ABSTRACT

SENTIMENT CLASSIFICATION WITH MULTIPLE CLASSIFIER SYSTEMS FOR TURKISH LANGUAGE

Mehmet NANĞIR - 2013

Sentiment analysis is to obtain individual information and inferences using natural language processing methods from raw data sources.

User reviews are valuable resource for commercial, social, political analysis and text mining. Consumer reviews, book reviews, social media analysis, political research, news reviews, movie reviews and stock market predictions can be given as examples fort he research and analysis topics in sentiment analysis.

With the explosive growth of social media and internet, the value of personal reviews is increased. The effect of the internet for the commerce changed the brand-consumer relationship significantly. Positive and negative experiences are not only between brand and consumers, but also they spread rapidly to the social environment. Analysis and evaluation of this data began to offer more important benefits for individuals and companies.

There are several studies in this area in literature for English language. This field was not investigated so much for Turkish language and there is not enough number of research studies. According to our literature survey, we only reached two studies for Turkish language. First study used a machine learning algorithm for a specific type and focused on a domain including single dataset [19]. In the second study, several machine learning algorithms are tested on three datasets from different domains. 85% accuracy was obtained with Naive Bayes machine learning algorithm [20].

In this thesis, multiple classifier machine learning algorithms have been applied for Turkish Language on different domains. As distinct from existing studies, a novel multiple classifier system (MCS) was designed by using three high-performance machine learning algorithms all together. In addition to this novel MCS approach, performance was increased by performing parameter optimization of machine learning algorithms. With this new approach, previous accuracy rate was increased to 86.13% accuracy. This accuracy rate revealed that this approach improves the performance and can be used in many studies.

Key Words: Sentiment Classification, Turkish, Naive Bayes, Support Vector Machines, Decision Tree, Multiple Classifier Systems, Parameter Optimization, Machine Learning, Natural Language Processing, Data Mining, Weka

ÖZET

TÜRK DİLİ İÇİN ÇOKLU SINIFLANDIRICI YÖNTEMLER İLE DUYGU SINIFLANDIRMA

Mehmet NANĞIR - 2013

Duygu analizi, doğal dil işleme yöntemlerinin kullanılarak kaynaklarda yer alan ham veriden kişisel bilgi ve çıkarımların elde edilmesidir.

Kullanıcı yorumları; ticari, sosyal, siyasi analizler ve metin madenciliği için çok değerli bir kaynaktır. Duygu analizinin araştırma ve inceleme alanına giren konulara; tüketici yorumları, kitap yorumları, sosyal medya analizi, siyasi araştırmalar, haber yorumları, film değerlendirmeleri ve borsa tahminleri örnek olarak verilebilir.

Son zamanlarda internet ve sosyal medya kullanımının artması, kişisel değerlendirmeleri önemli bir konuma getirdi. Internet kullanımının ticarete etkisi, marka-tüketici ilişkisini de önemli ölçüde değiştirdi. Olumlu ve olumsuz deneyimler artık marka ile tüketici arasında kalmıyor, sosyal çevreye hızla yayılıyor. Bu verinin analizi ve değerlendirilmesi, gerek birey gerekse şirketler için gittikçe daha fazla önemli kazançlar sunmaya başladı.

Bu alanda genel olarak İngilizce için çeşitli çalışmalar literatürde mevcuttur. Bu konu, Türk dili için henüz derinlemesine incelenmemiş ve yeterli sayıda araştırmanın yapılmadığı bir konudur. Yapılan literatür taramasında, Türk dili için gerçekleştirilen sadece iki çalışmaya ulaşabildik. İlk çalışma, sadece bir alan üzerine yoğunlaşıp, tek tipte veri seti üzerinde sadece belirli tipte bir makine öğrenmesi algoritması kullanmıştır [19]. İkinci çalışmada ise üç farklı veri seti üzerinde birden fazla makine öğrenmesi tek tek denenmiş ve Naive Bayes isimli makine öğrenmesi yöntemi ile yaklaşık olarak % 85 doğruluk oranı elde edilmiştir [20].

Bu tez çalışması kapsamında, Türk dili için farklı veri kümeleri üzerinde çoklu sınıflandırıcı makine öğrenmesi algoritmaları uygulanmıştır. Daha önce uygulanan çalışmalardan farklı olarak, performansı yüksek üç tane makine öğrenmesi algoritması birlikte kullanılarak özgün bir çoklu sınıflandırıcı makine öğrenmesi algoritması tasarlanmıştır. Bu özgün sınıflandırıcı yaklaşımının yanı sıra, makine öğrenmesi algoritmalarının parametre optimizasyonu gerçekleştirilerek performans arttırılmıştır. Bu yeni yaklaşım sayesinde, daha önce tek sınıflandırıcı ile elde edilen doğruluk oranı % 86,13'lük bir doğruluk oranına yükseltilmiştir. Bu doğruluk oranı, yeni yaklaşımın performansı iyileştirdiğini ve birçok çalışmada kullanılabileceğini ortaya koymuştur.

Anahtar Kelimeler: Duygu Sınıflandırma, Türkçe, Naive Bayes, Karar Destek Makineleri, Karar Ağacı, Çoklu Sınıflandırıcı Sistemler, Parametre Optimizasyonu, Makine Öğrenmesi, Doğal Dil İşleme, Veri Madenciliği, Weka