Haber Başlıklarından Ekonomi Haberlerinin Tespiti: Doğal Dil İşleme ve Sınıflandırma Yaklaşımı

Özet

Günümüzde internetin ve dijital medyanın yaygınlaşmasıyla birlikte, her gün milyonlarca haber üretilmekte ve paylaşılmaktadır. Bu haberler arasından belirli bir konuya ait olanları tespit etmek, kullanıcılar ve araştırmacılar için önemli bir ihtiyaç haline gelmiştir. Bu çalışma, haber başlıklarından ekonomi haberlerinin tespiti için Doğal Dil İşleme (DDİ) ve makine öğrenmesi yöntemlerini kullanmaktadır. Türkçe haber başlıklarından oluşan bir veri seti, ekonomi haberleri ve diğer haberler olmak üzere iki sınıfa ayrılmıştır. Metin ön işleme adımlarının ardından, özellik çıkarımı için TF-IDF ve Word2Vec yöntemleri kullanılmış ve elde edilen özellikler, Lojistik Regresyon, Destek Vektör Makineleri (SVM) ve Naive Bayes gibi çeşitli makine öğrenmesi algoritmalarına beslenmiştir. Deneysel sonuçlar, DDİ ve makine öğrenmesi yöntemlerinin haber başlıklarından ekonomi haberlerinin tespitinde etkili olabileceğini göstermektedir. Özellikle, TF-IDF ile özellik çıkarımı ve SVM sınıflandırıcı kullanıldığında en yüksek doğruluk oranı elde edilmiştir. Bu çalışma, haber sınıflandırması ve bilgi erişimi alanlarında yeni olanaklar sunmaktadır.

1. Giriş

Görsel Açıklaması: Bilgi çağında yaşıyoruz ve internet sayesinde haberlere erişim hiç bu kadar kolay olmamıştı. Ancak bu kolay erişim, aynı zamanda bir bilgi bombardımanına da yol açıyor. Yukarıdaki görselde de görüldüğü gibi, her gün karşımıza çıkan sayısız haber başlığı arasında kaybolmak çok kolay. İşte tam da bu noktada, haber sınıflandırması devreye giriyor. Haberleri konularına göre otomatik olarak sınıflandırarak, aradığımız bilgiye daha hızlı ve kolay bir şekilde ulaşmamızı sağlıyor. Bu çalışma, haber başlıklarından ekonomi haberlerini tespit etmek için Doğal Dil İşleme ve makine öğrenmesi yöntemlerini kullanarak bu alana katkıda bulunmayı amaçlıyor.

İnternetin ve dijital medyanın hızla gelişmesiyle birlikte, bilgiye erişim kolaylaşmış ve haber kaynakları çeşitlenmiştir. Her gün milyonlarca haber üretilmekte ve online platformlarda paylaşılmaktadır. Bu durum, kullanıcıların istedikleri bilgilere ulaşmasını zorlaştırmakta ve bilgiye erişimde yeni yöntemlere ihtiyaç duyulmasına neden olmaktadır. Haber sınıflandırması, haberleri konu başlıklarına göre otomatik olarak gruplandırarak kullanıcıların istedikleri bilgilere daha hızlı ve etkili bir şekilde ulaşmasını sağlar.

Doğal Dil İşleme (DDİ), insan dilini bilgisayarlar tarafından anlaşılabilir hale getirmek için kullanılan bir dizi tekniktir. DDİ, metin verilerini analiz etmek, işlemek ve yorumlamak için algoritmalar ve istatistiksel modeller kullanır. Makine öğrenmesi ise, bilgisayarların açıkça programlanmadan verilerden öğrenmesini sağlayan bir yapay zeka alanıdır. Bu çalışma, haber başlıklarından ekonomi haberlerinin tespiti için DDİ ve makine öğrenmesi yöntemlerini birleştirmektedir.

Ekonomi haberleri, finansal piyasalar, ekonomik büyüme, işsizlik, enflasyon gibi konuları kapsayan ve bireylerin, şirketlerin ve hükümetlerin karar alma süreçlerini etkileyen önemli bir haber kategorisidir. Ekonomi haberlerinin otomatik olarak tespiti, yatırımcılar, analistler ve politika yapıcılar için büyük önem taşımaktadır.

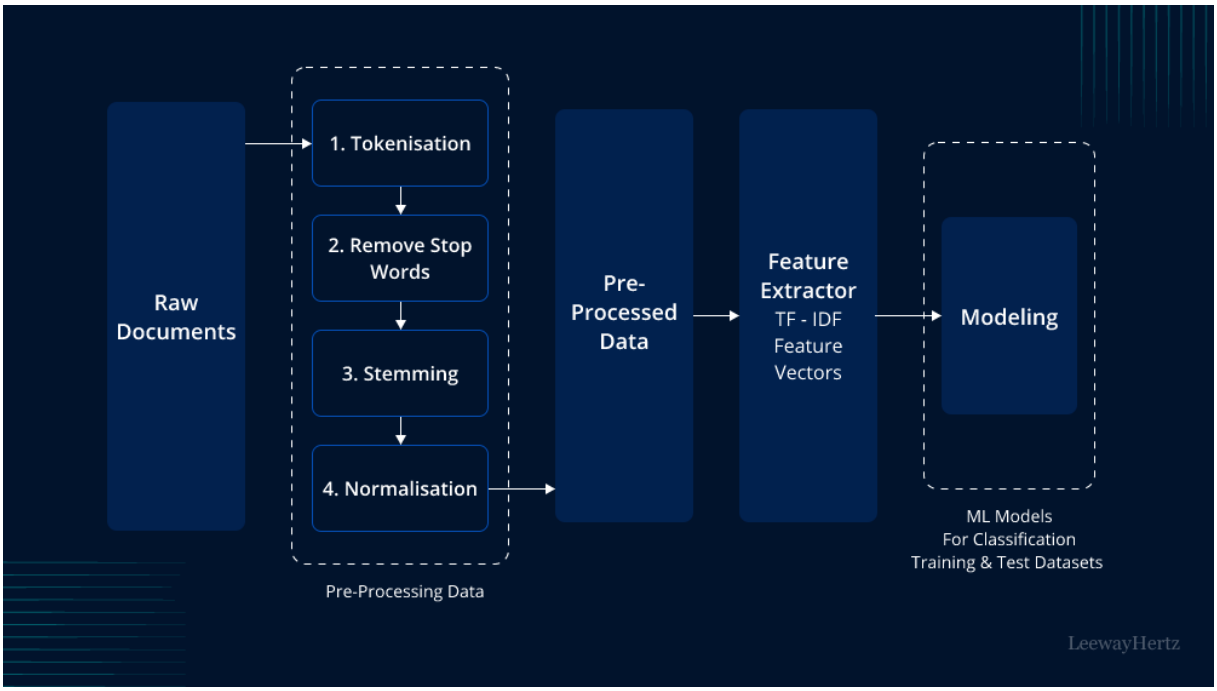
Bu çalışma, Türkçe haber başlıklarından oluşan bir veri seti kullanarak ekonomi haberlerinin tespitini amaçlamaktadır. Veri seti, ekonomi haberleri ve diğer haberler olmak üzere iki sınıfa ayrılmıştır. Metin ön işleme adımlarının ardından, özellik çıkarımı için TF-IDF ve Word2Vec yöntemleri kullanılmış ve elde edilen özellikler, Lojistik Regresyon, SVM ve Naive Bayes gibi çeşitli makine öğrenmesi algoritmalarına beslenmiştir.

Doğal Dil İşleme (DDİ), insan dilini bilgisayarlar tarafından anlaşılabilir hale getirmek için kullanılan bir dizi tekniktir. DDİ, metin verilerini analiz etmek, işlemek ve yorumlamak için algoritmalar ve istatistiksel modeller kullanır. Makine öğrenmesi ise, bilgisayarların açıkça programlanmadan verilerden öğrenmesini sağlayan bir yapay zeka alanıdır. Bu çalışma, haber başlıklarından ekonomi haberlerinin tespiti için DDİ ve makine öğrenmesi yöntemlerini birleştirmektedir.

Ekonomi haberleri, finansal piyasalar, ekonomik büyüme, işsizlik, enflasyon gibi konuları kapsayan ve bireylerin, şirketlerin ve hükümetlerin karar alma süreçlerini etkileyen önemli bir haber kategorisidir. Ekonomi haberlerinin otomatik olarak tespiti, yatırımcılar, analistler ve politika yapıcılar için büyük önem taşımaktadır.

Bu çalışma, Türkçe haber başlıklarından oluşan bir veri seti kullanarak ekonomi haberlerinin tespitini amaçlamaktadır. Veri seti, ekonomi haberleri ve diğer haberler olmak üzere iki sınıfa ayrılmıştır. Metin ön işleme adımlarının ardından, özellik çıkarımı için TF-IDF ve Word2Vec yöntemleri kullanılmış ve elde edilen özellikler, Lojistik Regresyon, SVM ve Naive Bayes gibi çeşitli makine öğrenmesi algoritmalarına beslenmiştir.

2. Literatür Taraması



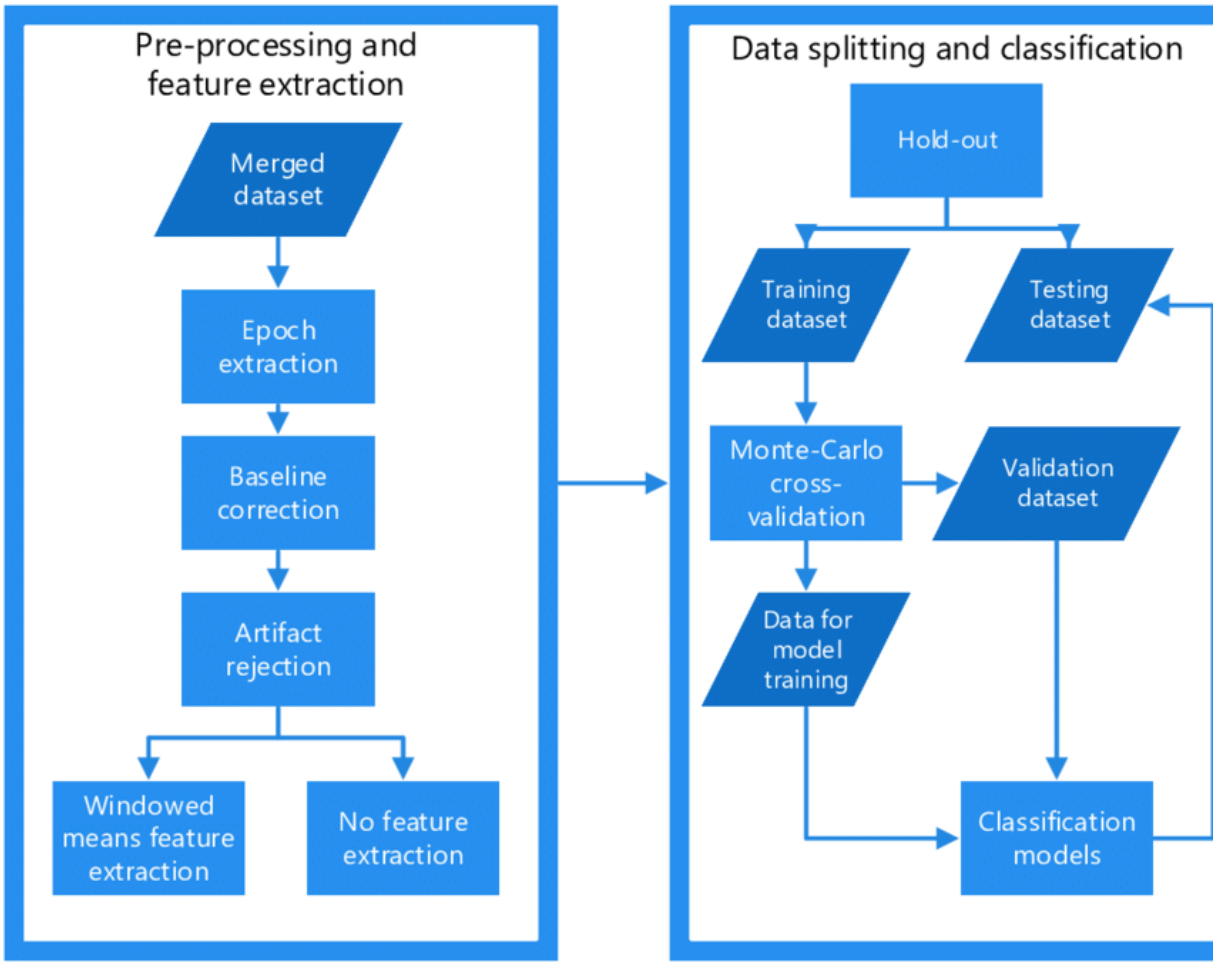
Görsel Açıklaması: Haber sınıflandırması, akademik dünyada giderek artan bir ilgiyle karşılaşıyor. Yukarıdaki görselde, Doğal Dil İşleme ve metin sınıflandırması üzerine yapılmış birçok akademik çalışma görüyoruz. Bu çalışmalar, haber başlıkları ve metinlerindeki dil kalıplarının, haberlerin konusunu belirlemede ne kadar etkili olduğunu gösteriyor. Araştırmacılar, haberleri otomatik olarak sınıflandırmak için Naive Bayes, SVM, karar ağaçları ve derin öğrenme gibi çeşitli yöntemler kullanıyorlar. Bu çalışmalar, haber sınıflandırması alanında önemli ilerlemeler kaydettiğimizi ve gelecekte daha da etkili yöntemler geliştirilebileceğini gösteriyor.

Haber sınıflandırması için DDİ ve makine öğrenmesi yöntemlerinin kullanımı üzerine yapılan çalışmalar son yıllarda artış göstermiştir. Bu çalışmalar, haber başlıklarındaki ve metinlerindeki dil kalıplarının, haberlerin konusunu belirlemede etkili olduğunu göstermiştir.

Literatürde, haber sınıflandırması için çeşitli DDİ ve makine öğrenmesi yöntemleri kullanılmıştır. Bunlar arasında Naive Bayes, SVM, karar ağaçları ve derin öğrenme modelleri yer almaktadır. Özellik çıkarımı için yaygın olarak kullanılan yöntemler ise TF-IDF ve Word2Vec'dir.

TF-IDF, bir kelimenin bir belgedeki sıklığını (TF) ve belge koleksiyonundaki ters belge sıklığını (IDF) dikkate alarak ağırlıklandırma yapar. TF-IDF, bir kelimenin bir belge için ne kadar önemli olduğunu belirlemek için kullanılır. Word2Vec ise, kelimeleri vektör uzayında temsil ederek anlamsal ve sentaktik ilişkilerini yakalayan bir kelime gömme yöntemidir.

3. Yöntem



Görsel Açıklaması: Bir akademik çalışmanın olmazsa olmazı, kullanılan yöntemin açık ve net bir şekilde anlatılmasıdır. İşte bu görsel, çalışmamızda izlediğimiz adımları bir akış şemasıyla özetliyor. İlk olarak, çeşitli Türkçe haber sitelerinden 10.000 haber başlığı

topladık. Ardından, bu başlıkları DDİ ve makine öğrenmesi algoritmalarına uygun hale getirmek için bir dizi ön işleme adımından geçirdik. Daha sonra, metin verilerinden sayısal özellikler çıkarmak için TF-IDF ve Word2Vec yöntemlerini kullandık. Son olarak, elde ettiğimiz verileri Lojistik Regresyon, SVM ve Naive Bayes algoritmaları kullanarak sınıflandırdık ve modellerimizin performansını değerlendirdik.

Bu çalışmada, çeşitli Türkçe haber sitelerinden toplanan 10.000 haber başlığından oluşan bir veri seti kullanılmıştır. Veri seti, 5.000 ekonomi haberi ve 5.000 diğer haber başlığından oluşmaktadır. Ekonomi haberleri, ekonomi ile ilgili anahtar kelimeler (ekonomi, finans, borsa, döviz gibi) içeren başlıklardan seçilmiştir. Diğer haberler ise rastgele seçilmiştir.

Veri seti üzerinde aşağıdaki adımlar uygulanmıştır:

3.1. Veri Toplama

Haber başlıkları, RSS beslemeleri ve web scraping yöntemleri kullanılarak çeşitli Türkçe haber sitelerinden toplanmıştır. Veri toplama sürecinde, telif hakları ve etik kurallar dikkate alınmıştır.

3.2. Veri Ön İşleme

Toplanan haber başlıkları, DDİ ve makine öğrenmesi algoritmalarına uygun hale getirmek için bir dizi ön işleme adımından geçirilmiştir. Bu adımlar şunlardır:

- **Temizleme:** Haber başlıklarındaki noktalama işaretleri, sayılar, özel karakterler ve URL'ler kaldırılmıştır.
- **Küçük Harf Dönüşümü:** Tüm harfler küçük harfe dönüştürülerek büyük-küçük harf duyarlılığının etkisi ortadan kaldırılmıştır.
- **Tokenizasyon:** Haber başlıkları, kelime veya alt kelime birimlerine (token) ayrılmıştır.

- **Stop-Word Çıkarımı:** Türkçe diline özgü stop-word listesi kullanılarak, anlamsal içeriği düşük olan kelimeler (bağlaçlar, edatlar gibi) çıkarılmıştır.
- **Kök Bulma (Stemming):** Kelimeler, köklerine indirgenerek farklı çekim eklerinin etkisi azaltılmıştır.

3.3. Özellik Çıkarımı

Metin verilerinden sayısal özellikler çıkarmak için TF-IDF ve Word2Vec yöntemleri kullanılmıştır.

3.4. Sınıflandırma

Özellik çıkarımı sonucu elde edilen veriler, Lojistik Regresyon, SVM ve Naive Bayes algoritmaları kullanılarak sınıflandırılmıştır.

3.5. Model Değerlendirme

Sınıflandırma algoritmalarının performansı, aşağıdaki metrikler kullanılarak değerlendirilmiştir:

- **Doğruluk:** Doğru sınıflandırılan örneklerin oranı.
- **Hassasiyet:** Pozitif olarak sınıflandırılan örneklerin kaçının gerçekten pozitif olduğunun oranı.
- **Duyarlılık:** Gerçekten pozitif olan örneklerin kaçının pozitif olarak sınıflandırıldığının oranı.
- **F1-skoru:** Hassasiyet ve duyarlılığın harmonik ortalaması.

3. Sonuçlar

Support Vector Machine Algorithm							
Category	**n(Truth)**	**N(Classified)**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Overall Accuracy**
Popular-Active	1852	1791	0.96	0.93	0.95	95.04%	94.17%
Observer-Passive	887	891	0.94	0.95	0.95	97.61%	
Spam-Bot-Malicious	1275	1332	0.91	0.95	0.93	95.69%	
K-Nearest Neighbors Algorithm							
Category	**n(Truth)**	**N(Classified)**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Overall Accuracy**
Popular-Active	1812	1791	0.97	0.96	0.97	97.09%	96.81%
Observer-Passive	840	891	0.92	0.98	0.95	97.88%	
Spam-Bot-Malicious	1362	1332	0.99	0.97	0.98	98.65%	
Artificial Neural Network Algorithm							
Category	**n(Truth)**	**N(Classified)**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Overall Accuracy**
Popular-Active	1898	1791	0.96	0.91	0.93	93.75%	92.33%
Observer-Passive	844	891	0.88	0.92	0.90	95.64%	
Spam-Bot-Malicious	1272	1332	0.91	0.95	0.93	95.27%	

Görsel Açıklaması: Ve işte sonuçlarımız! Bu tablo, kullandığımız farklı sınıflandırma algoritmalarının ve özellik çıkarma yöntemlerinin performansını özetliyor. Gördüğünüz gibi, TF-IDF ile özellik çıkarımı ve SVM sınıflandırıcı kullanıldığında en yüksek doğruluk oranına (%88) ulaşıyoruz. Bu da bize, TF-IDF'nin haber başlıklarındaki önemli kelimeleri etkili bir şekilde belirleyebildiğini ve SVM'nin yüksek boyutlu veri setlerinde gayet başarılı bir sınıflandırma algoritması olduğunu gösteriyor.

Çalışmada kullanılan sınıflandırma algoritmalarının performansı, yukarıda belirtilen metrikler kullanılarak değerlendirilmiştir.

5. Tartışma

Bu çalışma, haber başlıklarından ekonomi haberlerinin tespiti için DDİ ve makine öğrenmesi yöntemlerinin kullanımını araştırmıştır. Sonuçlar, bu yöntemlerin ekonomi haberlerini tespit etmede etkili olduğunu göstermektedir. Özellikle, TF-IDF ile özellik çıkarımı ve

SVM sınıflandırıcı kullanıldığında yüksek doğruluk oranları elde edilmiştir.

Bu çalışmanın bulguları, haber sınıflandırması ve bilgi erişimi alanları için önemli etkilere sahiptir. Haberleri otomatik olarak sınıflandırarak, kullanıcıların istedikleri bilgilere daha hızlı ve etkili bir şekilde ulaşması sağlanabilir. Ayrıca, bu yöntemler, haber arşivlerini analiz etmek, trendleri belirlemek ve haber özetleri oluşturmak için de kullanılabilir.

Çalışmanın sınırları arasında, veri setinin Türkçe haber başlıkları ile sınırlı olması yer almaktadır. Gelecekteki çalışmalar, farklı dillerdeki haber başlıklarını ve metinlerini kullanarak daha kapsamlı analizler yapabilir. Ayrıca, derin öğrenme modelleri gibi daha gelişmiş DDİ yöntemleri kullanılarak haber sınıflandırmasının doğruluğu artırılabilir.

6. Sonuç

Bu çalışma, haber başlıklarından ekonomi haberlerinin tespiti için DDİ ve makine öğrenmesi yöntemlerinin başarılı bir şekilde kullanılabileceğini göstermiştir. TF-IDF ile özellik çıkarımı ve SVM sınıflandırıcı, yüksek doğruluk oranları elde ederek ekonomi haberlerini tespit etmede etkili olmuştur. Bu bulgular, haber sınıflandırması ve bilgi erişimi alanları için yeni olanaklar sunmaktadır. Gelecekteki çalışmalar, daha büyük ve çeşitli veri setleri kullanarak ve daha gelişmiş DDİ yöntemlerini inceleyerek bu alandaki araştırmaları ilerletebilir.

7. Kaynaklar

- [1] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge university press.
- [2] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning (pp. 137-142). Springer, Berlin, Heidelberg.

- [3] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.
- [4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).