Orclever Proceedings of Research and Development
https://doi.org/10.56038/oprd.v5i1.581
OR CLEVER Science & Research Group

Konferans Bildirisi

Koşullu Varyasyonel Otokodlayıcılar ile Metin Sınıflandırma Performansını Artırmaya Yönelik Sentetik Veri Üretimi

Ömer Faruk Cebeci[1]*, Mehmet Fatih Amasyali[2]

[1] Yıldız Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Bilgisayar Mühendisliği Bölümü
Orcid ID: https://orcid.org/0000-0001-8174-2046, e-mail: faruk.cebeci@std.yildiz.edu.tr
[2] Yıldız Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Bilgisayar Mühendisliği Bölümü
Orcid ID: https://orcid.org/0000-0002-0404-5973, e-mail: amasyali@yildiz.edu.tr
* Sorumlu Yazar: faruk.cebeci@std.yildiz.edu.tr

Received 29 October 2024
Received in revised form 28 November 2024
In final form 12 December 2024

4th International Conference on Design, Research and Development (RDCONF 2024)
December 19 - 20, 2024

Reference: Cebeci, Ö. F., & Amasyali, M. F. (2024). Synthetic data generation for enhancing text classification performance using conditional variational autoencoders. Orclever Proceedings of Research and Development, 5(1), 498-514.

Özet

Bu çalışma, veri sayısının düşük veya veri kaynağının sınırlı olduğu durumlarda metin sınıflandırma performansını artırmak amacıyla, Koşullu Varyasyonel Otokodlayıcı modeliyle sentetik veri üretiminin sınıflandırma performansına etkisini incelemektedir. Farklı sınıf sayısına sahip veri setleri üzerinde gerçekleştirilen deneylerde, Koşullu Varyasyonel Otokodlayıcı modelleri kullanılarak iki farklı yöntemle sentetik veriler üretilmiştir. İlk yöntem, gauss dağılımından yapılan örnekleme ile başlatılan gürültüden cümle üretmeyi hedeflemiştir. İkinci yöntem ise gerçek cümlelerin ilk yarısı verilerek modelin cümlenin geri kalanını tamamlaması yoluyla sentetik veriler üretmeyi sağlamıştır. Her iki yöntemle üretilen sentetik veriler, sınıflandırma modellerinde orijinal eğitim kümesine çeşitli oranlarda eklenerek modellerin performans değişimleri gözlemlenmiştir. Her iki yöntemle üretilen sentetik veriler sınıflandırma modellerinin performansını anlamlı ölçüde artırmıştır. Bununla birlikte, sınıflandırma modellerinin eğitiminde kullanılan verilerin sayısı arttıkça sentetik verilerin katkısının azaldığı görülmüştür. Bu sonuçlar,

https://doi.org/10.56038/oprd.v5i1.581

veri kısıtlılığının yaşandığı metin sınıflandırma görevlerinde sentetik veri üretiminin ve kullanımının etkili bir strateji olabileceğini göstermektedir.

Anahtar: *Metin sınıflandırma, Varyasyonel Otokodlayıcı, yapay metin üretimi*

Orclever Proceedings of Research and Development
https://doi.org/10.56038/oprd.v5i1.581
OR CLEVER Science & Research Group

Conference Article

Synthetic Data Generation for Enhancing Text Classification Performance Using Conditional Variational Autoencoders

Abstract

This study investigates the effect of generating synthetic data using a Conditional Variational Autoencoder (CVAE) model on classification performance in scenarios where the amount of available data is limited or the data sources are constrained. Experiments were conducted on datasets with varying numbers of classes, where synthetic data were produced through two different methods using CVAE models. The first method aimed to generate sentences from noise, initiated by sampling from a Gaussian distribution. The second method involved providing the first half of a real sentence to the model, which then completed the remaining half to produce synthetic data. The synthetic datasets generated by both methods were integrated into the original training sets at various ratios, and the resulting changes in classification performance were observed. Both synthetic data generation approaches significantly improved the classification performance. However, as the amount of data used to train the classifiers increased, the marginal benefit of incorporating synthetic data decreased. These findings suggest that producing and utilizing synthetic data can be an effective strategy in text classification tasks that suffer from data scarcity..

Keywords: *Text classification, Variational Autoencoders, synthetic text generation*

https://doi.org/10.56038/oprd.v5i1.581

1.Giriş

Doğal dil işleme alanında yapılan çalışmaların başarısı model mimarilerine olduğu kadar eğitimde kullanılan veri miktarına ve kalitesine de bağlıdır. Veri büyüklüğünün yetersiz kaldığı durumlarda modellerin genelleme ve öğrenme kabiliyeti azalmakta, eğitildikleri görev için sundukları performanslar sınırlanmaktadır. Veri kaynaklarının İngilizceye göre sınırlı olduğu diller için bu sorun daha da kritik hale gelmektedir. Türkçe gibi dillerde hem ham veri setlerinin hem de etiketlenmiş veri setlerinin yetersizliği, hem Türkçe'nin yapısının doğru şekilde modellenmesi sınırlamakta hem de Türkçe dilindeki doğal dil işleme çalışmalarının gelişimini ciddi şekilde kısıtlamaktadır. Böylece Türkçe doğal dil işleme uygulamalarının kapsamı daralmakta, sınırlı verilerle eğitilen modeller çeşitli görevlerde yetersiz kalabilmektedir.

Metin sınıflandırma, doğal dil işleme çalışmalarının temel görevlerinden biri olup duygu analizi, istenmeyen e-posta tespiti, konu sınıflandırma ve sosyal medya içeriklerinin analizi gibi çeşitli alanlarda aktif olarak kullanılmaktadır. Özellikle sosyal medya platformlarında günlük olarak üretilen büyük miktardaki metin verisinin otomatik olarak sınıflandırılması, toplumsal olayların izlenmesi ve pazarlama stratejilerinin belirlenmesi gibi önemli konular için kritik bir rol oynar. Metin sınıflandırma için etkili modellerin geliştirilmesinde yeterli miktar ve kalitede etiketlenmiş veri kümelerinin olması çok önemlidir. Türkçe dilinde bu tür büyük ve kaliteli veri setlerinin sınırlı olması, bu alanda geliştirilen uygulamaların önündeki engellerden biridir.

Metin verilerini artırmak için çeşitli yaklaşımlar bulunmaktadır. Geleneksel yöntemlerden birisi mevcut metin verisini genişletmek için cümle içindeki bazı kelimelerin eş anlamlılarıyla yer değiştirilerek yeni cümleler üretilmesidir. Bunun gibi kelime düzeyinde gerçekleştirilen farklı veri artırma yöntemleri de literatürde yer almaktadır [1],[2],[3]. Veri setinde anlamı bozmadan farklı cümleler elde etmek için kullanılan bir başka yöntem ise geri-çeviri [4] yöntemidir. Bu yöntemle dilden dile başarılı çeviri yapan modellerden faydalanılarak cümle çeşitliliği artırılabilmektedir. Bu yöntemler, veri yetersizliğinden kaynaklanan engelleri aşmak ve dil modellerinin performansını artırmak için güçlü bir alternatif olarak değerlendirilmektedir.

Son yıllarda, derin öğrenme tabanlı üretici modeller, metin veri artırma alanında önemli bir ilgi görmektedir. Bu modellerden biri olan Varyasyonel Otokodlayıcılar (Variational Autoencoders-VAE), giriş verilerini daha düşük boyutlu bir gizli uzaya kodlayan ve bu uzaydan örneklemeler yaparak yeni veriler üretebilen olasılıksal üretici modellerdir. VAE'ler, karmaşık veri dağılımlarını öğrenebilme ve bu dağılımlardan anlamlı ve çeşitli

https://doi.org/10.56038/oprd.v5i1.581

örnekler üretebilme yetenekleri sayesinde çeşitli görevlerde etkili bir şekilde kullanılmaktadır [5]. VAE'ler kullanılarak farklı nitelikleri (duygu, zaman) kontrol edilebilen metin verileri üretmek mümkün olmaktadır [6]. Metin üretiminde VAE kullanılmasına bir diğer örnek ise girdi olarak verilen hikayeleri tamamlayan modeller gösterilebilir [7].

Bu çalışmada, metin sınıflandırma modellerinin performansını artırmak ve Türkçe dilinde veri yetersizliği sorununu ele almak amacıyla, kullanılan veri setlerinin her biri için Koşullu Varyasyonel Otokodlayıcı (Conditional VAE-CVAE) modeli eğitilmiştir. CVAE modeli, eğitim sırasında metin verileriyle birlikte sınıf etiketlerini de kullanarak her bir sınıfın özelliklerini öğrenmektedir. Bu sayede istenilen sınıfa ait yeni ve anlamlı metinler oluşturmak mümkün hale gelmektedir. Farklı sınıf sayısına sahip veri setleri için bu model kullanılarak sentetik veriler üreterek, orijinal veri setlerinin büyüklüğü ve çeşitliliği artırılmıştır. Üretilen sentetik metinler, orijinal veri setine farklı oranlarda eklenerek metin sınıflandırma modellerinin eğitiminde kullanılmıştır. Çalışmanın temel amacı, sentetik verilerin metin sınıflandırma modellerinin performansına olan etkisini kapsamlı bir şekilde değerlendirmektir. Yapılan deneyler sonucunda, sentetik verilerin eklenmesinin sınıflandırma başarısını iyileştirdiği ve veri kaynağının sınırlı olduğu görevlerde veri artırma yöntemi olarak etkili bir seçenek olduğu gözlemlenmiştir. Bu çalışma Türkçe dilinde metin üretimi yapabilen ve sınıf etiketleriyle koşullandırılmış bir CVAE modelinin geliştirilmesini, eğitilmesini ve sentetik metin verisi üretim süreçlerini içermektedir. Gürültüden metin üretimi ve cümle tamamlama gibi iki farklı metin üretme yöntemi kullanılarak üretilen sentetik metinlerin, kalitesi ve sınıflandırma performansına etkisi karşılaştırılmıştır. Sonuçlar, sentetik verilerin metin sınıflandırma modellerinin eğitimine dahil edilmesinin performansı anlamlı ölçüde artırdığını göstermektedir. Bu çalışma, Türkçe doğal dil işleme alanına katkı sağlamayı ve kaynak yetersizliği bulunan veri setlerinde veri artırma yöntemlerinin etkinliğini ortaya koymayı hedeflemektedir.

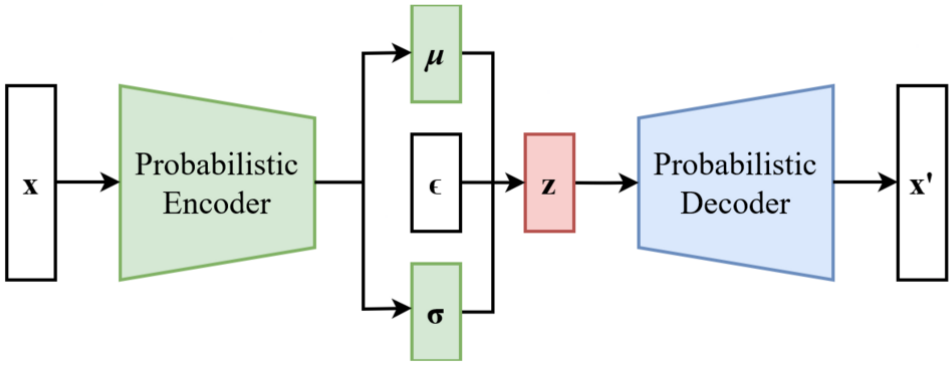
2.Materyal ve Yöntem

2.1.Varyasyonel Otokodlayıcılar

Otokodlayıcılar, bir giriş verisini daha düşük boyutlu bir temsile sıkıştıran ve bu temsili kullanarak orijinal veriyi yeniden oluşturmaya çalışan yapay sinir ağı modelleridir. Temel olarak bir kodlayıcı ve bir kod çözücü bileşeninden oluşurlar. Kodlayıcı, giriş verisini gizli bir uzaya kodlarken; kod çözücü gizli uzay temsilinden giriş verisini yeniden oluşturmaya çalışır. Modelin hedefi giriş verisi ile yeniden oluşturulan veri arasındaki farkı en aza indirmektir.

https://doi.org/10.56038/oprd.v5i1.581

Varyasyonel Otokodlayıcılar, geleneksel otokodlayıcılardan farklı olarak gizli uzaydaki temsilleri olasılıksal bir yaklaşımla oluşturmaya çalışırlar. VAE'lerde, kodlayıcı giriş verisini alarak gizli uzayda bir olasılık dağılımına dönüştürür. Bu dağılım ortalama (μ) ve standart sapma (σ) parametreleriyle ifade edilir. Kod çözücü ise bu dağılımdan örneklenen temsilleri kullanarak veriyi yeniden oluşturur [8]. (Şekil 1)



Şekil 1: VAE Mimarisi [https://www.wikiwand.com/en/Variational_autoencoder]

VAE'ler, otokodlayıcılara kıyasla yeni veri örnekleri üretebilme becerisine sahiptir. Gizli uzaydaki olasılıksal yaklaşım, modelin çeşitlilik içeren yeni veriler üretmesine olanak tanır. Normal dağılımdan örnekleme yaparak, eğitim verisine benzer ancak ondan farklı örnekler oluşturulabilir.

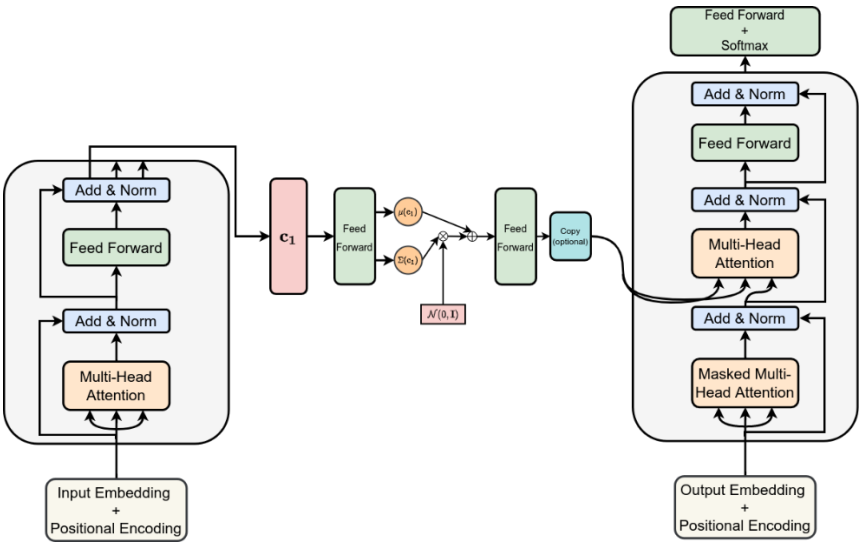
2.2.Koşullu Varyasyonel Otokodlayıcılar

Koşullu Varyasyonel Otokodlayıcılar, geleneksel Varyasyonel Otokodlayıcılardan farklı olarak, modelin belirli bir koşula bağlı olarak veri üretmesini sağlar. Bu yaklaşım, üretilen verilerin istenilen özelliklere veya sınıfa ait olmasını mümkün kılar. Özellikle kaynak yetersizliği veya sınıf dengesizliği bulunan veri setlerinde belirli sınıfların örnek sayısını artırarak veriden kaynaklı dezavantajı ortadan kaldırabilir. CVAE'lerde, VAE'lerin temel yapısına ek olarak koşul bilgisi de dahil edilir. Kodlayıcı aşamasında, giriş verisi ve koşul bilgisi birlikte işlenerek gizli uzaydaki temsiller oluşturulur, bu temsiller hem giriş verisinin hem de koşul bilgisinin özelliklerini içerir. Kod çözücü aşamasında ise temsiller ve koşul bilgisi kullanılarak veri yeniden oluşturulur. Bu yapı, modelin belirli bir sınıfa veya özelliğe ait çıktılar üretmesini mümkün kılar.

2.3.Dönüştürücü (Transformer) Tabanlı CVAE

https://doi.org/10.56038/oprd.v5i1.581

Geleneksel olarak, metin verisiyle çalışan VAE'lerin iç yapılarında Yinelemeli Sinir Ağları (RNN) [9] veya Uzun Kısa Süreli Bellek (LSTM) [10] kullanılmaktadır. Ancak, RNN ve LSTM tabanlı modellerin eğitim sırasında paralel çalışmaya olanak tanımaması, kaybolan gradyan (vanishing gradient) sorunları yaşanması ve uzun cümlelerde kelimeler arasındaki bağlantıların ve bağlamın yeterince öğrenilememesi gibi sorunlarla karşılaşılmaktadır. Bu çalışmada kodlayıcı ve kod çözücü bileşenlerinde dönüştürücü mimarisini kullanan bir CVAE modeli üzerinde çalışılmıştır [11]. Dönüştürücü tabanlı CVAE modeli, eğitim aşamasında girdileri paralel olarak işleyebilmekte ve öz dikkat (self-attention) mekanizması sayesinde kelimelerin bağlamlarını daha iyi anlayabilmektedir. Böylece RNN ve LSTM tabanlı modellerin sınırlarına takılmamaktadır. CVAE modelinde bulunan kodlayıcı yapısı, girdiyi *(x)* ve koşul bilgisini *(c)* gizli bir temsil uzayına *(z)* gauss dağılımı olarak dönüştürmektedir ($q_\phi(z \mid x, c)$) Bu dağılım sadece ortalama ve varyans değerleri ile ifade edilebilen, kovaryans matrisi diyagonal olan bir matristir. Modelin kod çözücü bileşeni de dönüştürücü tabanlı olup, gizli uzaydaki temsilleri (z) ve koşul bilgisini (c) kullanarak çıktı üretmektedir (Şekil 2). Eğitimin amacı yeniden üretme hatası (reconstruction loss) ve Kullback-Leibler (KL) hata değerlerinden oluşan ELBO'yu maksimize etmektir. Yeniden üretme hatası, z ve c değerleri verilen kod çözücünün ürettiği çıktı ile girdi (x) arasındaki "negatif log-likelihood" değeridir. KL hata değeri ise gizli uzaydaki temsillerin dağılımı ile gauss dağılımı arasındaki farkı belirten hata değeridir.



Şekil 2: CVAE Mimarisi [11]

https://doi.org/10.56038/oprd.v5i1.581

2.4.Veri Setleri

Bu çalışmada, CVAE modelini eğitmek için TSNews [12] isimli Türkçe haber metinlerinden oluşan bir veri seti kullanılmıştır. TSNews veri seti, farklı kategorilere ayrılmış haber metinlerinde oluşmakta ve Türkçe doğal dil işleme çalışmalarında kullanılabilecek zengin bir kaynak sunmaktadır. Modelin performansının farklı sınıf sayılarındaki değişimini de incelemek amacıyla, veri setinden iki farklı alt örneklem oluşturulmuştur.

İlk örneklem yalnızca politika ve spor kategorilerindeki metinlerden oluşturulmuştur. Bu örneklem, modelin iki sınıflı metin üretim başarımını ve sınıflandırma performansını değerlendirmek için hazırlanmıştır. Eğitim verisi olarak 2 milyon, test verisi olarak ise 400 bin örnek kullanılmıştır. İkinci örneklem ise, sınıf sayısının arttığı durumlarda modelin performansını test etmeyi amaçlamıştır. Bu örneklem, spor, politika, teknoloji ve sağlık kategorilerindeki metinlerden oluşturulmuş olup, 3 milyon eğitim ve 600 bin test örneği içermektedir.

Veri setlerinin hazırlanması sırasında uzun haber metinleri, cümle düzeyinde parçalanmış ve birimlerine ayrılmıştır. Örneklem seçiminde, birimlerine ayırma sonrası 32 birim (token) veya daha kısa olanlardan seçim yapılmıştır.

Her iki veri seti için sınıflandırma performanslarının ölçümünde kullanılması amacıyla kategori başına bin eğitim ve 10 bin test örneği olacak şekilde, verilerin çakışmadığı birer yan veri seti hazırlanmıştır.

2.5.CVAE Modeli Eğitimi ve Sentetik Verilerin Üretilmesi

Metin verilerini yanında koşul bilgisi olacak şekilde modellemek ve sentetik veri üretmek için tasarlanan CVAE modeli, Türkçe haber metinlerinden oluşan, bir tanesi iki sınıftan, diğeri dört sınıftan örnek içeren iki farklı veri seti üzerinde eğitilmiştir. Veri setlerinin yapısal gereksinimlerini ve modelin öğrenme kapasitesi göz önünde bulundurularak hiperparametreler seçilmiş ve deneysel olarak optimize edilmiştir.

Modele girdi olarak verilecek metinler için maksimum uzunluk 32 birim ile sınırlandırılmıştır. Veri setleri hazırlanırken de bu sınırlamaya dikkat edildiği için, tüm örnekler kesilmeye uğramadan modele girdi olarak verilmiştir. Giriş verilerinin sayısal temsilleri için kelime vektörlerinin boyutu 256 olarak belirlenmiştir. Kodlayıcı ve kod çözücü katmanları 6 katmandan oluşan bir yapı ile tasarlanmıştır. Her katmanda bulunan çoklu öz dikkat (multi-head self-attention) mekanizması için başlık (head) sayısı 3 olarak seçilmiştir. Modellerin gizli uzay (latent space) boyutu 256 olarak belirlenmiştir. Modelin eğitim sürecinde, farklı öğrenme oranları (learning rate) denenerek en başarılı sonuç veren 0.0001 değeriyle modeller eğitilmiştir. Her iki veri seti için ayrı ayrı olmak

https://doi.org/10.56038/oprd.v5i1.581

üzere iki farklı model 40'ar çevrim (epoch) eğitilmiştir. Eğitim sırasında kullanılan hata fonksiyonu, Varyasyonel Otokodlayıcılar için standart olarak kullanılan Evidence Lower Bound (ELBO) kaybıdır. Bu hata fonksiyonu, modelin hem giriş verisini doğru bir şekilde yeniden üretebilmesini hem de gizli uzaydaki temsillerin gauss dağılımına yakınsamasını sağlamaktadır.

Modellerin eğitimi bittikten sonra iki farklı yöntemle sentetik veriler üretilmiştir. İlk yöntemde, gizli uzay temsilinin boyutuna uygun olacak şekilde Gauss dağılımından örnekleme yapılarak bir temsil (z) oluşturulur. Bu temsile, üretilecek verinin ait olması istenilen sınıfın koşul bilgisi (c) eklenir. Ardından, oluşturulan z temsili ve başlangıç birimi ([START]) ile başlatılan hedef bir dizi kullanılarak modelin kod çözücü bileşeni tarafından yinelemeli bir üretim süreci gerçekleştirilir. Her adımda model, bir sonraki birim için sözlük üzerinde bir olasılık dağılımı üretir ve bu dağılımdan en iyi-k yöntemiyle k = 5 değeri kullanılarak en yüksek olasılığa sahip beş birim seçilir. Modelin verdiği olasılıklara göre bu beş birimden bir tanesi rastgele seçilerek hedef dizinin sonuna eklenir. Bu süreç, dizinin önceden belirlenen maksimum uzunluğa ulaşmasına veya bir bitiş birimi ([END]) üretilmesine kadar devam eder. Üretilen birim dizisi metne dönüştürülerek istenilen sınıfa ait sentetik cümle hazırlanmış olur. İkinci yöntemde ise, sınıflandırma modellerinin eğitiminde kullanılacak orijinal metinler kullanılmıştır. Modelin kod çözücü bileşenine verilen girdide ilk yöntemde sadece başlangıç birimi kullanılmıştır. Bu yöntemde orijinal metinler birimlerine ayrılarak ikiye bölünmüş ve ilk yarısı kod çözücüye verilmiştir. Böylece modelin sıfırdan bir örnek üretimi yerine, gerçek bir cümlenin yarısını nasıl tamamladığı ölçülmüştür.

Bahsedilen yöntemlerden ilkiyle, iki veri setindeki eğitim örneklerinin sayısı kadar sentetik örnek oluşturulmuştur. Yöntemlerden ikincisi için, iki veri setindeki eğitim örneklerinin her biri için yarısından üretme yöntemi uygulanarak sentetik örnekler oluşturulmuştur.

2.6.Metin Sınıflandırma Modellerinin Eğitilmesi ve Performanslarının Kıyaslanması

Metin sınıflandırma performanslarının ölçülmesi için, metinlerin BERT [13] tabanlı temsilleri çıkarılmış ve bu temsiller kullanılarak çeşitli makine öğrenimi modelleri eğitilmiş ve test edilmiştir. Sınıflandırma için farklı algoritmaların performanslarını değerlendirebilmek amacıyla CatBoost, Extra Trees ve Naive Bayes modelleri seçilerek eğitilmiştir. Farklı modellerin tercih edilmesi, performans değişimlerinin algoritmanın türünün sonuçlara yapacağı etkiyi azaltması nedeniyledir.

Sınıflandırma modelleri, iki veri seti için, iki farklı yöntemle üretilen sentetik veriler kullanılarak eğitilmiştir. Eğitim süreçleri, modellerin hem yalnızca orijinal verilerle hem

https://doi.org/10.56038/oprd.v5i1.581

de sentetik verilerle desteklenmiş verilerle olan performanslarını kıyaslamak için tasarlanmıştır. Modellerin eğitim sürecinde farklı örnek sayıları için; ilk olarak yalnızca orijinal veriyle eğitimler yapılmış, başarımlar baz doğruluk oranları olarak kaydedilmiştir. Ardından, orijinal verinin üzerine sırasıyla %25, %50, %75 ve %100 oranlarında sentetik veri eklenerek modeller eğitilmiştir. Bu çalışmaların her biri 5 kez tekrarlanarak sonuçların daha güvenilir bir şekilde değerlendirilmesi sağlanmıştır.

3.Sonuçlar

Sentetik metin üretimi için eğitilen CVAE modeli kullanılarak iki farklı yöntemle sentetik metin verileri üretilmiştir. Bu yöntemler, materyal ve yöntemler bölümünde anlatılan gürültüden metin üretimi ve cümle tamamlama yöntemleridir.

Gürültüden üretim yöntemi, sınıfların genel özelliklerini yansıtan ancak dilbilgisel hatalar ve anlamsız ifadeler içeren örnekler üretmiştir. Üretilen örnekler Tablo 1'de görülebilir. Tablo 1'de görülen cümleler, gürültüden üretme yönteminin metin bağlamını yeterince öğrenemediğini, böylece anlamlı ve uygun yapıda cümleler üretmediğini göstermektedir. Bununla birlikte, üretilen örnekler sınıfa özgü temel temaları ve kelime gruplarını içerdiği için sınıflandırma modellerinden alınacak performansa olumlu yönde etki edebileceği düşünülmektedir.

Tablo 1: TSNews-4 Sınıflı Veri Seti İle Eğitilen CVAE Modelinin Gürültüden Ürettiği Sentetik Örnekler

Kategori	Üretilen Örnek
Sağlık	araştırmanın sonuçları, yapılan açıklamada şu ifadelere.'' ın ile.
	özellikle çocuğun, çocuğun gelişimindeki önemli değişiklikler, kendi kendine özgü bir sorun olmaktan kaynaklandığını vurgulayan op.
	bu pozisyonda, mide, bağırsak ve bağırsaklar, bağırsak, mide, mide ve bağırsak gibi bir hastalıkdır.
	bu nedenle tedavi yöntemleri ile karşılaşıyoruz.
Politika	bu yüzden de bu konuda mutluyum.. gibiler.
	türkiye ' de ilk kez ortaya çıktı..i vete.
	bu nedenle sosyal demokrat olarak görev alacak.!!.,da gibi
	biz de,'sen'ya göre'' ', türkiye'diyoruz, türkiye.
Teknoloji	cep telefonunun internet sitesinde, apple ' ın bu kez piyasaya çıkacağını duyurdu. ve.
	yapılan bu araştırmaya göre, iphone 6s ' in iphone 6 ' yı piyasaya çıkaracağı ortaya çıktı. oldu.
	bu sayede bu konuda bir şey yapmak, bu kadar çok farklı bir şey olacak ve çok boyutlu olacaktır.
	bu arada google ' ın geçtiğimiz dönemde ve'nin ve.
Spor	dün gece, 30 dakika sonra, galatasaray ' a karşı ilk kez bu kez bir kez tekrar oynayacak.
	bu yüzden futbol adına teşekkür ediyorum.nden olaraki!
	bunun için türkiye'de bir hukuk ve demokrasi var ama bu tür söylemlere ihtiyaç var.
	galatasaray, bu sezon galatasaray ile beşiktaş ' ın resmi internet sitesinden, beşiktaş derbisinde de görev aldı..nda

https://doi.org/10.56038/oprd.v5i1.581

Cümle tamamlama yönteminde ise yapı olarak Türkçeye daha uygun cümleler üretilmiştir. Üretilen örnekler Tablo 2'de görülebilir. Ancak bu yöntemle de üretilen cümlelerde anlam kopukluğu ve bağlam dışı ifadeler gözlemlenmektedir.
Her iki yöntemle üretilen metinlerin genel kalitesi yüksek olmasa da ait oldukları sınıfların genel temalarına uygun kelimeler bulundurmaları nedeniyle sınıflandırma modellerinin performansını artırmada faydalı olabilecekleri değerlendirilmektedir. Özellikle az veri bulunan durumlarda, bu tür sentetik verilerin etkili olabileceği düşünülmektedir.

Tablo 2: TSNews-4 Sınıflı Veri Seti İle Eğitilen CVAE Modelinin Cümle Tamamlama Yöntemi İle Ürettiği Sentetik Örnekler

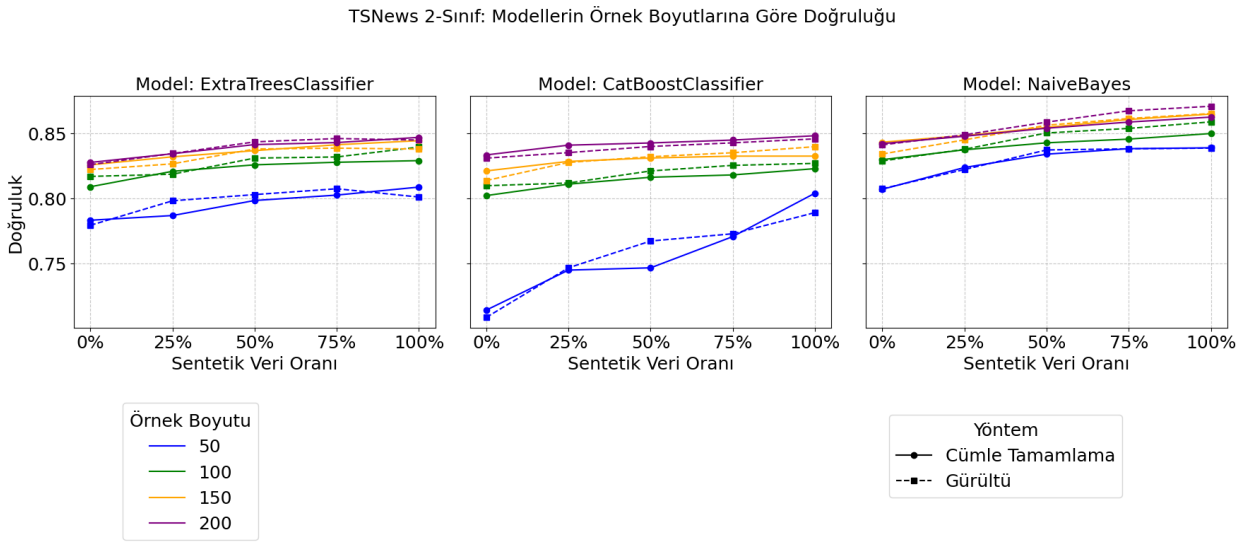
Kategori	**Orijinal Örnek**	**Üretilen Örnek**
Sağlık	dolaşım sisteminizdeki hızlanma böbreklere de yansıdığından ihtiyaç giderme sayınızda artacak.	dolaşım sisteminizdeki hızlanma böbreklere her türlü besin olarak zarar veriyor, bunu çok önemli değildir.
	lif oranı yüksek olduğu için kabızlık şikâyetlerinizi azaltır ve tokluk hissi verir.	lif oranı yüksek olduğu için kabızlık şikâyeti size nasıl davranır, nasıl çalışırız?
Politika	zararların giderilmesini, kayıpların tazmin edilmesini sağlayacağız.	zararların giderilmesini, kayıpların yanı sıra, 1. 5 lt. 5 litreye sahip çıktı.! seti
	devraldıkları ülke böyle bir ülke değildi.	devraldıkları ülke böyle olacak. için.
Teknoloji	yeni seride uygulanan tasarımla bu kez yağmur damlası gibi hafif ıslak ortamlarda da müzik dinlenmesi mümkün oluyor.	yeni seride uygulanan tasarımla bu kez yağmur damlası bir şey daha çok iyi olur diye düşünüyorum.
	cihazda 2 gb ram bulunuyorken, depolama seçenekleri olarak 16 / 32 / 64 ve arttırılabilir hafıza olarak göze çarpıyor.	cihazda 2 gb ram bulunuyorken, depolama seçenekleri olarak 16 tane değil o kadar çok şey yapmıyorlar.
Spor	fenerbahçe'de sezona hızlı başlayan ancak ikinci yarıda performansı düşen luis nani, euro 2016 için yeni bir sayfa açtı.	fenerbahçe ' de sezona hızlı başlayan ancak ikinci yarıda performansı düşen lu bizim için çok önemli bir şey.
	mori, topa vurmak isterken, bursaspor savunması meşin yuvarlağı tehlike bölgesinden uzaklaştırdı.	mori, topa vurmak isterken, bursaspor ile de aynı sezon içinde yapılacak ilk bir hafta daha oynayacağını söyledi.

Bu bölümde, daha önceden eğitilen CVAE modelleri kullanılarak, iki farklı veri üretim yaklaşımı olan "cümle tamamlama" ve "gürültüden üretme" yöntemleriyle üretilen sentetik verilerin, metin sınıflandırma performansına etkileri değerlendirilmiştir. Değerlendirme; sınıf sayısı, örnek boyutu ve eğitimde eklenen sentetik veri oranlarına göre yapılan kıyaslamalarla detaylandırılmıştır. Kıyaslamalar, her bir deneyin 5 kez tekrar çalıştırılarak alınan ortalamasıyla yapılarak sonuçların güvenilirliği artırılmıştır.

https://doi.org/10.56038/oprd.v5i1.581

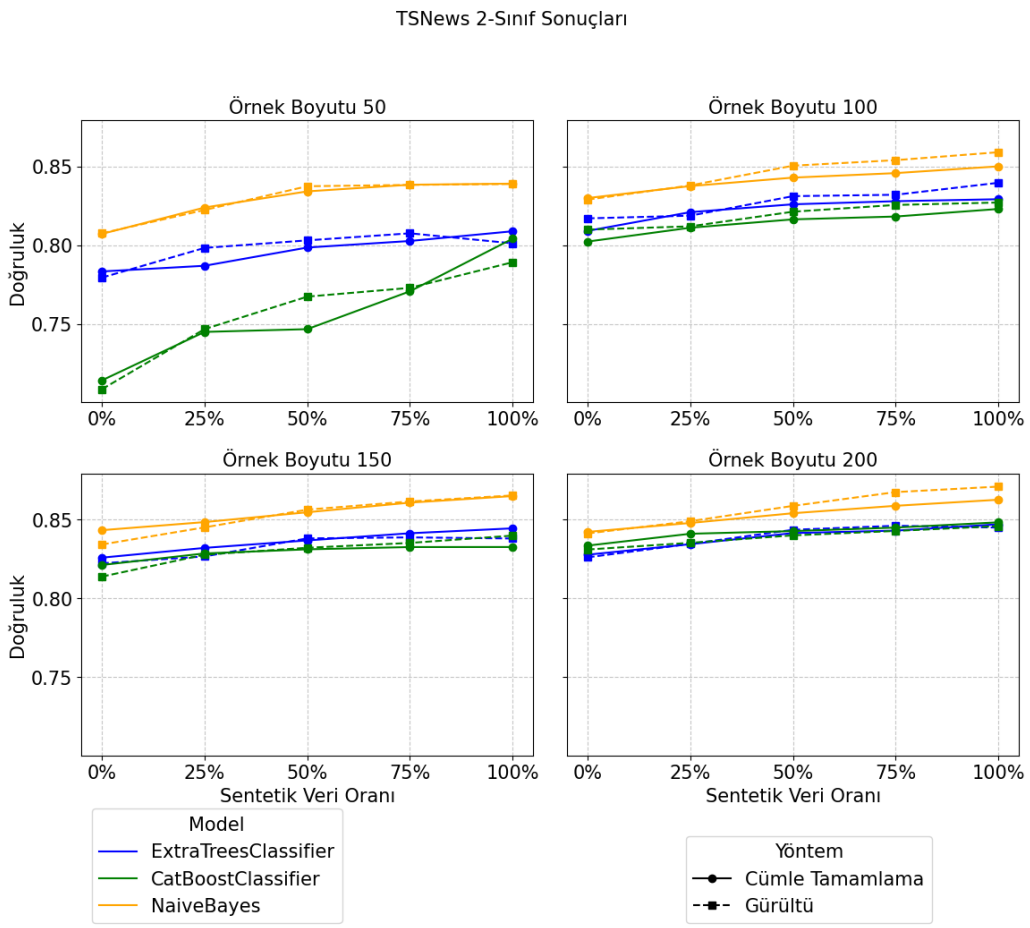
İki sınıflı veri setinde, her iki yöntem de baz doğruluk değerine kıyasla anlamlı performans artışı sağlamıştır. Özellikle eklenen sentetik verinin düşük oranlarda olduğu (%25 ve %50) durumda, elde edilen doğruluk artışlarının daha anlamlı olduğu, ancak eklenen veri oranı arttıkça bu artışın daha sınırlı bir seviyeye ulaştığı gözlemlenmiştir. Bu durum, düşük oranlarda eklenen sentetik verilerin modelin genelleme kapasitesini artırdığını, ancak eklenen sentetik veri oranı arttığında katkının azaldığını göstermektedir.

Şekil 3'te, TSNews 2-Sınıf veri setinde, eğitilen her bir modelin, farklı sentetik veri üretme yöntemleri (cümle tamamlama ve gürültüden üretme) kullanılarak, orijinal eğitim kümesine eklenen farklı oranlardaki sentetik verinin (%25, %50, %75, %100) doğruluk değerine etkisi, farklı örnek sayılarında (50, 100, 150, 200) incelenmiştir. Şekil 4'te ise farklı örnek sayıları için, eğitilen modellerin (ExtraTreesClassifier, CatBoostClassifier, NaiveBayes) ve farklı sentetik veri üretme yöntemlerinin, eklenen sentetik veri oranlarına göre modelin doğruluk değeri üzerindeki etkisi karşılaştırılmıştır.



Şekil 3: TSNews 2-Sınıf Verisi ve Sentetik Verilerin Sınıflandırma Başarıları (5 tekrarın ortalaması)

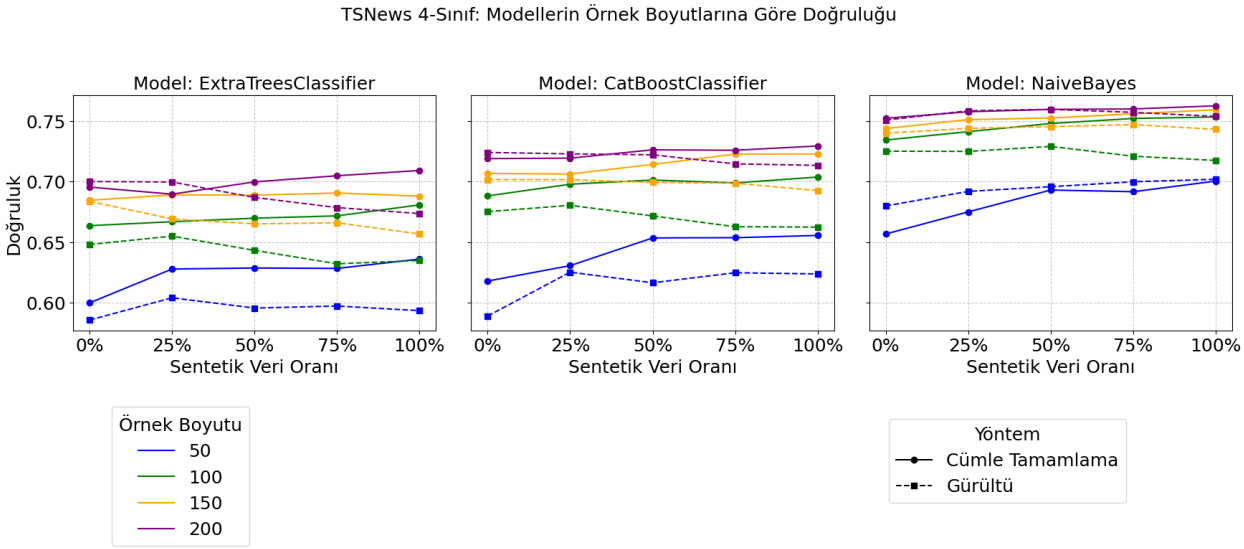
https://doi.org/10.56038/oprd.v5i1.581



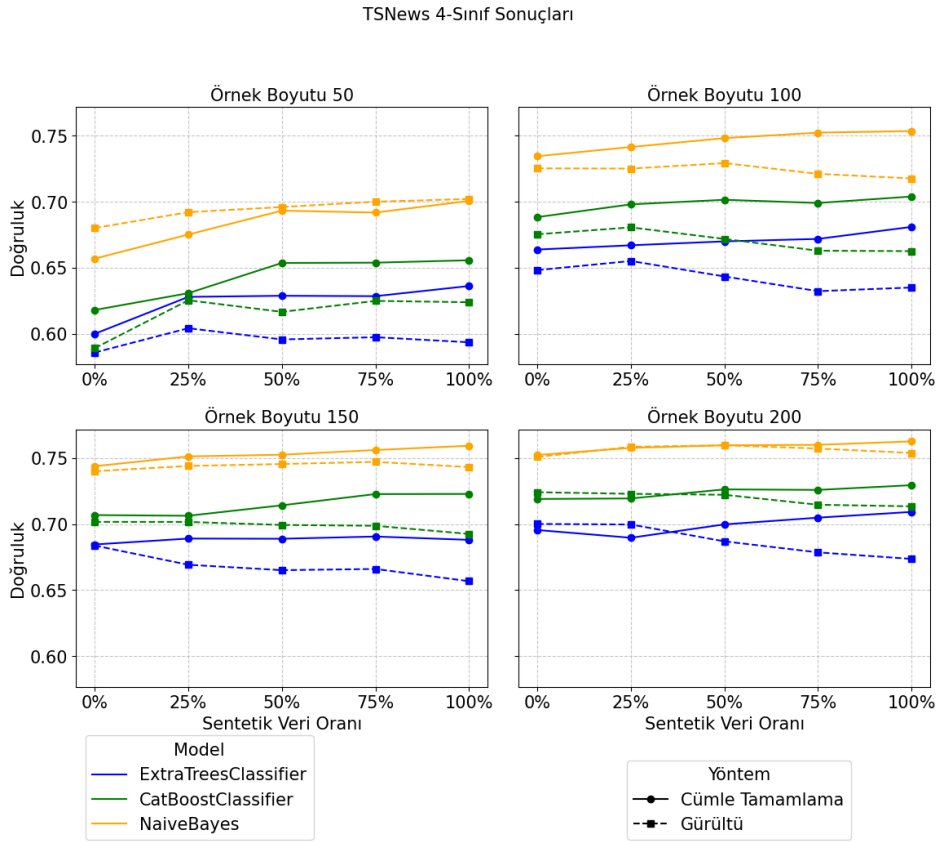
Şekil 4: TSNews 2-Sınıf Verisi ve Sentetik Verilerin Sınıflandırma Başarıları (5 tekrarın ortalaması)

Şekil 5'te, TSNews 4-Sınıf veri setinde, her bir modelin farklı sentetik veri üretme yöntemleri (cümle tamamlama ve gürültüden üretme) kullanılarak, orijinal eğitim kümesine eklenen farklı oranlardaki sentetik verinin (%25, %50, %75, %100) doğruluk değerine etkisi, farklı örnek sayılarında (50, 100, 150, 200) incelenmiştir. Şekil 6'da ise farklı örnek sayıları için modellerin (ExtraTreesClassifier, CatBoostClassifier, NaiveBayes) ve farklı sentetik veri üretme yöntemlerinin, eklenen sentetik veri oranlarına göre modelin doğruluk değeri üzerindeki etkisi karşılaştırılmıştır.

https://doi.org/10.56038/oprd.v5i1.581



Şekil 5: TSNews 4-Sınıf Verisi ve Sentetik Verilerin Sınıflandırma Başarıları (5 tekrarın ortalaması)



Şekil 6: TSNews 4-Sınıf Verisi ve Sentetik Verilerin Sınıflandırma Başarıları (5 tekrarın ortalaması)

https://doi.org/10.56038/oprd.v5i1.581

Dört sınıflı veri setinde, sonuçların daha farklı bir yapıda olduğu görülmektedir. Cümle tamamlama yöntemi ile üretilen sentetik veriler, özellikle düşük örnek sayılarında anlamlı bir performans artışı sağlamış, ancak örnek sayısı arttıkça yöntemin performans üzerindeki olumlu etkisi azalmıştır. Bu veri setinde gürültüden üretme yöntemi düşük örnek sayılarında bir miktar fayda sağlasa da örnek boyutu büyüdükçe modelin performansını düşürdüğü görülmektedir. Bu durum, CVAE modelinin yüksek sınıflı eğitiminde bu yöntemle ürettiği sentetik verilerin bağlamı yeterince temsil edememesiyle ilişkilendirilebilir.

Hem iki hem de dört sınıflı veri setlerinde, örnek boyutlarının artmasıyla birlikte sentetik verilerin sınıflandırma performanslarına olan katkısının azaldığı görülmektedir. Bu sonuçlar, veri setinin büyümesine bağlı olarak sentetik verilerin katkılarının sınırlı hale geldiğini göstermektedir.

4.Tartışma

Bu çalışmada, Türkçe metin verileriyle Koşullu Varyasyonel Otokodlayıcı modeli eğitilerek, bu modelden sentetik metin verilerinin üretilmesi ve bu verilerin metin sınıflandırma modellerinin performansına etkisinin incelenmesi amaçlanmıştır. Veri yetersizliğinin problem yarattığı Türkçe doğal dil işleme görevlerinde, özellikle sınıflandırma performansının geliştirilmesi için sentetik verileri kullanmanın etkili bir yaklaşım olabileceği gösterilmiştir. Elde edilen bulgular hem iki sınıflı hem de dört sınıflı veri setlerinde sentetik verilerin metin sınıflandırma problemlerinde performans artışı sağlayabildiğini ortaya koymaktadır.

İki sınıflı veri setinde elde edilen sonuçlar hem gürültüden üretme hem de cümle tamamlama yöntemleriyle üretilen sentetik verilerin eğitim verilerine eklenmesiyle, sadece orijinal veri setiyle eğitilen modellerin performansını anlamlı şekilde yükselttiğini göstermiştir. Sentetik verinin eğitim verisine eklenme oranı arttıkça elde edilen kazanımların bir noktadan sonra sınırlı kaldığı gözlenmiştir. Böylece, az miktarda dahi olsa iyi üretilmiş sentetik verinin sınıflandırma performanslarına katkı sağlayabildiği görülmüştür.

Dört sınıflı veri setinde ise sonuçlar daha farklı bir dinamik sergilemiştir. Cümle tamamlama yöntemi, eğitim verisindeki örnek sayısının az olduğu koşullarda önemli bir katkı sunarken, veri miktarı arttıkça etkinin azaldığı görülmüştür. Gürültüden üretme yöntemi ise özellikle yüksek sınıf sayıları ve büyük veri miktarlarında istenen performans kazanımını sağlayamamıştır. Bu durum, üretici modelin daha karmaşık veri dağılımlarını ve sınıf yapısını temsil etmede zorlandığını ortaya koymaktadır. Buradan

https://doi.org/10.56038/oprd.v5i1.581

hareketle, farklı koşullar (veri setindeki sınıf sayısı, veri miktarı vb.) için farklı sentetik veri üretim stratejilerinin gerekebileceği söylenebilir.

Her iki veri setinde de veri miktarı arttıkça sentetik verilerin katkısının sınırlı hale gelmesi, sentetik verinin özellikle az veri bulunan durumlarda daha etkili bir çözüm olduğunu göstermektedir. Az kaynaklı dillerde veya veri toplamanın güç olduğu problemlerde, eğitim verisine eklenen sentetik veri modellerin genelleme kapasitesini artırarak performanslarını yükseltebilmektedir. Ancak, büyük veri setlerinin bulunduğu durumlarda sentetik verilerin sağlayacağı katkı sınırlı kalabilir. Bu nedenle, sentetik veri üretiminin katkısı hem mevcut veri miktarına hem de hedeflenen görevlerin niteliğine bağlı olarak değişebilmektedir.

Özetle, bu çalışma Türkçe gibi az kaynaklı bir dilde sentetik metin üretiminin metin sınıflandırma performansına olumlu katkı sunabileceğini göstermiştir. CVAE tabanlı yaklaşım, istenen sınıflara ait yeni örnekler oluşturarak model genellemesini güçlendirebilmekte ve gerekirse veri dengesizliği problemini hafifletebilmektedir. Gelecekteki çalışmalarda, daha gelişmiş dil modelleri veya daha farklı örnek üretme yöntemleri gibi süreçlerle sentetik verilerin kalitesini ve etkililiğini artırmak mümkün olabilecektir. Böylece, Türkçe de dahil olmak üzere az kaynaklı dillerdeki doğal dil işleme uygulamaları daha geniş kapsamlı, bazı sınırlardan arınmış ve yüksek performanslı hale getirilebilir.

5.Teşekkür

Bu çalışma Yıldız Teknik Üniversitesi Bilimsel Araştırma Projeleri Koordinatörlüğü tarafından FBA-2024-6070 proje numarasıyla desteklenmiştir.

Referanslar

[1] X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” 2016.

[2] Z. Xie, S. I. Wang, J. Li, D. Lvy, A. Nie, D. Jurafsky, and A. Y. Ng, “Data noising as smoothing in neural network language models,” 2017.

[3] J. Wei and K. Zou, “EDA: Easy data augmentation techniques for boosting performance on text classification tasks,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6382–6388.

[4] R. Sennrich, B. Haddow, and A. Birch, “Improving neural machine translation models with monolingual data,” 2016.

[5] HU, Zhiting, et al. Toward controlled generation of text. In: *International conference on machine learning*. PMLR, 2017. p. 1587-1596.

[6] T. Zhao, R. Zhao, and M. Eskenazi, "Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders," *arXiv (Cornell University)*, Jan. 2017, doi: 10.48550/arxiv.1703.10960.

[7] T. Wang and X. Wan, "T-cvae: Transformer-based conditioned variational autoencoder for story completion," in Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 5233–5239.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014.

[9] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," in Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 10–21.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] Bilici, M. Şafak; Amasyali, Mehmet Fatih. Transformers as neural augmentors: Class conditional sentence generation via variational Bayes. *arXiv preprint arXiv:2205.09391*, 2022.

[12] Sezer, T. (2021). TS TimeLine News Category Dataset (Version 001) [Data set]. TS Corpus. https://doi.org/10.57672/P23D-B492

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186