**Karakter N-Gram Tabanlı**

**Doküman Sınıflandırma**

[**Nedret Gegeoğlu**](mailto:nedretgegeoglu@windowslive.com) **- 150202114**

[**Bora Acar**](mailto:boraacar95@gmail.com) **– 130202085**

**Kocaeli Üniversitesi**

**Bilgisayar Mühendisliği**

1. **– 2018**
2. **Problem Tanımı:**

Bizden makine öğrenmesinde “Naive Bayes” sınıflandırma yöntemi kullanarak verilen 1150 haber içeren veri setini metin bazında 2-gram ve 3-gram tabanında parçalara ayırarak eğitimden geçirmemiz istendi.

1. **Yapılan Araştırmalar:**

İlk olarak karakter tabanlı ayrıştırma yöntemi olan “N-Gram” metodu incelendi. Veri seti üzerinde n-gram tabanlı ayrıştırma uygulamaktan önce veri işleme ön adımları gerçekleştirildi.

* Veri işlemede ilk olarak bize verilen veri setindeki her haberi tek tek satırlarına ayrıştırdık.
* Ayrıştırılan metini kullanılan özel karakterlerinden temizledik. Örneğin; “.,!,?,’,-,\_” gibi karakterleri metinden çıkartarak makine öğrenmesinin daha verimli olmasını hedefledik.
* Metinden özel karakterler temizlendikten sonra tüm metin üzerindeki kelimeleri“Nüve” kütüphanesi yardımı ile Türkçe köklerine ayırdık.
* Son olarak Türkçe kök bulma, düzenleme işlemi sonrasında kelimeler 2 ve 3-gramlara ayrıldı.

Projedeki bir diğer sorun “Naive Bayes” sınıflandırma mantığının bu projeye nasıl uygulanacağıydı. Bunun için; “naive bayes” çalışma mantığı incelendi.

Teknik olarak “naive bayes” sınıflandırma yönteminde kategorilere ait koşullu olasılıklar söz konusudur. Bu koşullu olasılıklar üzerinden çıkarım yapılarak en yüksek gerçekleşme ihtimali ele alınır ve sonuç olarak yansıtılır. Buradaki sorun bizim elimizde 5 farklı haber kategorisi olması ve her kategorinin 2-gram ve 3-gram şeklinde kendi kategorilerine ayrılmasıydı. Bizde bu sorunu çözmek için; Haber kategorilerinin kendi aralarında gerçekleşme olasılıklarını (doğal eğilimi) hesapladıktan sonra ve bu kategoriler içerisindeki tüm haberlere ait 2 gram ve 3 gram frekansları hesaplandıktan sonra çıkan sonuçları toplayıp sonucu o şekilde etkilemeye karar verdik. Örneğin “Ekonomi x” isimli habere ait:

2grama ait 3 verimiz 3 grama ait 2 verimiz olsun.

2gram 3gram

An:55 lar: 50

Ra:60 ken:55

Ve:50

2 gram toplam frekans değerimiz: 60+55+50=165 ve 3gram toplam frekans değerimiz: 50+55=105 olur. Bunun sonucunda veriyi etkileyecek 2gram düzeyi:

Toplam 2 gram frekansı / (2gram +3gram frekans değeri)

: 2gram :165 / 270 = 0,61

: 3 gram ise : 105 / 270 = 0,39 olur.

1. **Tasarım:**

**3.1. Akış Şeması:**



1. **Genel Yapı:**

Projeyi C# dilinde “VS” ortamı üzerinde hazırladık. Kullanılan metotlar ve amaçları:

İlk olarak kaç haber okunacak ise bu haberlerin %75i eğitim, kalan %25i ise test verisi olacak şekilde düzenledik. Bunu oluştururken rastgele haberler alan bir dizi hazırladık.

Haberler oluşturulduktan sonra sırasıyla veri ön işleme adımları uygulandı, kelimeler köklerine ayrıldı ve ayrılan kelimeler “substr” string metodu ile n-gramlara ayrıldı. Bu işlem esnasında özel karakterler temizlendi ve “ “ boşluk karakterleri,”\_” karakterine dönüştürüldü.

N-Gram kelimeleri düzenlendikten sonra verilerin frekans değerleri hesaplandı ve değerleri 50 ve üzeri olan n-gramlar “gramx\_50” dizisinde yer aldı. Birbirinden farklı n-gramlarıda her n-gramdan bir adet olacak şekilde “farklı50\_x” (x=2 veya 3) dizisinde sakladık.

Veriyi alıp düzenleme işlemleri bu aşamada sona erdi ve veri eğitme işlemlerine başladık. İlk olarak:

“p\_oran\_fonk” isimli metotta her haber kategorisinin doğal eğilimleri hesaplandı. Örneğin:

Ekonomiden toplam gram frekansı 600 ve magazinden toplam gram frekansı 400. Bu durumda test ettiğimiz haberin ekonomi olma ihtimali 600/ (600+400) = %60 dır.

Doğal eğilim hesaplandıktan sonra her grama ait koşullu olasılık ihtimalleri hesaplandı. Hesaplama metodu olarak örneğin:

Ekonomide geçen “ar” gramının toplam frekansı / (toplam 2gram frekansları + birbirinden farklı geçen tüm gramlar) bu işlem dahilinde oluşan ihtimal “ar” gramının ekonomide haberleri içerisinde geçme ihtimalidir.

Her n-gram her haber türüne göre tek tek incelendikten sonra eğitim işlemimiz tamamlanır ve test işlemine başlarız. Test işleminde ilk olarak:

Test edilmesi gereken n-gramlar alındıktan sonra eğitim modelimizde o grama ait veri var mı bunu kontrol ederiz. Eğer veri yoksa o test verisine “1” verilir. Bunun sebebi koşullu olasılık oluşturma esnasında çarpım işlemleri olacağı için eğitimde bulunmayan verinin test sonucuna etki etmemesi istenmesidir.

Daha sonra her gramın bulunduğu haber isimleri listelenir ve aynı haber ismi tekrarları kaldırılarak haber isimlerini birer kez görüntüleriz.

Test verisinde bulunan n-gramın hangi kategoriye ait olduğunu anlamak için uyguladığımız test şu şekildedir:

Test verisinde bulunan gramların eğitim verisinde bulunduğu tespit edilmişse bulunan koşullu olasılık değerlerini alır. Örneğin x haberi için aldığımız 50üzeri frekanslı 2 gramlar:

“an” 2gramı koşullu olasılığı: 0,005,

“na” 2 gramı koşullu olasılığı = “0,01”

ise bu değeri alır sonrasında bu değeri her kategorinin doğal eğilim sonucu ile çarparız. Sonuç olarak hangi kategori ihtimali daha yüksek ise “naive bayes” n-gramların o kategoriye ait olduğunu gösterir.

Burada bize göre dikkat edilmesi gereken en önemli özellik haber kategorilerinin doğal eğilimleridir. Örneğin; x test verisi için genel olarak bulunan tüm ngramlar kelime yatkınlığı olarak spor kategorisine yakın olsalar ve spor kelimelerine benzeselerde, bizim eğittiğimiz veride doğal eğilim olarak spor kategorisi düşük ise test ettiğimiz veri muhtemelen başka bir kategoriye etiketlenecektir. Bunu çözmek için gerçekten yüksek ve dengeli miktarda verilere sahip olmalıyız ve bu veriler ayırt edici özellikte olmalıdır.

Son olarak projede “precision, recall ve f-measure” değerlerini hesapladık.

Precision sonucu: Teste tabi tuttuğumuz verilerdeki doğru eşleşmelerin, teste tabi tutulan tüm verilere oranı olarak hesapladık. Örneğin: 10 adet 2 gram teste tabi tutuldu ve 2 adet 2gram eşleşmesi doğru oldu.

Precision : 2/10=0,2

Recall sonucu: Teste tabi tuttuğumuz verilerdeki doğru sonuçların , teste tabi tuttuğumuz verilerdeki beklenen doğruya oranı. Örneğin 3 haberlik veri test ettik ve 1 haberi doğru tespit ettik.

Recall: 1/3= 0,33

F-Measure: Precision sonucu ve recall sonucunun harmonik ortalamasıdır.

F-Measure:

(2 \* ( precision \* recall ) ) / (precision + recall)

1. **Referanslar:**

[1.] N-Gram Prensipleri:<http://bilgisayarkavramlari.sadievrenseker.com/2011/04/23/n-gram/>

[2.]Naive-Bayes Çalışma Prensibi: <https://kodedu.com/2014/05/naive-bayes-siniflandirma-algoritmasi/>

[3.] Precision, Recall ve F-Score Kavramları: <http://bilgisayarkavramlari.sadievrenseker.com/2010/10/01/bilgi-getirimi-ve-cikarimi-information-retrieval-and-extraction/>