Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması

Hatice Nizam[1], Saliha Sıla Akın[2]

[1]İstanbul Üniversitesi Bilgisayar Mühendisliği Bölümü, İstanbul

[2] ERS Turizm Yazılım Şirketi, Bilgisayar Mühendisi, Antalya

haticenizam@outlook.com,sila.akin@hoteladvisor.net

Özet: Bu çalışmada, makine öğrenmesi yöntemlerinden denetimli öğrenme yaklaşımı kullanılarak sosyal medyada duygu analizi çalışması yapılmıştır. Denetimli öğrenme yaklaşımında özellik olarak tweetlerdeki tüm kelimelerin seçildiği eğiticisiz yöntem kullanılmıştır. Tweetler makine öğrenmesi yöntemlerinden unigram özelliğine göre analiz edilmiştir. Bazı gıda firmalarının çeşitli ürünlerine ait yapılan yorumlardan oluşturulan veri seti Twitter üzerinden el yordamı ile elde edilmiştir. Tweetler pozitif, negatif ve nötr olarak işaretlenerek 3 sınıfta toplanmıştır. Çalışmada pozitif, negatif ve nötr sınıftaki veri dağılımının Weka kütüphanesinde yer alan Naive Bayes (NB), Random Forest (RF), Sequential Minimal Optimization (SMO), Decision Tree (J48) ve 1-Nearest Neighbors (IB1) sınıflandırma algoritmalarının gösterdikleri başarım sonuçlarına etkisi incelenmiştir. Elde edilen deneysel sonuçlarda sınıflar arası uygun dağılım gösteren dengeli veri setinin (B veri seti) dengesiz veri setine (A veri seti) göre daha iyi performans sonuçlarının alındığı gözlemlenmiştir. En iyi performans gösteren sınıflandırma algoritması %72.33 ortalama doğruluk başarı oranıyla SMO olmuştur.

Anahtar Kelimeler: Duygu Analizi, Metin Sınıflandırma, Makine Öğrenmesi, Denetimli Öğrenme Yaklaşımı, Veri Seti Seçimi.

Abstract: In this study, an analysis of sentiment on social media using the supervised learning approach (one of the machine learning methods) is given. In the supervised learning approach, the unsupervised method is used as the characteristic in which all words of tweets are selected. Tweets are analyzed according to unigram property. Data set (created with the comments on several products of some food companies) is acquired manually via Twitter. Tweets are gathered into three groups by ticking up them as positive, negative and neuter. Then, the effect of the data distribution of tweets grouped as positive, negative and neuter are examined according to the success results of the classification algorithms at the Weka Library - Naive Bayes(NB), Random Forest (RF), Sequential Minimal Optimization (SMO), Decision Tree (J48) and 1-Nearest Neighbors (IB1). By experimental results, it is observed that the balanced data set (B data set) having proper distribution among classes give better performance results than that of the unbalanced data set (A data set). The SMO classification algorithm gives the highest performance with 72.33% average accuracy hit ratio.

Key Words: Sentiment Analysis, Text Classification, Machine Learning Supervised Learning Approach, Data Set Selection.

1.GİRİŞ

Duygu analizi (aynı zamanda duygu madenciliği, duygu sınıflandırma, fikir madenciliği, öznellik analizi, eleştiri madenciliği ya da değerlendirme çıkarma ve bazı durumlarda düşünce sınıflandırma) metinde hesaplamalı fikir değerlendirilmesi, duygu ve öznellik ile ilgilidir. Duygu analizi ile belli bir konu veya hedefe göre bir konuşmacı ya da yazarın görüşünü tespit etmek amaçlanır. Görüş yazarın düşünce, fikir ya da değerlendirmesini, duygusal durumunu (yazarın yazma anında nasıl hissettiği) veya amaçlanan duygusal iletişimi (yazar okuyucuyu nasıl etkilemek istemesi) ifade edebilir [10].

Duygu analizi çalışmaları doğal dil işleme, makine öğrenmesi, hesaplamalı dilbilim, sembolik teknikler

gibi yaklaşımları kullanır. Makine öğrenmesi, verilen bir problemi ortamdan edindiği bilgiye göre modelleyen Yapay Zekâ disiplininin bir alt dalıdır. Makine öğrenmesi teknikleri denetimli ve denetimsiz öğrenme metotlarından oluşur. Denetimli öğrenme, önceden gözlemlenmiş ve sonuçları bilinen (etiketlenmiş) verileri kullanarak bu verileri ve sonuçlarını kapsayan bir fonksiyon oluşturmayı amaçlayan makine öğrenimi metodudur. Denetimsiz öğrenme, etiketlenmemiş verideki gizli yapıyı bulma işlemidir. Yani, veriler arasında var olan ama gözle görülmeyen bağıntının açığa çıkarılması işlemidir.

Makine öğrenmesi yönteminin denetimli öğrenme tekniğiyle yapılan birçok çalışma literatürde mevcuttur. Bunlardan duygu analizi alanındaki ilk çalışmalardan biri olan Pang, Lee ve Vaithyanatham tarafından 2002 yılında İngilizce metinler için yaptıkları film yorumlarını analiz çalışması ve buna benzer bir çalışma olan Go, Bhayani ve Huang tarafından 2009 yılında Twitter mesajlarının sınıflandırılması çalışması verilebilir. Her iki çalışmada da metinler pozitif ve negatif olmak üzere 2 sınıfa ayrılmıştır [14] [15].

Bu çalışmada sosyal medyada duygu analizi çalışması yapılmıştır. Sosyal medya aracı olarak Twitter seçilmiştir. Twitter'ın seçilmesinin nedeni popüler olması, erişim kolaylığı ve çeşitliliğidir. Ayrıca tweetlerin 140 karakterle kısıtlı olması veriyi analiz etmede kolaylık sağlamaktadır. Avantajlarının yanında dezavantajları da vardır. Kendine ait bir jargonunun olması ve yazım hatalarıyla sıklıkla karşılaşılıyor olması morfolojik çözümlemeyi zorlaştırmaktadır. Veri seti olarak bazı gıda firmalarının çeşitli ürünlerine ait tweetler kullanılmıştır. Veri seti pozitif, negatif ve nötr olarak üç sınıftan oluşmaktadır. Bu çalışmada, sınıflardaki veri dağılımları nasıl olmalıdır, veri dağılımlarının sınıflandırma algoritmalarının performanslarına bir etkisi var mıdır sorularının cevabı aranmıştır.

Duygu analizi ile ilgili çeşitli çalışmalar yapılmıştır, çalışmamızın ikinci bölümünde yer alan literatür kısmında bunlardan bahsedilmiştir. Üçüncü bölümde veri seti, veri özellikleri, kullanılan algoritmalar, model başarım ölçütleri ve kappa istatistiği ayrıntılı bir şekilde tanıtılmıştır. Dördüncü bölümde sonuçların değerlendirilmesi ve tartışılması, beşinci bölümde sonuç ve son bölümde kaynakçaya yer verilmiştir.

2.LİTERATÜR

Duygu Analizi alanında yapılmış olan ilk temel çalışmalardan biri Pang, Lee ve Vaithyanatham tarafından 2002 yılında yapılan çalışmadır. Bu çalışmada unigram, bi-gram, Part of Speach (POS) ve ikili birleşimler gibi makine öğrenmesi yöntemleri kullanılarak metinler duygusal açıdan pozitif veya negatif olarak sınıflandırılmıştır. Bu çalışma İngilizce metinler için yapılmış ve film yorumları veri seti olarak kullanılmıştır. Veri seti çeşitli makine öğrenmesi yöntemlerinin Naive Bayes, Maximum Entropi ve SVM algoritmalarındaki başarımları elde edilmiştir. Yapılan çalışma sonucunda duygu analizi açısından sınıflandırmada en iyi sonucu unigram özelliğine göre makine öğrenmesi yöntemi vermiştir. Sınıflandırma algoritmalarından en iyi performansı SVM (%82.9) göstermiştir [14].

Go, Bhayani ve Huang, otomatik olarak twitter mesajlarını sınıflandırmak için bir yaklaşım öne sürmüşler ve mesajları pozitif ve negatif olarak 2 sınıfta sınıflandırmışlardır. Çalışmalarındaki amaç, Twitter kullanıcıların ve şirketlerin tweetleri kendi oturumlarından sınıflandırabilmelerini sağlamaktır. Diğer bir deyişle Twitter üzerinde uzaktan denetimli öğrenmeyi kullanarak duygu analizi yapmaktır. 800.000 pozitif tweet, 800.000 negatif tweet olmak üzere 1.600.000 tweetten oluşan veri kümesini incelemişlerdir. Bu çalışmayı İngilizce metinler için yapmışlar ve tweetleri unigram, bigram ve unigram ile bigramı birleştirerek analiz etmişlerdir. Elde ettikleri deneysel sonuçlara göre unigram sonucu %82.2 doğruluk oranıyla SVM algoritması, bi-gram sonucu %81.6 doğruluk oranıyla Naive Bayes algoritması, unigram ile bi-gramın birlikte kullanımı sonucu %83 doğruluk oranıyla Maximum Entropi algoritması en iyi performansı göstermiştir [15].

Çetin ve Amasyalı, Türkçe metinlerde duygu analizinde geleneksel 2 yöntemin ve eğiticili 6 yöntemin performansları NB, RF, SMO, J48 ve IB1 algoritmaları kullanılarak 2 veri kümesi üzerinde karşılaştırılmıştır. Veri kümeleri Telekom sektöründeki A ve B olmak üzere özel iki şirkete ait Twitter gönderilerinden oluşmaktadır. Her iki veri kümesinde de 6000'er örnek bulunmaktadır. Bu kümeler olumlu, olumsuz ve nötr olmak üzere 3 ayrı sınıfa el yordamı ile ayrılmıştır. Sınıf dağılımları eşit olacak şekilde veri kümeleri eşit boyutlu eğitim ve test kümelerine ayrılmıştır. A ve B veri kümeleri üzerinde terim olarak kelime köklerinin ve karakter n-gramları (2,3 ve 4 n-gram) kullanıldığında, 5 algoritmanın 8 terim ağırlıklandırma yöntemi ile elde edilen sonuçlar incelenmiştir. Metin temsilinde karakter n-gramlarının, kelime köklerine göre daha başarılı ve terim ağırlıklandırma da eğiticili yöntemlerin geleneksel eğiticisiz yöntemlere göre daha başarılı

ve içerdikleri az sayıda özellik sayısı sebebiyle daha kolay ve hızlı modellenebildiği sonucu deneysel sonuçlarla elde edilmiştir. En iyi performans gösteren sınıflandırma algoritması SMO'dur ve elde edilen sonuçlar literatürde İngilizce için yapılan çalışmalar ile paralellik göstermiştir [12].

B. İbrahim Sevindi, duygu analizinde Türkçe film yorumlarının duygu kutupları çeşitli yöntemler kullanılarak belirlenmeye çalışılmış ve bu yöntemler karşılaştırılmıştır. Bu tez çalışmasında makine öğrenmesi ve sözlük tabanlı yaklaşımlar kullanılarak karşılaştırılmıştır. Makine öğrenmesi yönteminde kullanılan sınıflandırıcılar C4.5 karar ağacı, KNN, Naive Bayes ve SVM'dir. Kullanılan öznitelikler 1, 2 ve 3 uzunluktaki n-gramlardır. Makine öğrenmesi yaklaşımları için yapılan denemelerde, en iyi sonuç 0,8258 F-skor değeri ile SVM sınıflandırıcısından alınmıştır. Bu sonuca 1 uzunluktaki n-gramların ekleriyle birlikte kullanıldığı ve etkisiz kelimelerin elendiği durumda ulaşılmıştır. Sözlük tabanlı yaklaşımda alınan en iyi sonuç 0,5969 F-skor değeri ile terim skor sınırı kullanılmadan ve terimlerin kutup bilgisi yerine skor bilgisi kullanılarak elde edilmiştir. Sonuç olarak Türkçe film yorumlarının duygu analizinde makine öğrenmesi yöntemleri daha başarılı sonuçlar üretmiştir [16].

3.DENEYLER

3.1.Veri Seti

Bu bölümde gıda sektöründeki farklı firmaların çeşitli ürünlerine ait tweetlerden oluşturulan dengeli ve dengesiz olmak üzere iki veri seti kullanılmıştır. Veri setlerinde bulunan tweetler el yordamı ile pozitif, negatif ve nötr olmak üzere üç sınıfa ayrılmıştır. Birinci veri seti, pozitif sınıfta 1113, negatif sınıfta 277 ve nötr sınıfta 610 veri olmak üzere toplam 2000, ikinci veri seti, pozitif sınıfta 257, negatif sınıfta 277 ve nötr sınıfta 290 veri olmak üzere toplam 824 veriden oluşmaktadır.

3.2.Veri Özellikleri

Bu bölümde, Türkçe tweetlerden oluşan iki veri seti kullanılmıştır. Tüm tweetler küçük harflere ve - ç, ğ, ı, ö, ş, ü - karakterleri - c, g, i, o, s, u - karakterlerine dönüştürülmüştür. Özellik değerlendirme metotlarından terim frekansı (TF) kullanılmıştır. TF (i,j) i. özelliğinin j sınıfında geçme sayısıdır. Bu çalışmada kelimeler özellik olarak alınmıştır.

3.3.Deneylerde Kullanılan Sınıflandırma Algoritmaları

Bütün deneyler 10-katlamalı çapraz geçerleme stratejisi Weka (versiyon 3.6) yazılımı kullanılarak yapılmıştır [17]. Veri setleri üzerinde NB, RF, SMO, J48 ve IB1 sınıflandırma algoritmaları uygulanmıştır.

3.3.1.Naive Bayes Sınıflandırma Algoritması

Naive Bayes sınıflandırıcı, olasılı sınıflandırma tekniklerinin en kısıtlayıcı uç yelpazesinde temsil edilir [1]. Sınıflandırılması gereken sınıflar (kümeler) ve örnek verilerin hangi sınıflara ait olduğu bellidir. Metin kategorizasyonu için çok etkili olduğu kanıtlanmıştır [2]. Bir Bayes yaklaşımı olarak, n boyutlu uzayda tanımlı olan X vektörü (x_1,..,x_2), m adet sınıf bulunan C_k (C_1,.,C_n) veri kümesinde son olasılığı maksimize eden bir sınıf etiketi C arar.

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i) \quad (1)$$

3.3.2. Random Forest Sınıflandırma Algoritması

Breiman tek bir karar ağacı üretmek yerine çok sayıda ve çok değişkenli ağaçların her birinin farklı eğitim kümeleriyle eğitilmesi sonucu ortaya çıkan kararların birleştirilmesini önerir. Bir sınıflandırıcı yerine birden çok sınıflandırıcı üreten ve sonrasında onların tahminlerinden alınan oylar ile yeni veriyi sınıflandıran öğrenme algoritmasıdır. Büyük veri tabanlarında eşsiz olarak çalışır ve dengesiz veri seti sınıfında hata dengeleme yöntemlerine sahiptir. Kaybolan verilerin büyük olasılığında doğruluk korunur ve kaybolan verilerin tahmin edilmesinde etkili bir metottur [3] [4].

3.3.3. SMO Sınıflandırma Algoritması

SMO, herhangi bir ekstra matris depolama olmadan ve tüm sayısal QP (Quatratic Programming) optimizasyon adımları kullanmadan SVM QP sorununu hızlı bir şekilde çözer [5]. Bu uygulama global olarak bütün kayıp değerleri yenisiyle değiştirir ve nominal öznitelikleri ikili olanlara dönüştürür. Ayrıca bütün öznitelikleri (attributes) önceden tanımlanmış değerlerle (default) normalize eder.

3.3.4. J48 Sınıflandırma Algoritması

J48, J. Ross Quinlan tarafından geliştirilen çok popüler C4.5 algoritması temeline dayanan bir karar ağacı algoritmasıdır. Karar ağaçları bir makine öğrenmesi algoritmasından bilgi temsil etmede klasik bir yoldur ve veri yapılarını ifade etmekte güçlü ve hızlı bir yol sunar. Bu algoritma verileri özyinelemeli olarak sınıflandırır. Bu işlem eğitim

verilerinin maksimum doğruluğunu sağlar ama verilerin sadece belirli davranış özelliklerini tanımlayan aşırı kurallar oluşturabilir [7].

3.3.5.IB1 Sınıflandırma Algoritması

IB1, en yakın komşu sınıflandırıcısı kullanır. Sınıflandırılmak istenen örneğe en yakın örneği bulmak için standartlaştırılmış Öklid mesafesini kullanır ve bu örnekle aynı sınıfın kestirimini yapar. Eğer birden çok örnek test örneğine aynı (en küçük) mesafeye sahipse, ilk bulunan kullanılır [8]. Öğrenme kümesindeki örneklere (y_i), sınıflandırılmak istenen örneğe (x_i) olan uzaklıklarına göre ağırlıklar verilerek Öklid mesafesi hesaplanır.

$$Similarity\ (x,y) = -\sqrt{\sum_{i=1}^{n} f(x_i, y_i)} \qquad (2)$$

3.4.Sınıflandırma Algoritmalarının Karşılaştırılmasında Kullanılan Kriterler

3.4.1.Model Başarım Ölçütleri

3.4.1.1. Doğruluk–Hata Oranı (Accuracy-Error Rate)

Model başarımının ölçülmesinde kullanılan en popüler ve basit yöntem, modele ait doğruluk oranıdır. Doğru sınıflandırılmış örnek sayısının (TP +TN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır. Hata oranı ise bu değerin 1'e tamlayanıdır. Diğer bir ifadeyle yanlış sınıflandırılmış örnek sayısının (FP+FN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır.

$$\text{Doğruluk} = \frac{(TP+TN)}{(TP+FP+FN+TN)} \qquad (3)$$

$$\text{Hata Oranı} = \frac{(FP+FN)}{(TP+FP+FN+TN)} \qquad (4)$$

3.4.1.2.Kesinlik (Precision)

Kesinlik, sınıfı 1 olarak tahmin edilmiş True Pozitif (TP) örnek sayısının, sınıfı 1 olarak tahmin edilmiş tüm örnek sayısına (TP+FP) oranıdır.

$$\text{Kesinlik} = \frac{TP}{(TP+FP)} \qquad (5)$$

3.4.1.3. Duyarlılık (Recall)

Doğru sınıflandırılmış pozitif örnek (TP) sayısının, toplam pozitif örnek sayısına (TP+FN) oranıdır.

$$\text{Duyarlılık} = \frac{TP}{(TP+FN)} \qquad (6)$$

3.4.1.4. F-Ölçütü (F-Measure)

Kesinlik ve duyarlılık ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli değildir. Her iki ölçütü beraber değerlendirmek daha doğru sonuçlar verir. Bunun için f-ölçütü (F) tanımlanmıştır. F-ölçütü, kesinlik (K) ve duyarlılığın (D) harmonik ortalamasıdır.

$$F = \frac{2DK}{(D+K)} \qquad (7)$$

3.4.2.Kappa İstatistiği

Gözlemciler arası varyasyon, iki veya daha fazla bağımsız gözlemciler tarafından aynı şeyi değerlendiriyor olduğu her durumda ölçülebilir [9]. Kappa katsayısı -1 ile +1 arasında değişir. Tam uyum söz konusu olduğunda K=1 olur. Gözlenen uyumun şansa bağlı uyuma eşit ya da ondan büyük olması durumunda K≥0 iken, gözlenen uyumun şansa bağlı uyumundan küçük olması durumunda K<0 olur. Kappa katsayısının yorumlanabilir aralığı 0 ile +1 arasında olup, negatif (K<0) değerlerinin güvenirlik açısından bir anlamı yoktur. 0.4 üzerinde bir kappa skoru makul bir anlaşmayı ifade eder [11]. Kappa değeri şu şekilde hesaplanır:

$$K = \frac{(P_o - P_c)}{(1-P_c)} \qquad (8)$$

(P_o kabul edilen oran, P_c kabul edilmesi beklenen oran)

4.Sonuçların Değerlendirilmesi ve Tartışılması

4.1 Deneysel Sonuçlar

Veri setinde kullanılacak olan tweetler makine öğrenmesi yönteminin denetimli öğrenme yaklaşımı kullanılarak Weka kütüphanesinde yer alan NB, RF, SMO, J48 ve IB1 sınıflandırma algoritmalarıyla model oluşturulmuştur. Test setinin içerdiği tweetler 3. bölümde anlatılan sınıflandırma algoritmaların karşılaştırılmasında kullanılan model başarım ölçütleri ve kappa istatistiği sonuçlarına göre sınıflandırma algoritmalarının başarımları ölçülmüştür.

Pozitif sınıfta 1113, negatif sınıfta 277 ve nötr sınıfta 610 veri olmak üzere toplam 2000 tweetten oluşturulmuş olan dengesiz veri seti üzerinde sınıflandırma algoritmalarının başarımları Tablo 1'de görülmektedir:

Tablo 1: Dengesiz veri seti (A veri seti) (A (Accuracy), P (Precision), R (Recall), F (F-Measure), K (Kappa Statistic) değerlerini temsil etmektedir)

Classifier	A (%)	P	R	F	K
Naive Bayes (NB)	59.30	0.58	0.59	0.58	0.28
Random Forest (RF)	60.50	0.60	0.60	0.55	0.21
Sequental Minimal Optimization (SMO)	66.40	0.65	0.66	0.65	0.38
Decision Tree (J48)	55.70	0.52	0.55	0.52	0.15
1-nearlest neighbors (IB1)	53.40	0.51	0.53	0.48	0.07

Doğruluk ölçütüne göre en iyi sonucu SMO algoritması göstermiş olup diğer algoritmalar bu ölçüte göre sırasıyla RF, NB, J48 ve IB1 şeklinde sıralanabilir.

Kesinlik ölçütüne göre en iyi SMO algoritması göstermiş olup diğer algoritmalar bu ölçüte göre sırasıyla RF, NB, J48 ve IB1 şeklinde sıralanabilir.

Duyarlılık ölçütüne göre en iyi sonucu SMO algoritması göstermiş olup diğer algoritmalar bu ölçüte göre sırasıyla RF, NB, J48 ve IB1 şeklinde sıralanabilir.

Kesinlik ve duyarlılık ölçütlerini beraber değerlendirmek için, her iki değerin harmonik ortalaması olan F-ölçütüne göre en iyi sonucu (SMO algoritması göstermiş olup diğer algoritmalar bu ölçüte göre sırasıyla NB, RF, J48 ve IB1 şeklinde sıralanabilir.

Deneysel sonuçlardaki başarımın düşük çıkmasının nedeni sınıflardaki veri dağılımının dengesizliğinden kaynaklanıyor olmasıdır. Pozitif sınıfta bulunan örnek sayısının negatif ve nötr sınıflarda bulunan örnek sayına oranı %55'tir. Bu veri madenciliği açısından uygun bir dağılım değildir [13]. Sonuçları istatistiksel yollarla bir öğrenme algoritması kullanarak elde etmek istersek kappa istatistiği uygun bir ölçüt olarak görülür. Burada kappa sonuçları gözleme dayalı uyumun şansa bağlı olarak gerçekleştiğini göstermektedir [11]. Kappa katsayısı 1 değerine yaklaştıkça gözlenen uyumun şans eseri gerçekleşmediğini ifade eder. Veriler arası dengesizliği ortadan kaldırdığımızda sonuçlar Tablo 2'de görülmektedir:

Tablo 2: Dengeli veri seti (B veri seti) (A (Accuracy), P (Precision), R (Recall), F (F-Measure), K (Kappa Statistic) değerlerini temsil etmektedir)

Classifier	A (%)	P	R	F	K
Naive Bayes (NB)	66.38	0.66	0.66	0.66	0.49
Random Forest (RF)	61.70	0.64	0.61	0.61	0.41
Sequental Minimal Optimization (SMO)	72.33	0.73	0.72	0.72	0.58
Decision Tree (J48)	65.16	0.65	0.65	0.65	0.47
1-nearlest neighbors (IB1)	51.09	0.64	0.51	0.45	0.26

Doğruluk ölçütüne göre en iyi sonucu SMO algoritması göstermiş olup diğer algoritmalar bu ölçüte göre sırasıyla NB, J48, RF ve IB1 şeklinde sıralanabilir.

Kesinlik ölçütüne göre en iyi sonucu SMO algoritması göstermiş olup diğer algoritmalar bu ölçüte göre sırasıyla NB, J48, IB1 ve RF şeklinde sıralanabilir.

Duyarlılık ölçütüne göre en iyi sonucu SMO algoritması göstermiş olup diğer algoritmalar bu ölçüte göre sırasıyla NB, J48, RF ve IB1 şeklinde sıralanabilir.

Kesinlik ve duyarlılık ölçütlerini beraber değerlendirmek için, her iki değerin harmonik ortalaması olan F-ölçütüne göre en iyi sonucu SMO algoritması göstermiş olup diğer algoritmalar bu ölçüte göre sırasıyla NB, J48, RF ve IB1 şeklinde sıralanabilir.

Genel olarak sonuçları değerlendirdiğimizde daha önceden elde ettiğimiz sonuçlara (Tablo 1) göre model başarım ölçütlerinin sonuçlarında artış meydana gelmiştir. En iyi performans gösteren sınıflandırma algoritması %72.33 ortalama doğruluk başarı oranıyla SMO'dur. IB1 algoritması dışında diğer algoritmalarda kappa istatistiği sonuçları gözlenen uyumun şans eseri gerçekleşmediği göstermektedir [11].

5.Sonuç

Bu çalışmada bazı gıda firmalarının çeşitli ürünlerine A ve B veri setlerinde duygu analizi çalışması yapılmıştır. Veri setleri üzerinde makine öğrenmesi tekniği ve makine öğrenmesi yöntemlerinden denetimli öğrenme yaklaşımı kullanılmıştır. A ve B veri seti üzerinde sınıflandırma algoritmalarının performans

karşılaştırılması model başarım ölçütleri ve kappa istatistik değerlerine göre yapılmıştır. Weka kütüphanesinde yer alan NB, RF, SMO, J48 ve IB1 sınıflandırma algoritmaları kullanılmıştır.

Model başarım ölçütleri ve kappa istatistiği sonuçları incelendiğinde B veri seti, A veri setine göre daha iyi performans göstermiştir. Sınıflardaki veri dağılımlarının sınıflandırma algoritmaları üzerindeki başarımını etkilediği görülmüştür. En iyi performansı gösteren sınıflandırma algoritması %72.33 ortalama doğruluk başarı oranıyla SMO'dur.

Bu çalışmada veri setinde geçen tüm kelimeler özellik olarak kullanılmış ve sınıflandırma algoritmaları üzerindeki performansları incelenmiştir. Tüm kelimelerin özellik olarak kullanılması boyut fazlalığını artırır. Pozitif, negatif ve nötr sınıflarında sadece bir kez geçen kelimeler veya her sınıfta eşit sayıda geçen kelimeler ayırt edici özellik olarak kullanılamaz. Bir kelimenin ayırt edici özellik olarak kullanılabilmesi için o kelimenin bulunduğu sınıftaki frekansının yüksek diğer sınıflardaki frekansının düşük olması gerekmektedir. Çetin ve Amasyalı, her bir kelimenin ağırlıklarının hesaplanıp sınıflardaki geçme sayısının hesaba katıldığı eğiticili terim ağırlıklandırma yönteminin eğiticisiz yöntemlere göre daha başarılı ve içerdikleri az sayıda özellik sayısı sebebiyle daha kolay ve hızlı modellenebildiği sonucunu elde etmişlerdir [12].

Bir sonraki çalışma olarak eğiticili terim ağırlıklandırma yönteminin yapılması planlanmaktadır.

6.Kaynakça

[1] Mehran Sahami (1996), Learning Limited Dependence Bayesian Classifiers.

[2] Dai, Wenyuan, et al. "Transfering naive bayes classifiers for text classification." Proceedings of the national conference on artificial intelligence. London; AAAI Press; MIT Press;1999, 2007.

[3] Leo Breiman and Adele Cutler, Random Forests, 2005.

[4] Leo Breiman,Machine Learning, 45, 5–32, 2001, Random Forests.

[5] John C. Platt (1998), Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.

[6] J.R, QUINLAN, Machine Learning 1: 81-106, 1986, Induction of Decision Trees.

[7] Laveena Sehgal, Neeraj Mohan, and Dr. Parvinder S. Sandhu (2012), Prediction of Function Based Software Using Decision Tree Approach.

[8] Aha, D. ve Kibler, D. (1991), "Instance-based learning algorithms", Machine Learning, vol. 6, Issue no. 1, January 1991.

[9] Anthony J. Viera, MD; Joanne M. Garrett, PhD (2005),Understanding Interobserver Agreement:The Kappa Statistic.

[10] Schrauwen, Sarah. "Machine learning approaches to sentiment analysis using the dutch netlog corpus." Machine Learning Approaches to Sentiment Analysis Using the Dutch Netlog Corpus (Antwerp, Belgium, 2010), CLiPS Technical Report Series, Computational Linguistics & Psycholinguistics (2010).

[11] Landis, J. Richard, and Gary G. Koch. "The measurement of observer agreement for categorical data." biometrics 33.1 (1977): 159-174.

[12] Cetin, M., and M. F. Amasyali. "Supervised and traditional term weighting methods for sentiment analysis." Signal Processing and Communications Applications Conference (SIU), 2013 21st. IEEE, 2013.

[13] Kılıçaslan, Yılmaz, Edip Serdar Güner, and Savaş Yıldırım. "Learning-based pronoun resolution for Turkish with a comparative evaluation." Computer Speech & Language 23.3 (2009): 311-331.

[14] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[15] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford (2009): 1-12.

[16] B.İbrahim Sevindi, "Türkçe Metinlerde Denetimli ve Sözlük Tabanlı Duygu Analizi Yaklaşımlarının Karşılaştırılması" Yüksek Lisans Tezi, 2013.

[17] www.cs.waikato.ac.nz/ml/weka/downloading/