Metinler için Derin Öğrenme ile Metin Madenciliği ve Belge Sınıflandırması

Proje Ekibi: Ahmet BURSALI, Hakan KIRANOĞLU

Danışman: Yard. Doç. Dr. Mete EMİNAĞAOĞLU

Amaç ve Kapsam

Bu proje metinler için Metin Madenciliği algoritmalarıyla birlikte derin öğrenme yöntemlerinin kullanılıp, belgelerin doğru biçimde kategorize edilmesini amaçlar. Amaç doğrultusunda yararlanılan disiplinler şunlardır:

- Derin Öğrenme
- Metin madenciliği
- Makine öğrenimi
- Görüntü İşleme

Hedefler

- Metinler için Metin Madenciliği ve Derin Yapay Sinir Ağları kullanarak Belge sınıflandırma yapmak.

- Proje ekibine Derin Öğrenme ve Metin Madenciliği alanında ilgili yaklaşımları görüp bunları ortak kullanabilme arasındaki bağlantıları kurabilme veya yeni yaklaşımlar geliştirebilme olanağı sağlamak.

Metin Madenciliği (Text Mining)

- Metin üzerinden yapısallaştırılmış (structured) veri elde etmeyi amaçlar.

- Genel kavramlar

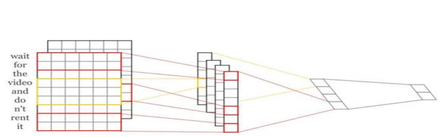
Tf X Idf : bir metinde geçen terimlerin çıkarılması ve bu terimlerin geçtiği miktara göre yapılan hesaplama

Kümeleme (Clustering) : Kümeleme ise bir veri kümesindeki bilgileri belirli kurallara göre gruplara ayırma işlemidir .

Sınıflandırma (Classification) : bir veri kümesi (data set) üzerinde tanımlı olan çeşitli sınıflar arasında veriyi dağıtmaktır.

Derin Öğrenme (Deep Learning)

- Dünyayı algılama ve anlamasına yönelik yapay zeka geliştirmede en popüler yaklaşımdır.
- Çok katmanlı Yapay sinir ağlarına dayalı öğrenme yapısına sahiptir.

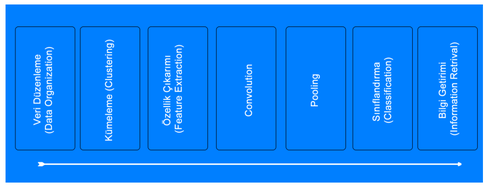


Analiz

Proje kapsamındaki veri seti :

- 4 farklı sınıf (C,E,S,W)
- Tf-idf ağırlık değerleri (normalize edilmiş şekilde)
- 17869 kayıt (cümleler, Türkçe fakat İngilizce terimler ile karışık şekilde, hatta bazılarında İngilizce sözcüklerden oluşan cümleler)
- 3804 attribute (sözcükler)
- Eğitim için kayıt sayısı 2200
- Test kayıt sayısı 1100

Tasarım



Proje Aşamaları

Uygulama

Bu projede;

- Her bir sınıf kendi içinde gruplanıp sadece o sınıfa ait olan kayıtların satır-sütunları yer değiştirdi(transpoze işlemi)

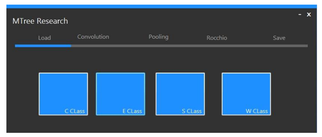
C:1122 kayıt , transpoze edilmiş yeni veri matrisi: 3408 x 1122

E: 515 kayıt , transpoze edilmiş yeni veri matrisi: 3408 x 515

S:363 kayıt , transpoze edilmiş yeni veri matrisi: 3408x363

W: 200 kayıt , transpoze edilmiş yeni veri matrisi: 3408 x 20

- Veri matrislerinin her biri için çeşitli clustering algoritmaları ayrı ayrı çalıştırılarak her bir sınıfa ait aday tek boyutlu filtreler belirlendi.
- Derin Öğrenme algoritmalarından Convolutional Neural Networks - CNN – uygulandı.
- Son aşama da ise sonuç bölümde adı geçen algoritmalar uygulandı.
- Proje de yazılımsal olarak C# ortamında sonuçlar incelendi.



Yazılımdan Arayüzünden Örnek Ekran Görüntüsü

Sonuç ve Öneriler

Model	Yöntem 1	Yöntem 2	Yeni Yaklaşım 1	Yeni Yaklaşım 2
Rocchio Classifier (with Cosine sim.)	0,72	0,756	0,754	0,764
k-NN (k=1,8 Euclidean distance)	0,674	0,782	0,764	0,714
Simple Naive Bayes (Weka)	0,740	0,782	0,698	0,760
Discriminative Multinomial Naive Bayes	0,742	0,805	0,832	0,799
Functional Tree (Weka)	0,724	0,742	0,831	0,775
Random Forest (Weka)	0,668	0,744	0,876	0,854
Simple Logistic Regression (Weka)	0,763	0,814	0,804	0,795
Multinomial Logistic Regression (Weka)	0,766	0,816	0,763	0,772
J48 (Decision Tree, Weka version of C4.5) (Weka)	0,693	0,735	0,794	0,729

Tablo 1: Önceden yapılmış çalışmalar (Yöntem 1 ve Yöntem 2) ve Proje kapsamında geliştirilen yaklaşımlardan (Yaklaşım 1 ve Yaklaşım 2) elde edilen sonuçlar

- Yaklaşım 1: Kümeleme aşamasında Mtree
- Yaklaşım 2: Kümeleme aşamasında Kmeans
- Yöntem 1 ve Yöntem 2: Klasik Metin Madenciliği algoritmaları kombinasyonları

Projenin daha da geliştirilmesi için

- Yaklaşım aşamalarının iyileştirmesi
- Benchmark verilerinde yaklaşımın tutarlılığını daha kapsamlı test etmek olabilir.

Kaynakça

Kim, Yoon (2014). "Convolutional Neural Networks for Sentence Classification." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar.* Lai, Siwei et al. (2015). "Recurrent Convolutional Neural Networks for Text Classification." In: *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI). Austin Texas, USA.*

Mihăescu M.C, Burdescu D. D. Using M Tree Data Structure as Unsupervised Classification Method Informatica 36 (2012) 153-160

Öztemel, E. (2003), Yapay Sinir Ağları, Papatya Yayın Evi.

Zhang X, Zhao J, LeCun Y. (2015). Character-level Convolutional Networks for Text Classification. Neural Information Processing Systems.