T.C.

KONYA TEKNİK ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

GÜNEY ÇİN DENİZİ POLİTİKALARI ÜZERİNE TWİTTER VERİLERİYLE DUYGU ANALİZİ

Khondoker Zahidul Hossain CHOYAN

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ

Aralık-2021
KONYA

TEZ KABUL VE ONAYI

Khondoker Zahidul Hossain tarafından hazırlanan "Güney Çin Denizi Politikaları Üzerine Twitter Verileriyle Duygu Analizi" adlı tez çalışması 14/12/2021 tarihinde aşağıdaki jüri tarafından oy birliği / oy çokluğu ile Konya Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri

Başkan
Dr. Öğr. Üyesi Ahmet Cevahir ÇINAR

Danışman
Dr. Öğr. Üyesi Ersin KAYA

Üye
Dr. Öğr. Üyesi Semiye DEMİRCAN

Yukarıdaki sonucu onaylarım.

Prof. Dr. Saadettin Erhan Kesen
Enstitü Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Khondoker Zahidul Hossain CHOYAN

Tarih: 31.10.2021

ÖZET

YÜKSEK LİSANS TEZİ

GÜNEY ÇİN DENİZİ POLİTİKALARI ÜZERİNE TWİTTER VERİLERİYLE DUYGU ANALİZİ

Khondoker Zahidul Hossain CHOYAN

Konya Teknik Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Ersin Kaya

2021, 63 Sayfa

Jüri
Dr. Öğr. Üyesi Ersin KAYA
Dr. Öğr. Üyesi Ahmet Cevahir ÇINAR
Dr. Öğr. Üyesi Semiye DEMİRCAN

Duygu Analizi, metni analiz etme ve metin verilerini otomatik olarak sınıflandırarak yazarın belirli bir konuya yönelik tutumunu belirleme sürecidir. Bir belgede, bir cümlede veya bir özellik düzeyinde ifade edilen görüşün olumlu, olumsuz veya tarafsız olup olmadığını terpit etme işlemidir. Üretilen ve işlenen sosyal ağ verilerinin ölçeği, Büyük Veri çağında katlanarak artıyor. Twitter şu anda tek başına saniyede 6000 tweet üretiyor. Bu tezde, özel bir Twitter veri seti üzerinde üç farklı açık kaynaklı Python duygu analiz aracı uygulanmıştır. Geçen senenin bir olayını gözlemlemeye çalıştık ve Twitter'daki tepkisini analiz ettik. Analizden, farklı ülkelerden insanların genel duygularının ne olduğunu bulmaya çalıştık. Buradaki ana odak noktamız, Çin'in yakın komşularını doğrudan veya dolaylı olarak etkileyen farklı Çin dış politika önlemleri üzerinden ülke bazlı kamuoyu analizini modellemektir. Analizimiz, sadece Çin'in komşu ülkelerinin değil, ABD ve İngiltere'nin de Güney Çin Denizi bölgesinde meydana gelen olaylarla çok daha fazla ilgilendiğini gösterdi. Ayrıca, tüm ülkeleri ifade eden en olumsuz duygu, ABD'nin olumlu diplomatik etkisine sahip olduğunu gösterdi.

Anahtar Kelimeler: duygu analizi, doğal dil işleme, makine öğrenme, twitter

ABSTRACT

MS THESIS

SENTIMENT ANALYSIS WITH TWITTER DATA ON SOUTH CHINA SEA POLITICS

Khondoker Zahidul Hossain CHOYAN

Konya Technical University
Institute of Graduate Studies
Department of Computer Engineering

Advisor: Asst. Prof. Dr. Ersin KAYA

2021, 63 Pages

Jury
Asst. Prof. Dr. Ersin KAYA
Asst. Prof. Dr. Ahmet Cevahir ÇINAR
Asst. Prof. Dr. Semiye DEMİRCAN

Sentiment Analysis is the process of analyzing text and determine the author's attitude towards a particular topic by automatically classifying textual data. Understanding whether the expressed opinion in a document, a sentence, or a feature level is positive, negative, or neutral. The scale of social network data generated and processed is increasing exponentially in the Big Data era. Twitter currently produces 6000 tweets per second alone. Because of this huge amount of open data availability, sentiment analysis over social media data has been hugely popular to understand recent marketing trends, political shape-shifting, etc. In this thesis, three different open-source Python sentiment analysis tool was applied on a custom Twitter dataset. We tried to observe an event from last year and analyzed its reaction on Twitter. From the analysis, analyzed what was the general sentiment of the people from different countries. Our main focus here is to model country-based analysis of public opinion over different Chinese foreign policy measures which directly or indirectly affect China’s near neighbors. Our analysis showed that not only China's neighboring countries but also the USA and the UK were very much more interested in the events happening around the South China Sea region. Also, the most negative sentiment expressing all countries has positive US diplomatic influence.

Keywords: sentiment analysis, machine learning, natural language processing, twitter

ÖNSÖZ

Bu çalışmanın gerçekleştirilmesinde, değerli bilgilerini benimle paylaşan sayın danışman hocam Dr. Öğr. Üyesi ERSİN KAYA'ya teşekkür ediyorum. Eğitim yolculuğum boyunca beni bu tarihe kadar destekleyen diğer tüm üniversite öğretmenlerime teşekkür ediyorum. Ayrıca arkadaşlarım ve ailem ilgimi kaybettiğim zamanlarda beni motive ederek bana destek oldular. Her zaman bana çok destek veren aileme özel olarak teşekkür ediyorum.

Khondoker Zahidul Hossain CHOYAN
KONYA-2021

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Kısaltmalar

NLP	Natural Language Processing
ML	Machine Learning
NN	Neural Network
SVM	Support Vector Machine
tf-idf	Term Frequency-Inverse Document Frequency
NLTK	Natural Language Toolkit
API	Application Programming Interface
REST	Representational State Transfer
VADER	Valence Aware Dictionary for Sentiment Reasoning
POS	Part of Speech
DNN	Deep Neural Network
FBIC	Formal Bilateral Influence Capacity

1. GİRİŞ

Son 30 yıldır ABD'nin egemen olduğu Soğuk Savaş sonrası uluslararası sistem, ABD ekonomisinin göreceli durgunluğuyla sarsıldı. Bu arada Çin hızla küresel bir siyasi güç olarak ortaya çıktı ve Amerika'nın Asya-Pasifik bölgesindeki hakimiyetine karşı mücadele etmeye başladı. Çin 2010 mali krizini atlatmayı başardı ve ekonomik bir güç merkezi olarak ortaya çıktı. Ancak Pekin'in küresel gelişimi de dünya çapında gerginlikler yarattı. Olağanüstü ekonomik büyümesi ve aktif diplomasisi zaten Doğu Asya'yı şekillendiriyor ve böylece gelecek on yıllar Çin'in gücü ve nüfuzunda daha da büyük artışlar görülecektir (Ikenberry, 2008).

Çıkarlarını korumak için giderek daha fazla ekonomik ve politik ilişkiye zorlanan uzun tedarik zincirleri, Çin'i modern bir imparatorluk haline getirdi. Çin Güney Çin Denizi'nden Hindistan'a ve Orta Asya'dan Avrupa'ya doğrudan faaliyet alanını genişletmek için ekonomik, politik ve askeri kaldıraçlarını kullanmaktadır (Baker, 2020).

Çin için kritik zorluk, Çin'in yükselişinin uluslararası korkularının Batı dünyasında artmasıdır. ABD ile olan güç dengesi, Washington artık Çin'in yükselişini göz ardı edebileceği noktada değildir (Baker, 2020). Çin'e olan bağımlılığını çoktan kırdı, çünkü teknoloji şirketleri Çin'den ayrılmaya çağırıyor. ABD Donanması da Çin çevresi boyunca sağlam bir operasyonel tempo sürdürüyor. Avrupa Birliği bile şu anda Çin'in ekonomik ve düzenleyici erişiminden endişe duyuyor ve kritik sektörlerdeki Çin yatırımlarını azaltmak için düzenleyici önlemler alıyor (T. Liu ve Woo, 2018).

Pekin, çevresi etrafındaki konumunu güçlendirmek için daha hızlı hareket ediyor ve komşu ülkeleri etrafındaki etkisini artırmaya çalışıyor. Ancak bu Güney Asya ülkeleri için, oldukça büyük ve son derece zengin bir Çin etnik grubunun varlığı ve Çin'in ekonomileri için büyümesine olan artan bağımlılığı, onları Çin ile ilişkileri konusunda son derece temkinli olmaya zorladı. Şu ana kadar Asya Pasifik bölgesinde hiçbir hükümet doğrudan Çin karşıtı politika benimsemedi (Han ve Paul, 2020). Ancak Malezya, Endonezya ve Filipinler'de bazı yerlerde Çin karşıtı ayaklanmalar meydana geldi.

Çin hükümeti, komşu ülkelerdeki bu gerginliği hafifletmek için Asya Pasifik bölgesinde "iyi bir komşu politikası"nı teşvik ediyor. Çin, bu ülkelerle olan ticareti artırarak ve ticaret fazlasının Çin ile olmasına izin vererek, kendisini vazgeçilmez bir

ticaret ortağı kurmuştur. Bunun yanı sıra Çin bu ülkelerle çeşitli bölgesel iş birliği mekanizma girmiştir.

Bu tezde, Çin'in çeşitli dış politika önlemlerinin komşu ülkeler arasındaki kamuoyu üzerinde nasıl etkileri olduğunu analiz edildi. Yapısal konu Modellemesini kullanıldı ve Twitter API'sinden topladığımız tweetlerde duygu analizi yapıldı. Buradaki odak noktamız Çin'in yakın komşularını doğrudan veya dolaylı olarak etkileyen çeşitli Çin dış politika önlemleri hakkında ülke kamuoyu analizini modellemektir. Ayrıca, bu önlemlerin sadece yakın komşuları üzerinde bir etkisi olup olmadığını veya ilgili tüm tarafların bu konularla ilgili görüşlerini ifade edip etmediklerini ve nasıl tepki verdiklerini araştırıldı.

2. İLGİLİ ÇALIŞMALAR

Günümüzde devam eden teknolojik gelişmeler, çeşitli biçimlerde büyük miktarda veri üretmemize ve depolamamıza izin vermektedir. Bazıları sosyal medyada yer alırken, diğerleri sağlık sektörü veya siyasi kurumlar gibi şirketler veya devlet kurumları tarafından veri tabanlarında saklanmaktadır. Bu veri tabanları genellikle daha biçilmez bilgiler içerir. Ancak bu veri tabanlarının büyüklüğü, onlardan bilgi ve bilgileri el ile çıkarmayı imkansız hale getirir. Bu nedenle yapılandırılmış bilgi ve ilişkileri çıkarmak için bu verilerin otomatik olarak işlenmesine büyük ihtiyaç vardır. Metinsel veriler söz konusu olduğunda, bu yapılandırılmış bilgiler örneğin duygu, adlandırılmış varlıklar, tarihler vb. olabilir.

Doğal dil işleme ve metin madenciliği metinden bilgilerin otomatik çıkarması için mümkün araçlar sağlayan alanlardır. Metin otomatik sınıflandırma ve işleme hakkında yayınlanmış literatürü çok geniştir. Ama yöntemlerin çoğu bir ortak noktaya sahiptir: eğitim için etiketli verilere büyük ölçüde güvenilmesidir. Daha önce de belirtildiği gibi, araştırma ve mühendislik amaçlı büyük veri tabanlarına erişmek çok kolay ve ucuzdur. Bununla birlikte, bu verilerin bir alt kümesini etiketlemek, genellikle manuel olarak yapılması gereken zaman alıcı ve sıkıcı bir işlemdir. X verilerinin ek bir miktarını manuel olarak etiketlemek için harcanan zaman ve çaba ile Y sınıflandırma doğruluğunda elde edilen kazanç arasında bir uzaklaşma vardır. Etiketlenmemiş verilerin büyük miktarlarının çoğu alanda kolayca erişilebilir olduğu ve genellikle etiketleme maliyetlerine kıyasla doğruluk kazancının düşük olduğu göz önüne alındığında, öğrenme algoritmalarını geliştirmek için etiketlenmemiş verilerin nasıl kullanılacağı sorusu ortaya çıkmaktadır.

Bu sorunu çözmek için bugüne kadar birçok denemeler yapildi. Örneğin Amazon Mechanical Turk gibi platformlar, büyük miktarda veriyi etiketlemek uzere çalışanları işe almak için kullanılır. Bunu yapmak için denetimsiz ve yarı denetimli birkaç algoritma geliştirdiler (Ghahramani, 2004; Ko ve Seo, 2000).

Daha önceki duygu çalışmaları öncelikle anlatılarla (Wiebe, 1994, 1990) veya gazete metinleriyle ilgiliyken (Bautin ve ark., 2008; Riloff ve Wiebe, 2003; Wilson ve ark., 2005), kısa bir süre sonra, sosyal medyanın insanların davranışlarını araştırmak için çok daha iyi bir zemin sağladığı anlaşıldı. Das ve Chen, çevrimiçi forumlardan kullanıcı görüşlerini almaya çalışan ilk kişiler arasındaydı (Das ve Chen, 2001). Bu amaçla, üç

polarite (almak, satmak ve boş) sınıfına sahip bir ekonomik sohbet panosundan 500 mesajdan oluşan bir koleksiyonu manuel olarak etiketlendi. Daha sonra bu veriler üzerinde beş farklı sınıflandırıcı eğitildi. Eğitimli sistemleri kullanarak, yazarlar kalan 25.000 forum gönderisinin anlamsal yönünü sınıflandırdılar ve bu parçacıkların kutuplarına göre hisse senedi fiyatlarını tahmin etmeye çalıştılar.

İlk büyük corpora'nın tanıtılmasıyla fikir madenciliği alanında gerçek bir atılım gerçekleşti. Bu konuda önemli katkılarda bulunan Pang ve Lee (Pang ve Lee, 2004, 2005), yıldız derecelendirmeleriyle birlikte yaklaşık 2.000 film incelemesinden oluşan bir veri seti yayınladılar; Hu ve Liu, manuel olarak etiketlenmiş bir dizi Amazon ve C|Net ürün yorumu sunmuştu (Hu ve Liu, 2004); Thomas ve ark. otomatik olarak bir kongre tartışmaları koleksiyonunu etiketletlemişti (Thomas ve ark., 2006); ve son olarak, Wiebe ve ark. (2005), 535 haber makalesinden oluşan manuel olarak açıklamalı bir duygu korpus geliştirmiştir (Wilson ve ark., 2005).

Park ve Kim, sözlük tabanlı bir yaklaşım kullanarak eş anlamlılar sözlüğü oluşturmak için bir yöntem önermişlerdir (Parks ve Kim, 2016). Onlar bir eş anlamlılar sözlüğü oluşturmak için üç çevrimiçi sözlük kullandı ve yalnızca sözlüğün güvenilirliğini artıran sözlükte birlikte bulunan kelimeleri saklamışlardır. Sözlüğün genişletilmesi, tohum kelimelerinin eşanlamlıları ve zıtlıkları ile yapılır. Genişletilmiş eş anlamlıları, sınıflandırma görevinin doğruluğunu artırır. Tf-idf yöntemlerini kullanarak tohum kelimesini seçilmişti. Ayrıca bu yaklaşımın zaman alıcı bir yaklaşım olduğunu belirtilmişti.

Agarwal ve ark., tweetleri olumlu, olumsuz ve tarafsız duygular olarak sınıflandırmak için bir sistem sunmuştur (Apoorv Agarwal ve ark., 2011). Görevin iki versiyonunu denemişlerdir. İlk görev, olumlu ve olumsuz tweetleri tanımlamanın ikili bir sınıflandırmasıdır. İkincisi ise üç kategoriyi birbirinden ayırmaya yönelik 3 yollu bir görevidir. Ayrıca, üç farklı modeli karşılaştırdılar: bir unigram modeli, 100 özel yapım özniteliği kullanan bir model ve bir ağaç çekirdeği tabanlı model. Yazarlar, özellik tabanlı modelin temel modelden daha iyi performans gösteremediğini bulmuştur. Ancak çekirdek tabanlı model her ikisinden de daha iyi performans göstermiştir. Bir özellik analizinin ardından, sınıflandırma görevi için en değerli olan kelimelerin konuşma bölümü etiketleriyle olan polarite olasılıklarının kombinasyonunu bulmuşturlardır. Bu klasik doğal dil özellikleri, ifadeler veya hashtagler gibi Twitter'a özgü özelliklerden daha iyi

performans göstermişlerdir. Modellerini manuel olarak açıklamalı 5127 tweet üzerinde eğitmişlerdir. Ön işleme boyunca, kendi tasarladıkları bir ifade sözlüğü ve bir kısaltma sözlüğü kullanmışlardır. Tweetler arasındaki benzerliği hesaplamak için olası tüm alt ağaçlar karşılaştırmıştır. İki yönlü sınıflandırma görevi için, çekirdek tabanlı yöntemi kullanarak ortalama %71,35 ve üç yönlü sınıflandırma görevi için %60,60'lık bir ortalama test doğruluğu rapor etmişlerdir. Diğer iki ana model daha düşük doğruluk puanları vermiştir. Modellerin kombinasyonlarını kullanarak yazarlar, iki yönlü görevde %74.61'lik bir test doğruluğu ve üç yönlü görevde %60.83'lük bir doğruluk elde edebilmiştir.

Wang ve ark.'nın çalışmasında duygu etiketli Twitter verileri 7 duygu sınıfı için toplanmıştır (sevinç, üzüntü, öfke, sevgi, korku, şükran ve sürpriz). Duygu kelimelerinin ilk listesi, duygu hashtaglerinin sözcüksel varyantları ile genişletmişlerdir. Belirsizliği azaltmak için diğer kavramlarla veya alanlarla ilişkili duygu sözcükleri listeden çıkarılmıştır. Elde edilen veri kümesinin tweetlerinin %93.16'sının ilgili duygu ile ilgili olduğu gösterilmiştir. Veri kümesi, n-gram özelliklerinin farklı kombinasyonları, konuşma parçası etiketleme ve WordNet efekti gibi önceden tanımlanmış birkaç sözcüksel kaynakla eğitilmiştir. Bu çalışmanın amacı, kısa sürede otomatik olarak açıklamalı bir veri kümesi oluşturmak ve farklı boyutlarda veri kümeleri için sınıflandırma performansını ölçmektir. Eğitim verilerinin boyutunu 1000'den 2 milyona çıkararak %22.16 doğruluk kazancı elde edildiği gösterilmiştir (Wang ve ark., 2012).

Başka bir çalışmada, uzaktan denetimli öğrenme için tweetlerde emoji kullanımı analiz edilmiştir (Go ve ark., 2009). Daha spesifik olarak, ifadeler, olumlu ve olumsuz duyguları temsil eden eğitim verilerini çıkarmak için gürültülü etiketler olarak kullanılmıştır. Yazarlar, farklı sınıflandırıcılar, yani Naive Bayes, Maksimum Entropi ve SVM'lerin yanı sıra unigramlar, bigramlar, ikisinin kombinasyonu ve konuşma parçası etiketlerinden oluşan farklı özellik kümeleriyle deney yapmışlardır. En iyi sonuç olarak, hem unigramlar hem de bigramlar üzerinde eğitilmiş %83'lük bir doğruluk rapor etmişlerdir. Eğitim verileri, eşit sayıda pozitif ve negatif tweet içeren 1.6 M tweetlerden oluşmuştur.

Kouloumpis ve ark., eğitim veri kümesi için etiketlerin tanımlanmasıyla ilgili zorlukları tartışırlar. Tweetlerdeki hashtagleri kullanarak eğitim etiketlerini oluşturmak için bir yöntem öneriyorlar. Edinburgh Twitter bütüncesinde 1000'den fazla kez görünen

hashtaglerla tweetleri çıkarak bunları duygularına göre üç grup (pozitif, negatif, nötr hashtagler) olarak manuel olarak gruplandırdılar. Bu hashtag açıklamalı veri setinin yanı sıra, Go ve ark.'in ifade ettiği veri kümesinden 19000 tweet ekleyerek ilk veri kümelerini geliştirmek için kullandılar (Kouloumpis ve ark., 2011).

Ayrıca, Barbosa ve Feng, tweetlerde duygu analizi yapmak için unigram gibi klasik öznetiklerle ek olarak "meta-bilgi" kullanımını göstermiştir. Bu meta-bilgi, konuşma parçası etiketlerinin yanı sıra daha önce sağlanan sözlüğün öznelliğinden çıkarılan öznellik ve kutupluluğu içeriyormuştur (Riloff ve Wiebe, 2003). Bu nedenle, iki aşamalı bir sınıflandırma yöntemi önermişlerdir. İlk olarak, öznel ve nesnel tweetleri ayırt etmek için bir sınıflandırıcı eğitmişlerdir. Daha sonra olumlu ve olumsuz duyguları ayırt etmek için başka bir sınıflandırıcıyı eğitmişlerdir (Barbosa ve Feng, 2010).

Boiy ve Moens İngilizce, Fransızca ve Hollanda dilinde yazılmış blog, makaleler ve forum metinlerinde duygu analizi ile ilgili makine öğrenme deneylerini sunmuşlardır. Cui ve ark. twitter duygu analiz problemini duygu sembolleri, düzensiz kelimeler biçimleri ve kombine noktalamalar analiziyle çözmeye çalışmışlardır. Narr ve ark., tamamen yeni bir tweet setinden eğitim verileri oluşturmak için gürültülü etiketler olarak emojileri kullanmışlardır. Veriler üzerinde bir Bayes sınıflandırıcısı eğitmişler ve ardından 4 açıklamalı dilde 10000'den fazla tweeti değerlendirmişler. (Boiy ve Moens, 2009; Cui ve ark., 2011; Narr ve ark., 2012).

Kısa geçmişine rağmen, Twitter'ın duygu analizi, NLP araştırmacılarından büyük ilgi gördü. Ancak birkaç istisna dışında, mevcut çalışmaların çoğu ya İngilizce mesajlarla ya da otomatik olarak çevrilen mikrobloglarla ilgiliydi (Araque ve ark., 2015; Basile ve Nissim, 2013; Bosco ve ark., 2013; Cesteros ve ark., 2015).

Güçlü siyasi partiler her zaman Twitter'i kendi ihtiyaçları için kullanmaya çalıştılar. İktidar partileri muhalefet görüşlerinin etkisini bastırmaya çalışırken, muhalefet partileri destekçileriyle ilişki kurmaya ve güvenilirliklerini artırmaya çalıştılar. Tunus, Libya ve Mısır hükümetlerini deviren son “Arap Baharı" etkinliği tamamen Twitter'da üretildi (Beaumont, 2011). Bu yüzden bu devrimler “Twitter Devrimi”olarak da bilinir. İnsanlar protestoculara katılmak ve Twitter üzerinden muazzam destek vermek için sokağa çıkmıştı.

El-Hatib, ABD-Çin Ticaret Savaşı ve büyük şehirle ilgili olaylar üzerine bir çalışması yapmıştı (AlKhatib ve ark., 2020). Ancak bu ABD'nin büyük şehirlerindeki olaylarla ilgiliydi. Yusufiarto, Jakarta İslam endeksindeki makroekonomik değişkeni de etkileyen devam eden ticaret savaşı fenomeninin sonucunu bulmak için harika bir çalışma yapmıştı (Yusfiarto ve Pambekti, 2020). ABD'nin Huawei yasağı, Khun ve Thant'ın çalışmalarında güzelce resmedilmişti (Khun ve Thant, 2019). Dünyanın farklı bölgelerindeki polarite varyasyonlarını görselleştirmek için küresel haritalar kullanmışlardır. Agar'ın Brexit tweetlerinde jeo-uzamsal dağılım analizi üzerine çalışması da duygu analizi üzerinde mükemmel bir çalışmaydı (Amit Agarwal ve ark., 2018). Önce tweetlerin tam olarak nerede oluşturulduğunu belirlemek için tweetlerden coğrafi kod kullanabildiler. Ancak 2017'den bu yana Twitter, kullanıcıların coğrafi verilerini API'den vermiyorlar. Bu nedenle, şu anda bir tweetin coğrafi konumunu tam olarak tanımlamak mümkün değildir.

3. MATERYAL VE METODLAR

2013'ten bu yana, doğal dil işlemenin çeşitli uygulamaları yaşamın çeşitli yönlerine hakim olmaya başladı. Tüm dev şirketler akıllı asistan sistemlerini kurmaya başladı. Google, Google asistan ile, Apple Siri ile, Bixby ile Samsung, Alexa ile Amazon bu akıllı ev sistemler pazarını yakalamaya çalıştılar. Bundan önce bile, Google daha akıllı arama motoru önerileri getirdi. Şimdi elimizde daha sofistike teknolojiler var. Klavyeler akıllı metinler ve otomatik düzeltme ile her gün daha akıllı hale geliyor. Grammarly-adlı, yapay zeka tabanlı bir yazma asistanı, çeşitli dilbilgisi düzeltmeleri, ton algılama ve özlülük algılama sağlar. Bütün bunlar doğal dil işlemenin son gelişmeleriyle mümkün olmuştur.

Modern web teknolojisinin ilerlemesiyle, sosyal medyanın ve blogların yükselişi de hızla arttı. Şu an dünya çapında milyarlarca insan birbirine bağlıdır. İnsanlar günlük aktivitelerini, duygularını, çeşitli konulardaki konumlarını paylaşırlar ve böylece hayal gücümüzün ötesinde çok sayıda veri üretirler. İncelemelerin, blogların, makalelerin, derecelendirmelerin ve diğer çevrimiçi ifade biçimlerinin çoğalmasıyla, çevrimiçi görüş bir tür sanal para birimine dönüştü. Dev şirketler şimdi bu para birimini toplamak ve bu sürekli değişen sanal dünyadaki konumlarına hakim olmak için sessiz bir savaştadır. Bir kullanıcı hakkında mümkün olduğunca fazla veri toplamak için kendi ekosistemleri oluşturmaya çalışıyorlar. Gürültü filtreleme, konuşmayı anlama, ilgili içeriği tanımlama ve doğru şekilde işleme sürecini otomatikleştirmek için birçok şirket şu anda duygu analizi alanına doğru ilerliyor.

Duygu analizi, doğal dil işlemenin başka bir uygulamasıdır. Doğal dil işleme, metin analizi, hesaplamalı dilbilimin ve biyometrilerin sistematik olarak tanımlanması, özütlenmesi, ölçülmesi ve çalışılmasının yanısıra, duygusal durumları ve öznel bilgileri incelemesini ifade eder. Duygu analizinin en popüler görevlerinden biri, belirli bir belgenin polaritesinin, cümlenin veya özellik seviyesinin belirlenmesidir. Belgedeki ifade edilen görüşün, bir cümlenin veya özellik seviyesinin olumlu, negatif veya nötr olduğunu anlamaktır.

Sosyal medya platformlarına uygulanan duygu analizi, iş, spor ve siyaset gibi çok çeşitli alanlardaki önemi nedeniyle son zamanlarda akademik çevrelerde artan bir ilgi görmektedir. Birçok çalışma, hisse senedi fiyatları, film gişe gelirleri ve siyasi seçimler gibi sosyal olguların sosyal medya verilerine yansıdığını iddia ediyor. Bu platformlarda

ifade edilen görüşler dolaylı olarak kamuoyunu değerlendirmek için kullanılabilir (Asur ve Huberman, 2010; Bollen ve ark., 2011; Gayo-Avello, 2013).

Twitter, büyük miktarda kamuya açık veri sağlayan en yaygın kullanılan mikroblog sitesidir. Bir tweet metninin ne kadar uzun olabileceği konusunda bir karakter sınırı vardır. Bu nedenle, uğraşılacak daha az ayrıntılı metin olduğu için tweetler belirli bir görüşü ifade etmede çok kesindir.

Bu tezde Çin'in son yıllarda aldığı çeşitli dış ve iç tedbirlerin komşu ülkelerin algısını nasıl etkilediği araştırılacak. Son yıllarda Çin'in gelişmekte olan ülkelerin çeşitli projelerine büyük yatırımlar yaparak sınırının dışında önlemler aldığını görüldü. Kendi iş alanlarında çok sayıda ülkeyi birbirine bağlamak için Kuşak ve Yol girişimini başlattılar. Bu girişime bağlı ülkeler, mal ihracatı ve ithalatı için birbirleriyle çok daha kolay bağlantı kurabilirler. Bunun yanı sıra Çin, ABD hükümetinin Güney Çin Denizi'ndeki olası ablukalarının üstesinden gelmek için birden fazla farklı yol açmaya çalışıyor.

Tüm bu görünüşte farklı ama bağlantılı olaylar, Çin'in komşu ülkelerini büyük ölçüde etkiledi. Bu ülkeler Çin üzerine baskı yapabilmek için ABD için de önemlidir. Bu tezde, Çin'in komşu ülkelerinin bu duygusu ve bu ülkelerin nüfusunun bu bölgede meydana gelen çeşitli olaylara nasıl tepki verdiği analiz edilecektir. Tepkiyi anlamak için Twitter'dan gelen tweetler bir hashtag kullanılarak toplandı. Bir tweet birkaç hashtag ve web adresi içerdiğinden, önceden işlenmesi gerekiyordu. Temizledikten sonra temiz metin duygu analizi yapmaya hazırdı. Duygu analizi yapmak için VADER ve Flair kütüphaneleri kullanıldı ve sonuçlar arasındaki farklarını karşılaştırıldı. Daha sonra farklı ülkelerdeki Twitter kullanıcılarının genel algıları ayıklanmış ve bu sonuç Resmi İkili Etki Kapasitesi (FBIC) endeksinin analizi ile karşılaştırılmıştır.

3.1. Twitter

Twitter en popüler sosyal medya platformlarından biridir. Bu platform dünyanın her yerinden kullanıcıları içerir ve şu an dünya çapında 187 milyon günlük aktif kullanıcıya sahiptir (Tankovska, 2021). Diğer istatistiklere göre, bu platformda her gün 500 milyondan fazla tweet yayınlanıyor (David, 2020). 18 yaşından büyük herhangi bir kişi Twitter'da bir hesap açabilir ve görüşlerini yazabilir. Bu görüşlere tweet denir. En başından beri, Twitter kullanıcılar tweetlerinde sadece 140 karakter yazabilecekleriyle sınırlıydı. Facebook ve Google Plus gibi diğer platformlardan farklı olarak kullanıcılara

görüşlerini mümkün olan en kısa sürede ifade etmelerini sağlayan bazı kısıtlamalar getirdi. Son zamanlarda bu kurallardan bazılarını biraz rahatlattılar, ancak yine de geçmiş tweet verileri çoğunlukla 140 karakter biçiminde.

Twitter'da herhangi bir kullanıcı diğer kullanıcıları ve görüşlerini düzenli olarak takip edebilir ve onlar tarafından motive edilebilir. Retweet mekanizması, bir kullanıcıya tek bir düğmeye basarak herhangi bir tweeti kopyalayıp yapıştırmak için kolay bir yöntem sunar. Bu nedenle, benzer fikirleri kullanıcıların kendi takipçilerine herhangi bir gecikme olmadan paylaşmanın en hızlı yoludur.

Twitter'in bir diğer önemli yönü, burada her tweetin herkese açık olması ve yayınlandıktan sonra hiçbir tweetin değiştirilememesidir. Bu, Twitter'a tarihsel bir bakış açısından bazı önemli özellikler kazandırdı. Kısa, özlü yayın sistemi, araştırmacıların belirli bir sosyal veya politik olayın zaman çizelgesini senkronize etmelerine yardımcı olan mikro günlüklere bir bakış sağlamıştır. Bu yüzden Twitter adını Bir mikro blog sitesi olarak kazanmıştır. Her vatandaş, tüm dünyada belirli bir olayı gösterme ve paylaşma yeteneğe sahiptir. Twitter ayrıca görüntüleri ve videoları tweetlerle paylaşmanın bir yolunu sunar. Bu da tweetleri çok daha etkili hale getirir.

Herhangi bir tweetin en önemli kısmı hashtagleridir. Her hashtag "#" sembolü ile başlayan bir kelimedir. Bu kelimeler tıklanabilir ve üzerlerine tıklamak, belirli bir terimin veya kelimenin kullanıldığı tweetlerin geçmiş verilerini sağlar. Bu, herhangi bir sosyal medya etkinliğine derinlemesine dalmak isteyen herkes için büyük bir keşif alanı sağlar. Twitter ayrıca her kullanıcıya o bölgedeki o anda trend olan hashtagleri görüntüler. Herhangi bir kullanıcının yakınlarda meydana gelen herhangi bir olayı keşfetmesine yardımcı olur. Temel kullanımda, bu kategorizasyon konulara dayanmaktadır. Ancak bu hashtagler, alaycılık veya ironi gibi tweetin ters yönünü eklemek için de kullanılabilir. Bu tür etiketler mesajın polaritesini tersine çevirebilir. Şu anda en çok kullanılan hashtagler trend konularında özetlenmiştir.

Twitter her yerde yaygın olarak kabul edildiğinden ve dünyanın dört bir yanından insanların dünya çapında neler olup bittiğine dair bir fikir edinmek için hizmetini kullandıkları için, çeşitli şirketler için bir altın madeni haline gelmiştir. Tweetleri analiz etmek, şirketlere müşteri çıkarlarını ve ihtiyaçlarını anlamak için dolaylı bir fırsat sunar.

Doğru adımlar zamanında atılırsa, onlara rakip şirketlerine göre büyük bir avantaj sağlayabilir.

Twitter'ın doğrudan ticari kullanımının yanı sıra dolaylı bir kullanımı da vardır. Örneğin şirketler, markalar veya ürünler hakkında kişisel mesajları aramak ve toplamak için Twitter kullanılır. Twitter internet üzerinden herkes için erişilebilir durumdadır. Bu nedenle ünlü haber sitelerine göre bu en hızlı kitle iletişim kanalıdır. Gerçek zamanlı güncellemeler, şirketler tarafından farklı faydalı amaçlar için kullanılabilir.

Twitter'daki tüm bilgilerden, bir şirket için sadece küçük bir kısım önemlidir. Tweetlerde bulunan bilgiler, kullanıcının kişisel ilgisine bağlıdır. Bu nedenle tweetler, gerçekler gibi nesnel içeriğe ve fikirler gibi öznel içeriklere sahip olabilir. Bu tür bilgiler toplanabilir ve şirket için bir varlık oluşturabilir.

Bu olaylar, Twitter'ın yerel ve uluslararası olayları anlamanın en önemli sosyal faktörlerinden biri haline geldiğini ve insan psikolojisi üzerindeki etkisinin zayıflatılamayacağını açıkça göstermiştir. Sosyal rızanın geliştirilmesinde ve kitleden açık kamuoyu görüşlerinin oluşturulmasında kilit bir faktör olarak kendini kanıtlamıştır.

3.2. Duygu Analizi

Duygular her zaman insanların ifadesinin önemli bir parçasıdır. Insanların bir konuyu tartışırken konuya yönelik duygularını gizlemesi zordur. İnsanlar yüz yüze konuştuklarında, konuya yönelik duyguları çeşitli bedensel hareketlerle (Melzer ve ark., 2019), göz teması (Jongerius ve ark., 2020) vb. yoluyla kendilerini gösterirler. Benzer şekilde, insanlar görüşlerini yazarak ifade ettiğinde, sözcük organizasyonları bu görüşe karşı tutumlarına göre değişir.

Duygu analizi, insanların görüşlerinin, duygularının, ve tutumlarının incelendiği ve analiz edildiği bir bilgi alanıdır (B. Liu, 2012). Ancak bu tanım, günümüzde duygu analizinin kapsadığı geniş alanı kavramak için yeterli değildir. Doğal dil işleme, bilgisayar dilbilimi ve metin analizi gibi birkaç farklı araştırma alanını ifade eder. Bunun yanı sıra, günümüzde metinlerin doğasını tanımlamak ve orada ifade edilen genel tutumu tanımlamak için çeşitli makine öğrenme yöntemleri de kullanılmaktadır.

Duygu analizlerde, üzüntü, sevinç, mutluluk gibi farklı duygular arasındaki farklılıklar tespit edilir. Diğer tarafta görüş tabanlı analizde ise belirli metinlerin olumlu ya da olumsuz yanlarını tespit etmeye çalışılır. Bu metinler, herhangi bir film veya kitap inceleme web sitesinden veya kullanıcıların belirli konuları tartıştığı bazı çevrimiçi forumlardan olabilir, ayrıca haber veya blog makaleleri, Facebook mesajları veya Twitter tweetleri olabilir. Ayrıca, bir belgenin, bir araştırma kağıdının veya bir kitabın genel duygularını belirlemeye çalıştığımız diğer bazı uygulamalar var. Görüş analizi ile ilgili tüm problemler, iki farklı etiketli metin sağlandığı ve metnin türünü tanımlamaya çalıştığımız standart sınıflandırma sorunu olarak da formüle edilebilir.

Bir ürünün incelemesini doğrudan bir görüş olarak kabul edebiliriz. Çoğu durumda, bu görüşler tek taraflı olarak olumlu/olumsuz ve tarafsız olarak ayırt edilebilir. Ancak insanların bu görüşleri sınıflandırmasının zor olduğu bazı durumlar da bile vardır. Dolaylı görüşlerde, ilgili nesneler için görüş farklı şekilde ayrıştırılabilir. Benzerlik veya iki nesne arasındaki fark, dolaylı bir görüş ile tanımlanabilir. Örneğin, bir durumda A ürünü, B ürünü kadar iyi olabilir. Başka bir durumda A, diğer B ürününden daha iyi olabilir. İlk durumda, görüş her iki ürün için de olumlu. İkinci durumda, A için olumlu ve B için olumsuzdur. Bu cümlelerdeki görüşler de tersine çevrilebilir. Bu olay mesajdan çıkarılması gereken dört farklı dolaylı karşılaştırmalı görüşe yol açar. Ayrıca, dolaylı görüşler, esasen aynı görüşü ifade eden birçok farklı kelime ile resmileştirilebilir. Örneğin, "A, B'den daha iyi performans gösteriyor" cümlesi, "B, A'dan daha fazla çarpıyor" cümlesi ile aynı görüşü ifade eder. Her iki cümle de A için olumlu, B için olumsuz duygulara yol açar, ancak tamamen farklı bir şekilde ifade edilir. Bu nedenle dolaylı bir karşılaştırmalı görüşün çıkarılması zordur. Doğrudan görüşe dönersek, bu Liu tarafından beşli olarak resmileştirilen beş özellik ile karakterize edilir (B. Liu, 2010). Beşli denklemde aşağıda belirtilmiştir

$$opinion = (o_j, f_{jk}, oo_{ijkl}, h_i, t_l) \qquad 3.2.1$$

Burada o_j belirli bir nesnedir; f_{jk}, O_J nesnesinin k özelliğidir; h_i bir görüş sahibidir; t_l zaman ve o_{oijkl}, O_J nesnesinin f_{jk} özelliği hakkında h_i sahibinin gerçek görüşüdür.

Bazı cümlelerin gerçek polaritesini veya o_{oijkl}'sini belirlemek, beşli beş özelliğin en zor görevidir. Bir mesajın polaritesi, elektrik polaritesine benzer şekilde pozitif veya

negatif olup olmadığıdır. Bu duygu özneldir, çünkü farklı insanlar güçlü veya zayıf bir görüş olarak gördükleri için farklı zihinsel ölçeklere sahiptir. Bu nedenle, bir cümlenin bir başkası için olumlu ve bir başkası tarafından tarafsız olarak etiketlendiği ortaya çıkabilir. En yaygın modeller görüşleri iki veya üç seviyeye ayırır: olumlu, olumsuz ve ek olarak nötr. Bununla birlikte, Wilson'un yaptığı gibi daha ayrıntılı bir ölçek kullanmak ta mümkündür (Wilson ve ark., 2004).

Duygu analizi, pazarlama, politika, psikoloji vb. gibi farklı bağlamlarda da kullanılmaktadır. Trendler, popüler ürünleri, halkın tepkisini, hedef reklamları analiz etmek için kullanılabilir. Elimizdeki açık veri miktarı ve çeşitli yerlerden veri toplamanın ne kadar kolay olduğu nedeniyle bu mümkün olmuştur. Bazı açık kaynak verileri, Google Veri Kümesi Arama, UCI Makine Öğrenimi deposu ve benzerlerinden de toplanabilir. Bu verilerle neler yapabileceğimiz ve nasıl yapacağımız konusunda yeni nesil imkanların kapısını açtı.

Her gün ürettiğimiz büyük miktarda veri nedeniyle duygu analizi olasılığı çok büyük olsa da, kendi dezavantajları da var. İnsanların fikirlerini ifade etme şekli evrensel değildir. Tüm dünyada 7000'den fazla konuşulan dil vardır. Tüm dünyada 7000'den fazla konuşulan dil var (Ethnologue, 2021). Her dilin kendi özellikleri ve yapısı vardır. Dolayısıyla, her dili ayrı ayrı analiz etmek için kullanabileceğimiz tek bir ortak algoritma yoktur. İngilizce gibi tek bir dili düşünsek bile, herkes benzer şekilde konuşmaz veya yazmaz. Görüşler, bir dilin çeşitli karmaşıklıklarını kullanarak farklı biçimlerde ifade edilebilir. Karmaşık cümleler, bunları ayrı ayrı ifade etmek için anlaşılması gereken farklı cümlelerden oluşur, bu da onları sınıflandırmayı zorlaştırır.

Daha önce, metni analiz etmek ve iç mesajı anlamak için daha uzun bir metin kaynağı uygulanmıştır. Örneğin, duygu analizi eğitimi için en popüler veri kümelerinden biri IMDB film inceleme veri kümesidir. Bu veri kümesinde eğitim için 25.000 film incelemesi ve test için 25.000 inceleme bulunmaktadır. Ayrıca her film için 1 ile 10 arasında derecelendirme içerir. Değerlendirme sistemi, inceleme ile birlikte, yazılı incelemenin duygunun bir göstergesidir. İnceleme 10 üzerinden 10 ise, incelemenin olumlu olduğu açıktır. Diğer taraftan eğer bu derece düşük ise, inceleme de negatif olması gerekir. İncelemeye eşlik eden bu derecelendirme, incelemelerin veri kümesini manuel olarak etiketlemeden sınıflandırılabileceği anlamına gelir.

Film incelemeleri genellikle duygu analizi yapmak için iyi bir veri kümesi sağlar. Çünkü metinler nispeten uzundur ve sınıflandırılması kolaydır. Veri kümesi film incelemeleriyle sınırlı olduğundan, her senaryo için ideal değildir. Çünkü tweetler genel bir film incelemesinden daha karmaşık bir metindir. Twitter gönderilerinin 140 karakter limiti olduğu için, haberlerle ilgili yayınlar veya burada ifade edilen görüşler doğal olarak çok kısadır. Bu nedenle Twitter kullanıcıları, genellikle kullandığımız dilden çok farklı olan bu sınırın üstesinden gelmek için çeşitli kısaltılmış ve argo dillerini kullanır. RT (retweets için), # (hashtag), @ (birinden bahsetmek için), vb. gibi twitter için bazı özel kelimeler vardır. Bu noktalardan bir tweetin ürettiği duygu miktarının 200 kelimeden çok daha az olduğu anlaşılabilir. Bir film incelemesi ayrıca bir tweet'ten çok daha fazla düşünülür. Twitter kullanıcıları da genel olarak yazım konusunda çok dikkatli değiller.

Duygu analizi, nesnenin tanımlanması, özelliklerin çıkarılması ve görüş yöneliminin bulunması gibi çeşitli zorluklarla başa çıkmaktadır. Duygu analizi sınıflandırma görevini 3 adımda gerçekleştirir:

- Belge seviyesi
- Cümle seviyesi
- Özellik seviyesi veya En Boy seviyesi

3.2.1 Belge Düzeyi Sınıflandırması

Bu seviyenin analizi, bir belgede ifade edilen duygunun olumlu ya da olumsuz olduğunu belirlemektir. Örneğin, ürün incelemeleri göz önüne alındığında, sistem genel duygu polaritesini değerlendirir (Segaran, 2007). Belge düzeyinde analiz, bir metnin tek bir hedefe karşı duygusallık ifade ettiğini varsayar. Bu genellikle ürün yorumları, film yorumları, restoran yorumları, vb. için doğru olsa da, muhtemelen bir belgenin birden fazla hedefi eleştirdiği durumlar için geçerli değildir.

3.2.2 Cümle Düzeyi Sınıflandırması

Bu seviyenin analizi, bir cümlede ifade edilen görüşlerin olumlu, olumsuz veya tarafsız olup olmadığını belirlemektir. Cümle seviyesi analizi iki şekilde yapılabilir. Bir yol ise bu tür analizleri etiketlerin pozitif, negatif ve nötr olduğu 3 yönlü bir sınıflandırma görevi olarak görmektir. İkinci yol ise kanaatli metinleri kanaatsiz metinlerden ayırmak için önce cümlelerdeki öznelliği tespit etmektir. Ondar sonra bu öznel metinleri iki etiketten (olumlu veya olumsuz) biriyle sınıflandırmaktır.

Cümle düzeyi analizinin zorluğu, her bir cümlenin semantik ve söz dizimsel olarak metnin diğer bölümleriyle bağlantılı olmasıdır. Bu nedenle, bu görev hem yerel hem de genel bağlamsal bilgi gerektirir. Yang, ürün incelemesini cümle düzeyinde analiz eder ve bu zorluğu başarıyla ele almıştır (Yang ve Cardie, 2014).

3.2.3 Özellik Seviyesi Sınıflandırması

Hem belge seviyesi hem de cümle seviyesi analizleri, insanların tam olarak neyi sevdiğini ve neyi sevmediğini keşfetmez. Özellik sınıflandırması daha ince taneli analiz yapar. Dil yapılarına (belgeler, paragraflar, cümleler, cümleler veya ifadeler) bakmak yerine, aspect level doğrudan görüşün kendisine bakar. Bir görüşün bir duygudan (olumlu veya olumsuz) ve bir hedeften (görüş)oluştuğu fikrine dayanır.

Özellik düzeyinde veya en boy düzeyinde, bir nesnenin çeşitli özelliklerinin analizi yapılır. Örneğin bir müşterinin bir Samsung cep telefonu satın aldığını varsayalım. Müşteri alınan cep telefonunun kamera kalitesinin makul olduğunu ancak cep telefonunun ses kalitesinin olmadığını gözlemledi. Bu nedenle, analiz etmek için, bir varlık yönü seviyesi analizinin çeşitli yönleri gerçekleştirilir. Bu ince taneli analizin görevleri üç katlıdır: (1) hedefin özelliklerini çıkarmak, (2) özellik açısından polariteyi belirlemek, (3) genel değerlendirmeyi özetlemek (Hu ve Liu, 2004). En-boy seviyesi duygu analizi, diğer analiz seviyelerine kıyasla en zor görevlerden biridir.

Ayrıca, araştırma cümle düzeyinde (Wilson ve ark., 2005), madde seviyesi (Wilson ve ark., 2004) veya kelime seviyesinde (Wilson ve ark., 2005) de yapılabilmişti. Bazı çalışmalar, birden fazla hedefin karşılaştırıldığı karşılaştırmalı görüşleri de analiz etmişti (Jindal ve Liu, 2006).

Twitter duygu analizi belge düzeyinde analiz olarak kabul edilebilir. Twitter en fazla 280 karakter yazmayı izin verdiğinden her bir tweet durumu çok kısa olma eğilimindedir. Genellikle, bir tweet sadece bir basit cümle veya sadece birkaç kelime içerir. Bu nedenle, Twitter duygu analizi, diğer analiz seviyelerinde kullanılan çok çeşitli stratejiler de kullanır.

Duygu analizi çalışmaları, veri ön işleme, özellik seçimi ve sınıflandırma, ardından verilerin kutuplarını bulmaktan oluşur. Veri ön işleme, tokenleme, durdurma sözleşmesi, kelimelerin kökü bulma, kök çözümleme, vb. kelimeleri durduran, herhangi

bir görüş tutmayan kelimelerdir (ben, var, içinde vb.). Bu nedenle bunları kaldırmak faydalıdır. Kelimelerin kök bulma, kelimelerin değişken formlarını, yardım etmelerine yardımcı olmak gibi temel formlarına dönüştürme görevidir.

3.3 Karmaşık Olay İşleme

Karmaşık olay işleme veya kısaca CEP, verilerin neredeyse gerçek zamanlı analizine duyulan ihtiyaç tarafından geliştirilen gelişmekte olan bir ağ teknolojisidir. CEP ile birden fazla kaynaktan gelen büyük miktarda olay veri akışı etkili bir şekilde analiz edilebilir. Hareket halindeki verileri hareket ettirir, izler ve analiz eder ve neredeyse gerçek zamanlı olarak karar verir. Tipik ilişkisel veri tabanı sistemleri sorgu odaklı olsa da, CEP olay odaklı bir yaklaşım sağlar. Bu yaklaşım, yüksek olay veri hızları, sürekli sorgular ve milisaniye gecikme süresi ile karakterize eder. Bu özellikler normal ilişkisel veri tabanlarının kullanılmasını pratik hale getirir (Wu ve ark., 2006).

Olay odaklı uygulamalar, anlamlı kalıpları, ilişkileri ve veri soyutlamalarını tanımlamak amacıyla birçok olayı işleyen bir teknoloji olan CEP'i kullanır. Geleneksel olarak sorgular, sonlu bir sonuç elde etmek için verilere karşı işlenir. Bitişik sorgular, verilerin sonsuz bir veri akışında tetiklendiği ayakta sorgulardır. Bu sorgu, verileri gerçek zamanlı olarak işler. Bu teknik, yaklaşan büyük veri odak alanı ile faydalı görünüyor. Gittikçe daha fazla veri üreten büyük şirketler tarafından desteklenmektedir. Olay odaklı uygulamalara duyulan ihtiyaç, veriler üzerinde gerçek zamanlı olarak hareket etmek için finansal hizmetler, sağlık hizmetleri, imalat, petrol ve gaz, ulaşım, kamu hizmetleri ve web analitiği gibi birçok sektörden gelmektedir (Wu ve ark., 2006).

CEP'nin uygulanmasının olası bir örneği, bir petrol ve gaz şirketi ile gösterilebilir. Büyük bir petrol ve gaz şirketi için, sondaj sırasında petrol, gaz veya malzeme israf etmemek çok önemlidir. Açıkçası israf her zaman kötüdür, bunu önlemek faydalı olacaktır. Hasat sırasında platform, yağı çıkarmak için çok fazla sabit ve yazılım kullanır. Petrol çıkarma sürecinde, frekanslar üzerindeki farkındalık potansiyel sorunlara işaret edebilir. Makineleri gerçek zamanlı olarak izlemek, sorunları erken bir aşamada tespit etmeye yardımcı olabilir. CEP, makinelerden gelen büyük miktarda veriyi gerçek zamanlı olarak izlemenize yardımcı olabilir. Sıklık, önceden tanımlanmış bir sürekli sorgu aracılığıyla tespit edilebilir. Bu nedenle gelecekteki sorunları önlemek ve çevre, malzeme ve para tasarrufu sağlamak için önlemler alınabilir (Cugola ve Margara, 2012)

Karmaşık olay işlemenin avantajları, bir duygu analizi uygulaması için de işe yarayabilir. CEP tekniği ile mesajlar, yayınlandıktan hemen sonra sürekli bir akışta alınabilir. Daha sonra, tweetler alındıktan hemen sonra işlenebilir. Yüksek işlem hacmi, CEP gibi bir tekniği Twitter'dan gelen veri akışıyla çalışmak için mükemmel hale getirir.

3.4 Duygu Analizinde Özellik Çıkarma:

Özellik çıkarma, özelliklerin verilerden çıkarıldığı işlemdir. Özellikler olarak adlandırılan bu özellikler, verilerden gelen özelliklerdir, çünkü pratikte tüm giriş verileri sınıflandırmada kullanmak için çok büyüktür. Özellikler, orijinal verileri mümkün olduğunca tanımlamak için ayırt edici olmalıdır. Öte yandan, özellikler fazlalıkları ve verilerin yüksek boyutsallığını önlemek için alanı azaltmalıdır. Araştırmacılar, hangi özellik çıkarma yöntemlerinin iyi çalıştığını test etmek için duygu analizinde farklı özellikler önerdi ve denedi. Aşağıdaki özellikler tartışılmıştır: n-gram, tf-idf ölçüsü, terim varlığı ve konuşma parçası (Part-of-speech POS) etiketleme.

3.4.1 n-gram Özellikleri

n-gram yöntemi nispeten basit bir algoritmadır. Bir n-gram, metinsel veya sözlü bir kaynaktan gelen *n* öğelerinin bitişik bir dizisidir. Öğeler genellikle harfler veya kelimelerdir. Tüm belgelerdeki tüm kelimelerin sıklığını saymak, bir kelime sıklığı tablosuyla sonuçlanır. Yüksek sık kullanılan kelimelerin metinlerde görünme olasılığı daha yüksektir, bu nedenle bu kelimeler veri kümesini daha iyi tanımlar. Tüm belgeler arasında frekansı ölçmenin iki olası yolu vardır: özetlenen terim frekansı ve belge frekansı. Bu iki frekans ölçüsü sırasıyla frekans terimine ve varlık terimine dayanır. Frekans (TF(w, d)) terimi, denklem 3.4.1.1'de belirtilen d belgesinde w kelimesinin kaç kez gerçekleştiğidir frekans terimini hesaplarken, belgedeki tüm kelime oluşumları sayılır. Bu nedenle, sıklık terimi, [0−n] aralığında bir değer varsayabilir; burada n, belgedeki toplam kelime sayısıdır. Frekans (tp(w, d)) terimi, yalnızca W kelimesinin d belgesinde mevcut olup olmadığını kontrol eder ve bu da ikili bir değere neden olur. Bu ölçü denklem 3.4.1.2'de belirtilmiştir. Temel fark, kelimelerin bir mesajda nasıl sayıldığıdır.

$$tf(w,d) = |\{\, w \in d \,\}| \qquad 3.4.1.1$$

$$tp(w,d) = \begin{cases} 1 & \text{if } w \in d \\ 0 & \text{if } w \notin d \end{cases} \qquad 3.4.1.2$$

Toplanan Terim Frekansı (DF(w, D)), bir kelimenin tüm dönem frekanslarını w tüm belgelerde toplar d, formül denklem 3.4.1.3'te belirtilmiştir. Belge sıklığı, bir sözcüğün oluştuğu belge sayısını ifade eder. Bu, denklem 3.4.1.4'te resmileştirilir, burada belge frekansı tüm belgeler arasında W kelimesinin df(w, D) 'dir. genellikle belge kümesi daha resmi dilsel jargonda 'bütünce' olarak adlandırılır.

$$stf(w,\ D) = \sum_{d \in D} tf(w,d) \qquad 3.4.1.3$$

$$df(w,D) = |\{\ d\ \in\ D : w\ \in\ D\ \}| \qquad 3.4.1.4$$

Bununla birlikte, sık kullanılan kelimeler sınıflandırma için mutlaka iyi özellikler değildir. Oldukça sık bir kelimenin dağılımı sınıflar üzerinde eşit olarak dağıtılırsa, ayırt edici gücü düşüktür. Örneğin, "özel" kelimesi oldukça sık görülür, ancak olumlu, olumsuz ve tarafsız belgelerde sıklıkla görülür ve bu nedenle sınıflardan biri için ayrımcı değildir. Yeni vakaları sınıflandırırken, sınıflandırıcı sınıflardan birini rastgele seçebilir. Bu nedenle, bir özelliğin sınıflar arasında ne kadar iyi ayrım yaptığını dikkate alan bir özellik seçim yöntemi de kullanılır.

Daha yüksek mertebeden n-gramlara kadar uzanan metinler, tek uzunluklu öğeler olarak değil, n-uzunluklu öğeler olarak ayrılır. N = 2 (bigrams) durumunda, öğeler iki ardışık kelimeden oluşur. Set, orijinal metinde ardışık olan iki kelimenin tüm kombinasyonlarını içerir. Aynı şey n = 3 (trigramlar) ve daha yüksek n değerleri ve ayrıca harf tabanlı n-gramlar için de geçerlidir. Sezgisel olarak, bu yüksek mertebeden n-gramlar, ardışık kelimelerin ilişkisini, örneğin olumsuzlamada daha iyi yakalar gibi görünmektedir.

Son olarak, en değerli kelimelerden bir seçim yapılır. Yalnızca üst K en değerli n-gram özellikleri, ağırlık ölçütüne göre, özellik vektörünü oluşturmak için seçilir. Bu, verilerin boyutluğunu azaltır.

Literatürde, birkaç araştırmacı (Apoorv Agarwal ve ark., 2011; Go ve ark., 2009; Pak ve Paroubek, 2010; Pang ve ark., 2002) n-gramlara dayalı sınıflandırma modelleri uygulamıştır. Uni ve bigram varyantları en çok ve daha az ölçüde trigramlar uygulanır. Unigram ve bigramların karşılaştırılması, hangisinin daha iyi olduğu konusunda net bir cevap vermiyor. Liu, unigram teriminin varlığında, unigram teriminin sıklığından biraz

daha iyi sonuçlar elde etti. Bu nedenle, varlık teriminin sıklık teriminden daha iyi çalıştığı sonucuna varmışlardır. Ayrıca bigram terim varlığı ve uni ve bigram terim varlığı kombinasyonu ile testler yaptılar, ancak bu sonuçlar unigram terim varlığının sonuçlarından daha iyi performans göstermedi. Bu arada, Go ve ark. unigramlar ve bigramların kombinasyonu ile sırasıyla sadece unigramlardan veya sadece bigramlardan daha iyi sonuçlar elde edildi (Go ve ark., 2009). Ayrı unigramlar hala araştırmasında bigramlardan daha iyi çalıştı. Buna karşılık, Pak ve ark. bigramlar ile unigramlardan daha iyi doğruluklar elde edildi (Pak ve Paroubek, 2010). Bu, farklı uygulama ve kullandığı farklı etki alanından kaynaklanabilir. Pang ve ark. ve Git ve ark., Agarwal ve diğ. unigramların Twitter'a diğer n-gram modellerinden daha iyi uygulanabileceği sonucuna varma eğilimindedir. bigram, kelimeler arasındaki ilişkiyi unigramlardan daha iyi yakaladığı görülmektedir.

3.4.2 tf-idf Ölçüsü

tf-idf (term frequency-inverse document frequency) ölçüsü, bir kelimenin bir belge kümesindeki önemini yansıtan bir istatistiktir. Bu ölçü iki ayrı önlemden oluşur: sıklık terimi ve belge sıklığının tersi. Frekans terimi ve normal belge frekansı tartışıldı Önceki bölüm 3.4.1 ve formülleri sırasıyla denklem 3.4.1.1 ve 3.4.1.4'te belirtilmiştir. Ters belge sıklığı, tüm belgelerdeki bir kelimenin nadirliğini ölçmek için kullanılır. Ters belge sıklığının değeri ne kadar yüksek olursa, belge kümesindeki kelime o kadar nadirdir. Ters belge sıklığı, ıdf (w, D) bir word W tüm belgeler arasında d denklem 3.4.2.1'de gösterilmiştir. Bu, denklem 3.4.2.2'de gösterilen tf-idf ölçüsüne frekans terimi ile birlikte birleştirilebilir. Bu denklemde, tf-idf, bir belge kümesindeki d belgesindeki W kelimesinin tf-idf'idir (w, d, D) D.

$$idf(w, D) = \log \frac{|D|}{df(w,D)} \qquad 3.4.2.1$$

$$tf - idf\ (w, d, D) = tf(w, d).idf(w, D) \qquad 3.4.2.2$$

Terim sayısını bulmak için kullanılan “terim sıklığı” bütüncede meydana gelir. "Terim varlığı" aslında terimin cümlede gerçekleşip gerçekleşmediğini gösteren ikili değerli bir özellik vektörüdür. 1 terimin varlığını temsil eder ve 0 terimin yokluğunu temsil eder. Pang ve Lee, duygu analizinde "terim varlığı" nın " terim sıklığı” ndan daha önemli olduğunu göstermektedir (Pang ve Lee, 2009). Burada ayrıca, nadir kelimelerin

oluşumunun, sık kullanılan kelimelerin oluşumundan daha güvenilir bilgiler içerdiğini bulduk. Bu fenomen Hapax Legomena olarak bilinir.

Bir kelime bir belgede sıklıkla geçtiğinde ve diğer tüm belgelerde sık sık geçmediğinde, tf-idf değeri daha yüksektir. Twitter analizinde belgeler tweetlerdir, yani tweetin içinde yüksek frekanslı ve tüm tweetlerde düşük frekanslı kelimelerin yüksek tf-idf değerine sahip olduğu anlamına gelir. Bu kelimeler çok fazla tweet içermediği için sınıflandırma için kullanılacak en iyi kelimeler gibi görünmüyor. Bu sebeple tf-idf ölçüsü bu araştırmada öznitelik seçimi olarak değerlendirilmemiştir.

3.4.3 Konuşma Bölümleri Etiketleme (Parts of Speech Tagging - POS)

Bir konuşma parçası etiketleyicisi veya kısaca bir POS etiketleyicisi, belirli bir konuşma parçasına karşılık gelen bir metindeki bir kelimeyi işaretlemenin bir yöntemidir. Konuşmanın bir kısmı, daha sezgisel isimler olan kelime sınıfı veya sözcük kategorisi olarak da bilinir. Bir metin içinde, tüm kelimeler karşılık gelen sözcük kategorisine göre sınıflandırılır. Bunun arkasındaki fikir, bir cümlede yalnızca sınırlı bir kelime kümesinin, duygu-kelimeler olarak adlandırılan duyguyu göstermesidir. İngilizce dilinde sözcüksel kategorilerin örnekleri isim, fiil, zarf ve sıfattır. Bu sözcüksel kategorilerden bazıları, sıfatlar ve zarflar gibi daha sık duygu sözcükleri içerir. POS etiketleyicisinin bir sorunu, birden fazla sözcüksel kategoride görünebilen bir kelime olabilir, ancak yalnızca bir sözcüksel kategoride bir duyguyu ifade eder.

Sözcük kategorileri Tüm diller için aynı olmadığından, her dilin kendi POS etiketleyicisini kullanması gerekir. Bölüm 3.4.2.1'de dil tanımlama işlemi bu sorunu çözmüştür. Tanımlanan dil daha sonra POS etiketleme yönteminde kullanılabilir.

POS etiketleyicisinin modelin kalitesini iyileştirmediği sonucuna varıldı. Diğer makalelerde yazarlar aynı sonuca varmışlardır, (Bethard ve ark., 2006; Go ve ark., 2009; Kouloumpis ve ark., 2011). Bununla birlikte, bazı makaleler vardır, örneğin Pak ve ark. bu, POS etiketleyicisinin performansı gerçekten artırdığı sonucuna varır. Bazı POS etiketlerinin duygusal metinlerin iyi göstergeleri olduğunu iddia ettiler. Pak ve ark., Agarwal ve diğ. POS özelliklerinin eklenmesiyle daha iyi sonuçlar elde edildi. Ek olarak, Barbosa ve ark. bir POS etiketliği ile daha iyi sonuçlar elde etti, ancak araştırması elle oluşturulmuş önyargılı veri kümesine dayanıyordu (Barbosa ve Feng, 2010). Bu nedenle,

bu son araştırmanın sonuçları bu araştırma için daha az önemlidir. Turney, POS etiketleyicisini unigram modeliyle birlikte kullanmadığından, karşılaştırması gereken hiçbir sonuç yoktu (Yared, n.d.). POS etiketliğici hakkında n-gram modelinden daha fazla bir şey sonuçlandırmak çok daha zor görünüyor. POS-etiketliğici ile daha iyi sonuçlar elde eden bazı araştırmacılar var, diğerleri ise daha kötü sonuçlar elde etti. Film incelemeleri ile ilgili literatürde bile POS-etiketleyici diğer modellerden daha iyi performans göstermedi.

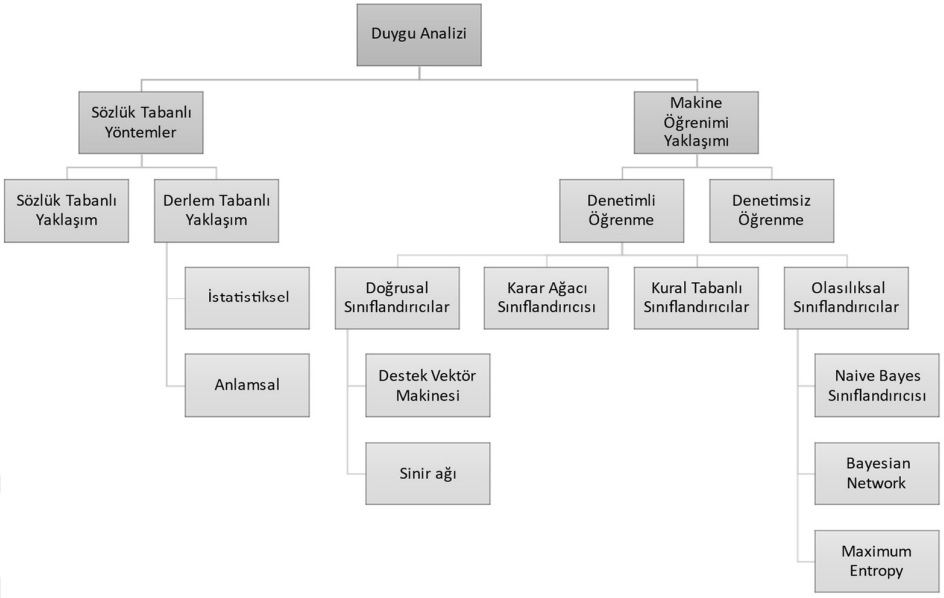
3.4.4 Olumsuzluğu anlama

Olumsuzlamalar, dilbilimin çok önemli bir parçasıdır çünkü diğer sözcüklerin kutuplarını da etkilerler. Olumsuzluk, "hayır", "yok", "değil" gibi kelimeleri içerir. Bir cümlede olumsuzluk geçtiğinde, bu terimden etkilenen kelime sırasını belirlemek önemlidir. Olumsuzlamanın kapsamı, yalnızca bir olumsuzlamadan önceki kelime ile sınırlandırılabilir veya olumsuzlamadan önceki diğer kelimelere kadar genişletilebilir.

Örneğin, "Bu cep telefonu güzel değil ama düzgün çalışıyor" cümlesinde, olumsuzlamanın kapsamı yalnızca olumsuzlamadan sonraki bir önceki kelimeyle sınırlıdır. Ancak başka bir cümlede “Pil uzun süre çalışmıyor” cümlesinin başlangıcına kadar olumsuzluk kapsamıdır. Bu örnekler, olumsuzlama kapsamının sabit olmadığını ve bağlaçlar, noktalama işaretleri ve Olumsuzlamanın Konuşma Bölümü vb. gibi farklı dilbilimsel özelliklere göre değiştiğini göstermektedir.

3.5 Duygu Analizi Metodlar

Duygu analizi yöntemleri, makine öğrenmesi tabanlı, sözlük tabanlı ve hibrit yöntemdir. Makine öğrenimi yönteminde, bir cümlenin polaritesinin zaten belirtildiği yerde etiketli veri kümesi kullanılır. Bu veri kümesinden, özelliği çıkarırız ve bu özellikler bilinmeyen giriş cümlesinin polaritesini sınıflandırmaya yardımcı olur. Makine öğrenimi yöntemleri denetimli öğrenme ve denetimsiz öğrenmeye ayrılmıştır.



Şekil 3.5.1 Duygu Analizi Yöntemleri

3.5.1 Denetimli öğrenme

Bu yaklaşım, modeli eğitmek için etiketli veriler mevcut olduğunda kullanılır. Denetimli öğrenmede iki adım kullanılır: birincisi modeli eğitmek, diğeri tahmin etmektir. Eğitim sırasında, etiketleri ile veri kümesi, bir çıktı olarak bir model veren sınıflandırma algoritmasına beslenir (Kaur ve Sabharwal, 2018). Bundan sonra, test verileri kategoriyi tahmin etmek için modele beslenir. Aşağıdaki gibi çeşitli denetimli sınıflandırma algoritması vardır:

Naive Bayes:

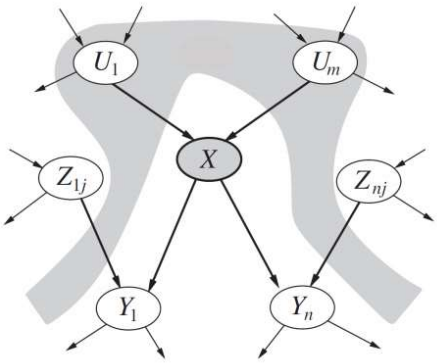
Bu olasılıksal bir sınıflandırma algoritmasıdır. Her kelimeyi bağımsız olarak görür, çünkü cümledeki terimin yerini dikkate almaz. Naive Bayes, bir etikete karşılık gelen her terimin olasılığını hesaplamak için Bayes teoremine dayanır.

$$p(\text{etiket}|\ \text{özellikleri}) = \frac{\text{p(etiket)} * \text{p(özellikleri}|\ \text{etiket)}}{p(\text{özellikleri})}$$

p(etiket), veri kümesindeki etiketin olasılığıdır. p(özellik|etiket) bir etiketle ilgili bir özelliğin olasılığıdır. p(özellik) bir özelliğin olasılığıdır. Geol ve ark., Naive Bayes'i kullanarak Twitter veri kümesinin sınıflandırmasını iyileştirdi. SentiWordNet Lexicon kullanarak pozitif ve negatif tweetlerin polarite skorunu sağladı (Goel ve ark., 2016).

Bayes Ağı:

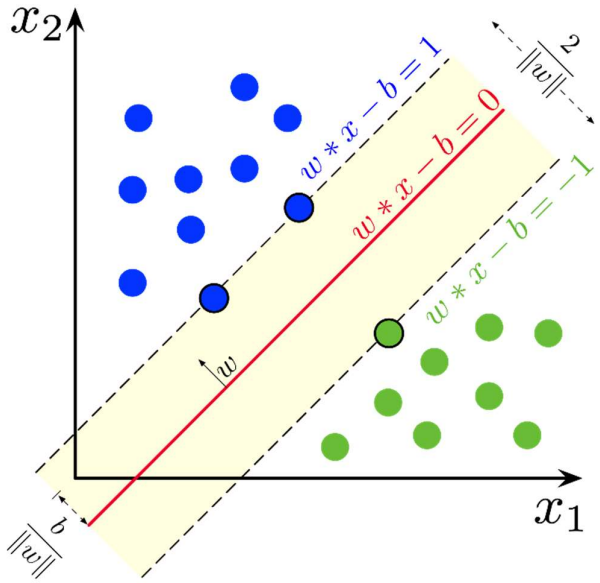
Naive Bayes sınıflandırıcısı her kelimeyi bağımsız olarak ele aldığından, kelimeler arasında anlamsal bir ilişki bulamaz, Oysa Bayes ağı bunu yapabilir. Bayes ağı, kelimelerin birbirine bağımlılığını güçlü bir şekilde dikkate alır. Bayes ağı, her düğümün kelimeyi bir değişken olarak temsil ettiği ve kenarların değişkenler arasındaki bağımlılığı temsil ettiği, döngüsel olmayan yönlendirilmiş bir grafik açısından bağımlılığı temsil eder. Al-Smadi et al. duygu sınıflandırıcı olarak Bayes ağlarını kullandı. Bu algoritmadan iyi sonuçlar buldular ve bazen diğer sınıflandırıcılardan daha iyi performans gösterdiler (Al-Smadi ve ark., 2019).



Şekil 3.5.2 Bayes ağı

Destek Vektör Makinesi (SVM):

SVM, ikili sınıflandırma problemlerini çözmek için ilk kez başlatılır. Farklı sınıflardan gelen veri noktaları arasındaki karar sınırlarını tanımlamak için ayırıcı görevi gören en iyi hiper düzlemleri belirlemeye odaklanır.

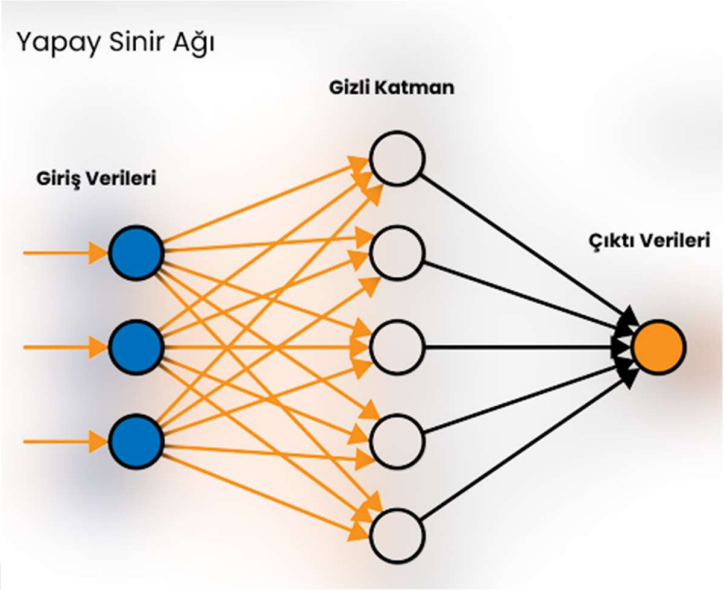


Şekil 3.5.3 Destek Vektör Makinesi

Şekilde gösterildiği gibi, farklı sınıfların iki destek vektörü arasındaki maksimum mesafeyi koruyabilen bir hiper düzlem seçilmelidir. SVM, doğrusal ve doğrusal olmayan sınıflandırma görevlerini yönetme yeteneğine sahiptir. Zainuddin ve ark. sınıflandırma için SVM kullanmışlardır. tf-idf, terim oluşumu, ikili oluşum gibi çeşitli ağırlıklandırma şemaları kullandılar. Boyutsallık azaltma ve gürültü giderme için kullanılan öznitelik seçimi olarak ki-kareyi kullanmışlardır. SVM ile ki-kare öznitelik seçiminin kullanılmasının doğruluğu artırdığını göstermiştir (Zainuddin ve Selamat, 2014).

Yapay Sinir Ağı:

Yapay Sinir Ağı (ANN) insan beyninin nöron yapısını taklit eder. Sinir ağı için temel birim nörondur. ANN bir giriş katmanı, gizli katman ve bir çıkış katmanı içerir. Bir vektör "a (i)" nörona girdi olarak verilir, vektör bir belgedeki bir kelimenin sıklığını gösterir. Fonksiyonu hesaplamak için kullanılan her nörona karşılık gelen bir " a" ağırlığı vardır. Sinir ağı kullanımı doğrusal fonksiyon x (i) = A. (a (i)) ' dir. X (i) işareti, sınıfı sınıflandırmak için kullanılır.



Şekil 3.5.4 Yapay Sinir Ağı

Yapay sinir ağlarında, modelin eğitimi iki aşamadan oluşur: ileri yayılma ve geri yayılma. İleri yayılımda giriş, rastgele sayılar olan ağırlıklar ile çarpılan nöronların giriş katmanında verilir. Fonksiyonlar, çıkış değerini 0 ile 1 arasında normalleştirmek için kullanılır. Daha sonra çıktı, hedef değerle karşılaştırılır, eğer iki değer arasında bir fark (hata) varsa, geriye doğru yayılma gerçekleştirilir. Geri yayılım sırasında giriş, hata değeri ile çarpılır, böylece ağırlık ayarlanabilir. Bu nedenle öğrenme hataya bağlıdır.

Vega ve Mendez-Vazquez, öğrenme süreci için rekabetçi ve Hebbian öğrenmeyi kullandığı dinamik bir derin öğrenme yöntem modeli önerdi. Temel yaklaşımı DNN ile karşılaştırdı ve DNN'in temel yöntemlerden daha iyi performans gösterdiğini gösterdi (Vega ve Mendez-Vazquez, 2016). Patil ve ark. bir evrişim sinir ağı (CNN) ile gizli anlamsal analiz (LSA) kullandığı bir teknik önerdi. LSA, kelimeleri vektörlere dönüştürmek için bir tekniktir. LSA'da ağırlıklandırma, tf-idf algoritması ile yapıldı. Bu model %87 doğruluk sağlamıştır (Patil ve ark., 2018).

Karar ağacı:

Karar Ağacı kullanmanın amacı, eğitim verilerinden çıkarılan basit karar kurallarını öğrenerek hedef değişkenin sınıfını veya değerini tahmin etmek için kullanılabilecek bir eğitim modeli oluşturmaktır. Karar Ağaçlarında, bir kayıt için bir sınıf etiketi tahmin etmek için ağacın kökünden başlar. Kök özniteliğinin değerlerini

kaydın özniteliği ile karşılaştırır. Karşılaştırma bazında, o değere karşılık gelen dalı takip eder ve bir sonraki düğüme gider.

Karar ağacındaki ana zorluk, hangi özniteliğin kök düğüm olarak seçileceğini bulmaktır. Bu, Bilgi kazancı ve Gini Endeksi gibi bazı istatistiksel yaklaşımlar kullanılarak çözülebilir. Bir karar ağacı, duygu analizi için iyi bir yöntemdir, çünkü aynı zamanda büyük miktarda veri üzerinde iyi bir sonuç sağlar. Yaygın olarak kullanılan karar ağacı algoritmaları CART, CHAID ve C5.0'dır.

Karar ağacı, eğitim verilerini hiyerarşik olarak böler. Verilerin bölünmesi için öznitelik değerindeki bir koşul kullanılır. Koşul, bir kelimenin var olup olmamasına bağlıdır. Bölme işlemi, terminal düğümler, sınıflandırma görevi için kullanılan az sayıda özelliği temsil edene kadar devam eder.

Kural Tabanlı Sınıflandırıcı:

Kural tabanlı sınıflandırıcı tarafından üretilen Model, bir dizi kural biçimindedir. Bu kurallara dayanarak, yeni bilgiler için bir tahmin yapılır (Xia ve ark., 2016). Kurallar her zaman öncül ve sonuç biçimindedir. Eğer (öncül) Sol tarafta ise koşulları temsil ederken, sağ taraf (sonuç) sınıfın tahminini temsil eder. Bir kural formu aşağıda görülebilir.

$$\{w1/\backslash w2/\backslash w3\} \rightarrow \{+|-\}$$

Bir kuraldaki kelime, aşağıda gösterilen duyguyu ifade eder.

$$\{Good\} \rightarrow \{+\}\ \{Bad\} \rightarrow \{-\}$$

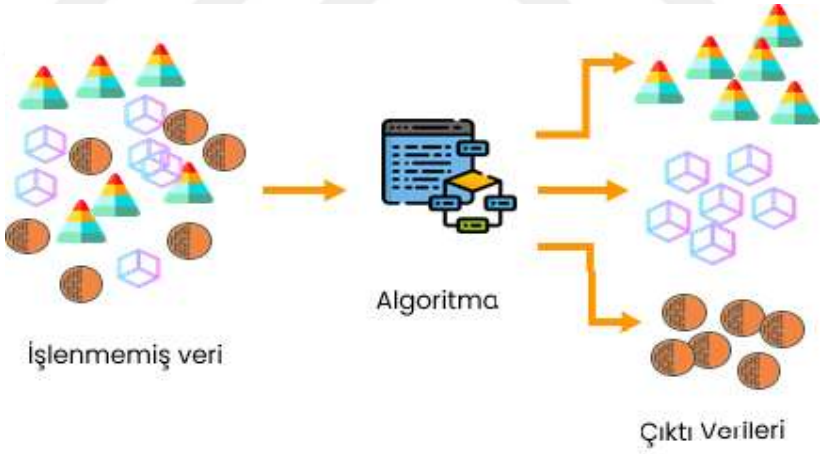
Metin sınıflandırmasında, IF bölümü, terim varlığı olabilecek özellikler kümesini temsil eder ve ardından bölüm, etiketi temsil eder. Kural tabanlı sınıflandırıcı, kuralı tanımlamak için iki terim kullanır: Güven ve Destek. Kuralla ilgili bir eğitim veri setindeki örnek sayısı destek tarafından tanımlanır. Bir özellik kümesi oluşursa bir etiketin koşullu olasılığı, bir kurala olan güven ile temsil edilir.

Buddeewong ve Kreesuradej Birliktelik Kuralı Tabanlı Metin sınıflandırıcı algoritması (ARTC) önerdiler. Çalışmada iki öğe kümesi oluşturdular: Biri diğer

sınıflarla örtüşmeyen kelimeler için, diğeri ise diğer sınıflarla örtüşen kelimeler için. Daha sonra sık kullanılan itemset yardımıyla kuralları oluşturdular. Kural oluşturma için Apriori algoritmasını kullanmış ve %95.08 doğruluk oranına ulaşmıştır (Buddeewong ve Kreesuradej, 2005).

3.5.2 Denetimsiz öğrenme:

Bu yöntem, etiketli verilerin güvenilirliğinin zor olduğu durumlarda kullanılır. Etiketlenmemiş verileri toplamak, etiketlenmiş verilere göre daha kolaydır. Cümle, her kategorinin anahtar kelime listelerine göre kategorize edilir. Etki alanına bağlı verileri analiz etmek için denetimsiz yaklaşımı kullanmak daha kolaydır. Unnisa ve ark., tweetlerin spektral kümeleme yaklaşımı kullanılarak pozitif ve negatif kümeye kümelendiği denetimsiz yaklaşımı kullanarak duygu analizi yaptı. Spektral kümeleme, Naive Bayes, SVM ve maksimum Entropiden daha iyi performans gösterir (Unnisa ve ark., 2016).



Şekil 3.5.2.1 Denetimsiz Öğrenme

3.5.3 Sözlük tabanlı yöntem:

Bu görüşü ifade eden kelimeler duyguları analiz etmek için en önemlisidir. Olumlu görüş, olumsuz görüşün bir varlık için istenmeyen bir etiket olduğu istenen etikettir. Lexicon, her kelimeyle bir polarite puanının ilişkili olduğu önceden tanımlanmış bir kelimenin bir koleksiyonudur. Bu, duyguları sınıflandırmak için en kolay yaklaşımdır. Bu sınıflandırıcı bir sözlükten yararlanır ve bir cümleyi kategorize etmek için kullanılan kelime eşleştirmesini gerçekleştirir. Bu sınıflandırma yaklaşımının performansı sözlüğün

büyüklüğüne bağlıdır. Alt bölümde açıklanan sözlük tabanlı yöntem altında kullanılan iki yaklaşım vardır.

Sözlük Tabanlı Yaklaşım:

Sözlük tabanlı yaklaşımda, bazı kelimeler bir tohum kelimesi olarak seçilir ve bu kelimeler kelime kümesinin boyutunu büyütmek için eşanlamlıyı bulmak için kullanılır. Çevrimiçi sözlükler boyutunu genişletmek için kullanılır. Tohum kelimeleri, bir bütüncede benzersiz ve önemli olan görüş kelimeleridir. Bu tohum kelimeleri ve yeni genişletilmiş kelimeler, duygu analizi yapmak için bir özellik olarak kullanılır. WordNet, SentiWordNet, SentiFul, SenticNet gibi çeşitli sözlükler vardır.

Bütünce tabanlı yaklaşım:

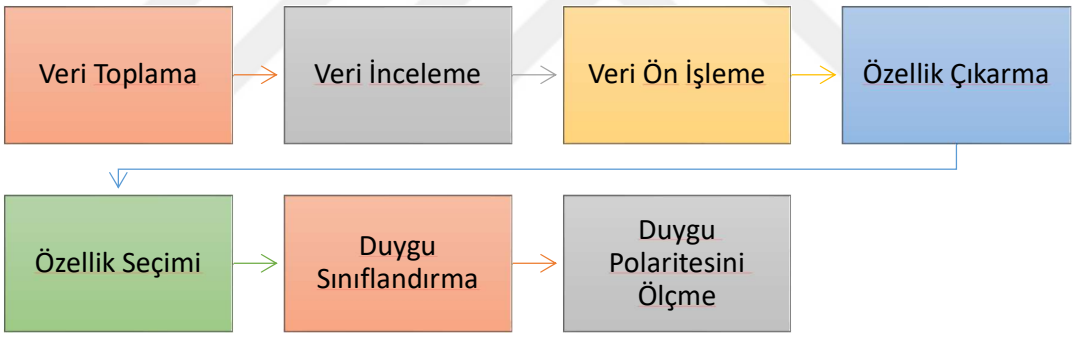
Bütünce tabanlı yaklaşımda, sadece bir kelimenin etiketini değil, aynı zamanda bağlam yönelimini de buluruz. Bu yaklaşımda, önce tohum kelimelerinin bir listesi hazırlanır ve daha sonra bu listelenen kelimelerin söz dizimsel deseni, bütünceden yeni öznel kelimeler oluşturmak için kullanılır. Söz dizimsel desen, birbirleriyle veya birlikte meydana gelen bir kelime anlamına gelir. Bu yaklaşım daha sonra iki şekilde çalışır:

- İstatistiksel temelli yaklaşım.
- Semantik temelli yaklaşım.

Yazar her iki sözlük tabanlı yaklaşımı da göstermektedir (Abdulla ve ark., 2013). SVM ile bütünce bazlı yaklaşımın ışık kaynaklı veriler için yüksek doğruluk sağladığını gözlemledi. Sözlüğü, sözlüğü doğruluğunu artması ile tabanlı yaklaşım da artmış olduğunu da ilan etti.

4. UYGULAMA

Veri setinin toplanmasından sonuçların bulunmasına kadar her bölümün birkaç farklı adımdan oluşan kendi süreci vardır. Bu bölümde, bu görevleri yürütmek için kullanılan süreç tartışılacaktır. İlk olarak bu veri setini Twitter API'sinden toplamak için kullanılan yöntem ele alınacaktır. Bu verileri toplamak için Twarc python kütüphane ile Twitter API kullanıldı. İkinci olarak, verilerin ön işleme tabi tutulması için burada kullanılan süreç açıklanacaktır. Bu adımda birkaç paket kullanıldı. Veri işleme ve analiz için Pandas python kütüphanesi kullanıldı. Pandas sayısal tabloları ve zaman serilerini işlemek için veri yapıları ve işlemler sunar. Twitter kullancıların konum bulması için Pycountry kütüphanesi kullanıldı. NLTK kütüphane kullanarak kök çözümleme yapıldı. Üçüncü olarak, işlenen verilerin farklı kütüphaneler ile nasıl kullanıldığı ve sonuçların karşılaştırılması anlatılacaktır.



Şekil 4.1 Uygulama Şeması

4.1 API kullanarak Twitter'dan Veri Akışı

Twitter, verilerine erişmek için farklı yöntemler sunar. Twitter ile iletişim kurmak için iki tür uygulama programlama arabirimi (API) sunulur: Twitter işlevlerine basit bir arabirim sağlayan bir REST API ve güçlü bir gerçek zamanlı API olan bir Akış API'si. Twitter'ın herkese açık verilerine erişim, REST API ile son derece sınırlıdır ve Twitter Streaming API için bu durum daha azdır. Twitter Akış API'sı, Twitter API'sinin daha yaratıcı kullanımları için tasarlanmıştır. İkisi de herkese açık tweetlere tam erişim sağlamaz. Twitter verilerine tam erişim için ticari paketler kullanılabilir. Ayrıca bir

akademik kurum veya üniversitede Yüksek Lisans öğrencileri, doktora adayları, doktora sonrası, öğretim üyeleri veya araştırma odaklı çalışanlar için başka bir Akademik Parça paketi de bulunmaktadır. Bu akademik takip paketi sadece akademik amaçlar için kullanılabilir ve bu paketi akademisyenlere belirli bir limite kadar ücretsiz olarak sağlamaktadırlar.

Twarc:

Twarc, Twitter API'sinden tweet toplamak için bir komut satırı aracı ve bir Python kütüphanedır. API istekleri sürecini basitleştirir ve ayrıca hız sınırını yönetir. Veriler tam olarak Twitter API'sinden döndürüldüğü gibidir, ancak bize kimlik doğrulamaya kolay erişim sağlar. Twarc, deneyimizde veri toplamak için kullanılan Twarc2 adlı güncellenmiş bir pakete sahiptir. Twarc'ı kullanmanın nedeni, akademik parça arama sonucuna erişimin Tweepy veya Twython gibi diğer popüler kütüphaneler tarafından sağlanmamasıdır. Twarc2'nin search_all işlevi, Twitter'ın veri havuzuna kolay erişim sağlar. Ayrıca Twitter verilerinin zaman çerçevesini sınırlayabilir ve toplamak istediğimiz tweet dilini tanımlayabiliriz. Veriler, CSV biçiminde kaydetmeden önce düzgün bir şekilde ayrıştırılması gereken satır yönelimli bir JSON nesnesi olarak gelir.

Seçilen Anahtar Kelime ve Zaman Çerçevesi:

Çalışmanın ilk aşamasında 'southchinasea' (*GüneyÇinDenizi*) kelimesini içeren Twarc ile 137.344 tweet toplandı. Güney Çin Denizi, Batı Pasifik Okyanusu'nun marjinal bir denizidir. Kuzeyde Güney Çin kıyıları, batıda Çinhindi Yarımadası, doğuda Tayvan ve kuzeybatı Filipinler adaları ve güneyde Borneo, doğu Sumatra ve Bangka Belitung Adaları ile çevrilidir. Yaklaşık 3.500.000 km2'lik bir alanı kaplamaktadır (*South China Sea - Wikipedia*, n.d.). Bu ekonomik ve jeostratejik öneme sahip muazzam bir bölgedir. Dünyanın deniz taşımacılığının üçte biri buradan geçer ve her yıl 3 trilyon ABD dolarının üzerinde ticaret taşır. Çin'in birçok komşu ülkesinin bu denizin etrafında olması onlar için en aktif faaliyet alanlarından biri olmuştur (Sinaga, 2015). Aynı zamanda bir coğrafi konum olduğu için, bu alandaki olaylar genellikle tweetlerde bu anahtar kelimeye sahiptir.

21 Nisan 2020'de “Amerika” adlı bir ABD savaş gemisi Malezya'nın tartışmalı sularına girdi. Aynı zamanda, bölgede bir Çin hükümet gemisi günlerdir arama sondajı yapan bir Malezya devlet petrol şirketi gemisini takip ediyor (Beech, 2020). Bölgede

büyük bir gerilim yarattı ve Twitter olayla ilgili çeşitli haberler ve görüşlerle dolup taşıyordu. Bu nedenle 20 Nisan 2020'den 24 Nisan 2021 tarihleri birincil veri kaynağı olarak seçildi ve bahsi geçen tweet sayıları toplandı.

Bu aşamada herhangi bir duygu analizi yapılmadı ve ham veriler tüm tweetleri içeren ayrı bir CSV dosyasına kaydedildi. Aşağıdaki tablo, doğrudan Twitter'dan topladığımız ilk alanları içermektedir.

Tablo 1: Twitter Veri Tanımlama

Alan adı	**Tanımlama**
Id	Bu Tweet için benzersiz tanımlayıcının tamsayı gösterimi
possibly_sensitive	Bu alan yalnızca bir tweet bir bağlantı içerdiğinde ortaya çıkar. Alanın anlamı tweet içeriğinin kendisiyle ilgili değildir, bunun yerine tweette yer alan URL'nin hassas içerik olarak tanımlanan içerik veya medya içerebileceğinin bir göstergesidir.
entities	Varlıklar bölümü, tweetlere dahil edilen ortak şeylerin dizilerini sağlar: hashtagler, kullanıcı sözleri, bağlantılar, hisse senedi kayıtları (semboller), Twitter anketleri ve ekli medya.
reply_settings	Bu tweete kimlerin yanıt verebileceğini gösterir.
conversation_id	Konuşmanın orijinal tweetinin tweet kimliği
author_id	Bu kullanıcının benzersiz tanımlayıcısı
public_metrics	Talep anındaki tweet için katılım metrikleri.

lang	Twitter tarafından algılanırsa tweetin dili. BCP47 dil etiketi olarak döndürüldü.
source	Kullanıcının tweetlediği uygulamanın adı.
created_at	Tweetin oluşturulma zamanı.
text	Tweetin içeriği.
author	Tweet yazarının kullanıcı nesnesi
referenced_tweets	Bu tweetin atıfta bulunduğu tweetlerin listesi.
Attachments	Bu tweette bulunan eklerin türünü (varsa) belirtir.
Geo	Kullanıcı, belirtmişse, bu tweette etiketlenen konumla ilgili ayrıntıları içerir.

4.2 Ön İşleme

Bölüm 3.1'de tartışıldığı gibi, Twitter'ı analiz etmek birçok zorluğu beraberinde getirir. Twitter verileri, içerdiği metin, emojiler, resimler, videolar, web bağlantıları, yazım hataları vb. gibi doğası gereği dağınıktır. Bir analiz yapabilmek için verilerin temizlenmesi gerekir.

4.2.1 Veri temizliği

Verileri temizlerken ilk eylem, her tweete benzersiz bir anahtar atamaktı. Bu analizde anahtar, kullanıcı "*id*" ve tweetin "*created_at*" tarihinin bir kombinasyonudur. İlk aşama, API'den döndürülen 'text' ve 'author' alanı olmayan tweetlerin kaldırılmasıydı. Böylece, Pandas 'dropna' işleviyle bu satırları veri kümesinden çıkarıldı.

İkinci adımda, veri kümesinden dışa aktarılan bazı kesin verilere ihtiyaç duyuldu. "author" ve "public_metrics" alanı bir JSON dizesiydi. Bu nedenle, 'author' alanından 'verified' ve 'location' bilgilerini ve 'public_metrics' alanından 'retweet_count' ve 'like_count' bilgilerini çıkarmanın bir yolunu bulmanın bir yolunu bulmak gerekiyordu. Bunun için JSON dizesini nesne benzeri bir sözdizimine dönüştürmek gerekiyordu. Python'un 'ast' modül, dizeyi JSON nesnesi olarak değerlendirmek ve karşılık gelen

alanları dönüştürülmüş değerleriyle değiştirmek için kullanıldı (*32.2. Ast — Abstract Syntax Trees — Python 3.6.14 Documentation*, n.d.). Dizeyi bir nesneye dönüştürdükten sonra, '*author*' ve '*public_metric*' alanından gerekli verilerimizi almak çok kolaydı.

Bu adımda, kullanıcıların konumlarını bulması gerekiyordu. Twitter tweetin gönderildiği yerden tam olarak coğrafi konumunu vermez, bunun yerine Twitter kullanıcısının konum ayarlarında bahsettiği konumu verir. Ancak bazen bir kullanıcı tam konumu belirtmez veya gizlilik nedeniyle yanlış konumu yazar. Her ne kadar insanların iyi bir yüzdesi konumlarını doğru bir şekilde belirtse de, bir Twitter kullanıcısının konumunu kullanıcı profilinden tam olarak belirlemek de zordur. Bu nedenle, tam kullanıcı ülkesini çıkarmak için bu verilere güvenmesi gerekiyor.

Elimizde kullanıcı konum verileri olmasına rağmen, herkes ikamet ettiği ülkeden bahsetmiyor. Bazı kullanıcılar Bölge veya eyalet adını veya eyalet adının bir kısmını yazar. Bu nedenle, bu konum alanından ülke adını bulmak başka bir zorluktur. Google'da arama yapmak ve söz konusu konumun menşe ülkesini bulmak mümkündür. Ancak 130 binden fazla kayıttan oluşan bir veri kümesi için çok zor ve zaman alıcı bir iştir. Görevi basitleştirmek için '*pycountry*' adlı bir Python kütüphane kullanıldı (Theune, 2016). Girdi olarak herhangi bir konumu alır ve o konumun ülkesini çıkarır. Bu kütüphanenin bir diğer önemli özelliği de girdi olarak bulanık konum alabilmesidir. Yani örneğin birisi bulunduğu yerin biyografisine "İtsanbul" yazarsa, kütüphane metni analiz edebilir ve girdinin aslında "İstanbul" olduğunu anlayabilir ve sonuç olarak Türkiye çıktısını alabilir. Bu kütüphaneyi mümkün olduğunca çok ülke konumu bulabildi. Bu işlemden sonra ülke konumunu bulamadığı her satırı bırakıldı. Bu operasyondan sonra elimizde sadece 22050 civarında tweet vardır.

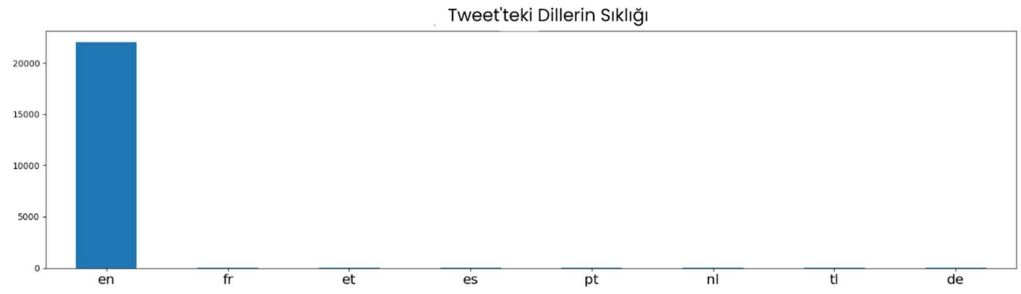
Bir sonraki adımda, artık gerekli olmayan sütun alanları bırakılır. Bu işlemden sonra aşağıdaki alanlar kalır.

Tablo 2: Güncellenmiş Twitter Veri Sütunları

Alan Adı	**Tanımlama**
id	Bu tweet için benzersiz tanımlayıcısı.
author_id	Bu kullanıcının benzersiz tanımlayıcısı

possibly_sensitive	Bu alan yalnızca bir tweet bir bağlantı içerdiğinde ortaya çıkar. Alanın anlamı tweet içeriğinin kendisiyle ilgili değildir, bunun yerine tweette yer alan URL'nin hassas içerik olarak tanımlanan içerik veya medya içerebileceğinin bir göstergesidir.
location	Twitter kullanıcısının menşe ülkesi
verified_user	Doğru olduğunda, kullanıcının doğrulanmış bir hesabı olduğunu gösterir
lang	Twitter tarafından algılanırsa Tweetin dili. BCP47 dil etiketi olarak döndürüldü.
retweet_count	Bu tweetin retweetlenme sayısı.
created_at	Tweetin oluşturulma zamanı.
Text	Tweetin içeriği.
like_count	Bu tweetin Twitter kullanıcıları tarafından yaklaşık olarak kaç kez beğenildiğini gösterir

Sadece İngilizce yazılmış tweetleri toplamaya çalışsak da yine de diğer dillerden tweetlerimiz vardı. Dil listesini bu şekilden bulabiliriz



Şekil 4.2.1.1 Tweetteki Dillerin Sıklığı

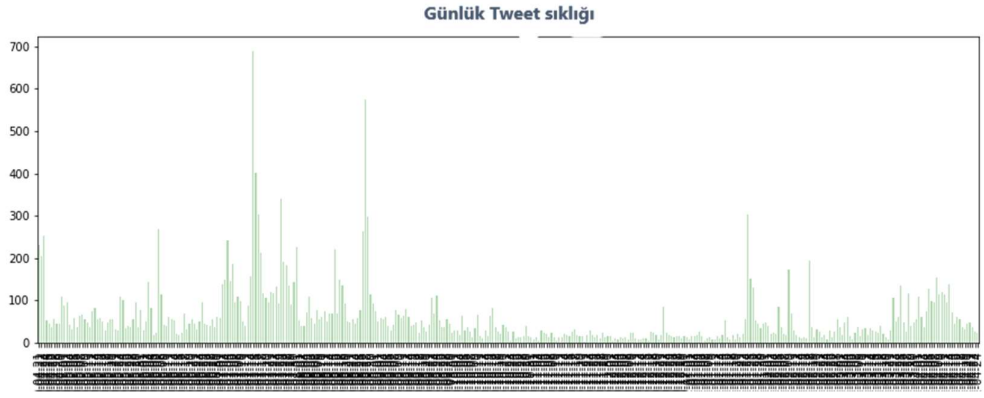
Diğer dillerden de az sayıda tweetimiz var gibi görünüyor. Dolayısıyla bu adımda sadece İngilizce yazılan tweetleri tutacağız. Bu son aşamadan sonra, şimdi analizi yapmak için 22034 tweetimiz var.

Tablo 3: Toplam Tweet Sayısı

	Tweet Sayısı
İlk Koleksiyon	137344
Menşe ülke toplandıktan sonra	22050
İngilizce olmayan tweetleri attıktan sonra	22015

4.2.2 İlk analiz

Matplotlib kullanarak günlük tweet sıklığını gözlemleyebiliriz. Grafikten, zaman geçtikçe tweet sayısının azaldığı kolayca görülebilir. Çünkü sosyal medyadaki dikkat süremiz çok düşüktür. (Firth ve ark., 2020).



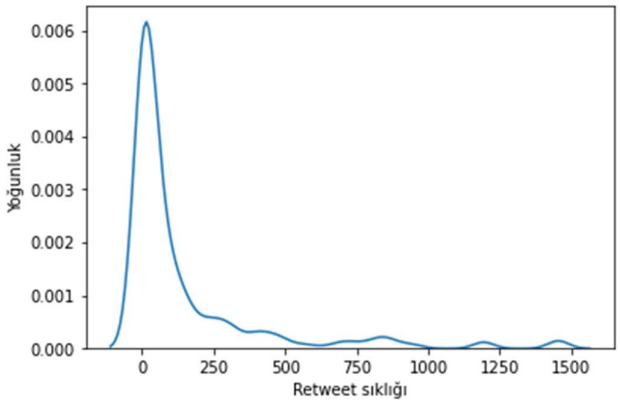
Şekil 4.2.2.1 Günlük Tweet sıklığı

Başlangıçta öğrenebileceğimiz başka bir şey de en popüler tweetleri aramaktır. Veri setinde aynı tweetlerin olması çok yaygın olduğu için, aynı metinden aynı tweetleri filtreledi. Ardından, *'retweet_count*' alanını kullanarak en popüler tweeti bulabiliriz. Veri kümemizdeki en popüler 5 tweet şekil 4'te görebiliriz.

Tablo 4: En popüler 5 tweet.

Tweet	Tweet Konumu	Retweet Sayısı
RT @SenRubioPress: Today's visa sanctions & entity listings are good first steps to hold #Beijing accountable for its aggression, environme…	Hong Kong	1454
RT @SolomonYue: 造反了! Rebellion! Let's put it in the "killing the chicken to scare the monkeys" context. It's monkeys' rebellion. #CCPChina…	Hong Kong	1193
RT @chinfo: Who wants to tell them? https://t.co/zuVnaYfR5t	India	945
RT @SenateForeign: Reports of a Chinese Coast Guard vessel sinking a Vietnamese fishing vessel and #China's other activities on illegally r…	Hong Kong	888
RT @JanJekielek: Big news: Australia rejects all #CCP / #China claims in the #SouthChinaSea.https://t.co/NfkgezF7M8	Hong Kong	881

Ayrıca '*retweet_count*' yoğunluğunu da çizebiliriz. Şekil 4.2.2.2 gösterdiği gibi genel olarak çöğü tweetler o kadar retweet olmadı. Ama 1454 kere retweet olan bir tweet de bu veri kümesinde var.



Şekil 4.2.2.2 Retweet Sayısı Yoğunluğu

4.2.3 Tweet ön-işlem

Bu analizin amacı için, bir retweet ve bir yinelenen tweet mutlaka aynı şey değildir. Yinelenen tweet, başka bir tweet ile aynı anahtara sahip olan ve bir botun aynı anda tweetleri gönderebilmesi nedeniyle oluşmuş olabilir. Bir retweet yolu, birden çok farklı kullanıcı arasında paylaşılan aynı tweettir ve her örnek için benzersiz bir anahtarla sonuçlanır. Retweetler veri kümesinde tutuldu, çünkü her biri birden fazla kullanıcının orijinal tweeti kabul ettiğini veya bu tweeti toplulukla tekrar paylaşacak kadar beğendiğini gösteriyor.

Bu analiz için aşağıdaki temizleme eylemleri gerçekleştirilmiştir:

- Her şeyi küçük harfe dönüştür,
- @mentions ile başlayan tüm sözleri kaldırın,
- Tüm web bağlantılarını kaldırılır, çünkü bağlantılardan fark edilebilir bir şey elde edilememelidir,
- Tüm #hashtaglerden "#" sembolünü kaldırın, ancak #hashtag ile ilişkili kelimeleri saklayın,
- Tüm tweetlerden "RT" veya retweet kimliğini kaldırın. Gerçek sembol herhangi bir faydalı bilgi eklemez. Ancak, daha önce de belirtildiği gibi, çoklu retweetlerin kümülatif etkisi dahil edildi,
- Tüm noktalama işaretlerini kaldırın. Noktalama işaretinin duygu sınıflandırmasında değeri yoktur,
- Tüm emojileri kaldırılır. Bu analizin amacı için emojiler dahil edilmeyecektir,
- Tüm durdurma sözcüklerini kaldırın. Durdurma sözcükleri, duygu sınıflandırmasında herhangi bir değer sağlamayan sözcüklerdir. Durma sözcükleri, "the", "of", "to" gibi yaygın sözcüklerdir,
- Kök Çözümleme - çekimli kelimeleri o kelimenin köküne indirger,

Farklı kütüphaneler, kök çözümleme için biraz farklı sonuçlar verir. Bu analiz için, kök çözümleme için NLTK kullanıldı. Temizlikten sonra ilk 5 tweet aşağıdaki gibi görünüyor:

Tablo 5: NLTK ile Temiz Metin

Metin	NLTK ile Temiz Metin
#China's most advanced amphibious assault ship expected to be deployed in disputed #SouthChinaSea\n\nThe newly commissioned Type 075 vessel, named the #Hainan, can carry dozens of helicopters and hundreds of troops\nhttps://t.co/cQQNlQI4kB	china advanced amphibious assault ship expect deployed dispute southchinasea newly commission type vessel name hainan carry dozen helicopter hundred troop
RT @USNavy: Making memories while serving side by side in the #SouthChinaSea.\n\nHard at work and underway, learn more about how a #USNavy Ch…	make memory serve southchinasea hard work underway learn usnavy
No com #Squawk 7600 detected #Airbus A330-343E 🛩 ascending over #SouthChinaSea at 37000ft at 436.6mph heading N with tail B-6102 #CCA614 #AirChina 780dfe #China #ADSB 🔔 #RaspberryPi https://t.co/24xbi5XRim	com squawk detect airbus ascend southchinasea head tail airchina china adsb raspberrypi
RT @ASHIM217: .@asthana_shashi @InsightGL @desertfox61I @BrigUsman @rspathania\nWhile we continue to fight #COVID19. #China continues to bui…	continue fight china continue
RT @IndoPac_Info: Xi commissions 3 ships for the #Chinese Navy\n\nA nuclear-powered submarine, Type 075 amphibious assault ship and a Type 05…	commission ship chinese navy submarine type amphibious assault ship type

4.2.4 Sınıflandırma yöntemleri

Twitter verilerinin duygu analizinin bir sonraki zorluğu, uygun bir sınıflandırma yöntemi seçmektir. Sınıflandırıcıya yönlendirmeden önce tweete birçok husus uygulanmalıdır. Duygu analizine başlamanın ilk adımı bir sınıflandırıcı seçmektir. Bu, her tweeti tutumuna göre sınıflandırabilir. Bu çalışmada, her tweet olumlu, olumsuz ve nötr olmak üzere üç kategoriden birine atanmıştır. Daha doğru sonuçlar elde etmek için aynı veri seti üzerinde üç farklı sınıflandırma ve duygusallık yöntemi uygulanmaktadır. Birincisi, "VADER" python kütüphanesi kullanmak ve ilinci olarak "Flair" python kütüphanesi kullanmaktadır.

VADER:

VADER (Valence Aware Dictionary for Sentiment Reasoning / Duygu Akıl Yürütme için Değer Duyarlı Sözlük), duygunun hem polaritesine (pozitif/negatif) hem de yoğunluğuna (gücüne) duyarlı olan metin duygu analizi için kullanılan bir modeldir. NLTK paketinde mevcuttur ve doğrudan etiketlenmemiş metin verilerine uygulanabilir.

VADER duygusal analizi, sözcüksel özellikleri duygu puanları olarak bilinen duygu yoğunluklarıyla eşleyen bir sözlüğe dayanır. Bir metnin duygu puanı, metindeki her kelimenin yoğunluğunun toplanmasıyla elde edilebilir.

Örneğin- 'Aşk', 'keyfini çıkarın', 'mutlu', 'beğen' gibi kelimelerin tümü olumlu bir duygu taşır. Ayrıca VADER, olumsuz bir ifade olarak "sevmedi" gibi bu kelimelerin temel bağlamını anlayacak kadar zekidir. Ayrıca, "ENJOY" gibi büyük harf ve noktalama işaretlerinin vurgusunu da anlar (Gilbert ve Hutto, 2014).

Flair:

Flair, tüm sözcük yerleştirmelerinin yanı sıra isteğe bağlı yerleştirme kombinasyonları için basit, birleşik bir arabirimdir. Bu arayüz, araştırmacıların daha sonra herhangi bir ek mühendislik çabası olmadan herhangi bir kelime yerleştirme türünü kullanabilecek tek bir model mimarisi oluşturmasına olanak tanır.

Deney kurma ve yürütme sürecini daha da basitleştirmek için Flair, standart NLP araştırma veri kümelerini indirmek ve bunları çerçeve için veri yapılarına okumak için kolaylık yöntemleri içerir. Ayrıca, tipik eğitim ve test iş akışlarını kolaylaştırmak için model eğitimi ve hiperparametre seçim rutinlerini içerir. Ek olarak, Flair ayrıca, kullanıcıların önceden eğitilmiş modelleri metinlerine uygulamalarına olanak tanıyan, büyüyen bir önceden eğitilmiş model listesiyle birlikte gelir.

Ayrıca bazı görevler için çok dilli modeller içerir. Bunlar, birden çok dilde metin için etiketleri tahmin edebilen "bir model, birçok dil" modelleridir. Örneğin, Flair, 12 dilde metin için evrensel PoS etiketlerini tahmin eden çok dilli konuşma bölümü etiketleme modelleri içerir.

5. SONUÇ VE ANALİZİ

Hazırlanan veri seti ve seçilen paketler ile tweetlere tek tek duygu analizi uygulanarak çeşitli grafiklerle sonuçlar hazırlanacaktır. Önce VADER ve Flair ile duygu analizi yapıp bunların sonuçları gösterildi. Son aşamada ise bu iki kütüphanenin sonuçları karşılaştırıldı.

5.1 VADER ile Duygu Analizi

VADER'in github'ına göre, VADER "Birden fazla bağımsız insan yargıç tarafından deneysel olarak doğrulanmıştır, VADER, özellikle mikroblog benzeri bağlamlara uygun bir "altın standart" duygu sözlüğü içerir." Vader önceden eğitilmiş bir modeldir. VADER şöyle bir çıktı veriyor:

{'neg': 0.0, 'neu': 0.436, 'pos': 0.564, 'compound': 0.3802}

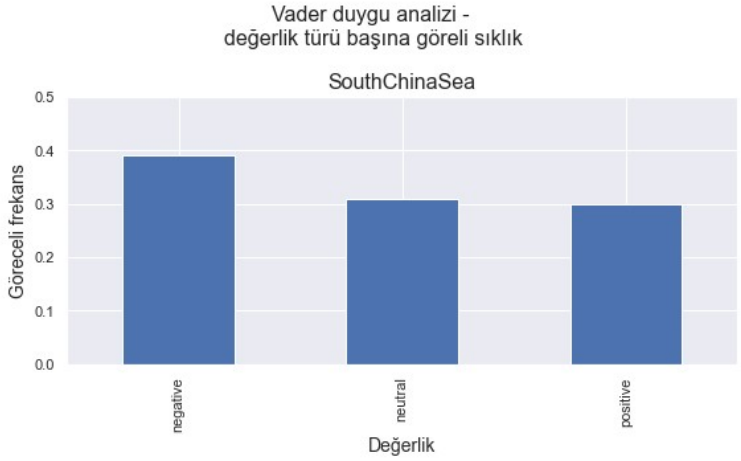
Negatif, nötr ve pozitif 0 ile 1 arasındaki puanlardır. Bileşik değer metnin genel duygusunu yansıtır. Negatif, nötr ve pozitif değerlerine göre hesaplanır. -1 (maksimum negatiflik) ile 1 (maksimum pozitiflik) arasında değişir.

VADER modeli ile Sentiment Analysis'i çalıştırdıktan sonra puanı `vader_sentiment` kolonuna kaydedildi. Ayrıca önceden eğitilmiş bir model sağladığı için, bu modelle sonucumuzu analiz edebilir ve duygumuzun yaklaşık doğruluk puanını de bulabilir. Tweetleri VADER'a göndermeden önce temizlememizin bir önemi var mı? VADER'ın kendisi yeterince iyi bir temizlik yapıyor mu? Her iki yaklaşımı da kullanarak tweetleri pozitif / nötr / negatif olarak sınıflandırarak ve ardından iki yöntemle elde edilen etiketlerin doğruluk_skoruna bakarak bu soruya cevap verebilir.

VADER'ın kendi kök çözümleme metodolojisi vardır. Bu, ham verilerle ön işleme gerek kalmadan çalışabileceği anlamına gelir. Ama tweetlerimiz önceden hazır hale geldi. Artık hem işlenmiş hem de ham veriler ile sonuç arasındaki farkı gözlemlemek için kullanılabilir. Scikit-learn'in doğruluk_metrik fonksyonu kullanılarak, duygu puanı arasındaki benzerlik bulundu.

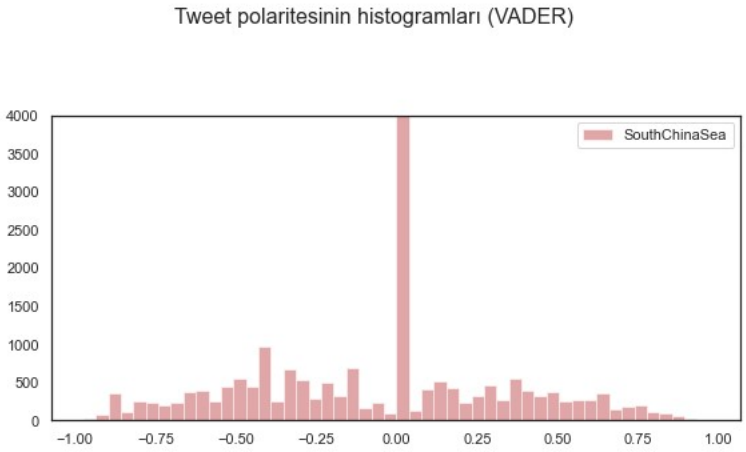
SouthChinaSea için temizlik vs vader tweet temizliği için doğruluk puanı: 0.9063

Bu, ham veya temizlenmiş verileri besleme kararının biraz düşünülmesi gerektiği anlamına gelir. Şekil 5.1.1'de değerlik türü başına göreli frekansı gösterildi.



Şekil 5.1.1 VADER Duygu Analizi - Göreceli Frekans

Şekil 5.1.1, tweetlerin çoğunun VADER kullanılarak olumsuz olarak sınıflandırıldığını göstermektedir. Ancak şekil 5.1.2'den çok sayıda tweet'in de tarafsız olarak sınıflandırıldığı açıktır.



Şekil 5.1.2 Tweet polaritesinin histogram.

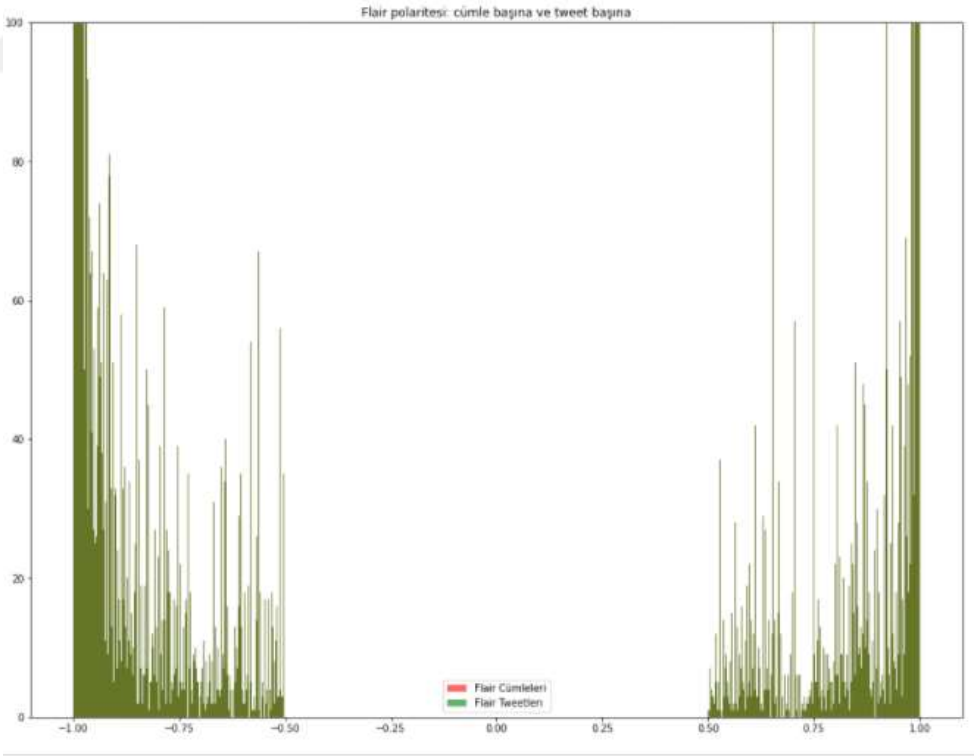
5.2 Flair ile Duygu Analizi

Bu adımda aşağıdakileri yöntemleri karşılaştırıldı:

- VADER,
- Cümle başına Flair,
- Tweet başına Flair

Veri kümesi etiketlenmediğinden, tahminleri bazı 'temel gerçeklerle' karşılaştırmanın bir yolu yoktur. Bu nedenle, her bir algoritmanın tahminlerini diğer ikisinden gelen tahminlerle karşılaştırmak gerekir.

Ancak bundan önce, Flair'i kullanmanın iki yolu için duygu etiketlemesini araştırmak ve karşılaştırmak gerekir. Öncelikle tweetin tamamını mı tahmin etmeliyiz yoksa cümlelere bölüp her bir cümle için tahminde bulunup ardından bir ortalama mı almalıyız bunu belirlemek gerekiyor. Resmi Flair belgelerinde bunun için net bir gösterge bulunmadığından, her iki yöntemin de araştırılması gerekmektedir. Ayrıca, Flair'in yeni sonucu, veri kümesinin yeni sütunlarına kaydedildi.

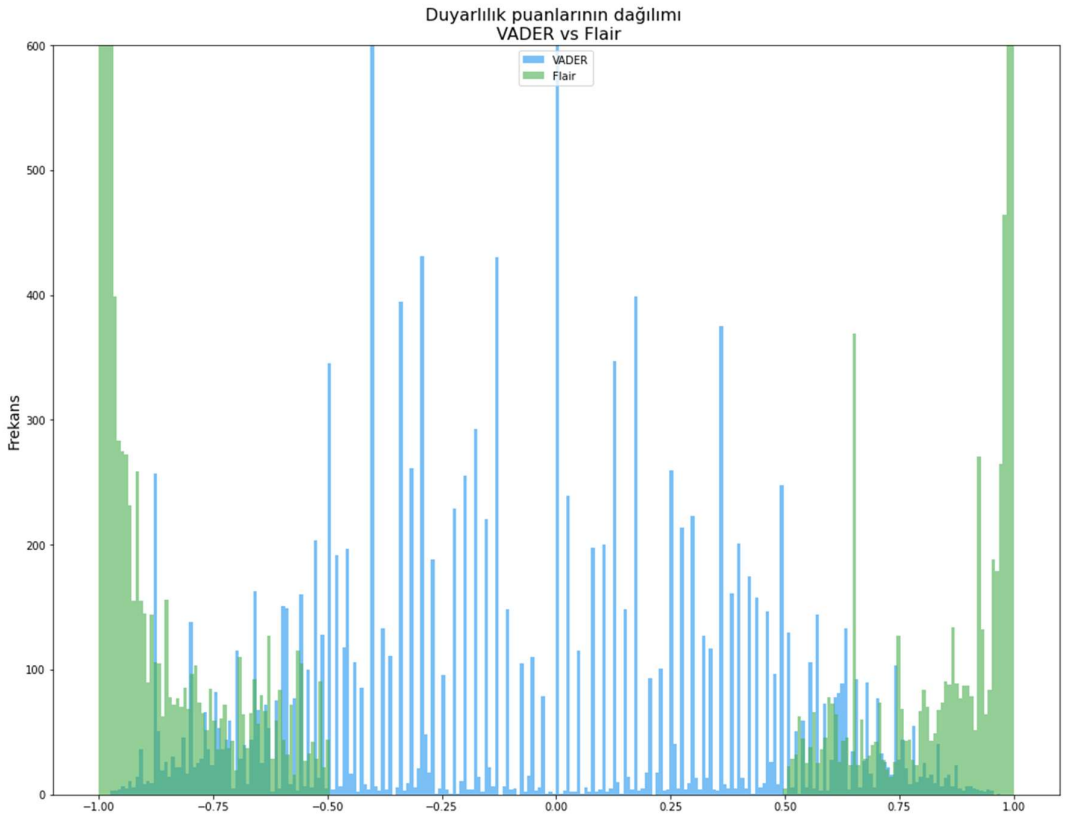


Şekil 5.2.1 Flair Polarite: Cümle başına vs Tweet Başına

Şekil 5.2.1'de, Flair ile bir tweetin polaritesini hesaplamak için kullanılan iki yöntemin tamamen aynı sonuçları verdiğini göstermektedir.

5.3 İki Duygu Analizi Kütüphanenin Karşılaştırılması:

İlk etapta Duygu Puanlarının dağıtımını gerçekleştirebiliriz.



Şekil 5.3.1 VADER ve Flair tahminleri arasındaki anlaşma

VADER ve Flair tarafından üretilen duygu dağılımı grafiğinden gözlemler:

VADER genel olarak:

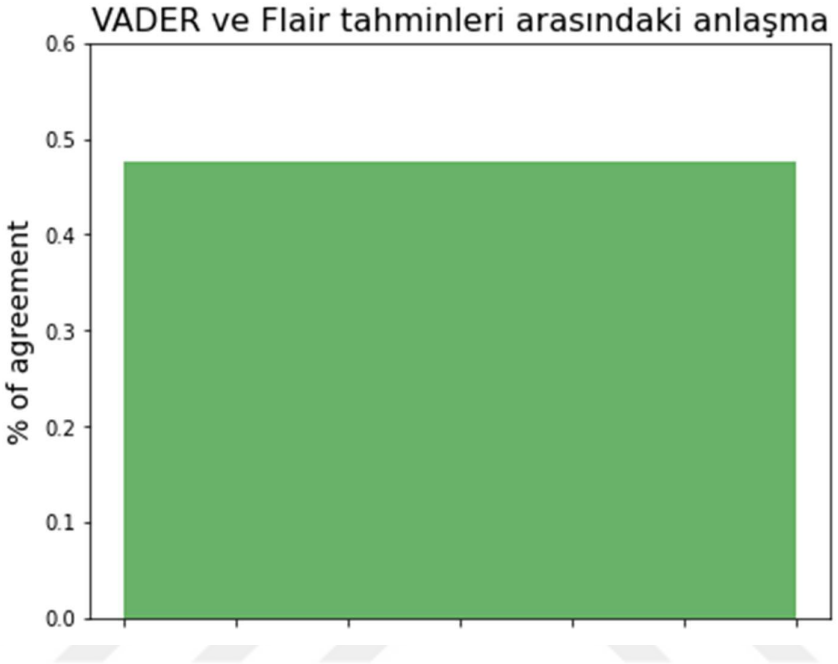
- Verilerin çoğunu nötr olarak sınıflandırın,
- VADER iki modlu bir dağılıma sahiptir.

Diğer yandan Flair:

- Nötr tercihi yoktur,
- VADER ile karşılaştırıldığında son derece polarize

Üçünün, tahmin edilen duygusallık değerlerinin ne kadar aşırı olduğu konusunda farklı olduğunu biliyoruz. Ancak, yoğunluk konusunda hemfikir olup olmadıklarına bakılmaksızın, bu duygunun yönü üzerinde anlaşıyorlar mı? Yani, TexBlob zayıf bir duygu ve Flair güçlü bir duygu öngörüyorsa, ikisi de olumsuz mu yoksa her ikisi de

olumlu mu? Veya biri -.2 (zayıf negatif) ve diğeri .9 (güçlü pozitif) tahmin ediyor mu? Yukarıdaki soruyu cevaplamak için, bir tweetin duygusallığını sınıflandırırken bu algoritmaların birbiriyle uyuşma yüzdesini görelim.



Şekil 5.3.2 VADER ve Flair tahminleri arasındaki anlaşma

Şekil 5.3.2, VADER ve Flair'in genel sonuç üzerinde %48'in üzerinde anlaşmaya sahip olduğunu göstermektedir. Bu, tweetlerin %48'inde hem Flair hem de VADER'ın aynı sonuçları sağladığı anlamına geliyor.

5.4 Veri Keşfi

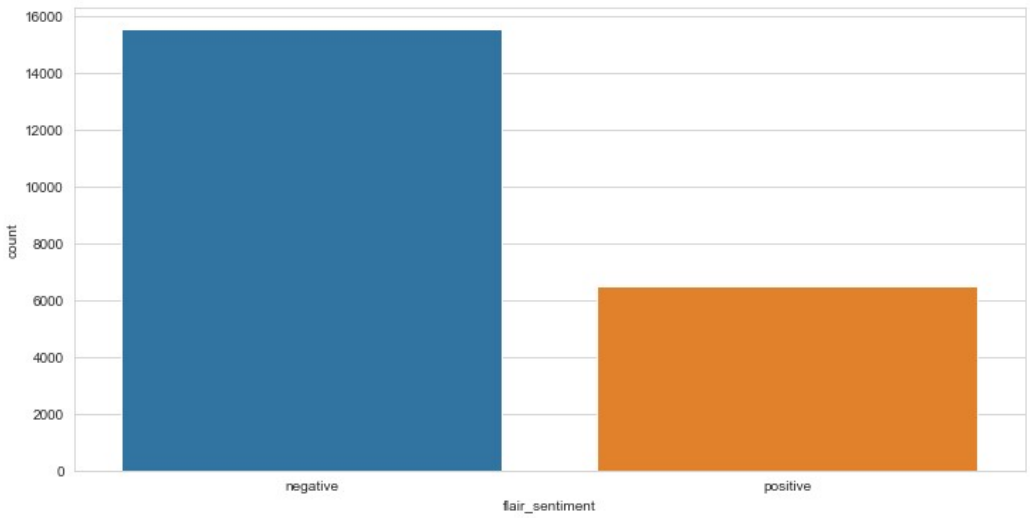
Duygu analizinden sonra, verilerin içinde ne olduğuna ve ayrıca ülke adlarıyla gruplandırarak bulabileceğimiz bazı korelasyonların neler olduğuna ilişkin ayrıntılara girildi.

Verilerden bulunan bazı veri duygusallığı:

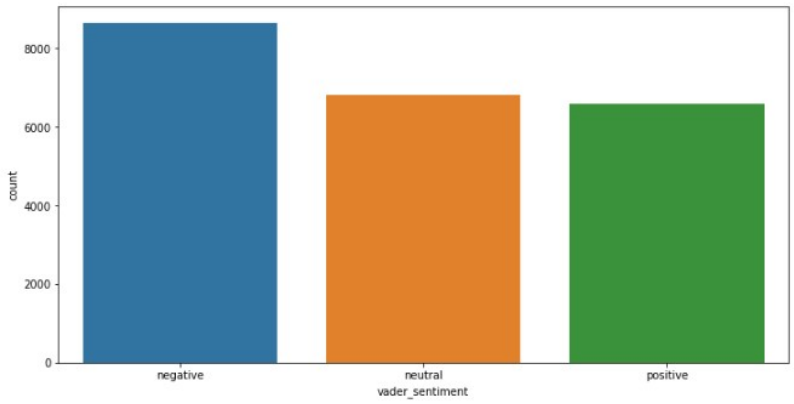
Tablo 4: Duygu dağılımı

Duygu	**Toplam Sayısı**
VADER – Pozitif	6583
VADER – Nötr	6795
VADER – Negatif	8637
Flair – Pozitif	6471
Flair – Negatif	15544

Bu verileri Çubuk Grafikte de görselleştirebiliriz



Şekil 5.4.1 Flair ile Duygu Dağılımı Sayım grafiği



Şekil 5.4.2 VADER ile Duygu Dağılımı Sayım grafiği

Şekil 5.4.1 ve 5.4.2, her iki algoritmanın da tweetlerin çoğunu negatif olarak tanımladığını açıkça göstermektedir. Özellikle Flair, tweetleri sınıflandırırken çok kutuplaştı ve Flair analizinden tarafsız tweet çıkmadı.

Artık tweetlerden en yaygın 30 kelimeyi arayabiliriz.

Şekil 5.4.3 Tüm Tweetlerde En Yaygın 30 Kelime

Şimdi hangi kelimelerin farklı senaryolarda kullanıldığına dair bazı derinlemesine veriler aşağıdaki gibidir

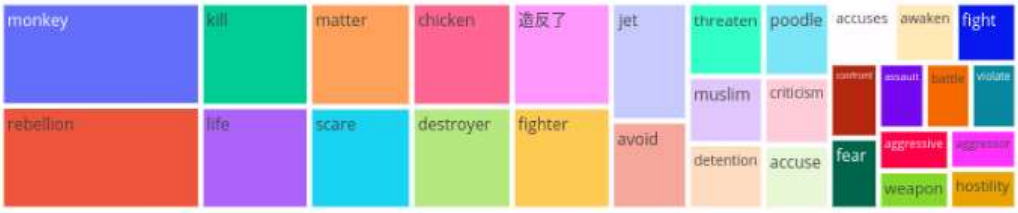
Şekil 5.4.4 Olumlu Tweetlerde En Çok Kullanılan 70 Kelime (Flair)

Şekil 5.4.5 Olumlu Tweetlerde En Çok Kullanılan 70 Kelime (VADER)

Şekil 5.4.6 Olumsuz Tweetlerde En Sık Kullanılan 70 Kelime (Flair)

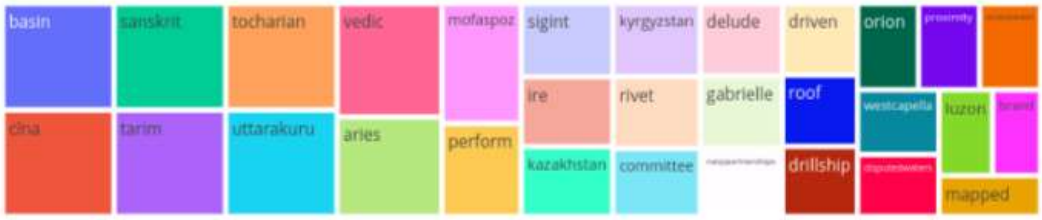
Şekil 5.4.7 Olumsuz Tweetlerde En Sık Kullanılan 70 Kelime (Vader)

Şekil 5.4.8 Tarafsız Tweetlerde En Sık Kullanılan 70 Kelime (Vader)

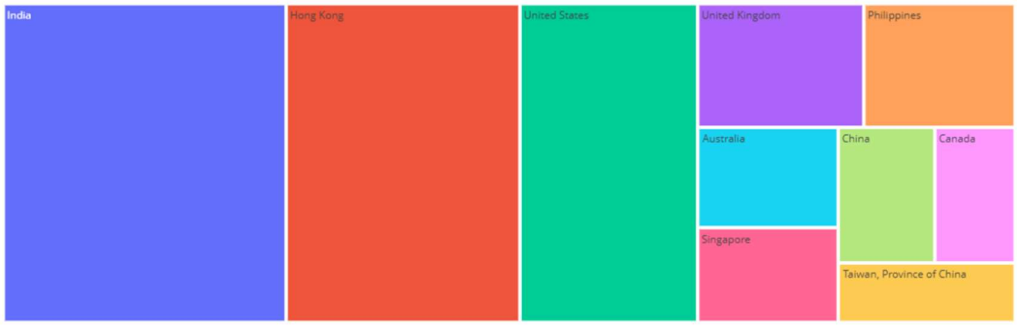


Şekil 5.4.9 Negatif tweetlerdeki ilk 30 benzersiz kelime (Vader)

Şekil 5.4.10 Olumlu tweetlerdeki ilk 30 benzersiz kelime (Vader)



Şekil 5.4.11 Tarafsız tweetlerdeki en iyi 30 benzersiz kelime (Vader)



Şekil 5.4.12 En benzersiz olumsuz tweetler 10 menşe ülkesi (Flair)

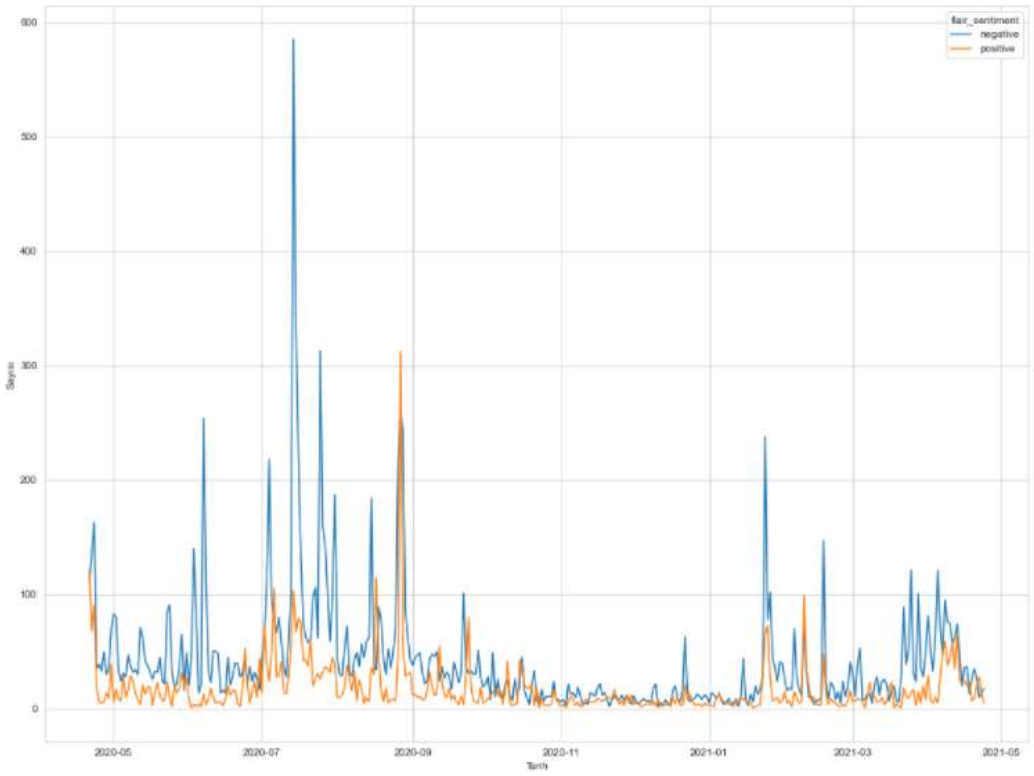
Şekil 5.4.13 En benzersiz olumsuz tweetler 10 menşe ülkesi (Vader)

Şekil 5.4.14 En benzersiz olumlu tweetler 10 menşe ülke (Flair)

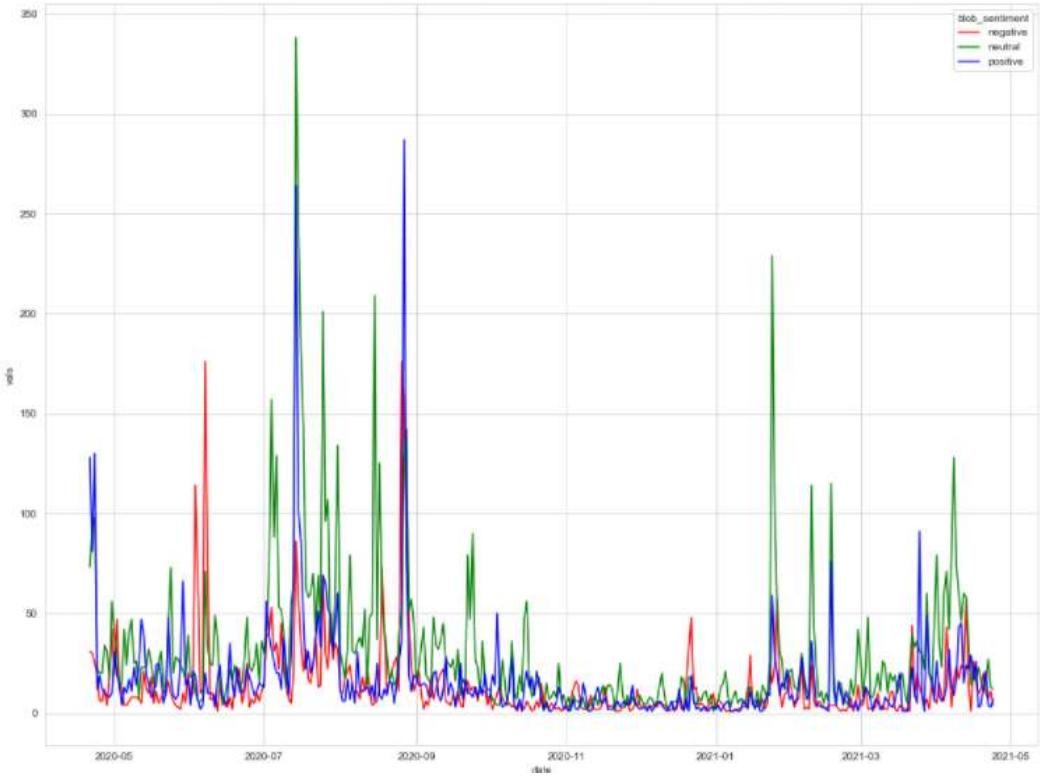
Şekil 5.4.15 En benzersiz olumlu tweetler 10 menşe ülke (Vader)

Şekil 5.4.16 En benzersiz tarafsız tweetler 10 menşe ülke (Vader)

Şimdi, zamanın ilerlemesi ile duygunun nasıl değiştiği gözlemlenebilir. Flair analizinden, olumsuz duygunun genel olarak olumlu duyguya açıkça hükmettiği açıktır. Ancak Vader analizi bazı farklı sonuçlar gösterdi. Hem olumsuz hem de olumlu duyguların zaman boyunca belirsiz olduğunu ifade eder.



Şekil 5.4.17 Twıtter Duygularının Zaman Temel Analizi:- Flair



Şekil 5.4.18 Twıtter Duygularının Zaman Temel Analizi:- VADER

Şekil 5.4.17 ve 5.4.18'den Temmuz 2020 ile Ağustos 2020 arasında bazı ani artışlar olduğunu görebiliriz. Bunun nedeni hem ABD hem de Çin'in bu bölgedeki bazı faaliyetleri tırmandırması (*China, U.S. Escalate Forces, Threats in South China Sea | World Report | US News*, n.d.). 27 Ağustos'ta ABD, Güney Çin Denizi'ndeki inşaat ve askeri çabalarla ilgili 24 Çinli şirketi ve kişiyi kara listeye aldı (*South China Sea: US Unveils First Sanctions Linked to Militarisation | South China Sea | The Guardian*, n.d.).

Artık ülke bazlı analizlere bakabiliriz. Bunun için, her ülkeden toplam tweet sayısını bulacağız, o ülkenin toplam duygusallığını bulacağız, ayrıca toplam beğeni sayısı ve toplam retweet sayısının toplamına bakacağız. Ayrıca onlara ortalama kutupluluğa dayalı genel bir duygu atayacağız. Bu işlemden sonra ilk 5 ülke için şöyle bir tablo buluyoruz.

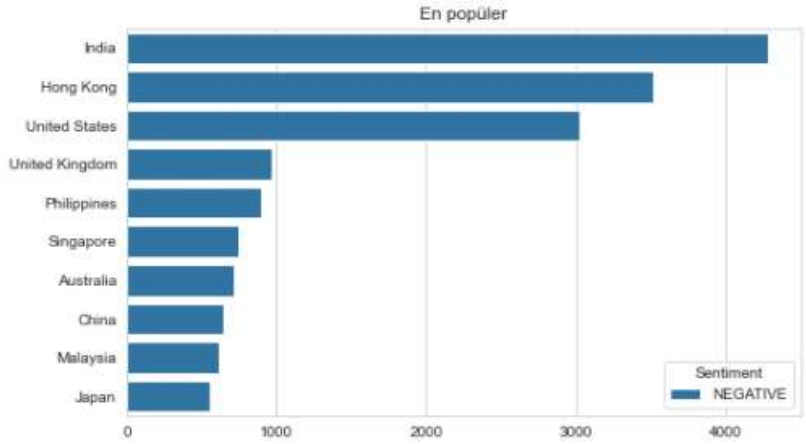
	country	Frequency	Avg_Polarity	Total_Likes	Total_Retweets	Sentiment
0	India	4289	-0.458990	4857	577228	NEGATIVE
1	Hong Kong	3517	-0.484585	1358	1279349	NEGATIVE
2	United States	3027	-0.296345	2175	431304	NEGATIVE
3	United Kingdom	965	-0.469735	400	71895	NEGATIVE
4	Philippines	894	-0.448965	612	56819	NEGATIVE

Şekil 5.4.19 En çok tweetlenen 5 ülkenin Ortalama Kutupluluğu (Flair)

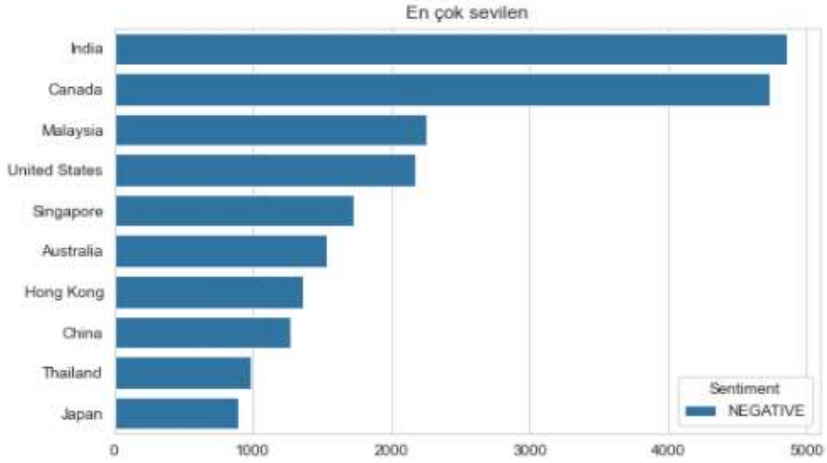
	country	Frequency	vader_polarity	Total_Likes	Total_Retweets	Sentiment
India	4289	-0.055389	4857	577228	NEGATIVE
Hong Kong	3517	-0.111378	1358	1279349	NEGATIVE
United States	3027	-0.052435	2175	431304	NEGATIVE
United Kingdom	965	-0.082971	400	71895	NEGATIVE
Philippines	894	-0.044801	612	56819	NEGATIVE

Şekil 5.4.20 1En çok tweetlenen 5 ülkenin Ortalama Kutupluluğu (Vader)

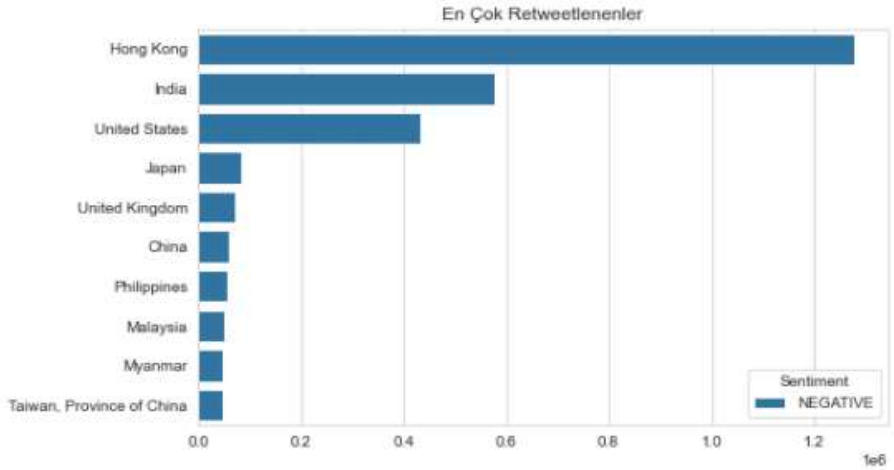
Bundan sonra en popüler, en çok beğenilen ve en çok retweet edilen tweetlerin menşe ülkesini görselleştirmek artık mümkündür.



Şekil 5.4.21 En popüler (Flair)



Şekil 5.4.22 En çok sevilen ülkeleri (Flair)



Şekil 5.4.23 En çok retweet edilen (Flair)

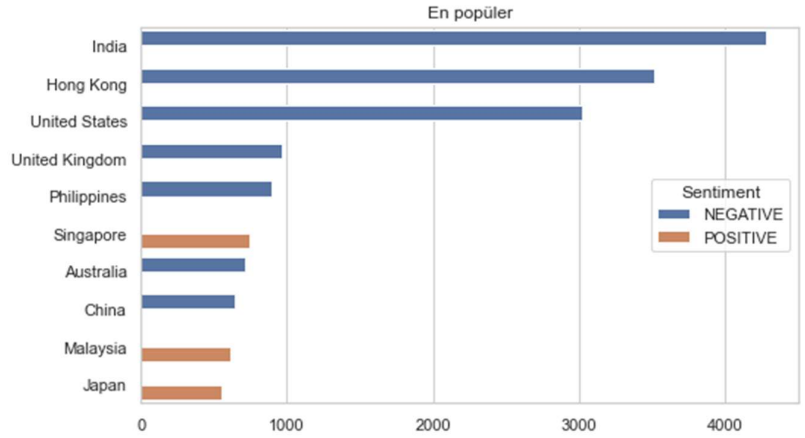


Figure 5.4.24 En Popüler (VADER)

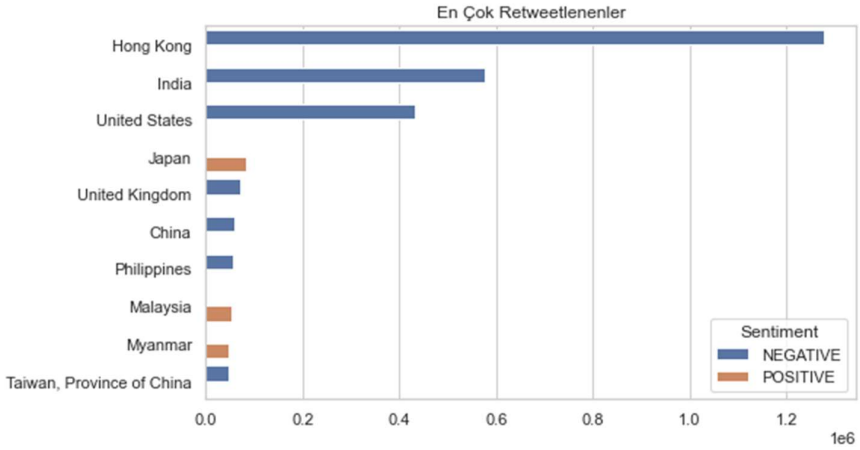


Figure 5.4.25 En Çok retweetlenenler (VADER)

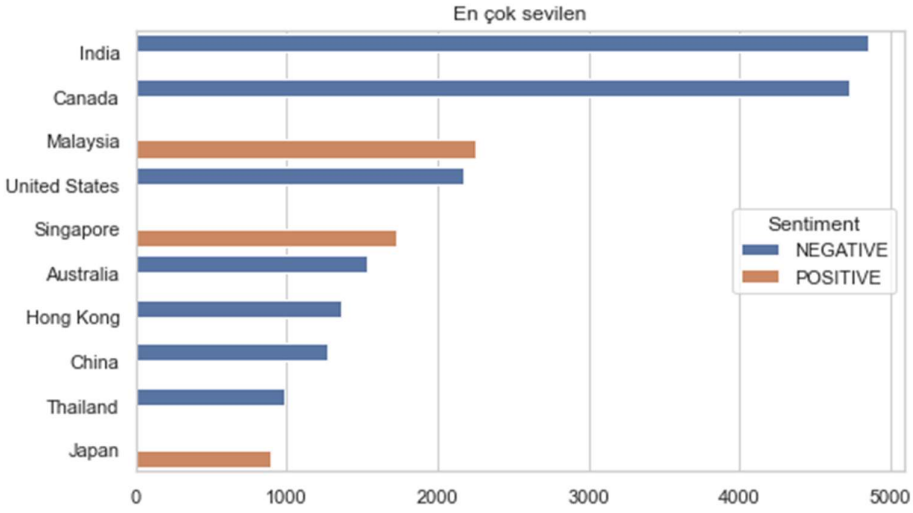


Figure 5.4.26 En Çok Sevilen

VADER ve Flair sonuçları arasında fark olsa da, Flair algoritmasının sonucunun ne kadar polarize olduğunu bir kez daha göstermektedir. VADER bu durumda yine çok daha nötr çıktısını gösterir.

Analizden, ilk 10 listesine sadece Çin'in komşu ülkelerinin değil, ABD ve İngiltere'nin de geldiği açıktır. Her iki ülke de Çin ve Güney Çin denizi ile sınır paylaşmıyor, ancak Twitter kullanıcıları bu uzak doğu bölgesinde faaliyette bulundu.

5.5 Sonuç

Atlantik Konseyi'nin yakın tarihli bir araştırma makalesi ile karşılaştırıldığında, bir başka net sonuç bulunabilir (Jonathan D. Moyer ve ark., 2021). ABD ve Çin'in dünya çapındaki küresel etkisini ölçtüler. Onlar Çin ve ABD'nin her bir ülke üzerindeki mevcut göreceli etkisini ölçtüler.

country	Frequency	Avg_Polarity	Total_Likes	Total_Retweets	Sentiment
India	4289	-0.055389	4857	577228	NEGATIVE
Hong Kong	3517	-0.111378	1358	1279349	NEGATIVE
United States	3027	-0.052435	2175	431304	NEGATIVE
United Kingdom	965	-0.082971	400	71895	NEGATIVE
Philippines	894	-0.044801	612	56819	NEGATIVE
Singapore	746	0.083060	1724	40115	POSITIVE
Australia	714	-0.084962	1537	39705	NEGATIVE
China	640	-0.001262	1272	59101	NEGATIVE
Malaysia	613	0.025612	2255	51986	POSITIVE
Japan	553	0.004338	892	83530	POSITIVE
Canada	504	-0.052049	4734	41409	NEGATIVE

Figure 5.5.1 Çoğu tweet kaynaklı ülkeler

Ayrıca ABD-Çin ve BK-Çin arasında büyük bir mesafe olmasına rağmen, her iki ülkenin de bu bölgeye ilgi duyduğunu gösteriyor. Şekil 5.5.1'deki diğer ülkeler, Güney Çin Denizi bölgesine çok daha yakındır.

Bulgudan kalan 8 ülkenin tamamı güçlü ABD etkisine sahip ve Güney Çin Denizi'ndeki çeşitli olaylar hakkında en çok seslerini yükselten ülkeler. Bunlardan 7 ülke, ticari faaliyetleri için düzenli olarak Güney Çin Denizi'ni kullanıyor.

FBIC Endeksi, bir ülkenin diğerine çok boyutlu asimetrik bağımlılığını ölçer. Biri bir ilişkideki "bant genişliğini" ölçen ve diğeri bir durumun diğerine olan "bağımlılığını" ölçen iki alt endeksin ürünüdür. Tedbir, ekonomik, politik ve güvenlikle ilgili değişkenleri temsil eden verileri içerir.

Jonathan D. Moyer ve ark. ABD ve Çin'in FBIC endeksini dünyadaki tüm ülkeler üzerinden ölçtü. Bu analizden elde edilen sonuç, bu tweet analizinin bulgusu ile karşılaştırıldı. Tablo 5.5.1'de gösteren her iki ülkenin FBIC endeksi ile genel tweet duyarlılığı arasında bir korelasyon olduğu tespit edildi.

Tablo 5.5.1

Ülke	En çok etkilenen ülke	Çin'e göre net ABD etki kapasitesi	Twitter Duygusu
Hindistan	Amerika Birleşik Devletleri	0.14	Negatif
Filipin	Amerika Birleşik Devletleri	0.19	Negatif
Birleşik Krallık	Amerika Birleşik Devletleri	0.23	Negatif
Malezya	Diğer	0.01	Pozitif
Singapur	Amerika Birleşik Devletleri	0.35	Pozitif
Avustralya	Amerika Birleşik Devletleri	0.22	Negatif
Japonya	Amerika Birleşik Devletleri	0.28	Negatif
Kanada	Amerika Birleşik Devletleri	0.54	Negatif

5.6 Gelecek Çalışmalar

Gelecekte, bu çalışmayı daha fazla hashtag üzerinde genişletmek ve daha fazla tweet toplamak istiyoruz. Bu özel anahtar kelime için çok sayıda tweet yönetebilmeme rağmen, tweetin belirli bir yerinin olmaması nedeniyle birçoğunu kaybettik. Bu araştırmayı daha da genişletmek için verileri manuel olarak aramalı ve yerleştirmeli veya daha fazla tweet toplamalıyız. Bunun yanı sıra, birlikte daha fazla veri toplanırsa, derin öğrenmeyi uygulamaya çalışabilir ve sonucu bu mevcut yaklaşımla karşılaştırabiliriz. Ayrıca, bu araştırma tamamen İngilizce tweetlere bağlı olduğundan, İngilizce olmayan çok sayıda tweeti görmezden geliyoruz. Farklı diller için, bu dillerde duygu analizini bulmaya yardımcı olabilecek bazı araçlar mevcuttur. Almanca veya Hintçe dilinde tweetleri bulmak ve sonucu analiz etmek bir ilgi deneyi olabilir.

KAYNAKLAR

32.2. ast — Abstract Syntax Trees — Python 3.6.14 documentation. (n.d.). Retrieved August 15, 2021, from https://docs.python.org/3.6/library/ast.html

Abdulla, N. A., Ahmed, N. A., Shehab, M. A., & Al-Ayyoub, M. (2013). Arabic sentiment analysis: Lexicon-based and corpus-based. *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, 1–6.

Agarwal, Amit, Singh, R., & Toshniwal, D. (2018). Geospatial sentiment analysis using twitter data for UK-EU referendum. *Journal of Information and Optimization Sciences*, *39*(1), 303–317.

Agarwal, Apoorv, Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011). Sentiment analysis of twitter data. *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 30–38.

Al-Smadi, M., Al-Ayyoub, M., Jararweh, Y., & Qawasmeh, O. (2019). Enhancing aspect-based sentiment analysis of Arabic hotels' reviews using morphological, syntactic and semantic features. *Information Processing & Management*, *56*(2), 308–319.

AlKhatib, M., El Barachi, M., AleAhmad, A., Oroumchian, F., & Shaalan, K. (2020). A sentiment reporting framework for major city events: Case study on the China-United States trade war. *Journal of Cleaner Production*, 121426.

Araque, O., Corcuera, I., Román, C., Iglesias, C. A., & Sanchez-Rada, J. F. (2015). Aspect Based Sentiment Analysis of Spanish Tweets. *TASS@ SEPLN*, 29–34.

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, *1*, 492–499.

Baker, R. (2020). *China's Rise as a Global Power Reaches Its Riskiest Point Yet*. https://www.realcleardefense.com/articles/2020/07/06/chinas_rise_as_a_global_power_reaches_its_riskiest_point_yet_115441.html

Barbosa, L., & Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. *Coling 2010: Posters*, 36–44.

Basile, V., & Nissim, M. (2013). Sentiment analysis on Italian tweets. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 100–107.

Bautin, M., Vijayarenu, L., & Skiena, S. (2008). International sentiment analysis for news and blogs. *ICWSM*.

Beaumont, P. (2011). *The truth about Twitter, Facebook and the uprisings in the Arab world*. Guardian.

Beech, H. (2020, April 21). *U.S. Warships Enter Disputed Waters of South China Sea as Tensions With China Escalate - The New York Times*. https://www.nytimes.com/2020/04/21/world/asia/coronavirus-south-china-sea-warships.html

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., & Jurafsky, D. (2006). Extracting opinion propositions and opinion holders using syntactic and lexical cues. In *Computing Attitude and Affect in Text: Theory and Applications* (pp. 125–141). Springer.

Boiy, E., & Moens, M.-F. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval*, *12*(5), 526–558.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Bosco, C., Patti, V., & Bolioli, A. (2013). Developing corpora for sentiment analysis:

The case of irony and senti-tut. *IEEE Intelligent Systems*, *28*(2), 55–63.
Buddeewong, S., & Kreesuradej, W. (2005). A new association rule-based text classifier algorithm. *17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05)*, 2-pp.
Cesteros, J. S., Almeida, A., & de Ipiña, D. L. (2015). DeustoTech Internet at TASS 2015: Sentiment Analysis and Polarity Classification in Spanish Tweets. *TASS@ SEPLN*, *2*, 23–28.
China, U.S. Escalate Forces, Threats in South China Sea | World Report | US News. (n.d.). Retrieved December 28, 2021, from https://www.usnews.com/news/world-report/articles/2020-07-20/china-us-escalate-forces-threats-in-south-china-sea
Cugola, G., & Margara, A. (2012). Processing flows of information: From data stream to complex event processing. *ACM Computing Surveys (CSUR)*, *44*(3), 1–62.
Cui, A., Zhang, M., Liu, Y., & Ma, S. (2011). Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. *Asia Information Retrieval Symposium*, 238–249.
Das, S., & Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, *35*, 43.
David, S. (2020). *The Number of tweets per day in 2020*. DSayce.
Davies, A., & Ghahramani, Z. (2011). Language-independent Bayesian sentiment mining of Twitter. *The 5th SNA-KDD Workshop*, *11*.
Firth, J. A., Torous, J., & Firth, J. (2020). Exploring the Impact of Internet Use on Memory and Attention Processes. In *International Journal of Environmental Research and Public Health* (Vol. 17, Issue 24). https://doi.org/10.3390/ijerph17249481
Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review*, *31*(6), 649–679.
Ghahramani, Z. (2004). Unsupervised learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3176*, 72–112. https://doi.org/10.1007/978-3-540-28650-9_5
Gilbert, C. H. E., & Hutto, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) Http://Comp. Social. Gatech. Edu/Papers/Icwsm14. Vader. Hutto. Pdf*, *81*, 82.
Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(12), 2009.
Goel, A., Gautam, J., & Kumar, S. (2016). Real time sentiment analysis of tweets using Naive Bayes. *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 257–261.
Han, Z., & Paul, T. V. (2020). China's Rise and Balance of Power Politics. *The Chinese Journal of International Politics*, *13*(1), 1–26.
Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177.
Ikenberry, G. J. (2008). The rise of China and the future of the West-Can the liberal system survive. *Foreign Aff.*, *87*, 23.
Jindal, N., & Liu, B. (2006). Mining comparative sentences and relations. *Aaai*, *22*(13311336), 9.
Jonathan D. Moyer, Collin J. Meise, Austin S. Matthews, David K. Bohl, & Dr. Mathew J. Burrows. (2021, June 16). *IN BRIEF: Fifteen takeaways from our new report*

measuring US and Chinese global influence - Atlantic Council. Atlantic Council. https://www.atlanticcouncil.org/in-depth-research-reports/report/in-brief-15-takeaways-from-our-new-report-measuring-us-and-chinese-global-influence/

Jongerius, C., Hessels, R. S., Romijn, J. A., Smets, E. M. A., & Hillen, M. A. (2020). The Measurement of Eye Contact in Human Interactions: A Scoping Review. *Journal of Nonverbal Behavior*, 1–27.

Kaur, J., & Sabharwal, M. (2018). Spam detection in online social networks using feed forward neural network. *RSRI Conference on Recent Trends in Science and Engineering*, *2*, 69–78.

Khun, N. H., & Thant, H. A. (2019). Visualization of Twitter Sentiment during the Period of US Banned Huawei. *2019 International Conference on Advanced Information Technologies (ICAIT)*, 274–279.

Ko, Y., & Seo, J. (2000). Automatic text categorization by unsupervised learning. *Aclweb.Org*, 453–459. https://doi.org/10.3115/990820.990886

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! *Fifth International AAAI Conference on Weblogs and Social Media*.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, *2*(2010), 627–666.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.

Liu, T., & Woo, W. T. (2018). Understanding the US-China trade war. *China Economic Journal*, *11*(3), 319–340.

Melzer, A., Shafir, T., & Tsachor, R. P. (2019). *How Do We Recognize Emotion From Movement? Specific Motor Components Contribute to the Recognition of Each Emotion*.

Narr, S., Hulfenhaus, M., & Albayrak, S. (2012). Language-independent twitter sentiment analysis. *Knowledge Discovery and Machine Learning (KDML), LWA*, 12–14.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *LREc*, *10*(2010), 1320–1326.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *ArXiv Preprint Cs/0409058*.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *ArXiv Preprint Cs/0506075*.

Pang, B., & Lee, L. (2009). Opinion mining and sentiment analysis. *Comput. Linguist*, *35*(2), 311–312.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *ArXiv Preprint Cs/0205070*.

Park, S., & Kim, Y. (2016). Building thesaurus lexicon using dictionary-based approach for sentiment classification. *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)*, 39–44.

Patil, S., Gune, A., & Nene, M. (2018). Convolutional neural networks for text categorization with latent semantic analysis. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS 2017*, 499–503. https://doi.org/10.1109/ICECDS.2017.8390217

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 105–112.

Segaran, T. (2007). *Programming collective intelligence: building smart web 2.0*

applications. " O'Reilly Media, Inc."
Sinaga, L. C. (2015). China's Assertive Foreign Policy in South China Sea Under XI Jinping: Its Impact on United States and Australian Foreign Policy. *Journal of ASEAN Studies*, *3*(2), 133–149. https://doi.org/10.21512/jas.v3i2.770
South China Sea: US unveils first sanctions linked to militarisation | South China Sea | The Guardian. (n.d.). Retrieved December 28, 2021, from https://www.theguardian.com/world/2020/aug/27/south-china-sea-us-unveils-first-sanctions-linked-to-militarisation
South China Sea - Wikipedia. (n.d.). Retrieved August 14, 2021, from https://en.wikipedia.org/wiki/South_China_Sea
Tankovska, H. (2021). *Leading countries based on number of Twitter users as of January 2021*. Statista.
Theune, C. (2016, October 23). *pycountry · PyPI*. https://pypi.org/project/pycountry/
Thomas, M., Pang, B., & Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *ArXiv Preprint Cs/0607062*.
Unnisa, M., Ameen, A., & Raziuddin, S. (2016). Opinion mining on Twitter data using unsupervised learning technique. *International Journal of Computer Applications*, *148*(12), 975–8887.
Vega, L., & Mendez-Vazquez, A. (2016). Dynamic neural networks for text classification. *2016 International Conference on Computational Intelligence and Applications (ICCIA)*, 6–11.
Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012). Harnessing twitter" big data" for automatic emotion identification. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 587–592.
Wiebe, J. M. (1994). Tracking point of view in narrative. *ArXiv Preprint Cmp-Lg/9407019*.
Wiebe, J. M. (1990). *Identifying subjective characters in narrative*. 401–406. https://doi.org/10.3115/997939.998008
Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 347–354.
Wilson, T., Wiebe, J., & Hwa, R. (2004). Just how mad are you? Finding strong and weak opinion clauses. *Aaai*, *4*, 761–769.
Wu, E., Diao, Y., & Rizvi, S. (2006). High-performance complex event processing over streams. *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, 407–418.
Xia, R., Xu, F., Yu, J., Qi, Y., & Cambria, E. (2016). Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing & Management*, *52*(1), 36–45.
Yang, B., & Cardie, C. (2014). Context-aware learning for sentence-level sentiment analysis with posterior regularization. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 325–335.
Yared, P. (n.d.). Why sentiment analysis is the future of ad optimization. *Електронний Ресурс]:–Режим Доступу URL: Www. Venturebeat. Com/2011/03/20*.
Yusfiarto, R., & Pambekti, G. T. (2020). EFFECT OF MACROECONOMIC VARIABLES ON JAKARTA ISLAMIC INDEX: EVIDENCE THE GLOBAL TRADE WAR PHENOMENON. *Media Ekonomi*, *27*(2), 119–132.

Zainuddin, N., & Selamat, A. (2014). Sentiment analysis using support vector machine. *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, 333–337.