

A MAPREDUCE BASED DISTRIBUTED COMBINED SENTIMENT ANALYSIS MODEL AND ITS APPLICATION

MAPREDUCE TABANLI BİR DAĞITIK KOMBİNE DUYGU ANALİZİ MODELİ VE UYGULAMASI

Fikriye ATAMAN* 

Van Yuzuncu Yil University, Department of Informatics, Van, Turkey.

H. Eray ÇELİK 

Van Yuzuncu Yil University, Department of Econometri, Van, Turkey.

*Corresponding Author: fataman@yyu.edu.tr

Geliş Tarihi / Received: 20.11.2020
Kabul Tarihi / Accepted: 26.12.2020

Araştırma Makalesi/Research Article
DOI: 10.38065/euroasiaorg.372

ABSTRACT

In this study, in order to eliminate the performance losses experienced in the processing of big data, a distributed combined model working on the Hadoop ecosystem was designed and developed. A new model was used by combining dictionary-based methods and machine learning-based methods which are commonly used in sentiment analysis. The combined model we developed has been programmed and implemented as both the distributed version on Hadoop architecture and the serial version on traditional programming architecture and performance results have been compared and reported. Parallel model on Hadoop Distributed File System, which we believe will contribute significantly to the literature, developed it in this study process, and used in big data analysis, has achieved a higher performance by significantly eliminating performance losses.

In addition, with this study, it is aimed to keep a perspective on the migration-migrant-refugee-immigrant problem, which concerns many countries of the world. Twitter users in European countries were selected as the target audience. It has been determined that the perceptions of Twitter users included in the analysis vary by country. The results of the study showed that the reflex and reactions to the immigrant problem can vary from country to country. It is thought that these results also provide important data to the researchers.

Keywords: Big data, Hadoop, Migration, Multi nomial naive bayes, Sentiment analysis.

ÖZET

Bu çalışmada, büyük verilerin işlenmesi aşamasında yaşanan performans kayıplarının giderilmesi amacıyla, Hadoop ekosistemi üzerinde çalışan MapReduce tabanlı dağıtık kombine bir duygu analizi modeli tasarlanarak geliştirilmiştir. Duygu analizi konusunda yaygın olarak kullanılan sözlük tabanlı yöntemler ve makine öğrenmesi tabanlı yöntemler birleştirilerek kombine yeni bir model sunulmaktadır. Geliştirilen kombine model, hem Hadoop mimarisinde dağıtık sürüm olarak, hem de geleneksel programlama mimarisinde seri sürüm olarak programlanarak uygulanmış ve başarımları karşılaştırılarak verilmiştir. Literatüre önemli ölçüde katkı sunacağını düşündüğümüz, bu çalışma sürecinde geliştirilen ve büyük veri analizinde kullanılan Hadoop Dağıtık Dosya Sistemi (HDDS) tabanlı paralel model ile performans kayıpları önemli ölçüde giderilerek daha yüksek bir performans elde edilmiştir.

Ayrıca, bu çalışma ile dünyanın birçok ülkesini ilgilendiren göç-göçmen-mülteci-sığınmacı sorununa bir bakış açısı geliştirilmesi hedeflenmiştir. Hedef kitle olarak Avrupa ülkelerindeki Twitter kullanıcıları seçilmiştir. Analiz sonuçları ile Twitter kullanıcılarının algılarının, ülkelere göre değişiklik gösterdiği tespit edilmiştir. Çalışma sonuçları, göçmen problemine gösterilen

refleksin ve tepkilerin ülkeden ülkeye değişebilmekte olduğunu göstermiştir. Elde edilen bu sonuçların konu ile ilgilenen bilim insanlarına önemli bir veri sunacağı düşünülmektedir.

Anahtar kelimeler: Büyük Veri, Çok Değişkenli Naif Bayes, Duygu Analizi, Göç, Hadoop.

1. GİRİŞ

Günümüz dünyasında, anlık üretilen verinin boyutu katlanarak artmaktadır. Bilgi üreten kaynakların artması sonucu, veri üretimi ve veri birikmesi çok hızlı bir şekilde gerçekleşmektedir. Devasa hızlarda üretilen ve biriken bu verilerin, mevcut geleneksel metot ve yöntemlerle incelenmesi çok zorlaşmaktadır. Bu durumda birikmiş veri yığınlarını incelemeye yönelik yeni çalışmalar başlatılmıştır. Ulaşım, ekonomi, sağlık, eğitim, güvenlik vb. birçok sektörde milyonlarca veri üretilmektedir. Büyük veri çağına girmiş olduğumuz bilinen bir gerçektir (Brown vd., 2011).

Dünya genelinde, sosyal medya kullanımında ciddi artışlar meydana gelmektedir. Dünya nüfusunun % 59'u internet kullanıcısı iken % 49'u aynı anda aktif sosyal medya kullanıcısıdır (Kemp, 2020). Sosyal ağların kullanımı her gün artmaktadır. Nisan 2020 döneminde, sosyal medya kullanıcı sayısı Nisan 2019 dönemine göre % 8 büyüyerek 3.81 milyara ulaşmıştır. Benzer şekilde Nisan 2020 itibari ile bir önceki yıla göre internet kullanıcı sayısı % 7 büyüyerek 4.57 milyara ulaşmıştır. Global ölçekte 5 milyar mobil kullanıcı mevcuttur (Kemp, 2020).

Sosyal ağ uygulamaları, bireylere mümkün olmayanı sunarak, kişiye duygu ve düşüncelerini ifade etme olanağı vermektedir. Twitter sosyal ağ uygulamalarından en çok kullanılanlardan biridir ve insanlara, olaylar, durumlar ve günlük hayattaki diğer konular üzerindeki yorumlarını paylaşarak, kendilerini ifade edebilecekleri bir ortam sunmaktadır. Anlık olarak yayınlanan binlerce tweet mesajını anlamlandıran ve değerli kılan şey ise insanların iletilerin içerisinde bıraktıkları duygular ve fikirlerdir. Duygu verilerini anlamlandırmaya çalışmak, bu verilerden anlamlı bilgiler üretmek, bu verileri analiz ederek bu veriler üzerinde kestirim ve çıkarsama yapmak, günümüzde oldukça önem kazanmaktadır. Bu kestirim ve çıkarsamalar, ilgili şirket, kurum vb. taraflar için ilgilendikleri problem üzerinde daha doğru karar vermeyi sağlayan önemli araçlar olmaktadır. Özellikle şirketler, kurumlar, siyasi partiler, dernek, vakıf ve organizasyonlar gibi tüzel kişilikler, müşteri memnuniyeti, pazarlama stratejileri, örgütsel bağlılık gibi konularda süreç yönetiminde bu kestirim ve çıkarsamalardan faydalanmaktadır.

1.1. Problem ve Motivasyon

Günümüzde, göç ve göç hareketleri tüm dünya için büyük bir sorun haline gelmiştir (Garrett, 2020; Gerşil, 2019). Ülkelerin göç ile ilgili kurum ve kuruluşları her ne kadar sosyal entegrasyon amacıyla yasal çerçevede sosyal eşitlik, hukukun üstünlüğü, ayrımcılık karşıtlığı ve toplumsal uzlaş gibi belirleyici ve denetleyici ilkeleri kabul etseler de pratikte göç, farklı sorunlara yol açarak, olumlu veya olumsuz birçok etki bırakmaktadır (Gerşil, 2019).

Göç hareketlerinin bu kadar arttığı günümüz dünyasında Avrupa kıtasının en çok tercih edildiği ve son 20 yılda hemen hemen her Avrupa ülkesinin çok yoğun göç aldığı bilinmektedir (Anonim, 2019c). Bu kadar geniş bir coğrafyada göç, göçmenlik, iltica ve sığınmacı kavramları ile ilgili kapsayıcı bir çalışma yapmak sosyal ağ verileri ile mümkün olabilmektedir.

Büyük verilerin geleneksel metin madenciliği yöntemleriyle işlenmesi oldukça zordur çünkü geleneksel metin madenciliği yöntemlerinin boyut, hız, bellek kullanımı gibi kısıtlayıcı birçok problemi olmaktadır. Büyük veri modellerinde ise bu problemler dağıtık ve paralel mimariler aracılığıyla en aza indirilerek verimli ve etkin bir altyapı sunulmaktadır. Bu çalışma kapsamında da Hadoop dağıtık dosya sistemi üzerinde yeni bir Duygu Analizi (DA) metodu geliştirilerek verilerin performans, hız ve verimlilik açısından daha iyi işlenmesi sağlanmaya çalışılmıştır. Çalışmanın hem

geleneksek hem de Hadoop modeli ayrı ayrı kurularak karşılaştırılması yapılmıştır. Ayrıca her ülkenin, tema olarak ele alınan göç-göçmen-mülteci-sığınmacı kavramlarına olan yaklaşımı duygu analizi aracılığıyla ortaya konmaya çalışılmıştır. Bu çalışma ile hem güncel bir probleme ışık tutulması hedeflenmiş hem de büyük verilerin işlenebileceği yeni bir metot geliştirilmiştir.

1.2. Duygu Analizi Nedir?

Fikir madenciliği olarak da adlandırılan Duygu Analizi (Pozzi vd., 2017); insanların ürünler, hizmetler, organizasyonlar, bireyler, konular ve olaylar veya onların özellikleri hakkında duygularını, değerlendirmelerini ve tutumlarını analiz eden bir çalışma alanıdır (Agarwal ve Mittal, 2016; Liu, 2012a; Poria vd., 2018; Satapathy vd., 2018). Duygu madenciliği, fikir analizi, fikir çıkarımı, fikir madenciliği ve buna benzer birçok başlık duygu analizi veya fikir madenciliği altında toplanmaktadır (Agarwal ve Mittal, 2016; Liu, 2012b). Duygu Analizi, elektronik ortamdaki yazılı metin içeriklerinde, makine öğrenmesi tabanlı veya istatistiksel yöntemlerle çeşitli analizler yapılarak insani duyguların tespit edilerek sınıflandırılması işlemidir (Poria vd., 2018).

1.3. Büyük Veri Nedir?

Büyük Veri, ölçeği, dağılımı, çeşitliliği ve güncelliği ile değerli kaynakların ortaya çıkarılmasına vesile olacak öngörüler sağlamak için yeni teknik mimariler ve analitik kullanımı gerektiren verilerdir (Buyya vd., 2016; Gupta ve Kohli, 2016; Ohlhorst, 2012).

Büyük veri bilimi; matematik alanında istatistik ve optimizasyon; bilgisayar bilimleri alanında algoritmalar, programlar, makine öğrenmesi, veri madenciliği; alan bilimlerinde ise işletme, görselleştirme vd. uygulamaların iç içe girdiği bir disiplinler arası çalışma alanıdır (Buyya vd., 2016; Chen vd., 2019). Büyük verinin özellikleri ile ilgili çok farklı tanımlamalar yapılmıştır. İlk özellikleri Laney (2001) tarafından 3V (Volume-Velocity-Variety) olarak tanımlanmıştır. Daha sonra bu tanıma, IBM tarafından Veracity-Sağlamlık özelliğini (Anonim, 2019a; Gantz ve Reinsel, 2012) , Microsoft ise Value ve Visibility özelliklerini katarak 6V olarak tanımlı genişletmişlerdir (Buyya vd., 2016).

1.4. Göç ile ilgili tanımlamalar

Bu başlıkta göç ile ilgili olan “Göçmen”, “Mülteci” ve “Sığınmacı” kavramlarının tanımlarına yer verilmektedir.

Birleşmiş Milletler, göçmeni, “sebepleri, gönüllü olup olmaması, göç yolları, düzenli veya düzensiz olması fark etmeksizin yabancı bir ülkede bir yıldan fazla ikamet eden bir birey” olarak tanımlar (Perruchoud ve Redpath, 2013)

1951 tarihli Mültecilerin Hukuki Durumuna Dair Cenevre Sözleşmesi’nde Mülteci; “*Irkı, dini, tabiiyeti, belirli bir sosyal gruba mensubiyeti ve siyasi görüşleri yüzünden haklı bir zulüm korkusu nedeniyle vatandaşı olduğu ülkenin dışında bulunan ve söz konusu korku yüzünden, ilgili ülkenin korumasından yararlanmak istemeyen kişi*” olarak ifade edilmektedir (Perruchoud ve Redpath, 2013).

Göç ile ilintili olan ve sıklıkla Mülteci terimi ile karıştırılan bir diğer kavram ise “Sığınmacı” terimidir. Uluslararası Göç Örgütünün hazırlamış olduğu Göç Terimleri Sözlüğünde “Sığınmacı”, zulüm veya ciddi zarardan korunmak amacıyla, kendi ülkesi dışında bir ülkede güvenlik arayışında olan ve ilgili ulusal ya da uluslararası belgeler çerçevesinde mültecilik statüsüne ilişkin yaptığı başvurunun sonucunu bekleyen kişi olarak ifade edilmiştir (Perruchoud ve Redpath, 2013).

2. MATERYAL VE YÖNTEM

2.1. Araştırma materyali

Bu çalışmada Twitter sosyal medya sitesinin geliştirici uygulama ara yüzünden “göçmen, göçmenler, mülteciler, mülteci ve sığınmacı” kavramları İngilizce taranarak filtreleme yapılmak suretiyle atılan ham tweetler çekilerek veri seti olarak hazırlanmıştır. Çalışmanın kapsamına dil olarak sadece İngilizce tweetler konu edinilmiştir. Tweetlerin çekilmesi esnasında tweetin atıldığı konum bilgileri için ülke kodu dikkate alınarak veriler setleri oluşturulmuştur. Hedef kitle olarak Avrupa ülkeleri seçilmiştir. Birleşmiş Milletlerin 2019 yılındaki son ülke nüfus verileri (Anonim, 2019b) ve dünya göçmen verileri (Anonim, 2019c) kaynak alınarak örneklem oluşturulmuştur. Bu çalışmada nüfusu 1 milyonun altında olan ve ülke göç yoğunluğu değeri 0.0693’ün altında olan ülkeler bulundukları bölgeye göre “diğer” başlığı altında toplanmıştır. Ayrıca Avrupa kıtasında kısmen yer alan Rusya ve Doğu Avrupa ülkeleri olan Ukrayna ve Beyaz Rusya için yeterince veri sağlanamadığından çalışma kapsamı dışında bırakılmıştır. İfade edilen belirleyici koşulları sağlayan, tüm Avrupa ülkeleri için göçmen yoğunluğu ile orantılı olarak tabakalı örneklem kullanılmıştır. Tweetler çekilerek toplamda 500000 tweetten oluşan veri setleri oluşturulmuştur. Baz alınan ülkelere ait nüfus verileri ve hesaplanan istatistikler Çizelge 2.1.’de verilmiştir.

2.2. Araştırma Yöntemi

Bu çalışmada, sözlük tabanlı ve makine öğrenmesi tabanlı yöntemler karşılaştırmalı olarak seri ve paralel (Hadoop) olarak uygulanmaktadır. Modelin proje dosyaları için Java SE Runtime Environment 1.8.0_161 sürümü kullanılmıştır. Paralel programlar için Hadoop kümesi kurulmuştur. Hadoop 3.1.1. kullanılarak kurulan küme beş makineden oluşmaktadır ve küme homojen bir yapıda olup 5 köle ve 1 ana makineden oluşmaktadır. İşletim sistemi olarak Linux Ubuntu 16.4 sürümü kullanılmıştır. Toplam 16 GB Ram ile küme desteklenmiştir.

2.2.1. Multinomial Naive Bayes (MNNB) Sınıflayıcısı

Multinomial Naive Bayes (MNNB) bir Naif Bayes sınıflayıcı çeşidi olup çok duygu sınıflamada sıkça kullanılmaktadır (Juan ve Ney, 2002; Kibriya vd., 2004; Schneider, 2004). Çok terimli veri setlerinde etkin bir şekilde kullanılmaktadır. Multinomial Naive Bayes metindeki kelimelerin dağılımını çok terimli olarak modeller. Naive Bayes sınıflayıcılarda olduğu gibi burada da bir özelliğin bir sınıfa ait olma olasılığı hesaplanır. Dolayısıyla sınıfa ait olma olasılığı en yüksek olan özellikler öznitelik olarak seçilir. Bir öznenin seçilmesini sağlayan en çok olabilirlik eşitlik 2.1’de görüldüğü gibi hesaplanmaktadır.

$$C_{\max} = \arg \max_{c \in C} P(c) \prod P(t_i | c) \quad (2.1)$$

Burada $P(c)$, bir dokümanın bir C sınıfına ait olma olasılığıdır. Eşitlik 2.2. ile hesaplanır. $P(t_i|c)$ ise t_i özelliğinin C sınıfına ait olmasının koşullu olasılığıdır. Eşitlik 2.3. ile hesaplanır;

Çizelge 2.1. Avrupa ülkeleri demografik verileri ve istatistikleri

	GN	ÜN	GY	NYGTS
Avrupa	82304539	747182815	0.111	500000
Beyaz Rusya	1069395	9452409	0.114	6326
Rusya Federasyonu	11640559	145872260	0.08	97615
Ukrayna	4964293	43993643	0.113	29440
Danimarka	722878	5771877	0.126	3863
Estonya	190242	1325649	0.144	888
İrlanda	833564	4882498	0.171	3268
Letonya	237266	1906740	0.125	1276
Norveç	867765	5378859	0.162	3600
İsveç	2005210	10036391	0.2	6717
Birleşik Krallık	9552110	67530161	0.142	44836
Hırvatistan	518083	4130299	0.126	2764
Yunanistan	1211382	10473452	0.116	7009
İtalya	6273722	60550092	0.104	40056
Portekiz	888162	10226178	0.087	6844
Sırbistan	820312	8772228	0.094	5871
Slovenya	253122	2078654	0.122	1391
İspanya	6104203	46736782	0.131	31276
Avusturya	1779857	8955108	0.199	5993
Belçika	1981919	11539326	0.172	7722
Fransa	8334875	65129731	0.128	43584
Almanya	13132146	83517046	0.158	55888
Hollanda	2282791	17097123	0.134	11442
İsviçre	2572029	8591361	0.3	5750
Diğer Kuzey				12523
Diğer Güney				12579
Diğer Batı				6754
Diğer Doğu				44725
Toplam				500000
NYGTS: Nüfus yoğunluğuna göre tweet sayısı				
GN: Göçmen Nüfusu				
ÜN: Ülke Nüfusu				
GY: Göçmen Yoğunluğu				

$$P(c) = \frac{N_c}{N} \quad (2.2.)$$

$$P(t|c) = \frac{n_{t;c} + 1}{N_c + |V|} \quad (2.3.)$$

MNNB Hadoop ortamında dağıtık modellenerek, MNNB için paralel dağıtık bir metot geliştirilmektedir. Kullanılan sözlük tabanlı ilk yöntem ile ikinci paralel yöntem birleştirilerek kombine bir model oluşturulmaktadır. Bu kombine modelde, sözlük tabanlı yöntemler ile verilerin etiketlenmesi sağlanmaktadır ve MNNB sınıflayıcının seri ve paralel olarak uygulanmasıyla, tweetlerin yansıtıkları duygunun sınıflanması sağlanmaktadır.

2.2.2. Duygu analizinde sözlük tabanlı teknikler

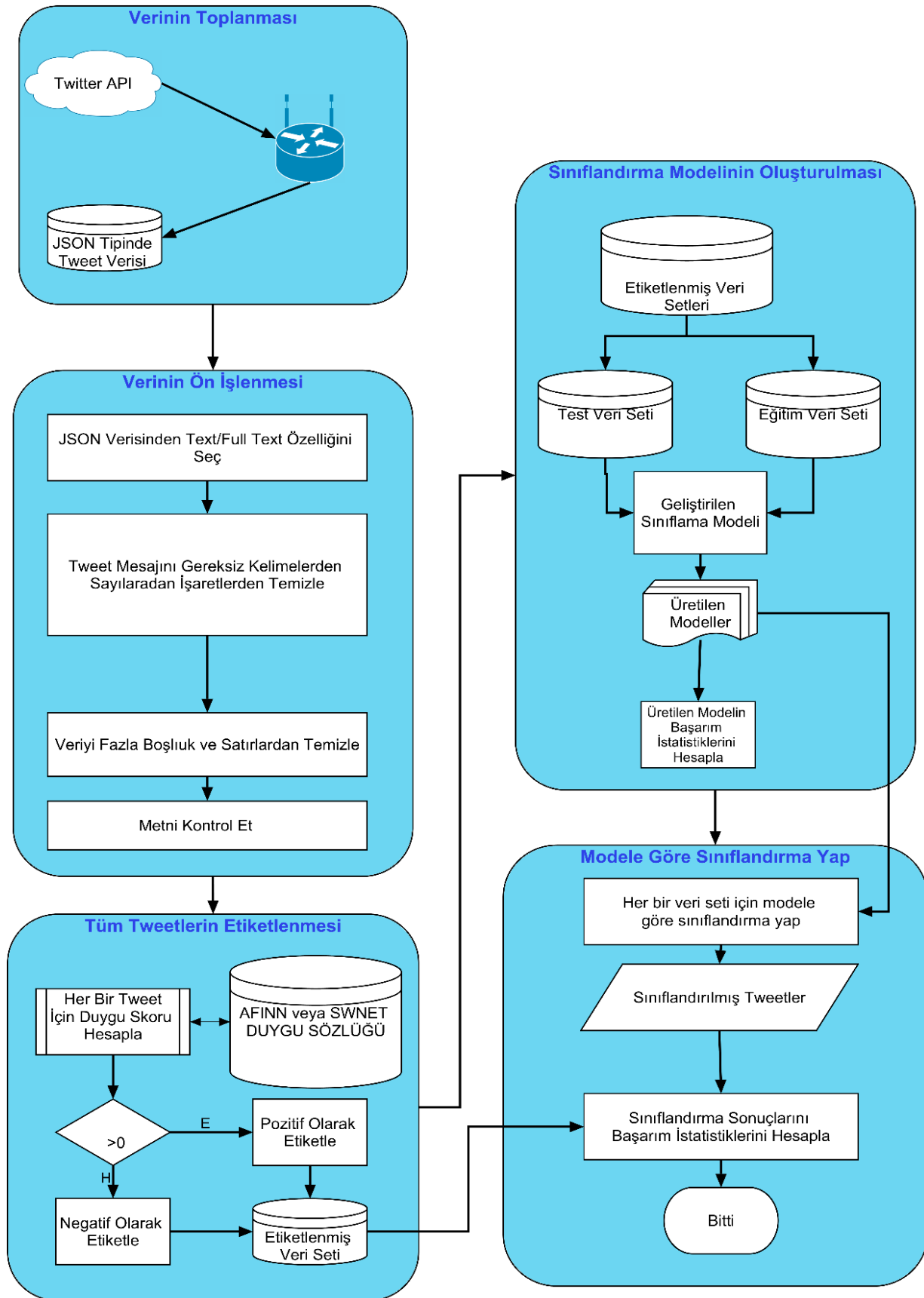
Bu çalışma kapsamında SentiWordNet 3 (Baccianella vd., 2010) ve AFINN (Nielsen, 2011) duygu sözlükleri veriyi etiketlemek amacıyla kullanılmıştır. Her iki sözlük için de paralel (Hadoop kümesinde) ve seri olarak metotlar geliştirilmiş ve uygulanmıştır.

2.2.3. Büyük Veri Platformunda Geliştirilen Analiz Modeli

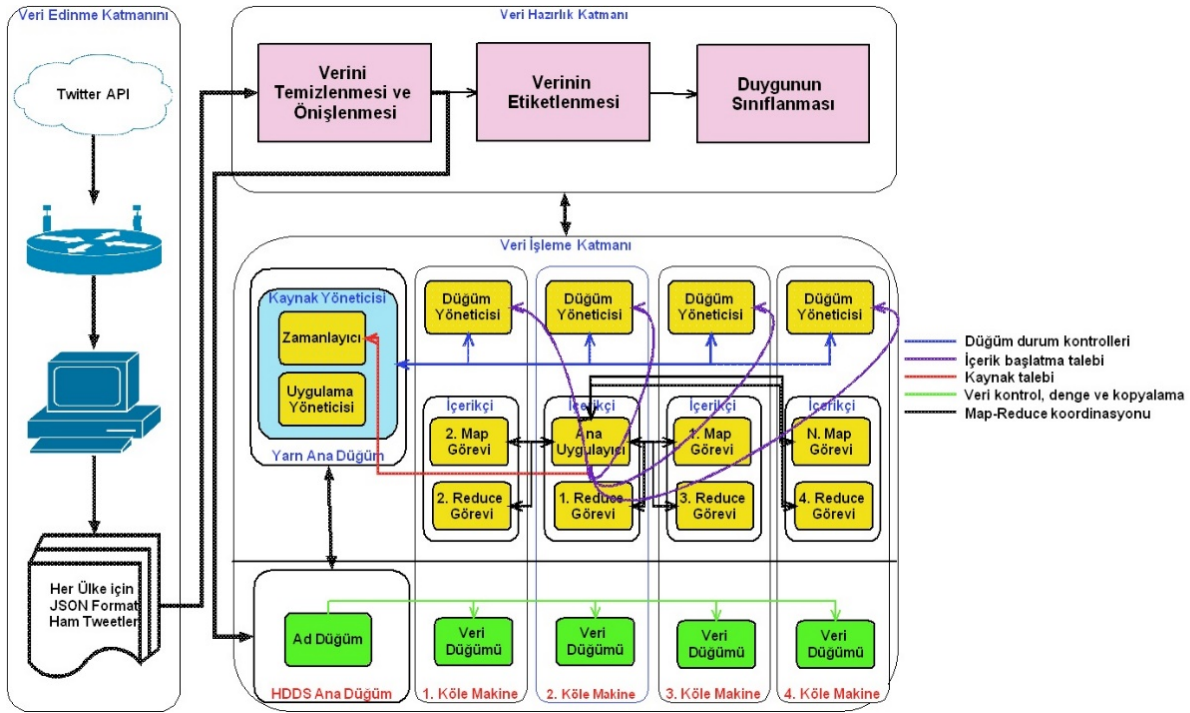
Hadoop mimarisi için geliştirilen yeni model MapReduce tabanlı çalışmaktadır ve paralel dağıtık bir yapı oluşturulmuştur. Bu oluşturulan yeni model ile veriler için yapılan işlemler dağıtık mimari aracılığıyla paralel çalıştırılarak maliyet kazancı elde edilmiştir. Geliştirilen tüm modelin çalışma şekli ve akış diyagramı Şekil 2.1.'de ve genel altyapı mimarisi ise Şekil 2.2.'de sunulmaktadır.

2.2.4. Sınıflandırma ve başarımların istatistiklerinin hesaplanması

Sınıflandırma işleminin başarımlar istatistikleri Çizelge 2.2.'te gösterilen hata matrisi değerlerine göre hesaplanmıştır. Hata matrisinin değerleri kullanılarak bazı istatistikler hesaplanabilmektedir (Bowes vd., 2014). Bunlar Recall-Geri Çağırma, Precision-Hassasiyet, F-Measure - F-Skoru, Accuracy - Doğruluk değerleridir. Recall tüm pozitif sınıflardan ne kadar doğru tahmin edildiğini ölçmektedir. Doğruluk ise tüm sınıflardan ne kadar doğru tahmin edildiğini belirtmektedir (Bowes vd., 2014; Solanki vd., 2020). Bu istatistiki metrik değerler aşağıda yazılan Eşitlik 2.4., Eşitlik 2.5., Eşitlik 2.6. ve Eşitlik 2.7. ile hesaplanmaktadır.



Şekil 2.1. Hadoop MapReduce ile Duygu Analizi işlem akış diyagramı.



Şekil 2.2. Kurulan modelin mimari altyapısı.

Çizelge 2.2. Hata matrisi gösterimi

		Gerçek Değerler (Etiketlenen)	
		Pozitif	Negatif
Tahmin	Pozitif	TP:Doğru Pozitif	FN:Yanlış Negatif
	Negatif	FP:Yanlış Pozitif	TN:Doğru Negatif

$$Presicion = p = \frac{TP}{TP + FP} \quad (2.4)$$

$$Recall = r = \frac{TP}{TP + FN} \quad (2.5)$$

$$F1 = 2 \times \frac{p \times r}{p + r} = \frac{2 \times TP}{2TP + FP + FN} \quad (2.6)$$

$$Accuracy = ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (2.7)$$

3. BULGULAR

3.1. SentiWordNet (SWNET) ile paralel/seri duygu etiketleme sonuçları

Paralel ve seri yöntemler arasında etiketleme sonuçları açısından bir fark olmayıp işlem performanslarında fark vardır. Etiketleme işlemleri sonucunda Diğer Doğu, Fransa, Hırvatistan, Hollanda, Portekiz, Slovenya, Yunanistan, İsveç ve İtalya ülkeleri göreceli olarak daha çok negatif olarak etiketlenmiştir. Bu ülkelere karşın Almanya, Avusturya, Belçika, Birleşik Krallık, Danimarka, Diğer Güney, Diğer Kuzey, Diğer Batı, Estonya, Letonya, Norveç, Sırbistan, İrlanda, İspanya, İsviçre ülkeleri daha çok pozitif olarak etiketlenenlerdir.

SWNET paralel-seri etiketleme sonuçlarının kümeleme sonuçlarına (Şekil 3.1.) bakıldığında ülkelerin temel olarak iki küme altında kendi aralarında kümelendiği gözlenmektedir. Almanya, Birleşik Krallık, Diğer Doğu, İtalya, İspanya ülkeleri bir küme altında bulunurken, diğer ülkeler bir küme altında bulunup kendi aralarında da alt kümeler ayrışmaktadır. Alt kümeler olarak Letonya, Hırvatistan, Slovenya, Estonya, Norveç, İrlanda, Danimarka aynı alt küme altında gözlenirken, diğer veri setlerinden Diğer Güney, Diğer Kuzey ve Hollanda bir alt küme ve Portekiz, Yunanistan, Diğer Batı, Belçika, İsveç, Avusturya, Sırbistan ve İsviçre başka bir alt küme olarak başka bir kümede gözlenmektedir.

3.2. AFINN ile paralel/seri etiketleme sonuçları

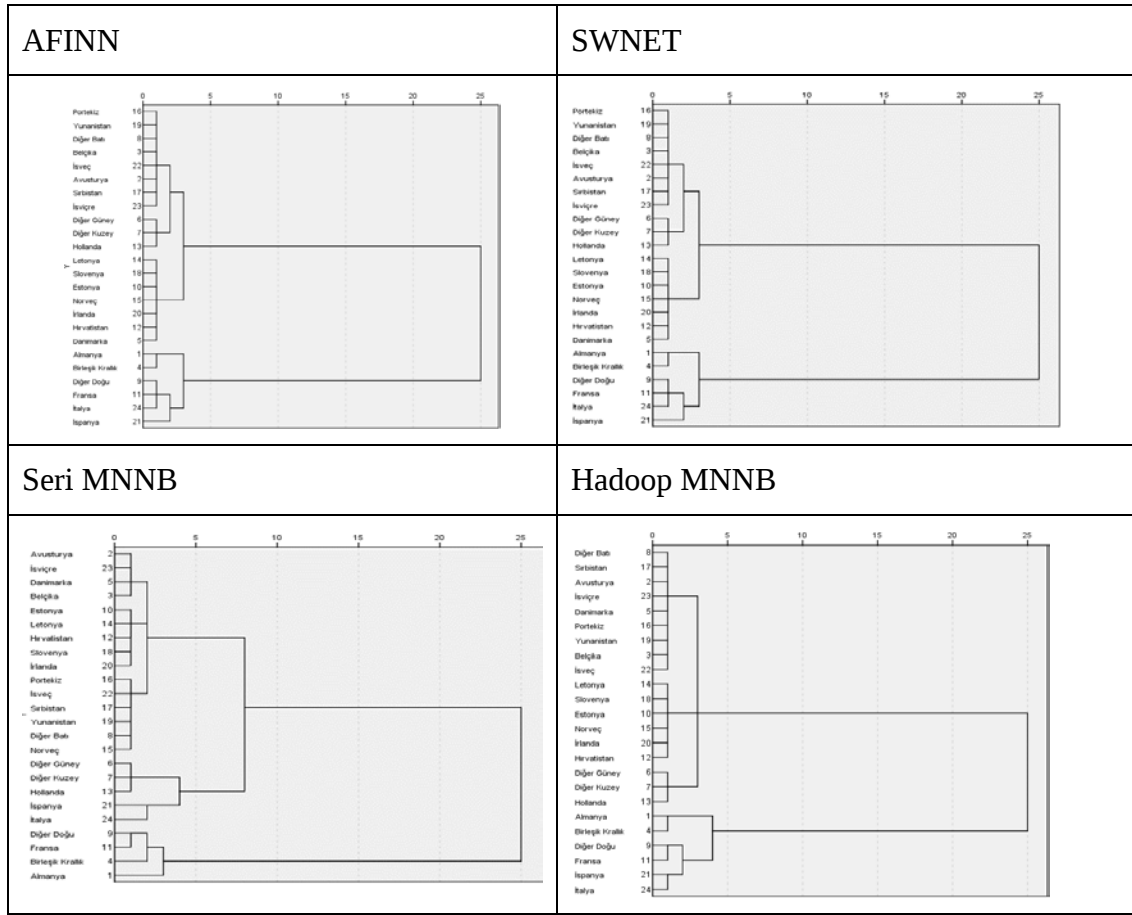
AFINN sözlüğü ile paralel ve seri etiketleme işlemleri arasında etiketleme sonuçları açısından bir fark yoktur. Ancak verinin işleme sürecinde zamandan performans kazancı elde edilmiştir. AFINN ile paralel ve seri etiketleme işleminin sonucunda sadece İsviçre veri seti daha çok pozitif etiketlenmektedir. Almanya, Avusturya, Belçika, Birleşik Krallık, Danimarka, Diğer Doğu, Diğer Güney, Diğer Kuzey, Diğer Batı, Estonya, Fransa, Letonya, Norveç, Hırvatistan, Hollanda, Portekiz, Slovenya, Yunanistan, İsveç, İtalya, Sırbistan, İrlanda ve İspanya ülkelerinin verisi ise daha çok negatif olarak etiketlenmektedir. Kümeleme sonuçlarına (Şekil 3.1.) bakıldığında ülkelerin 4 kategorik küme altında kümelendiği gözlenmektedir. Letonya, Hırvatistan, Slovenya, Estonya, Norveç, İrlanda, Danimarka bir küme altında görünürken, Diğer Güney, Diğer Kuzey ve Hollanda başka bir küme altında ve Portekiz, Yunanistan, Diğer Batı, Belçika, İsveç, Avusturya, Sırbistan ve İsviçre başka bir kümede gözlenmektedir. Almanya, Birleşik Krallık, Diğer Doğu, İtalya, İspanya ülkeleri bir küme altında bulunurken, Diğer Doğu, Fransa, İtalya ve Birleşik Krallık aynı kümede ve Almanya ile İspanya aynı kümede gözlenmektedir. Çizelge 3.1. ülkelerin AFINN ve SWNET sözlüğü ile paralel/seri olarak etiketlenmiş pozitif, negatif ve nötr tweet sayılarını göstermektedir.

Çizelge 3.1. AFINN ve SWNET Sözlükleri ile paralel/seri etiketleme sonuçları

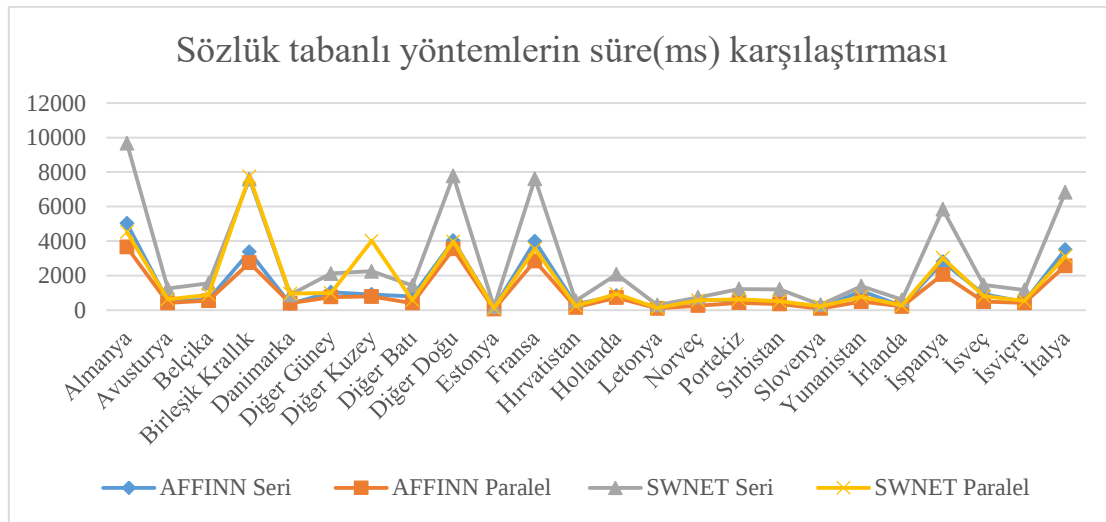
Veri Seti	AFINN Paralel			AFINN Seri			SWNET Seri ve Paralel		
	Pozitif	Negatif	Nötr	Pozitif	Negatif	Nötr	Pozitif	Negatif	Nötr
Almanya	12974	16914	26000	12974	16914	26000	24909	23842	7137
Avusturya	1502	1662	2829	1502	1662	2829	2740	2421	832
Belçika	2019	2067	3636	2019	2067	3636	3517	3319	886
Birleşik Krallık	11240	14080	19516	11240	14080	19516	20870	19775	4191
Danimarka	1222	1385	1783	1222	1385	1783	2069	1842	479
Diğer Güney	2673	4199	5707	2673	4199	5707	5397	5315	1867
Diğer Kuzey	2695	3931	5897	2695	3931	5897	5393	5374	1756
Diğer Batı	1324	1736	3694	1324	1736	3694	2764	2684	1306
Diğer Doğu	7980	12949	23796	7981	12949	23795	17086	18108	9531
Estonya	180	285	423	180	285	423	405	346	137
Fransa	8420	12277	22887	8420	12277	22887	17404	18018	8162
Hırvatistan	546	754	1464	546	754	1464	1136	1139	489
Hollanda	2404	3465	5573	2404	3464	5574	4803	4930	1709
Letonya	331	366	579	331	366	579	584	518	174
Norveç	848	1250	1502	848	1250	1502	1644	1598	358
Portekiz	1364	2258	3222	1364	2258	3222	2879	2942	1023
Sırbistan	1360	1829	2682	1360	1829	2682	2582	2515	774
Slovenya	291	493	607	291	493	607	581	624	186
Yunanistan	1389	2093	3527	1389	2093	3527	2961	3004	1044
İrlanda	796	1040	1432	796	1040	1432	1529	1403	336
İspanya	6865	8368	16043	6865	8368	16043	12648	12643	5985
İsveç	1857	2642	2218	1857	2642	2218	3103	3206	408
İsviçre	1582	1453	2715	1582	1453	2715	2807	2358	585
İtalya	7739	10595	21722	7739	10595	21722	16335	16579	7142

3.3. AFINN – SWNET Performans Sonuçları

Her iki sözlük ile paralel ve seri etiketleme işlemlerinde, performans açısından büyük bir fark olmuştur. Hadoop ortamında paralel olarak çalıştırılan metotlar ile veri boyutuna bağlı olarak iki kata kadar artan performans iyileşmesi sağlanmıştır. Veri boyutu arttıkça aradaki zaman farkı da artmaktadır. Bu durum, veri boyutu arttıkça performans kazancının da arttığı anlamına gelmektedir. Ancak hem seri hem paralel yöntemde AFINN sözlüğü ile SWNET ten daha hızlı işlem gerçekleşmiştir. Bu ise sözlüklerin yapısından kaynaklanmaktadır. Çizelge 3.2. ile SWNET ve AFINN sözlükleri ile seri ve paralel olarak kullanılan tüm metotlar için işlem süreleri mili saniye (ms) cinsinden performans sonuçları olarak gösterilmektedir. Tüm metotların performans karşılaştırması ise Şekil 3.2.'de gösterilmektedir.



Şekil 3.1. Ülkelerin paralel/seri duygu polarite sonuçlarının kümelenmesi.



Şekil 3.2. Sözlük tabanlı yöntemlerin performans karşılaştırması.

Çizelge 3.2. Sözlükler ile paralel-seri etiketleme işlemi performans sonuçları

Veri Seti	AFINN SS (ms)	AFINN PS (ms)	SWNET SS (ms)	SWNET PS (ms)
Almanya	5025	3653	9676	4523.5
Avusturya	510	419	1263	632.5
Belçika	613	547	1559	887.5
Birleşik Krallık	3377	2742.5	7595	7748
Danimarka	371	394	881	956
Diğer Güney	1043	760	2123	995.5
Diğer Kuzey	913	790	2255	4012
Diğer Batı	797	411	1431	578
Diğer Doğu	4025	3539.5	7772	3983
Estonya	127	53.5	189	97.5
Fransa	3987	2849.5	7619	3553
Hırvatistan	326	160.5	534	243
Hollanda	844	722.5	2078	935.5
Letonya	103	98.5	293	133
Norveç	577	262.5	740	568
Portekiz	593	424	1221	626.5
Sırbistan	480	354	1205	520.5
Slovenya	123	90.5	319	209.5
Yunanistan	1106	489	1394	775
İrlanda	245	214.5	601	264
İspanya	2805	2064	5847	3021
İsveç	922	488	1468	802.5
İsviçre	475	419.5	1169	529
İtalya	3513	2566.5	6834	3112.5
SS: Seri Süre	PS: Paralel Süre			

3.4. MNMB ile paralel duygu analizi sonuçları

Paralel MNMB için yapılan sınıflandırma işleminin sonucunda sadece İsviçre veri setinde daha çok pozitif tweet sınıflanmaktadır. Diğer Güney, Diğer Kuzey, Diğer Doğu, Estonya, Fransa, Hollanda, Norveç, Portekiz, Slovenya, Yunanistan ve İsveç ülkelere ait veri setlerinde negatif sınıflanmış tweet sayısı pozitif sınıflanmış tweet sayısının bir buçuk katından fazladır. Özellikle diğer güney, diğer doğu, Portekiz, Slovenya için bu oran yaklaşık olarak iki katına çıkmaktadır. Bu ülkelerden yapılan yorumların daha şiddetli bir şekilde göçmen karşıtlığı sergilediği yorumu yapılabileceğini göstermektedir. Hadoop MNMB ile sınıflanmış tweetlerin ülkelere göre dağılımı Çizelge 3.4. ile gösterilmektedir. Her bir veri setine ait, başarımlar istatistiklerine Çizelge 3.3.'te yer verilmektedir. Paralel MNMB sınıflayıcı iyi bir performansla başarılı bir sınıflama sağlamaktadır.

Kümelenme sonuçlarına (Şekil 3.1.) bakıldığında; Diğer Batı, Sırbistan, Avusturya, İsviçre, Danimarka, Portekiz, Yunanistan, İsveç ve Belçika bir küme altında toplanırken, Letonya, Slovenya, Estonya, Norveç, İrlanda ve Hırvatistan bir küme altında, Diğer Güney, Diğer Kuzey ve Hollanda bir küme altında, Almanya ve Birleşik Krallık bir küme altında, Diğer Doğu ile Fransa, İspanya ile İtalya başka bir alt küme altında olmak üzere farklı kümelerde kümelenebilmektedirler.

Çizelge 3.3. MNNB Hadoop ile DA başarımlar istatistikleri

Ülke	TP	FN	TN	FP	Doğruluk	Duyarlılık	Keskinlik	F1
Almanya	11750	482	16440	1229	0.943	0.961	0.905	0.932
Avusturya	1451	32	1634	49	0.974	0.978	0.967	0.973
Belçika	1961	50	2019	57	0.974	0.975	0.972	0.973
Birleşik Krallık	9954	422	13664	1286	0.933	0.959	0.885	0.921
Danimarka	1181	22	1364	40	0.976	0.982	0.967	0.974
Diğer Güney	2451	61	4136	227	0.958	0.976	0.915	0.945
Diğer Kuzey	2492	94	3839	205	0.955	0.964	0.924	0.943
Diğer Batı	1250	22	1717	74	0.969	0.983	0.944	0.963
Diğer Doğu	7182	307	12652	794	0.947	0.959	0.900	0.929
Estonya	177	1	284	3	0.991	0.994	0.983	0.989
Fransa	7573	399	11882	853	0.940	0.950	0.899	0.924
Hırvatistan	525	4	751	21	0.981	0.992	0.962	0.977
Hollanda	2261	86	3378	145	0.961	0.963	0.940	0.940
Letonya	329	3	363	2	0.993	0.991	0.994	0.992
Norveç	823	3	1249	23	0.988	0.996	0.973	0.984
Portekiz	1305	35	2225	58	0.974	0.974	0.966	0.966
Sırbistan	1301	26	1804	59	0.973	0.980	0.957	0.968
Slovenya	283	2	491	8	0.987	0.993	0.973	0.983
Yunanistan	1305	47	2046	84	0.962	0.965	0.940	0.952
İrlanda	776	12	1029	21	0.982	0.985	0.974	0.979
İspanya	6322	169	8204	541	0.953	0.974	0.921	0.947
İsveç	1735	31	2612	122	0.966	0.982	0.934	0.958
İsviçre	1546	49	1406	36	0.972	0.969	0.977	0.973
İtalya	7191	219	10398	550	0.958	0.970	0.929	0.949

TP: (Sınıf=1 Etiket=1) FN: (Sınıf=1 Etiket=0) TN: (Sınıf=0 Etiket=0) FP: (Sınıf=0 Etiket=1)

3.5. MNNB ile seri duygu analizi sonuçları

Model dosyası, seri yöntemde çok büyük boyutlu oluşturulmuştur. Paralel yöntemde, MapReduce işlemi ile veri paralel işlenirken model dosyası daha iyi derlenerek daha iyi bir öznitelik seçimi sağlanmaktadır. Seri modelde ise kuyruklardaki bellek yetersizliği ve data overflow problemi model dosyası için iyi bir öznitelik seçimini engellemiştir. Sonuç olarak seri model dosyasının yeterince iyi derlenememesi sonucunda seri yöntemdeki sınıflama istenilen başarıyı sağlayamamıştır (Çizelge 3.5.).

Seri analiz sonuçlarına bakıldığında verinin genel olarak pozitif sınıflandığı gözlenmiştir. Seri MNNB metodunun uygulandığı veri setleri için elde edilen performans süreleri ve sınıflama sonuçları Çizelge 3.4.'de ayrı ayrı gösterilmiştir. Kümelenme analizi sonuçlarına ise Şekil 3.1.'de yer verilmiştir.

Çizelge 3.4. MNNB model ve test performans sonuçları

Ülke Veri Seti	Performans Süreleri (ms)				Sınıflanan Tweet Sayıları			
	Seri Model	Hadoop Model	Seri Test	Hadoop Test	Seri Pozitif	Seri Negatif	Hadoop Pozitif	Hadoop Negatif
Almanya	28575284	52	4449	4337	29541	360	12232	17669
Avusturya	747850	35	549	1330	1574	1592	1483	1683
Belçika	833307	42	660	1455	1980	2107	2011	2076
Birleşik Krallık	17318823	110	3205	3408	21176	4150	10376	14950
Danimarka	511291	33	524	1341	1373	1234	1203	1404
Diğer Güney	2102072	45	890	1869	6865	10	2512	4363
Diğer Kuzey	2106952	42	929	2035	6300	330	2586	4044
Diğer Batı	362794	52	339	1024	2074	989	1272	1791
Diğer Doğu	12477417	45	2391	3079	20925	10	7489	13446
Estonya	25771	34	120	525	413	52	178	287
Fransa	13369101	62	2468	3201	19831	876	7972	12735
Hırvatistan	108155	51	205	731	1001	300	529	772
Hollanda	1641357	49	812	1661	5547	323	2347	3523
Letonya	47674	36	150	573	374	323	332	365
Norveç	382653	29	332	1076	2013	85	826	1272
Portekiz	655840	51	448	1247	3621	2	1340	2283
Sırbistan	589323	55	417	1193	2763	427	1327	1863
Slovenya	65426	39	124	624	780	4	285	499
Yunanistan	585360	40	415	1480	3015	467	1352	2130
İrlanda	237011	68	267	893	1324	514	788	1050
İspanya	8600638	54	2432	2671	8392	6844	6491	8745
İsveç	1579222	36	793	2849	4418	82	1766	2734
İsviçre	569698	31	564	1139	1349	1688	1595	1442
İtalya	9375264	36	2229	2774	13444	4914	7410	10948

3.6. MNNB ile paralel-seri duygu analizi performans sonuçları

Model dosyasının oluşturulması paralel modelde MapReduce işlemi ile yapılmaktadır. Paralel ve seri modele ait performans ölçümleri Çizelge 3.4.'de gösterildiği gibidir. Seri modelde, model dosyasının oluşması çok büyük verilerde (Almanya örneği) yaklaşık 500000 katı iken küçük verilerde (Estonya örneği) 700 katı kadar olmaktadır. Bu da sınıflama maliyetini ciddi oranda etkilemektedir.

Çizelge 3.5. MNNB Seri ile DA başarımlar istatistikleri

Ülke	TP	FN	TN	FP	Doğruluk	Duyarlılık	Keskinlik	F1
Almanya	12979	16562	360	0	0.446	0.439	1	0.611
Avusturya	1497	77	1589	3	0.975	0.951	0.998	0.974
Belçika	1972	8	2061	46	0.987	0.996	0.977	0.987
Birleşik Krallık	11240	9936	4150	0	0.608	0.531	1	0.694
Danimarka	1221	152	1234	0	0.942	0.890	1	0.941
Diğer Güney	2678	4187	10	0	0.391	0.390	1	0.561
Diğer Kuzey	2697	3603	330	0	0.457	0.428	1	0.600
Diğer Batı	1324	750	989	0	0.755	0.638	1	0.779
Diğer Doğu	7976	12949	10	0	0.382	0.381	1	0.552
Estonya	180	233	52	0	0.499	0.436	1	0.607
Fransa	8426	11405	876	0	0.449	0.425	1	0.596
Hırvatistan	546	455	300	0	0.650	0.546	1	0.706
Hollanda	2406	3141	323	0	0.465	0.434	1	0.605
Letonya	331	43	323	0	0.938	0.885	1	0.939
Norveç	846	1167	85	0	0.444	0.420	1	0.592
Portekiz	1363	2258	2	0	0.377	0.376	1	0.547
Sırbistan	1360	1403	427	0	0.560	0.492	1	0.660
Slovenya	291	489	4	0	0.376	0.373	1	0.543
Yunanistan	1389	1626	467	0	0.533	0.461	1	0.631
İrlanda	797	527	514	0	0.713	0.602	1	0.752
İspanya	6847	1545	6828	16	0.898	0.816	0.998	0.898
İsveç	1857	2561	0	82	0.431	0.420	1	0.592
İsviçre	1349	0	233	1455	0.923	1	0.853	0.921
İtalya	7737	5707	4910	4	0.689	0.576	0.999	0.730

TP: (Sınıf=1 Etiket=1) FN: (Sınıf=1 Etiket=0) TN: (Sınıf=0 Etiket=0) FP: (Sınıf=0 Etiket=1)

4. TARTIŞMA VE SONUÇ

Duygu analizi günümüzde oldukça önem kazanmış bir çalışma alanıdır. İnsanların nesneler, olay ve olgular hakkında ne düşündükleri merak edilen bir araştırma konusudur. Bu çalışmada da duygu analizi için büyük verilerin işlenmesini iyileştirilebilecek yeni bir model önerilmektedir. Önerilen model Hadoop dağıtık dosya sistemi üzerinde modellenerek kurulmuştur. Geliştirilen modelin, geleneksel programlama mimarisi ile seri modeli ve HDDS mimarisi ile paralel modeli kurularak uygulanmıştır. Uygulanan seri ve paralel modelin başarımlar ve performans karşılaştırması yapılmıştır. Seri model ve paralel model, iki aşamalı modeller olup ilk aşamada sözlük tabanlı etiketleme yapılmaktadır ve ikinci aşamada da MNNB sınıflayıcı ile duygu sınıflandırması yapılmaktadır.

Çalışma konusu olarak Avrupa ülkelerindeki Twitter sosyal ağ kullanıcılarının göç-göçmen-mülteci temasındaki duyguları analiz edilmiştir. Twitter mesajlarındaki metinler analiz edilerek her bir ülke için duygu analizi sonuçları çıkarılmıştır. Veri seti olarak yeni ve özgün bir veri seti oluşturularak, 2011-2020 tarihleri arasında gönderilen tweet metinleri çalışmaya konu edinilmiştir. Kullanıcı profillerindeki adres ve lokasyon bilgileri baz alınarak çeşitli filtrelemeler aracılığıyla veriler çekilerek her ülke için özgün bir veri seti oluşturulmuştur.

Sonuç olarak, bu çalışma kapsamında, Hadoop ortamında geliştirilen model ile etiketleme işleminde ciddi bir performans artışı sağlandığı görülmektedir. Sözlükler karşılaştırıldığında AFINN sözlüğü

veri işlenirken hem seri hem paralel modelde SentiWordNet sözlüğüne göre daha hızlı etiketleme yapmaktadır. Yapılan bu çalışma sonucunda, önerdiğimiz model performans ve başarımları olarak başarılı olduğu düşünülmektedir.

Paralel MNNB ve Seri MNNB modeli performans açısından karşılaştırılırsa, paralel model ile çok büyük bir performans kazancı sağlanmıştır. En küçük veri setinden en büyük veri setine doğru 13 kattan 83 kata kadar çıkan bir performans farkı gözlemlenmiştir. Dolayısıyla Hadoop mimarisinin büyük veri işlemede performans kazancı sağladığı bu çalışmada bir kez daha görülmüştür.

Araştırma konumuz kapsamında sadece göç ve ilişkili olduğu kavramların açıklaması ve araştırmaya dâhil edilen Avrupa ülkelerindeki göçmen algısı ele alındığından konunun daha detaylı açıklamasına girilmemiştir. Ancak yine de araştırmamızın uygulama ve analiz bölümünde elde edilen sonuçların konu ile ilgilenen Sosyoloji, Siyaset Bilimi, Uluslararası İlişkiler gibi farklı disiplinlerden gelen araştırmacılara kaynak teşkil edecek verilerin elde edilmesi açısından önemli olduğu düşünülmektedir.

Bu çalışmada, uygulanan tüm metotlardaki duygu analizi sonuçları birbirleri ile tutarlıdır. Hazırlanan veri setlerinin, sözlük tabanlı ve makine öğrenmesi tabanlı sınıflandırma işlemleri sonucunda pozitif olarak etiketlenen tek ülke İsviçre çıkmıştır. Diğer ülkeler ise artan oranlarla negatif olarak ön plana çıkmıştır. Bu sonuç ile ele aldığımız mülteci-göçmen probleminin gerçek bir problem olduğunun, sadece bir algıdan ibaret olmadığının bir kez daha altı çizilmiştir. Dolayısıyla bu çalışmanın, göçmen/ mülteci konusu üzerinde çalışacak diğer araştırmacılara bir katkı sağlayacağı düşünülmektedir. Ayrıca sonuç olarak, bu tez çalışması kapsamında önerilen model ile büyük verilerde DA için performans ve başarımları kazancı sağlanmıştır. Literatürde naif bayes ile yapılan sınıflandırma çalışmalarına bakıldığında başarımları oranı farklı öznitelik çıkarımı ve kullanımına bağlı olarak % 70 ile % 80 aralığında değişmektedir (Ardhana ve Cahyani, 2019; Delizo vd., 2020; Gokalp vd., 2020; Narayanan vd., 2013; Priya, 2020; Sunarya vd., 2019) ve farklı uygulamalar aracılığıyla yapılan naif bayes sınıflama çalışmalarında başarımları oranının % 70-% 85 aralığında sağlandığı gözlenmektedir (Abbas vd., 2019; Hossain vd., 2020; Laksono vd., 2019; Xu vd., 2020). Ayrıca naif bayes sınıflayıcılarda uygulanan farklı optimizasyon yöntemleriyle yapılan çalışmalarda doğrulama başarımları % 89-90'lara kadar bir iyileştirme sağlanmaktadır (Farisi vd., 2019; Rahman ve Akter, 2019; Santoso vd., 2020). Farklı kombine modellerin kullanılması ile ilgili yapılan çalışmalarda MNNB sınıflayıcı başarımları oranının % 70 altına düştüğü belirtilmektedir (Artissa vd., 2019; Asif vd., 2020). Bu çalışma kapsamında geliştirilen dağıtık kombine model ile başarımları oranları ortalaması % 95 üzerinde gerçekleşmektedir. Dolayısıyla öznitelik çıkarımı ve özellik seçimi gibi algoritmaya maliyet getiren ek işlemlerin bu model ile en aza indirilmesi sınıflama başarımları artırılabilir. Tüm kelimeler ele alınarak, her bir kelimenin duygu skoruna etkisinin göz ardı edilmeksizin sınıflandırmaya olan etkisi en yüksek değerde tutularak daha yüksek başarımlı bir sınıflama sağlanmıştır. Ayrıca oluşturulan veri setlerinin literatüre katkı sağlayacağı düşünülmektedir.

TEŞEKKÜRLER

Bu çalışma FDK-2019-7850 Numaralı proje ile Van Yüzüncü Yıl Üniversitesi Bilimsel Araştırma Projeleri Başkanlığı tarafından desteklenmiştir. [*] Sorumlu yazarın “Büyükveri Tabanlı Sistemlerde Duygu Analizine Yönelik Bir Uygulama: Avrupa'daki Twitter Kullanıcılarının Mülteci Algısı” isimli doktora tezinden üretilmiş yayındır.

5. KAYNAKLAR

- Abbas, M.; Memon, K. A. & Jamali, A. A. & Memon & S., Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. *IJCSNS*, **19**(3): 62.
- Agarwal, B. & Mittal, N. (2016). **Prominent feature extraction for sentiment analysis**. (Vol. 2). Springer, India, 21-45.
- Anonim. (2019a). The Four V's of Big Data. <https://www.ibmbigdatahub.com/infographic/four-vs-big-data> IBM, Erişim Tarihi: 10.04.2019.
- Anonim. (2019b). United Nations World Population Prospects 2019, Online Edition. Rev. 1. (Rev.1.). <https://population.un.org/wpp/Download/Standard/Population/> Erişim Tarihi: 07.03.2020
- Anonim. (2019c). United Nations, Department of Economic and Social Affairs, Population Division. <https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates17.asp>. Creative Commons, Erişim Tarihi: 07.03.2020.
- Ardhana, A. & Cahyani, D. (2019). Classification of Javanese Language Level on Articles Using Multinomial Naive Bayes and N-Gram Methods. *Journal of Physics: Conference Series*. 012049.
- Artissa, Y.; Asror, I. & Faraby, S. (2019). Personality Classification based on Facebook status text using Multinomial Naïve Bayes method. *Journal of Physics: Conference Series*. 012003.
- Asif, M.; Ishtiaq, A. & Ahmad, H. & Aljuaid, H. & Shah, J. (2020). Sentiment analysis of extremism in social media from textual information. *Telematics and Informatics*, **48**: 101345.
- Baccianella, S.; Esuli, A. & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *Lrec*. 2200-2204.
- Bowes, D.; Hall, T. & Gray, D. (2014). DConfusion: a technique to allow cross study performance evaluation of fault prediction studies. *Automated Software Engineering*, **21**(2): 287-313.
- Brown, B.; Chui, M. & Manyika, J. (2011). Are you ready for the era of 'big data'. *McKinsey Quarterly*, **4**(1): 24-35.
- Buyya, R.; Calheiros, R. N. & Dastjerdi, A. V. (2016). **Big data: Principles and Paradigms**. Morgan Kaufmann, USA, 3-223.
- Chen, K.; Seshadri, S. & Zhang, L.-J. (2019). Big Data–BigData 2019. San Diego, CA, USA: Springer International Publishing.
- Delizo, J. P. D.; Abisado, M. B. & De Los Trinos, M. I. P. (2020). Philippine Twitter Sentiments during Covid-19 Pandemic using Multinomial Naïve-Bayes. *International Journal*, **9**(1.3).
- Farisi, A. A.; Sibaroni, Y. & Al Faraby, S. (2019). Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier. *Journal of Physics: Conference Series*. 012024.
- Gantz, J. & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, **2007**(2012): 1-16.
- Garrett, T. M. (2020). The Security Apparatus, Federal Magistrate Courts, and Detention Centers as Simulacra: The Effects of Trump's Zero Tolerance Policy on Migrants and Refugees in the Rio Grande Valley. *Politics & Policy*.

- Gerşil, G. S. (2019). Küresel Göç ve Entegrasyon Sorunlarına Yönelik Değerlendirme. **International Mediterranean Social Sciences Congress (MECAS)**. 06.09.2019. 355.
- Gokalp, O.; Tasci, E. & Ugur, A. (2020). A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification. **Expert Systems with Applications**, **146**: 113176.
- Gupta, V. S. & Kohli, S. (2016). Twitter sentiment analysis in healthcare using Hadoop and R. **Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on**. 3766-3772.
- Hossain, E.; Sharif, O. & Hoque, M. M. (2020). Sentiment Polarity Detection on Bengali Book Reviews Using Multinomial Naive Bayes. **arXiv preprint arXiv:2007.02758**.
- Juan, A. & Ney, H. (2002). Reversing and Smoothing the Multinomial Naive Bayes Text Classifier. **PRIS**. 200-212.
- Kemp, S. (2020). Digital around the world in april 2020.
<https://wearesocial.com/blog/2020/04/digital-around-the-world-in-april-2020>. Erişim Tarihi: 01.05.2020
- Kibriya, A. M.; Frank, E. & Pfahringer, B. & Holmes, G. (2004). Multinomial naive bayes for text categorization revisited. **Australasian Joint Conference on Artificial Intelligence**. 488-499.
- Laksono, R. A.; Sungkono, K. R. & Sarno, R. & Wahyuni, C. S. (2019). Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naïve Bayes. **2019 12th International Conference on Information & Communication Technology and System (ICTS)**. 49-54.
- Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. **META group research note**, **6**(70): 1.
- Liu, B. (2012a). Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, **5**(1): 1-167.
- Liu, B. (2012b). **Sentiment analysis and opinion mining**. (Vol. 5), 1-167.
- Narayanan, V.; Arora, I. & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model. **International Conference on Intelligent Data Engineering and Automated Learning**. 194-201.
- Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. **arXiv preprint arXiv:1103.2903**.
- Ohlhorst, F. J. (2012). **Big Data Analytics: Turning Big Data into Big Money**. (Vol. 65). John Wiley & Sons, Canada, 1-30.
- Perruchoud, R. & Redpath, J. (2013). Göç terimleri sözlüğü. **Uluslararası Göç Örgütü, Türkçe Baskı**(31).
- Poria, S.; Hussain, A. & Cambria, E. (2018). **Multimodal Sentiment Analysis**. (Vol. 8). Springer, 1-84.
- Pozzi, F. A.; Fersini, E. & Messina, E. & Liu, B. (2017). Challenges of sentiment analysis in social networks: An overview. In *Sentiment analysis in social networks* (pp. 1-11): Elsevier.

- Priya, K. S. (2020). A Comparative Sentiment Analysis of Sentence Embedding Using Machine Learning Techniques. **2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)**. 493-496.
- Rahman, M. A. & Akter, Y. A. (2019). Topic Classification from Text Using Decision Tree, K-NN and Multinomial Naïve Bayes. **2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)**. 1-4.
- Santoso, H. A.; Rachmawanto, E. H. & Nugraha, A. & Nugroho, A. A. & Basuki, R. S. (2020). Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization. **Telkomnika**, **18**(2): 799-806.
- Satapathy, R.; Cambria, E. & Hussain, A. (2018). **Sentiment analysis in the bio-medical domain: techniques, tools, and applications**. (Vol. 7). Springer,
- Schneider, K. M. (2004). A new feature selection score for multinomial naive Bayes text classification based on KL-divergence. **Proceedings of the ACL Interactive Poster and Demonstration Sessions**.
- Solanki, S.; Verma, S. & Chahar, K. (2020). A Comparative Study of Information Retrieval Using Machine Learning. In *Advances in Computing and Intelligent Systems* (pp. 35-42): Springer.
- Sunarya, P. A.; Refianti, R. & Mutiara, A. B. & Octaviani, W. (2019). Comparison of Accuracy between Convolutional Neural Networks and Naïve Bayes Classifiers in Sentiment Analysis on Twitter.
- Xu, F.; Pan, Z. & Xia, R. (2020). E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework. **Information processing & management**: 102221.