Destek Vektör Makinesi Yöntemi ile Sosyal Medya Verileri Üzerine Bir Duygu Çözümleme Uygulaması

Evrim Kasaba[1], Engin Yıldıztepe[1]
[1]Dokuz Eylül Üniversitesi, İstatistik Bölümü, İzmir
evrimkasaba@gmail.com , engin.yildiztepe@deu.edu.tr

Özet: Duygu çözümleme, doğal dil işleme, istatistiksel öğrenme ve metin madenciliği yöntemlerinin kullanıldığı, popüler bir çalışma alanıdır. "Düşünce Çözümleme" ve "Fikir Madenciliği" olarak da isimlendirilmektedir**.** Duygu analizi bir metin sınıflandırma problemi olup popülerliği ve ticari getirileri sebebiyle günümüzde ilgi gören bir çalışma alanıdır. Son yıllarda sosyal ağlarda, web bloglarında, alışveriş sitelerinde belirtilen yorumların otomatik sınıflandırılması ihtiyacı bu alana olan ilgiyi arttırmıştır. Ticari şirketler, yapımcılar ve siyasiler gibi birçok kuruluş-kişi için isimlerinin ve olayların insanlarda hangi duyguyu oluşturduğu her zaman önemli bir bilgidir. İnternetin ve sosyal medyanın yaygınlaşmasından önce bu bilgiyi elde etmek oldukça zor ve masraflıydı. Ancak günümüzde insanların duygu ve düşüncelerini paylaşabildiği platformlar bu alandaki araştırmalar için önemli bir bilgi kaynağı haline gelmiştir. Duygu çözümlemede temel iki yaklaşım bulunmaktadır; sözlük tabanlı yaklaşım ve istatistiksel yaklaşım. Sözlük tabanlı yaklaşımlar, duygu çözümleme işlemlerinde anlamsal bir sözlük veritabanını kullanan yarı denetimli yaklaşımlardır. İstatistiksel veya makine öğrenmesi yaklaşımları ise etiketli eğitim verisi üzerinden öğrenen denetimli yöntemlerdir. Destek vektör makinesi de bu yöntemlerden biridir. Bu çalışmada, duygu çözümleme ve güncel kullanım alanları hakkında bilgi verilmiş ve destek vektör makinesi yöntemi kullanılarak yapılan uygulama tartışılmıştır. Uygulamada, Twitter verileri kullanılmıştır. Çalışma R istatistiksel programlama dili kullanılarak gerçekleştirilmiştir.

Anahtar Sözcükler: Duygu çözümleme, Destek vektör makinesi, Twitter

1. Duygu Çözümleme

Duygu analizi bir metin sınıflandırma problemi olup popülerliği ve ticari getirileri sebebiyle günümüzde ilgi gören bir çalışma alanıdır. Ticari şirketler, yapımcılar ve siyasiler gibi birçok kuruluş-kişi için isimlerinin ve olayların insanlarda hangi duyguyu oluşturduğu her zaman önemli bir bilgidir. Daha önceleri bu bilgileri elde etmek için anket, dilek ve şikayet kutuları gibi zaman alan ve masraflı yöntemler kullanılmaktaydı. Gelişen teknoloji ve yaygınlaşan internet (sosyal medya) kullanımıyla, bu bilgiler sosyal medya analizleri sayesinde elde edilebilmektedir. İnsanların düşüncelerini internet üzerinden herkes tarafından erişilebilen bir şekilde paylaşması sayesinde, sosyal medya, birçok kuruluş-kişi için önemli bir veri kaynağı haline gelmiştir.

Duygu çözümleme, kişilerin olaylar, hizmetler, ürünler, kurumlar hakkındaki yorumlarına göre duygu ve düşüncelerini belirlemeye çalışır. Duygu çözümlemede genellikle metinler olumlu, olumsuz veya nötr (yansız) olarak sınıflandırılır. Duygu çözümlemede temel iki yaklaşım bulunmaktadır. Sözlük tabanlı yaklaşımlar, duygu çözümleme işlemlerinde anlamsal bir sözlük veritabanını kullanan yarı denetimli yaklaşımlardır. İstatistiksel veya makine öğrenmesi yaklaşımları ise etiketli eğitim verisi üzerinden öğrenen denetimli yöntemlerdir. Destek vektör makinesi de bu yöntemlerden biridir.

2. Destek Vektör Makinesi

Sınıflandırma işlemi, benzer özellikteki nesnelerin önceden belirlenmiş alt gruplara ayrılması işlemidir. Destek vektör makinesi (DVM) (Support Vector Machine-SVM), sınıflandırma konusunda kullanılan oldukça basit ve etkili yöntemlerden birisidir. Amacı sınıfları birbirinden ayıracak optimal ayırma hiperdüzleminin elde edilmesidir. Başka bir ifadeyle, farklı sınıflara ait destek

vektörleri arasındaki uzaklığı maksimize etmektir. DVM güçlü istatistiksel teoriler üzerine inşa edilmiş bir makine öğrenmesi yöntemidir. İlk kez 1995 yılında Vapnik tarafından sınıflandırma ve regresyon tipi problem çözümleri için önerilmiştir. (Vapnik, 1995). Destek Vektör Makineleri iki durum için ele alınabilir; doğrusal destek vektör makineleri ve doğrusal olmayan destek vektör makineleri.

3. Uygulama

Bu çalışmada sosyal medya verileri üzerinde bir duygu analizi çalışması yapılmıştır. Uygulamada, mikro blog sitesi olarak tanımlanan Twitter verileri kullanılmıştır. Twitter'da cümlelerin 140 karakterle sınırlandırılması hem avantaj hem de dezavantaj sağlamaktadır. Karakter sınırı mesaja çok uzun metinler yazılmasını engellemekle kolaylık sağlamaktadır ancak bu durum mesajdan elde edilebilecek veri miktarını da azaltmaktadır. Ayrıca Twitter'da kullanılan bir jargonun olması, kısaltmalar ve yazım hataları çözümleme çalışmalarını zorlaştırmaktadır. Çalışmada, belirlenen etiket için Twitter API kullanılarak elde edilen mesajlar temizlenmiş ve kelimeler köklerine ayrılmıştır. Rasgele seçilen eğitim kümesi için sınıf etiketleri belirlenmiştir. Daha sonra DVM yöntemi ile sınıflandırma modeli elde edilmiştir. Kurulan model test verileri ile denenmiş ve sonuçlar tartışılmıştır. Çalışmada R istatistiksel programlama dili kullanılmıştır.

Kaynaklar

[1] Ayhan, S., Erdoğmuş, Ş., "Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü İçin Çekirdek Fonksiyonu Seçimi", **Eskişehir Osmangazi Üniversitesi İİBF Dergisi**, 9(1),(2004).

[2] Jurka, P., T., Collingwood, L., Boydstun, E., A., Grossman, E., Atteveldt, W., "RTextTools: A Supervised Learning Package for Text Classification"**, The R Journal**, 5(1), (2013).

[3] Kavzoğlu, T., Çölkesen, I., "Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi", **Harita Dergisi**, 144,(2010).

[4] Uçan, A., "Otomatik Duygu Sözlüğü Çevirimi Ve Duygu Analizinde Kullanımı", Yüksek Lisans Tezi, (2014).

[5] Vapnik, V., "The Nature of Statistical Learning Theory", Springer, New York (1995).

[6] Vapnik, V., Cortes, C.., "Support-Vector Networks**", Machine Learning**, 20, (1995).