dışı bırakılan tweet'lere yönelik de duygu tespitinde bulunmaktadır.
- Literatürde kullanılan kelimelerdeki fazladan harf tekrarlarının sabit bir sayıya indirgemek yerine WordNet sözlüğü üzerinden özyinelemeli (recursive) olarak kontrol edilerek anlamlı en sade hale dönüştürülmesini sağlamaktadır.
- TwitterSentiDetector, duygu analizi esnasında literatürde sıklıkla kullanılan unigram ve bigram'ların yanı sıra trigram ve four-gram'ların da duygu yükleri göz önünde bulundurularak daha kapsamlı duygu analizi yapmaktadır.
- TwitterSentiDetector, unigram analizi öncesinde kelimelerin temel hallerinin WordNet sözlüğü üzerinden elde edilerek isimler için çoğul eklerinden, fiiller için ise zaman çekimlerinden kaynaklanan kelime biçim değişikliklerini tespit ederek duygu sözlüklerinde tanımlı olan temel halleri üzerinden polarite hesabını yapmaktadır.

- TwitterSentiDetector, duygu tespitini ön tanımlı, kısıtlı pozitif ve negatif kelime listelerine göre değil; literatürde kabul görmüş duygu sözlükleri kullanarak yapmaktadır. Ayrıca birden çok duygu sözlüğü kullanarak daha tutarlı polarite değerlerinin elde edilmesini sağlanmaktadır.
- SenticNet sözlüğünün kaynak tanımlarını, RDF/XML olarak sunmasından ötürü arama performansının düşük olmasının üstesinden gelmek için TwitterSentiDetector, SenticNet kaynak tanımlarını ayrıştırarak NoSQL veritabanında depolamakta ve kaynakları, performans kazancı tez çalışması kapsamında kanıtlanmış olan NoSQL veritabanından sorgulamaktadır.
- Veri setlerinde nötr duygu etiketleriyle etiketlenen tweet'lerde pozitif ve negatif duygu yükü barındıran kelimelerin olmasından ötürü TwitterSentiDetector parametrik olarak hem üçlü (pozitif, negatif, nötr) hem de ikili (pozitif, negatif) duygu tespitinde bulunabilmektedir.

Gelecekte Türkçe duygu sözlüklerinin oluşturulması ve gelişmiş Türkçe doğal dil işleme kütüphanelerinin geliştirilmesi sonrasında tez çalışmasıyla öne sürülen Twitter tabanlı duygu analizi sisteminin kolaylıkla Türkçe'ye uyarlanabileceği düşünülmektedir. Gelecekteki bir geliştirme olarak, kişilik analizi yapan bir modül sisteme entegre edilerek tweet'ler analiz edilirken tweet sahibinin kullandığı dilin de göz önünde bulundurulmasıyla daha hassas bir duygu analizinin sağlanabileceği düşünülmektedir.

7. KAYNAKLAR

[1] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in *Proceedings of LREC'10, the Seventh International Conference on Language Resources and Evaluation*, 2010, pp. 1320–1326.

[2] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Proceeding COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 36–44.

[3] S. Prentice and E. Huffman, "Social Media's New Role in Emergency Management," in *Emergency Management and Robotics for Hazardous Environments*, Albuquerque, Mexico: Idaho National Laboratory, 2008, pp. 1–5.

[4] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010, pp. 178–185.

[5] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez, "On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 9, pp. 1799–1816, Sep. 2015.

[6] E. Glynn, B. Fitzgerald, and C. Exton, "Commercial adoption of open source software: An empirical study," in *2005 International Symposium on Empirical Software Engineering, ISESE 2005*, 2005, pp. 225–234.

[7] K. Dehinbo, P. Pretorius, and J. Dehinbo, "Strategic analysis towards deriving competitive advantage with the use of FOSS: The case of a South African university," in *Proceedings of the 9th International Conference on Information Technology, ITNG 2012*, 2012, pp. 335–342.

[8] J. L. Mindel, L. Mui, and S. Verma, "Open source software adoption in ASEAN member countries," in *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, 2007.

[9] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," in *CEAS 2010 - Seventh annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, 2010, vol. 6, pp. 12–21.

[10] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.

[11] J. DiGrazia, K. McKelvey, J. Bollen, and F. Rojas, "More tweets, more votes: Social media as a quantitative indicator of political behavior," *PLoS One*, vol. 8, no. 11, 2013.

[12] Twitter, "Twitter Reports First Quarter 2015 Results; Lowers Full-Year 2015 Expectations," *Twitter*, 2015. [Online]. Available: https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=909177. [Accessed: 28-Apr-2017].

[13] Statista, "Twitter: number of monthly active users 2015 | Statistic," *statista*, 2015. [Online]. Available: http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/. [Accessed: 28-Apr-2017].

[14] “Twitter Usage Statistics - Internet Live Stats,” *InternetLiveStats*, 2017. [Online]. Available: http://www.internetlivestats.com/twitter-statistics/. [Accessed: 28-Apr-2017].

[15] “Company | About,” *Twitter*, 2017. [Online]. Available: https://about.twitter.com/company. [Accessed: 28-Apr-2017].

[16] D. Sayce, “Number of tweets per day?,” 2016. [Online]. Available: http://www.dsayce.com/social-media/tweets-day/. [Accessed: 12-Apr-2017].

[17] M. E. Larsen, T. W. Boonstra, P. J. Batterham, B. ODea, C. Paris, and H. Christensen, “We Feel: Mapping Emotion on Twitter,” *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 4, pp. 1246–52, Jul. 2015.

[18] A. DuVander, “Which APIs Are Handling Billions of Requests Per Day? | ProgrammableWeb,” *Programmable Web*, 2012. [Online]. Available: http://www.programmableweb.com/news/which-apis-are-handling-billions-requests-day/2012/05/23. [Accessed: 28-Apr-2017].

[19] E. Haddi, X. Liu, and Y. Shi, “The Role of Text Pre-processing in Sentiment Analysis,” *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013.

[20] H. Saif, Y. He, M. Fernandez, and H. Alani, “Semantic Patterns for Sentiment Analysis of Twitter,” in *Proceedings of The 13th International Semantic Web Conference (ISWC)*, 2014, pp. 324–340.

[21] H. Tang, S. Tan, and X. Cheng, “A survey on sentiment detection of reviews,” *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10760–10773, 2009.

[22] P. Burnap, O. F. Rana, N. Avis, M. Williams, W. Housley, A. Edwards, J. Morgan, and L. Sloan, “Detecting tension in online communities with computational Twitter analysis,” *Technol. Forecast. Soc. Change*, vol. 95, pp. 96–108, Jun. 2015.

[23] D. Derks, A. E. R. Bos, and J. von Grumbkow, “Emoticons and social interaction on the Internet: the importance of social context,” *Comput. Human Behav.*, vol. 23, no. 1, pp. 842–849, 2007.

[24] T. Schnoebelen, “Do You Smile with Your Nose? Stylistic Variation in Twitter Emoticons,” *Univ. Pennsylvania Work. Pap. Linguist.*, vol. 18, no. 2, p. Article 14, 2012.

[25] J. Park, V. Barash, C. Fink, and M. Cha, “Emoticon Style : Interpreting Differences in Emoticons Across Cultures,” in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2013, pp. 466–475.

[26] T. Dimson, “Emojineering Part 1 : Machine Learning for Emoji Trends,” *Instagram*, 2015. [Online]. Available: https://engineering.instagram.com/emojineering-part-1-machine-learning-for-emoji-trendsmachine-learning-for-emoji-trends-7f5f9cb979ad. [Accessed: 07-May-2017].

[27] U. Pavalanathan and J. Eisenstein, “Emoticons vs. Emojis on Twitter: A Causal Inference Approach,” in *AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content*, 2016.

[28] P. K. Novak, J. Smailović, B. Sluban, and I. Mozetič, “Sentiment of emojis,” *PLoS One*, vol. 10, no. 12, 2015.

[29] A. R. Damasio, “Aphasia,” *N. Engl. J. Med.*, vol. 326, no. 8, pp. 531–539, 1992.

[30] L. Van Fossen, “The communicative use of iconic face drawings to express emotional and evaluative statements in persons with aphasia,” California State University, 2015.

[31] M. L. Rose, “The utility of arm and hand gestures in the treatment of aphasia,” *Adv. Speech Lang. Pathol.*, vol. 8, no. 2, pp. 92–109, Jan. 2006.

[32] E. Power and C. Code, “Waving not drowning: Utilising gesture in the treatment of aphasia,” *Adv. Speech-Language Pathol.*, vol. 8, no. 2, pp. 115–119, 2006.

[33] L. Lanyon and M. L. Rose, “Do the hands have it? The facilitation effects of arm and hand gesture on word retrieval in aphasia,” *Aphasiology*, vol. 23, no. 7–8, pp. 809–822, Jul. 2009.

[34] E. Daemen, P. Dadlani, J. Du, Y. Li, and P. Erik-paker, “Designing a Free Style , Indirect , and Interactive Storytelling Application for People with Aphasia,” in *INTERACT’07 Proceedings of the 11th IFIP TC 13 international conference on Human-computer interaction*, 2007, pp. 221–234.

[35] T. a. Rose, L. E. Worrall, L. M. Hickson, and T. C. Hoffmann, “Exploring the use of graphics in written health information for people with aphasia,” *Aphasiology*, vol. 25, no. 12, pp. 1579–1599, 2011.

[36] “Samsung Wemogee: A New Communication Tool for People with Language Disorders – Samsung Global Newsroom,” *Samsung*, 2017. [Online]. Available: https://news.samsung.com/global/samsung-wemogee-a-new-communication-tool-for-people-with-language-disorders. [Accessed: 07-May-2017].

[37] C. Whitelaw, C. Whitelaw, N. Garg, N. Garg, S. Argamon, and S. Argamon, “Using appraisal groups for sentiment analysis,” in *Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM ’05*, 2005, pp. 625–631.

[38] A. Abbasi, S. France, Z. Zhang, and H. Chen, “Selecting attributes for sentiment classification using feature relation networks,” *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 3, pp. 447–462, 2011.

[39] J. Spencer and G. Uchyigit, “Sentimentor: Sentiment analysis of twitter data,” in *Proceeding of The First International Workshop on Sentiment Discovery from Affective Data (SDAD 2012)*, 2012, vol. 917, pp. 56–66.

[40] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP ’02*, 2002, vol. 10, pp. 79–86.

[41] A. Montejo-Ráez, E. Martinez-Cámara, M. T. Martin-Valdivia, and L. A. Ureña-López, “Random Walk Weighting over SentiWordNet for Sentiment Polarity Detection on Twitter,” *Comput. Speech Lang.*, vol. 28, no. 1, pp. 93–107, 2014.

[42] R. Pandarachalil, S. Sendhilkumar, and G. S. Mahalakshmi, “Twitter Sentiment Analysis for Large-Scale Data: An Unsupervised Approach,” *Cognit. Comput.*, vol. 7, pp. 254–262, 2015.

[43] S. Rosenthal, A. Ritter, P. Nakov, and V. Stoyanov, “SemEval-2014 Task 9: Sentiment Analysis in Twitter,” *Proc. 8th Int. Work. Semant. Eval. (SemEval 2014)*, no. SemEval, pp. 73–80, 2014.

[44] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proceedings of Fifth International AAAI Conference on Weblogs and Social Media - ICWSM 11*, 2010, pp. 122–129.

[45] S. Shahheidari, H. Dong, and M. N. R. Bin Daud, "Twitter sentiment mining: A multi domain analysis," in *Proceedings of 2013 7th International Conference on Complex, Intelligent, and Software Intensive Systems, CISIS 2013*, 2013, pp. 144–149.

[46] H. Li, Y. Chen, H. Ji, S. Muresan, and D. Zheng, "Combining social cognitive theories with linguistic features for multi-genre sentiment analysis," in *Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC-2012)*, 2012, pp. 127–136.

[47] M. del P. Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. Á. Rodríguez-García, R. Valencia-García, Rodr&#xed, guez-Garc&#xed, M. &#xc1 A, ngel, Valencia-Garc&#xed, and R. A, "Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach," *Comput. Math. Methods Med.*, vol. 2017, pp. 1–9, 2017.

[48] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," *Processing*, vol. 150, no. 12, pp. 1–6, 2009.

[49] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 2011, pp. 30–38.

[50] H. Saif, Y. He, and H. Alani, "Alleviating data sparsity for twitter sentiment analysis," in *Proceedings of the WWW'12 Workshop on "Making Sense of Microposts,"* 2012, vol. 838, pp. 2–9.

[51] T. Singh and M. Kumari, "Role of Text Pre-processing in Twitter Sentiment Analysis," in *12th International Conference on Communication Networks (ICCN) / 12th International Conference on Data Mining and Warehousing (ICDMW) / 12th International Conference on Image and Signal Processing (ICISP)*, 2016, vol. 89, pp. 549–554.

[52] E. Frydenberg, "Coping research: Historical background, links with emotion, and new research directions on adaptive processes," *Aust. J. Psychol.*, vol. 66, no. 2, pp. 82–92, 2014.

[53] M. T. Ikram, N. A. Butt, and M. T. Afzal, "Open source software adoption evaluation through feature level sentiment analysis using Twitter data," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 24, no. 5, pp. 4481–4496, 2016.

[54] S. Liu, X. Cheng, F. Li, and F. Li, "TASC: Topic-Adaptive Sentiment Classification on Dynamic Tweets," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 6, pp. 1696–1709, Jun. 2015.

[55] A. T. Kabakus and R. Kara, "A Survey of Spam Detection Methods on Twitter," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 3, pp. 29–38, 2017.

[56] Z. M. Seward, "Twitter admits that as many as 23 million of its active users are automated," *Quartz*, 2014. [Online]. Available: http://qz.com/248063/twitter-admits-that-as-many-as-23-million-of-its-active-users-are-actually-bots/. [Accessed: 28-Apr-2017].

[57] L. Whitney, “Twitter says as many as 23 million accounts connect with automated services,” *CNET*, 2014. [Online]. Available: https://www.cnet.com/news/twitter-reveals-23-million-of-accounts-active/. [Accessed: 28-Apr-2017].

[58] D. Goodin, “Mystery attack drops avalanche of malicious messages on Twitter,” *Ars Technica*, 2014. [Online]. Available: http://arstechnica.com/security/2014/04/mystery-attack-drops-avalanche-of-malicious-messages-on-twitter/. [Accessed: 28-Apr-2017].

[59] X. Zhang, S. Zhu, and W. Liang, “Detecting Spam and Promoting Campaigns in the Twitter Social Network,” in *IEEE International Conference on Data Mining (ICDM 2012)*, 2012, pp. 1194–1199.

[60] A. H. Wang, “Don’t follow me: Spam detection in Twitter,” in *SECRYPT 2010 - Proceedings of the International Conference on Security and Cryptography*, 2010, pp. 1–10.

[61] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, “A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle,” *Proc. 50th Annu. Meet. Assoc. Comput. Linguist.*, no. July, pp. 115–120, 2012.

[62] J. Read and J. Carroll, “Weakly supervised techniques for domain-independent sentiment classification,” in *Proceeding of the 1st international CIKM workshop on Topicsentiment analysis for mass opinion TSA 09*, 2009, p. 45.

[63] D. J. S. Oliveira, P. H. de S. Bermejo, and P. A. dos Santos, “Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls,” *J. Inf. Technol. Polit.*, vol. 14, no. 1, pp. 34–45, 2017.

[64] Y. Bae and H. Lee, “Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 12, pp. 2521–2535, Dec. 2012.

[65] V. A. Kharde and S. S. Sonawane, “Sentiment Analysis of Twitter Data: A Survey of Techniques,” *Int. J. Comput. Appl.*, vol. 139, no. 11, pp. 975–8887, 2016.

[66] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, “Combining lexicon-based and learning-based methods for Twitter sentiment analysis,” *Int. J. Electron. Commun. Soft Comput. Sci. Eng.*, vol. 89, pp. 1–8, 2015.

[67] Y.-M. Li and T.-Y. Li, “Deriving market intelligence from microblogs,” *Decis. Support Syst.*, vol. 55, no. 1, pp. 206–217, Apr. 2013.

[68] N. K. Alex Cheng, Mark Evans, “Inside the Political Twittersphere,” *sysomos*, 2009. [Online]. Available: https://sysomos.com/inside-twitter/political-twittersphere. [Accessed: 28-Apr-2017].

[69] A. Ceron, L. Curini, and S. M. Iacus, “First- and second-level agenda setting in the Twittersphere: An application to the Italian political debate,” *J. Inf. Technol. Polit.*, vol. 13, no. 2, pp. 159–174, 2016.

[70] S. Rill, D. Reinel, J. Scheidt, and R. V. Zicari, “PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis,” *Knowledge-Based Syst.*, vol. 69, no. 1, pp. 24–33, 2014.

[71] A. Bermingham and A. F. Smeaton, “On Using Twitter to Monitor Political Sentiment and Predict Election Results,” *Psychology*, pp. 2–10, 2011.

[72] J. Chung and E. Mustafaraj, “Can collective sentiment expressed on twitter predict

political elections?," *Aaai*, pp. 1770–1771, 2011.

[73] Y. Mejova, P. Srinivasan, and B. Boynton, "GOP Primary Season on Twitter: Popular Political Sentiment in Social Media," in *The Sixth ACM International Conference*, 2013, p. 517.

[74] M. Ringsquandl and D. Petković, "Analyzing Political Sentiment on Twitter," *AAAI Spring Symp. Ser. Anal. Microtext*, pp. 40–47, 2013.

[75] S. Asur and B. A. Huberman, "Predicting the Future with Social Media," *Computing*, vol. 1, pp. 492–499, 2010.

[76] D. P. Chapman and G. S. Perry, "Depression as a Major Component of Public Health for Older Adults," *Prev. Chronic Dis.*, vol. 5, no. 1, p. A22, 2008.

[77] M. S. Reddy, "Depression: The Disorder and The Burden," *Indian J. Psychol. Med.*, vol. 32, no. 1, pp. 1–2, Jan. 2010.

[78] L. Mayoh, "Are you hiding depression behind a smile?," *bodyandsoul*, 2015. [Online]. Available: http://www.bodyandsoul.com.au/health/health+news/are+you+hiding+depression+behind+a+smiler,14985. [Accessed: 28-Apr-2017].

[79] D. D. A. M. Chandrasekaran, "Depression can hide behind a smile," *THE STAR*, 2014. [Online]. Available: http://www.thestar.com.my/Lifestyle/Health/2014/12/21/Depression-can-hide-behind-a-smile/. [Accessed: 28-Apr-2017].

[80] J. Cawte, "Aboriginal Recovery from Depression: Eleven Clinical Types," *Aborig. Isl. Health Work. J.*, vol. 5, no. 3, pp. 14–31, Sep. 1981.

[81] W. Bucci, *Psychoanalysis and Cognitive Science: A Multiple Code Theory*, 1st ed. New York, NY, USA: The Guilford Press, 1997.

[82] Z. Lothane, "The Power of the Spoken Word in Life, Psychiatry, and Psychoanalysis—A Contribution to Interpersonal Psychoanalysis," *Am. J. Psychoanal.*, vol. 67, no. 3, pp. 260–274, 2007.

[83] I. Modai, A. Bleich, and G. Cygielman, *Masked depression--an ambiguous entity.*, vol. 37, no. 4. 1982, pp. 235–240.

[84] F. Bravo-Marquez, M. Mendoza, and B. Poblete, "Combining strengths, emotions and polarities for boosting Twitter sentiment analysis," in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM '13)*, 2013, pp. 1–9.

[85] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 1, pp. 163–173, 2012.

[86] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter Sentiment Classification," in *Proceeding of HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 151–160.

[87] Y. Zhang, J. Tang, J. Sun, Y. Chen, and J. Rao, "MoodCast: Emotion prediction via dynamic continuous factor graph model," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2010, pp. 1193–1198.

[88] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li, "User-level Sentiment

Analysis Incorporating Social Networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 1397–1405.

[89] A. Kanavos, N. Nodarakis, S. Sioutas, A. Tsakalidis, D. Tsolis, and G. Tzimas, "Large Scale Implementations for Twitter Sentiment Classification," *Algorithms*, vol. 10, no. 1, p. 33, 2017.

[90] G. A. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[91] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010, pp. 2200–2204.

[92] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "WordNet: An on-line lexical database," *Int. J. Lexicogr.*, vol. 3, pp. 235–244, 1990.

[93] A. Esuli and F. Sebastiani, "SENTIWORDNET: A high-coverage lexical resource for opinion mining," in *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, 2006, pp. 417–422.

[94] F. S. Andrea Esuli, "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining," in *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, 2006, pp. 417–422.

[95] K. Denecke, "Using SentiWordNet for multilingual sentiment analysis," in *Proceedings of the 2008 IEEE 24th international conference on data engineering workshop*, 2008, pp. 507–512.

[96] T. Ogawa, Q. Ma, and M. Yoshikawa, "News bias analysis based on stakeholder mining," *IEICE Trans. Inf. Syst.*, vol. E94–D, no. 3, pp. 578–586, 2011.

[97] E. Cambria, C. Havasi, and A. Hussain, "SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis," in *The Twenty-Fifth Florida Artificial Intelligence Research Society Conference*, 2012, pp. 202–207.

[98] E. Cambria, D. Olsher, and K. Kwok, "SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis," in *The Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI-14)*, 2014, pp. 1515–1521.

[99] E. Cambria, T. Benson, C. Eckl, and A. Hussain, "Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10533–10543, 2012.

[100] A. H. Erik Cambria, *Sentic Computing - Techniques, Tools, and Applications*, 1st ed. Berlin, Germany: Springer Netherlands, 2012.

[101] B. McBride, "Jena: A semantic web toolkit," *IEEE Internet Comput.*, vol. 6, no. 6, pp. 55–58, 2002.

[102] M. M. Bradley and P. P. J. Lang, "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings," *Psychology*, vol. Technical, no. C-1, 1999.

[103] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," in *Proceedings of the ESWC2011 Workshop on "Making Sense of*

Microposts": Big things come in small packages, 2011, vol. 718, pp. 93–98.

[104] Twitter, "The Search API - Twitter Developers," *Twitter*, 2017. [Online]. Available: https://dev.twitter.com/rest/public/search. [Accessed: 28-Apr-2017].

[105] "API Rate Limits | Twitter Developers," *Twitter*, 2017. [Online]. Available: https://dev.twitter.com/rest/public/rate-limiting. [Accessed: 28-Apr-2017].

[106] R. N. Shepard, "Recognition memory for words, sentences, and pictures," *J. Verbal Learning Verbal Behav.*, vol. 6, no. 1, pp. 156–163, 1967.

[107] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak, "Exploiting Emoticons in Sentiment Analysis," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 2013.

[108] "List of emoticons," *Wikipedia*, 2017. [Online]. Available: https://en.wikipedia.org/wiki/List_of_emoticons. [Accessed: 28-Apr-2017].

[109] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, 2001.

[110] Emogi, "2015 Emoji Report," 2015.

[111] A. H. Huang, D. C. Yen, and X. Zhang, "Exploring the potential effects of emoticons," *Inf. Manag.*, vol. 45, no. 7, pp. 466–473, 2008.

[112] D. A. Huffaker and S. L. Calvert, "Gender, Identity, and Language Use in Teenage Blogs," *J. Comput. Commun.*, vol. 10, no. 2, 2005.

[113] A. Wolf, "Emotional Expression Online: Gender Differences in Emoticon Use," *CyberPsychology Behav.*, vol. 3, no. 5, pp. 827–833, 2000.

[114] X. Lu, W. Ai, X. Liu, Q. Li, N. Wang, G. Huang, and Q. Mei, "Learning from the Ubiquitous Language: an Empirical Analysis of Emoji Usage of Smartphone Users," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16*, 2016, pp. 770–780.

[115] "Oxford Dictionaries Word of the Year 2015 is…," *Oxford Dictionaries*, 2015. [Online]. Available: http://blog.oxforddictionaries.com/2015/11/word-of-the-year-2015-emoji/. [Accessed: 28-Apr-2017].

[116] "The most used emoji | TwitterData," *Twitter*, 2016. [Online]. Available: https://twitter.com/TwitterData/status/711950817478234113. [Accessed: 28-Apr-2017].

[117] T. Whitlock, "Emoji unicode characters for use on the web," 2015. [Online]. Available: http://apps.timwhitlock.info/emoji/tables/unicode. [Accessed: 28-Apr-2017].

[118] M. Davis and P. Edberg, "Unicode Emoji," 2016.

[119] J. Skeet, "UTF-8 vs Unicode," 2009. [Online]. Available: http://stackoverflow.com/a/643713/282855. [Accessed: 28-Apr-2017].

[120] H. Ahmedov, "What is the difference between ASCII and unicode characters & difference between UTF-8 and UTF-16?," 2015. [Online]. Available: https://www.quora.com/What-is-the-difference-between-ASCII-and-unicode-characters-difference-between-UTF-8-and-UTF-16/answer/Humoyun-Ahmedov?srid=3YLI. [Accessed: 28-Apr-2017].

[121] E. Kouloumpis, T. Wilson, and J. Moore, “Twitter Sentiment Analysis: The Good the Bad and the OMG!,” *Artif. Intell.*, pp. 538–541, 2011.

[122] Wiktionary, “Appendix:English internet slang - Wiktionary,” *Wiktionary*, 2017. [Online]. Available: https://en.wiktionary.org/wiki/Appendix:English_internet_slang. [Accessed: 28-Apr-2017].

[123] “Open American National Corpus,” *American National Corpus*, 2016. [Online]. Available: https://www.anc.org/data/oanc/download/. [Accessed: 28-Apr-2017].

[124] U. Quasthoff, M. Richter, and C. Biemann, “Corpus portal for search in monolingual corpora,” in *Proceedings of the fifth international conference on language resources and evaluation (LREC 2006)*, 2006, pp. 1799–1802.

[125] S. Batra and C. Tyagi, “Comparative analysis of relational and graph databases,” *IJSCE Int. J. Soft Comput. Eng.*, vol. 2, no. 2, pp. 509–512, 2012.

[126] L. Mastin, “Some Commonly Misspelled Words,” 2008. [Online]. Available: http://www.lukemastin.com/testing/spelling/cgi-bin/database.cgi?action=misspelled. [Accessed: 28-Apr-2017].

[127] “Commonly misspelled English words,” *Wikipedia*, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Commonly_misspelled_English_words. [Accessed: 28-Apr-2017].

[128] R. Nordquist, “The 201 Most Commonly Misspelled Words in English,” *About.com*, 2008. [Online]. Available: http://grammar.about.com/od/words/a/misspelled200.htm. [Accessed: 28-Apr-2017].

[129] “100 Most Often Misspelled Words in English,” *LoveToKnow, Corp*, 2008. [Online]. Available: http://grammar.yourdictionary.com/spelling-and-word-lists/misspelled.html. [Accessed: 28-Apr-2017].

[130] R. Rabbany K., M. Takaffoli, and O. Zaïane R., “Social network analysis and mining to support the assessment of on-line student participation,” *ACM SIGKDD Explor. Newsl.*, vol. 13, no. 2, pp. 20–29, 2012.

[131] H. J. Jain, M. S. Bewoor, and S. H. Patil, “Context Sensitive Text Summarization Using K Means Clustering Algorithm,” *Int. J. Soft Comput. Eng.*, vol. 2, no. 2, pp. 301–304, 2012.

[132] A. T. Kabakuş, “Anlamsal Web Tabanlı Bir Doğal Dil Soru-Cevap Sistemi,” Gazi University Graduate School Of Natural And Applied Sciences, 2014.

[133] S. Matos, D. Campos, and J. L. Oliveira, “Vector-space models and terminologies in gene normalization and document classification,” in *Proceedings of BioCreative III Workshop*, 2010.

[134] D. Zhang, “Inconsistency: The good, the bad, and the ugly,” in *2009 IEEE International Conference on Information Reuse & Integration*, 2009, pp. 182–187.

[135] M. McCord and M. Chuah, “Spam Detection on Twitter Using Traditional Classifiers,” in *Autonomic and Trusted Computing - 8th International Conference (ATC 2011)*, 2011, vol. 6906 LNCS, pp. 175–186.

[136] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, “Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach,” in

Proceedings of the 20th ACM international conference on Information and knowledge management, 2011, pp. 1031–1040.

[137] I. G. Councill, R. McDonald, and L. Velikovich, "What's great and what's not: learning to classify the scope of negation for improved sentiment analysis," in *Proceedings of the ACL Workshop on Negation and Speculation in Natural Language Processing Uppsala Sweden*, 2010, no. July, p. 51.

[138] F. Wogenstein, J. Drescher, D. Reinel, S. Rill, and J. Scheidt, "Evaluation of an algorithm for aspect-based opinion mining using a lexicon-based approach," in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '13*, 2013, pp. 1–8.

[139] A. C. E. S. Lima, L. N. de Castro, and J. M. Corchado, "A polarity analysis framework for Twitter messages," *Appl. Math. Comput.*, vol. 270, pp. 756–767, Nov. 2015.

[140] S. Liu, W. Zhu, N. Xu, F. Li, X. Cheng, Y. Liu, and Y. Wang, "Co-training and visualizing sentiment evolvement for tweet events," in *Proceeding WWW '13 Companion Proceedings of the 22nd International Conference on World Wide Web*, 2013, pp. 105–106.

[141] T. Wilson, T. Wilson, J. Wiebe, J. Wiebe, R. Hwa, and R. Hwa, "Just how mad are you? Finding strong and weak opinion clauses," *Proc. Natl. Conf. Artif. Intell.*, pp. 761–769, 2004.

[142] V. Mani Kiran Peddinti and P. Chintalapoodi, "Domain Adaptation in Sentiment Analysis of Twitter," in *AAAIWS'11-05 Proceedings of the 5th AAAI Conference on Analyzing Microtext*, 2011, pp. 44–49.

[143] A. Aue and M. Gamon, "Customizing Sentiment Classifiers to New Domains: a Case Study," *Proc. Recent Adv. Nat. Lang. Process. RANLP*, vol. 49, no. 2, pp. 207–18, 2005.

[144] TechStacks, "Who uses Redis?," *TechStacks*, 2016. [Online]. Available: http://techstacks.io/tech/redis. [Accessed: 28-Apr-2017].

[145] A. T. Kabakus and R. Kara, "A performance evaluation of in-memory databases," *J. King Saud Univ. - Comput. Inf. Sci.*, 2016.

[146] V. Abramova, J. Bernardino, and P. Furtado, "Which NoSQL Database? A Performance Overview," *Open J. Databases*, vol. 1, no. 2, pp. 17–24, 2014.

[147] B. Stone, "Google Adds Live Updates to Results," *The New York Times*, 2009. [Online]. Available: http://www.nytimes.com/2009/12/08/technology/companies/08google.html. [Accessed: 28-Apr-2017].

[148] H. Drucker, D. Wu, and V. N. Vapnik, "Support Vector Machines for Spam Categorization," *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 1048–1054, 1999.

[149] E. Blanzieri and A. Bryl, "A Survey of Learning-Based Techniques of Email Spam Filtering," *Artif. Intell. Rev.*, vol. 29, no. 1, pp. 63–92, 2008.

[150] J. Song, S. Lee, and J. Kim, "Spam Filtering in Twitter using Sender-Receiver Relationship," in *RAID'11 Proceedings of the 14th international conference on Recent Advances in Intrusion Detection*, 2011, pp. 301–317.

[151] "Reporting Spam on Twitter," *Twitter*, 2017. [Online]. Available: https://support.twitter.com/articles/64986. [Accessed: 28-Apr-2017].

[152] D. Boyd and J. Heer, "Profiles as Conversation: Networked Identity Performance on Friendster," in *HICSS '06 Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, 2006, vol. 3.

[153] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social Phishing," *Commun. ACM*, vol. 50, no. 10, pp. 94–100, 2007.

[154] K. Lee, B. D. Eoff, and J. Caverlee, "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter," in *The Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-11)*, 2011, pp. 185–192.

[155] C. M. Zhang and V. Paxson, "Detecting and Analyzing Automated Activity on Twitter," in *PAM'11 Proceedings of the 12th international conference on Passive and active measurement*, 2011, vol. 6579 LNCS, pp. 102–111.

[156] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@spam: The Underground on 140 Characters or Less," in *Proceedings of the 17th ACM Conference on Computer and Communications Security*, 2010, pp. 27–37.

[157] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "COMPA: Detecting Compromised Accounts on Social Networks," in *20th Annual Network and Distributed System Security Symposium (NDSS 2013)*, 2013.

[158] C. Chen, J. Zhang, X. Chen, Y. Xiang, and W. Zhou, "6 Million Spam Tweets: A Large Ground Truth for Timely Twitter Spam Detection," in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 7065–7070.

[159] D. K. McGrath and M. Gupta, "Behind Phishing: An Examination of Phisher Modi Operandi," in *LEET'08 Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, 2008, p. 4.

[160] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is Tweeting on Twitter: Human, Bot, or Cyborg?," in *26th Annual Computer Security Applications Conference (ACSAC 2010)*, 2010, pp. 21–30.

[161] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru, "Phi.sh/$oCiaL: The phishing landscape through short URLs," in *The 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS 2011)*, 2011.

[162] F. Klien and M. Strohmaier, "Short Links Under Attack: Geographical Analysis of Spam in a URL Shortener Network," in *23th ACM Conference on Hypertext and Social Media (HT 2012)*, 2012, pp. 83–87.

[163] D. Antoniades, E. Athanasopoulos, and T. Karagiannis, "we.b: The web of short URLs," in *WWW '11 Proceedings of the 20th international conference on World wide web*, 2011, pp. 715–724.

[164] Z. Qian, Z. M. Mao, Y. Xie, and F. Yu, "On Network-level Clusters for Spam Detection," in *The 17th Annual Network and Distributed System Security Symposium (NDSS '10)*, 2010.

[165] Y. Xie, F. Yu, and R. Panigrahy, "Spamming Botnet: Signatures and Characteristics," in *ACM SIGCOMM 2008*, 2008.

[166] "The Twitter Rules," *Twitter*, 2017. [Online]. Available: https://support.twitter.com/articles/18311. [Accessed: 28-Apr-2017].

[167] S. Tata and J. M. Patel, “Estimating the selectivity of tf-idf based cosine similarity predicates,” *ACM SIGMOD Rec.*, vol. 36, no. 2, pp. 75–80, 2007.

[168] S. Shum, N. Dehak, R. Dehak, and J. R. Glass, “Unsupervised Speaker Adaptation based on the Cosine Similarity for Text-Independent Speaker Verification,” in *Odyssey: The Speaker and Language Recognition Workshop 2010*, 2010.

[169] M. Bilenko and R. J. Mooney, “Adaptive duplicate detection using learnable string similarity measures,” in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003, pp. 39–48.

[170] J. Echeverría and S. Zhou, “The `Star Wars’ botnet with >350k Twitter bots,” London, UK, 2017.

[171] P. Domingos and M. Pazzani, “On the Optimality of the Simple Bayesian Classifier under Zero-One Loss,” *Mach. Learn.*, vol. 29, pp. 103–130, 1997.

[172] J. L. Hellerstein, T. S. Jayram, and I. Rish, “Recognizing end-user transactions in performance management,” vol. 1, pp. 596–602, 2000.

[173] T. M. Mitchell, *Machine Learning*. Maidenhead, UK: McGraw-Hill, 1997.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı	: Abdullah Talha KABAKUŞ
Doğum Tarihi ve Yeri	: 27.09.1989, Erzurum
Yabancı Dili	: İngilizce
E-posta	: talhakabakus@gmail.com

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Doktora	Elektrik-Elektronik ve Bilgisayar Müh.	Düzce Üniversitesi	2017
Y. Lisans	Bilgisayar Müh.	Gazi Üniversitesi	2014
Lisans	Bilgisayar Müh.	Çankaya Üniversitesi	2010
Lise	Fen	Özel Yavuz Selim Lisesi	2006

TEZ ÇALIŞMASINDAN YAPILAN YAYINLAR

SCI Kapsamındaki Yayınlar

- "TwitterSentiDetector - A Domain-Independent Twitter Sentiment Analyzer", Kabakus A. T., Kara R., Information Systems and Operational Research (INFOR), (Değerlendirmede - Rev.2 aşamasında).
- "TwitterSpamDetector - A spam detection framework for Twitter", Kabakus A. T., Kara R., Turkish Journal of Electrical Engineering and Computer Sciences, (Değerlendirmede).

Diğer Uluslararası İndekslerde Taranan Yayınlar

- “A Survey of Spam Detection Methods on Twitter”, Kabakus A. T., Kara R., (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 3, 2017.
- “A performance evaluation of in-memory databases”, Kabakus A. T., Kara R., Journal of King Saud University - Computer and Information Sciences, In Press, doi: http://doi.org/10.1016/j.jksuci.2016.06.007, 2016.
- “Sentence Level Spell Check Framework”, Kabakus A. T., Kara R., The IIOAB Journal, Vol. 8, No. 3, 36-41, 2017.