Makine Öğrenmesi İle Duygu Analizinde Veri Seti Performansı

Hatice NİZAM İstanbul Üniversitesi Bilgisayar Mühendisliği Bölümü

haticenizam@outlook.com

Saliha Sıla AKIN ERS Turizm Yazılım Şirketi, Bilgisayar Mühendisi

sila.akin@hoteladvisor.net

Sunum Planı

- Duygu analizi nedir ve neden ihtiyaç duyulur?
- Literatürdeki çalışmalar
- Yapılan çalışma
- Deneyler
- Deneysel sonuçlar
- Tartışma ve Sonuç

Duygu Analizi

- Duygu analizi, duygu ve öznellikle ilgili hesaplamalı fikir değerlendirmesinin yapıldığı bir alandır.

- Belirli bir konu veya hedefin özelliğine göre metinler olumlu, olumsuz ya da tarafsız içeriğe sahip olup olmadığına göre analiz edilir.

Duygu Analizine Neden İhtiyaç Duyulur?

- İnternet'teki verilerin hızlı bir şekilde artması bir konu veya hedefi manuel olarak takip etmeyi imkansız hale getirmektedir.

- Günümüzde sosyal medya platformlarının popülerliği giderek artmaktadır. İnternet kullanımının hızlı bir şekilde artmasıyla sosyal medyayı takip eden kişiler herhangi bir konu hakkında görüşlerini bu platformlar aracılığıyla duyururlar.

- Medya takibi yapan kişi veya kurumlar metinleri pozitif, negatif veya nötr olarak sınıflandırmak için duygu analizine ihtiyaç duymaktadırlar.

Literatürdeki Bazı Çalışmalar

- Duygu analizinde günümüze kadar yapılan çalışmalar genellikle İngilizce metinler üzerinde pozitif ve negatif olmak üzere iki veya üç sınıfta incelemeler yapıldığı gözlenmektedir.
- Bunlara örnek;

	Kullanılan Yöntemler	Konu	Sınıf Sayısı	Alınan Sonuçlar
Pang, Lee & Vaithyanatham (2002)	• Uni-gram • Bi-gram • POS • İkili kombinasyonları	İngilizce Film Yorumları	2 Sınıf (Pozitif ve Negatif)	%82,9 SVM
Go, Bhayani & Huang (2009)	• Uni-gram • bi-gram	İngilizce Twitter Mesajları	2 Sınıf (Pozitif ve Negatif)	%83 Maximum Entropi
Çetin ve Amasyalı (2013)	• N-gramlar • Kelime Kökleri	Türkçe Twitter Mesajları	3 Sınıf (Pozitif, Negatif ve Nötr)	SMO
B. İbrahim Sevindi (2013)	• Makine Öğrenmesi • Sözcük Tabanlı	Türkçe Film Yorumları	2 Sınıf (Pozitif ve Negatif)	SVM

Yapılan Çalışma

- Türkçe metinler için yapılmış olan duygu analizi çalışmalarında elde edilen başarılar, İngilizce metinlerinkine göre düşüktür.
- Sınıflardaki veri dağılımlarının sınıflandırma algoritmalarındaki başarı oranlarına etkisinin olup olmadığını cevaplayan bir araştırma yoktur.
- Yaptığımız çalışmada, Türkçe metinlerin duygu analizinde nasıl bir performans gösterdiği incelenmiş, sınıflardaki veri dağılımları nasıl olmalı ve veri dağılımlarının sınıflandırma algoritmalarının performanslarına bir etkisinin olup olmadığı sorularının cevabı aranarak literatüre katkıda bulunmak amaçlanmıştır.
- Çalışmamızda sosyal medya aracı olan Twitter seçilmiş ve bazı gıda firmalarının çeşitli ürünlerine ait yapılan yorumlar üç sınıfa manuel olarak ayrılarak analiz edilmiştir.

Deneyler

Kullanılan Veri Seti:

- Gıda sektöründeki farklı firmaların çeşitli ürünlerine ait tweetlerden oluşturulan dengeli ve dengesiz olmak üzere iki veri seti kullanılmıştır.
- Veri setlerinde bulunan tweetler manuel olarak **Tablo 1**'de görüldüğü gibi pozitif, negatif ve nötr olmak üzere üç sınıfa ayrılmıştır:

	A Veri Seti (Dengesiz)	B Veri Seti (Dengeli)
Pozitif	1113	257
Negatif	277	277
Nötr	610	290

Tablo 1: Veri Setleri

Deneyler

Veri Özellikleri:

- Tweetlerde yer alan tüm harfler 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.

- Yapılan çalışmamızda her bir kelime özellik olarak alınmıştır.

Deneyler

- Duygu analizi çalışmaları doğal dil işleme, makine öğrenmesi, hesaplamalı dilbilim, sembolik teknikler gibi yaklaşımları kullanır. Yaptığımız çalışmada, makine öğrenmesi yönteminin denetimli öğrenme tekniği kullanılmıştır.
- Bütün deneyler 10-katlamalı çapraz geçerleme stratejisi ile Weka (versiyon 3.6) yazılımı kullanılarak yapılmıştır.

Kullanılan Sınıflandırma Algoritmaları

- Naive Bayes (NB),
- Random Forest (RF),
- Sequential Minimal Optimization (SMO),
- Decision Tree (J48),
- 1-Nearest Neighbors (IB1)

Deneyler

Sınıflandırma Algoritmalarının Karşılaştırılmasında Kullanılan Kriterler

1. Model Başarım Ölçütleri

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

1.2. Kesinlik (Precision)

1.3. Duyarlılık (Recall)

1.4. F-Ölçütü (F-Measure)

2. Kappa İstatistiği

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

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

Kappa aralığı	Gözlemlenen Uyum
<0	Hiç uyuşma olmaması
0.0 - 0.20	Önemsiz uyuşma olması
0.21 - 0.40	Orta derecede uyuşma olması
0.41 - 0.60	Ekseriyetle uyuşma olması
0.61 - 0.80	Önemli derecede uyuşma olması
0.81 - 1.00	Nereeyse mükemmel uyuşma olması

Deneysel Sonuçlar

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

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

Deneysel Sonuçlar

- A veri setinde (dengesiz veri seti) 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 (Kılıçaslan, Güner, Yıldırım, 2009).

- 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 (Landis, J. Richard, Gary G. Koch, 1977). Kappa katsayısı 1 değerine yaklaştıkça gözlenen uyumun şans eseri gerçekleşmediğini ifade eder.

- Veriler arası dengesizliği ortadan kaldırdığımızda sınıflandırma algoritmalarındaki model başarım ölçütleri ve kappa istatistiği sonuçlarının arttığı gözlemlenmiştir.

Tartışma ve Sonuç

- Model başarım ölçütleri ve kappa istatistiği sonuçları incelendiğinde dengeli veri seti, dengesiz veri setine göre daha iyi performans göstermiştir.
- Her iki veri setinde de en iyi performansı gösteren sınıflandırma algoritması SMO'dur ve dengeli veri setinde %72.33 ortalama doğruluk başarı oranı göstermiştir.
- Sınıflardaki veri dağılımlarının sınıflandırma algoritmaları üzerindeki başarımı etkilediği görülmüştür.

Tartışma ve Sonuç

- Tüm kelimelerin özellik olarak kullanılması boyut fazlalığını artırır. Pozitif, negatif ve nötr sınıflarda 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.

- Bir sonraki çalışma olarak eğiticili terim ağırlıklandırma yönteminin yapılması planlanmaktadır.

TEŞEKKÜRLER...