|  |
| --- |
| **YILDIZ TEKNİK ÜNİVERSİTESİ – BİLGİSAYAR MÜHENDİSLİĞİ** |
| **Doğal Dil İşlemeye Kavramsal Bir Bakış Projesi** |
| **Duygu-Düşünce Analizinde Farklı Özellik Kümeleri ve Sınıflandırma Algoritmalarının Kararlarının Birleştirilmesi** |

|  |
| --- |
| Melike Nur Mermer - 15501010  14.06.2017 |

**Giriş:**

Duygu-düşünce analizi, bir konuda bilgi sahibi olan kişilerin yazılı olarak ifade ettiği paylaşımlar kullanılarak o konu hakkındaki kanaatlerin otomatik olarak çıkartılıp olumlu, olumsuz veya tarafsız olarak sınıflandırılmasıdır. Bu çalışmalarda kaynak olarak roman, hikâye, blog yazıları gibi metinler kullanılabileceği gibi kişilerin sosyal medya üzerinden yaptığı paylaşımlar da kullanılabilir. Kişilerin Twitter üzerinden yapmış olduğu paylaşımlar olan tweetler de uzun zamandır duygu analizi için kaynak metin olarak kullanılmaktadır. Bu çalışmada bir telekomünikasyon şirketi hakkında yazılan olumlu ve olumsuz yorumları içeren twitter paylaşımları kaynak metin olarak kullanılmıştır. 3 farklı özellik kümesi ile ifade edilen metinler 2 farklı sınıflandırma algoritması ile sınıflandırılmış ve öncelikle tekil sınıflandırıcıların ardından bu sınıflandırıcıların kararlarının birleştirilmesi ile elde edilen toplulukların başarısı incelenmiştir.

**Veri kümesi ve Özellik çıkarma yöntemleri:**

Üzerinde çalışılacak veri seti1 bir telekomünikasyon firması ile ilgili olumlu ve olumsuz görüşleri içermektedir. Popüler bir sosyal medya aracı olan Twitter üzerinden atılan tweetlerden elde edilen yorumlar kaynak doküman olarak kullanılmıştır. Kaynakta 756 tane olumlu, 1287 tane olumsuz olmak üzere toplamda 2043 yorum bulunmaktadır. Olumlu ve olumsuz yorumlara örnekler aşağıdaki tabloda verilmektedir.

|  |  |
| --- | --- |
| **Olumlu yorumlar** | **Olumsuz yorumlar** |
| 32 GB Gold paket gayet iyi :) | Dogru bilgiyi alabilmek icin ne yapmam lazim illa kavga illa siddet off |
| iyi ki iPad 2'yi almışım. cam gibi lig tv her yerde :) teşekkürler | Operatörler internet paketlerini ücretlendirme konusunda çok acımasız. |
| yaptığınız kampanya süper hiç bitmesinnn | Dolmabahce tunelindeki 3G sorununa cozum lutfen |

Kaynak olarak kullanılacak metinlerin ön işlemlerden geçirilip özelliklerinin çıkarılması için text2arff2 aracı kullanılmıştır. Metinler bu araçla 3 farklı özellik kümesi ile ifade edilmiştir:

1. Type: Metinde geçen isim, fiil, sıfat, zamir, edat, bağlaç, ünlem, yankı, sayı, soru, zaman ve kısaltma kelimelerinin geçme sayılarının çıkartılmasıdır.
2. Unigram: Metinde geçen tüm kelimelerin tek başına geçme durumlarının binary olarak ifade edilmesidir.
3. Word 2-grams: Metinde geçen kelime ikililerinin birlikte geçme durumlarının binary olarak ifade edilmesidir.

1 Çetin M., Amasyalı, M. F., "Active learning for Turkish sentiment analysis.", *Innovations in Intelligent Systems and Applications (INISTA) IEEE International Symposium*, 2013.

2 Amasyali, M. F., Davletov, F., Torayew, A. I., ve Çiftçi, Ü., “Text2arff: Automatic feature extraction software for Turkish texts”, *Signal Processing and Communications Applications Conference (SIU),* 2010.

**Sınıflandırma algoritmaları ve Tekil sınıflandırıcıların başarıları:**

Özellik matrisleri ile ifade edilen dokümanlar çeşitli yöntemler kullanılarak sınıflandırılabilir. Bu çalışmada 3 farklı özellik korpusu kullanılarak Rastgele Orman ve Destek vektör makineleri ile sınıflandırma yapılmıştır. Her bir özellik korpusu ile eğitilen tekil sınıflandırıcıların her birinin başarısı aşağıdaki tabloda verilmektedir. Tablodaki değerler 10-fold cross validation ile elde edilen başarı oranlarının ortalamasını göstermektedir.

|  |  |  |
| --- | --- | --- |
| **Sınıflandırıcı** | **Özellik kümesi** | **Başarı(%)** |
| Rastgele Orman | Type | 65,54 |
| Unigram | **71,71** |
| Word 2-gram | 69,85 |
| Destek Vektör Makineleri | Type | 63,53 |
| Unigram | 71,07 |
| Word 2-gram | **72,89** |

**Kararları birleştirme yöntemleri ve Toplulukların başarıları:**

Tekil sınıflandırıcıların kararları oylama kuralı ve toplam kuralı ile birleştirilmiştir. Bu yöntemlerin yanı sıra veri kümesi eğitim-validasyon-test olarak ayrılıp her iki yöntemde validasyon kümesindeki başarı oranlarına göre karar ağırlıklandırılması gerçekleştirilmiştir.

Öncelikle her bir sınıflandırıcı için farklı özellik korpuslarıyla elde edilen kararların birleştirilmesi ardından tüm sınıflandırıcı ve özellik korpusları (2x3=6 farklı sınıflandırma) ile elde edilen kararların birleştirilmesinin sonuçları aşağıdaki tabloda verilmektedir. Tablodaki değerler 10-fold cross validation ile elde edilen başarı oranlarının ortalamasını göstermektedir.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Oylama kuralı** | | **Toplam kuralı** | |
| Sabit | Ağırlıklı | Sabit | Ağırlıklı |
| **Rastgele Orman** | 71,26 | 72,25 | 71,03 | **72,60** |
| **Destek Vektör Makineleri** | 71,62 | 72,94 | 73,19 | **74,80** |
| **Rastgele Orman + Destek Vektör Makineleri** | 73,37 | 73,42 | 71,76 | **74,46** |

**Sonuç:**

Duygu-düşünce sınıflandırması için rastgele orman ve destek vektör makinesi sınıflandırıcılarına verilen type, unigram ve word 2-gram özellik korpuslarıyla gerçekleştirilen sınıflandırma tahminlerinin birleştirilmesi sonucunda elde edilen sınıflandırma başarılarının tekil sınıflandırıcılardan daha yüksek olduğu görülmüştür. Farklı birleştirme yöntemlerinden en başarılı yöntemin ağırlıklı toplam kuralı olduğu görülmektedir.