Data Augmentation with LLMs

Yusuf Enes Kurt
Bilgisayar Mühendisliği
Yıldız Teknik Üniversitesi
enes.kurt1@std.yildiz.edu.tr

***Öz*–Bu çalışmada, metin sınıflandırma görevlerinde veri çeşitliliğini artırmak amacıyla büyük dil modelleri (LLM) kullanılarak paraphrase tabanlı veri üretimi incelenmiştir. Veri setleri üzerinde yürütülen deneylerde, eldeki eğitim ve test verileri farklı oranlarda artırılarak MLPClassifier modeliyle eğitilmiştir. Metinleri sayısal vektör uzayında temsil etmek için gömme modeli kullanılmış, böylece metinlerin anlamsal benzerlikleri korunarak çeşitli sınıflandırma görevlerine uyarlanmıştır.**
Proje kapsamında PEGASUS ve T5 tabanlı paraphrase modellerinden yararlanılarak özgün cümlelerin anlamını koruyan ancak farklı ifade biçimlerine sahip yeni cümleler üretilmiştir. Eğitim ve test veri kümesindeki bu artışların doğruluğa etkisi gözlemlenmiş ve yorumlanmıştır.

Anahtar Kelimeler – Metin Arttırma (Text Augmentation), Metin Gömme (Text Embedding), Büyük Dil Modelleri (LLM), Ensemble, Paraphrase

GitHub Linki: *https://github.com/YEnesK/Data-Augmentation-with-LLMs*

I. GİRİŞ

Makine öğrenmesi projelerinde veri miktarı ve çeşitliliği, model başarısını doğrudan etkilemektedir. Veri arttırma (data augmentation) yöntemleri, eldeki veriyi yapay olarak zenginleştirerek modelin daha iyi genelleme yapmasını sağlar. Görüntü verileri için döndürme, çevirme veya parlaklık değiştirme gibi teknikler sıklıkla kullanılırken, metin tabanlı projelerde de "text augmentation" yöntemleri öne çıkmaktadır. Metin için veri arttırma, cümlelerin anlam bütünlüğünü korumakla birlikte çeşitli ekleme, silme veya dönüştürme işlemleriyle veriyi çeşitlendirir. Böylece modelin farklı varyasyonları öğrenmesi ve daha güçlü bir şekilde genellemesi hedeflenir. Makine öğrenmesinde gerçek hayat problemleriyle uğraşırken çoğu zaman uygun ve kaliteli veri anında bulunamadığı için bu tür veri arttırma yaklaşımları önem kazanmaktadır.

Projede, text augmentation tekniklerinden yararlanarak özellikle paraphrase üretebilen LLM'ler ile mevcut eğitim ve test verimizde çeşitli dönüşümler yapılmıştır. LLM'lerin yardımıyla orijinal cümlelerin içerik ve anlamını koruyarak alternatif ifadeler üretilmiş; böylece veri setindeki örnek sayısı ve çeşitliliği artırılmıştır. Bu sayede model, eğitimi sırasında daha farklı cümle yapıları ve kelime kullanımlarıyla karşılaşmış; yapılan deneyler sonucunda elde edilen doğruluk (accuracy) değerlerindeki değişimler gözlemlenmiştir. Proje kapsamında hem eğitim hem de test veri seti, paraphrase yöntemleriyle genişletilerek daha zengin bir eğitim süreci kurgulanmıştır.

II. DATASET

Projede iki adet dataset kullanılmıştır. Her iki dataset de text classification türündedir. İlki film yorumlarının toplandığı IMDB datasetidir. Dataset'te her satır için bir adet text ve label bulunmaktadır. Pozitif ve negatif yorum olmak üzere iki tür label vardır. İkinci dataset ise AG News'tir. AG News, 1 milyondan fazla haber makalesini barındıran büyük bir haber koleksiyonudur 2004'ten bu yana çalışan akademik haber arama motoru ComeToMyHead tarafından, 1 yıl içinde 2000'den fazla kaynaktan derlenen bu makaleler, dünya, spor, iş ve bilim/teknoloji (world, sports, business, sci/tech) gibi dört ana başlık altında sınıflandırılır. Her iki dataset'te de küçültme yapılmıştır

Datasetleri load etme, kesme, kaydetme, labelları eşit oranda yapma gibi işlemler dataset.ipynb kod dosyasında yapılmıştır.

III. TEXT EMBEDDING

Text embedding, metinleri bilgisayarların işleyebileceği biçimde sayısal vektörlere dönüştürerek anlamsal içeriği yakalamayı amaçlayan bir temsildir. Bu yöntem, insan dilini makinelere tanıtıp metin sınıflandırma, benzerlik analizi ve semantik arama gibi uygulamalarda kullanılır. Metin veya kelimelerin vektör uzayında yakın konumlanması, benzer kavramların veya anlamların paylaşıldığına işaret ederek doğal dil işleme süreçlerini kolaylaştırır. Bu projemizde kullanılan temsil modeli all-MiniLM-L6-v2'dir. Bu model, cümle ve kısa paragrafları 384 boyutlu bir vektör uzayında temsil eden kompakt ama güçlü bir modeldir Model, 1 milyardan fazla cümle çiftinden oluşan büyük bir veri kümesi üzerinde contrastive learning yaklaşımıyla eğitilmiştir. Bu sayede anlamsal benzerlik, bilgi erişimi, kümelendirme ve semantik arama gibi görevlerde etkili sonuçlar elde etmeye olanak tanır.

Orijinal ve augmente edilmiş text datasetinin gömme işlemi embedding.ipynb kod dosyasında yapılmıştır.

IV. MODEL TRAINING

Model eğitimi için MLP Classifier ML algoritması kullanılmıştır. MLPClassifier, bir Yapay Sinir Ağı modeli olan Çok Katmanlı Algılayıcı (Multi-Layer Perceptron) yaklaşımını uygular. Özellikle sınıflandırma ve regresyon gibi gözetimli

öğrenme problemlerinde sıklıkla kullanılır. Model, katmanlar arasındaki ağırlıkları güncellemek için geri yayılım (backpropagation) sürecine dayanır ve her gizli katman, verideki karmaşık ilişkileri yakalamaya yardımcı olur. Bu sayede, IMDb gibi film yorumlarını sınıflandırma, AG News gibi haber içeriği sınıflandırma görevlerinde etkili bir performans sergiler.

Embedding verileri ile eğitim, test evaluation, ensemble evaluation işlemleri train.ipynb kod dosyasında yapılmıştır.

V. DATA AUGMENTATION

Projemiz kapsamında, train veri seti 2, 3 ve 5 katına, test veri seti ise 4 ve 6 katına çoğaltılmıştır. Bu çoğaltma işlemi için "text augmentation" yöntemleri kullanılarak, aynı anlamı taşıyan ancak farklı biçimlerde ifade edilen cümleler üretildi.

Burada PEGASUS ve T5 tabanlı modellerden yararlanıldı. tuner007/pegasus_paraphrase, Google Research ekibince geliştirilen PEGASUS modelinin paraphrase amacıyla ince ayar (fine-tuning) yapılmış sürümüdür. humarin/chatgpt_paraphraser_on_T5_base modeli, ChatGPT'den alınan paraphrase dataset kullanılarak T5-base üzerinde transfer öğrenme yaklaşımıyla eğitilmiş ve yüksek kalitede paraphrase üretebilme potansiyeline sahiptir. mrm8488/t5-small-finetuned-quora-for-paraphrasing ise Quora soru eşleştirme veri setiyle T5-small üzerinde ince ayar yapılarak soru ve cümlelerin yeniden ifade edilmesini hedefler.

Bu modeller sayesinde, proje kapsamında hem eğitim hem de test veri setlerine yeni ve anlamca tutarlı ifadeler eklenmiştir. Böylece model, daha fazla ve çeşitli örneğe maruz kalarak farklı ifade biçimlerini öğrenebilmiş; sonuç olarak elde edilen doğruluk (accuracy) değerlerinde önemli değişimler gözlemlenmiştir.

LLMlerin yüklenmesi ve bu modeller aracılığıyla yapılan data augmentation işlemleri augmentation.ipynb kod dosyasında yapılmıştır.

VI. SONUÇLAR VE YORUM

		test 0.5k		test 2k		test 3k		2k test ensemble		3k test ensemble
imdb llm1		accuracy		accuracy		accuracy		accuracy		accuracy
train 2k		77.2		69.1		68.5		70.2		69.6
news llm1										
train 2k		86.6		86.0		85.9		87.6		87.4
imdb llm2										
train 2k		77.2		72.7		72.5		78.2		76.6
news llm2										
train 2k		86.6		86.3		86.4		86.2		87.4
imdb llm3										
train 2k		77.2		73.5		73.3		74.6		74.6
news llm3										
train 2k		86.6		87.9		87.5		88.0		87.6

Tablo. 1 Test verisinin augmente edilmesiyle doğrulukta meydana gelen değişimler ve ensemble doğrulukları

		train 2k		train 4k		train 6k		train 10k
imdb llm1		accuracy		accuracy		accuracy		accuracy
test 0.5k		77.2		75.0		77.0		77.4
news llm1								
test 0.5k		86.6		88.0		88.0		89.2
imdb llm2								
test 0.5k		77.2		76.8		76.6		76.0
news llm2								
test 0.5k		86.6		87.0		87.4		88.0
imdb llm3								
test 0.5k		77.2		78.0		78.2		77.6
news llm3								
test 0.5k		86.6		86.8		87.6		87.2

Tablo. 2 Train verisinin augmente edilmesiyle doğrulukta meydana gelen değişimler

IMDB ve News veri setleri üzerinde gerçekleştirilen deneylerde üç farklı büyük dil modeli (LLM1, LLM2, LLM3) paraphrase tabanlı veri artırımıyla eğitilmiştir. Veri boyutları arttıkça modellerin doğruluk oranları genel olarak yükselse de özellikle IMDB veri setinde bu artış her zaman düzenli bir çizgide ilerlememektedir. IMDb'de, küçük eğitim ve test kümelerinde elde edilen yüksek başarılar, veri boyutunun büyümesiyle yer yer dalgalanma göstermektedir. Bu durum, modelin duygusal içerikli yorumlarla ilgili belirli bir doygunluğa ulaşması veya paraphrase verilerinin yararlılığının sınırlı kalmasıyla açıklanabilir.

News veri setindeyse ek ve daha çeşitli veri örnekleri, modelin genellikle istikrarlı biçimde gelişmesini sağlar. Haber metinlerinin kapsamı ve çeşitliliği, paraphrased veriden daha fazla yararlanmaya olanak tanıdığı için eğitim veri boyutunun artışıyla doğruluk oranlarında nispeten düzenli bir yükseliş görülür. Bununla birlikte orijinal veri ve augmente edilmiş verilerin birleştirilmesiyle oluşan "ensemble" (çoğunluk kararı) yöntemi bazı senaryolarda belirgin iyileşmeler sağlamaktadır.

Genel olarak IMDb veri seti belirli bir noktadan sonra ek verinin getirisinden tam yararlanmayabilirken News veri seti büyüyen eğitim verisiyle performansta daha istikrarlı kazanımlar elde etmektedir. Sonuçlar; paraphrase temelli veri artırımı yaklaşımının veri setinin doğası ve görev türüyle yakından ilişkili olduğunu, ayrıca ensemble yönteminin özellikle belirli modeller ve veri boyutları için önemli artılar sunabildiğini göstermektedir.

KAYNAKLAR

[1] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 142–150, June 2011.H. Simpson, *Dumb Robots*, 3rd ed., Springfield: UOS Press, 2004, pp.6-9.

[2] X. Zhang, J. J. Zhao, and Y. LeCun, "Character-level Convolutional Networks for Text Classification," in NIPS, 2015.B. Simpson, et al, "Title of paper goes here if known," unpublished.

[3] V. Vorobev and M. Kuznetsov, "A paraphrasing model based on ChatGPT paraphrases," 2023.Y. Yorozu, M. Hirano, K. Oka, and Y.

Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” *IEEE Translated J. Magn. Japan*, vol. 2, pp. 740-741, August 1987 [*Digest 9th Annual Conf. Magnetics Japan*, p. 301, 1982].

[4] Sentence-Transformers, “all-MiniLM-L6-v2,” Hugging Face, 2023, https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.J. K. Author, “Title of paper,” in *Unabbreviated Name of Conf.*, City of Conf., Abbrev. State (if given), year, pp *xx-xxx.*

[5] Tuner007, “pegasus_paraphrase,” Hugging Face, 2023, https://huggingface.co/tuner007/pegasus_paraphrase.

[6] Mrm8488, “t5-small-finetuned-quora-for-paraphrasing,” Hugging Face, 2023,https://huggingface.co/mrm8488/t5-small-finetuned-quora-for-paraphrasing.