ISSN 2224-087X. Електроніка та інформаційні технології. 2024. Випуск 28. С. 61–72
Electronics and information technologies. 2024. Issue 28. P. 61–72

УДК 004 DOI: https://doi.org/10.30970/eli.28.6

INFLUENCE OF DATA AUGMENTATION ON NAMED ENTITY RECOGNITION USING TRANSFORMER-BASED MODELS

B. Pavlyshenko, I. Drozdov

Ivan Franko National University of Lviv
50 Drahomanova St., UA-79005 Lviv, Ukraine
bohdan.pavlyshenko@lnu.edu.ua, ihor.drozdov@lnu.edu.ua

Transformer-based models have demonstrated their effectiveness for natural language processing tasks. Training these models requires huge amounts of textual data. The creation of a high-quality dataset demands substantial resources dedicated to the collection, processing, and annotation of data. Also, building a large dataset for less commonly used languages or domains presents a significant challenge due to the inadequacy of available information for forming a comprehensive dataset. Data augmentation is one of the approaches to generating synthetic information, which helps increase the initial dataset size and enhance model performance.

The main goal of this article is to explore the possibilities of using data augmentation to enhance the capabilities of popular transformer-based models: BERT, ALBERT, DistilBERT, and RoBERTa. The study used one of the most popular datasets for named entity recognition research - CoNLL 2003. During the experiments, reduced versions of the initial dataset were created: down to 20%, 10%, and 5%, with different approaches to sentence selection in these datasets. Word-level augmenters were used for data augmentation: antonym augmentation, synonym augmentation, and word embeddings and their combinations. The experiments were conducted on identical equipment to obtain comparable results. The evaluation of results is based on the F1 score. The results demonstrated the effectiveness of data augmentation for small datasets, where significant improvements were achieved. With larger datasets, the impact of augmentation decreases.

Keywords: named entity recognition, natural language processing, augmentation, BERT, ALBERT, DistilBERT, RoBERTa.

Introduction.

During the last decade, the amount of information has increased dramatically. A significant amount of this information is texts: books, articles, news, and social media messages. As an area of research, Natural Language Processing (NLP) has growing challenges in processing this amount of data and extracting valuable information. Named Entity Recognition (NER) is one of the key tasks aiming to understand texts and extract specific categories of information like person names, locations, organization, date and time, etc. As noted in [1,2], NER is one of the fundamental sub-tasks for multiple NLP tasks like text understanding, translation, text summarization, etc. Hence, the effective extraction of named entities provides clues for more effective text understanding and processing.

The approaches to NER have evolved significantly through the last decades. Based on [1-4] hundreds of different approaches were introduced from simple rule-based approaches up to the current state-of-the-art models based on neural networks and transformers architecture. The first NER models utilized rule-based approaches, unsupervised learning, feature-based

supervised learning approaches, and the most recent trends with deep learning approaches. Vaswani et al. [5] introduced a new model, named Transformer. One of the key benefits of the new architecture is – the self-attention mechanism, which gives the possibility for the model to extract complex patterns from huge text corpora without supervision. Moreover, the new architecture shows the great possibility for parallelization during the training and prediction process. Based on transformers, BERT (bi-directional transformers for language understanding) was introduced and showed new state-of-the-art results for multiple tasks in the NLP area [6]. Despite such great results, the BERT model is undertrained, and building more effective transformer-based models is a very important direction of the research. As a result, models RoBERTa [7], DistilBERT [8], and ALBERT [9] were introduced, mostly, with minor differences and the goal of improving BERT. Each of them has its advantages and disadvantages like complexity, time required to train, size of the initial dataset, etc.

Despite the advantages of these models, for effective training and fine-tuning, they require a large amount of data to effectively solve specific tasks. As noted in [10,11], fine-tuning of the Large Language models can impact on model performance. The current state of research and practical implementation of NER significantly depends on the quality of the datasets, especially for specialized domains and languages with a limited amount of text information and high-quality datasets available. On the other hand, building high-quality datasets for NER could be costly and time-intensive way, especially in fast-changing environments like social networks. Thus, approaches, which allow to spend less resources to build datasets could be very useful. One of the approaches is to extend the dataset by synthetic, context-dependent data.

Data augmentation (DA) presents a promising solution for the artificial creation of synthetic data based on prior knowledge about the problem domain, limited labeled data, etc [12, 13]. Data augmentation helps to extend a dataset and increase its diversity without extending it with new data. Approaches to augment data could differ between areas of research and domain area but could be grouped based on the scope of application: character level, word level, sentence level, and document level [14,15]. Hence, choosing appropriate methods could be a challenging problem – the effect could be negative in some cases [16,17]. Despite the rising popularity of DA usage in the NLP area, this area of research doesn't have enough attention.

The scope of this paper is to research the influence of different data augmentation approaches on transformer-based models BERT, RoBERTa, ALBERT, and DistilBERT.

Methods and materials.

During this research, the CoNLL 2003 [18] dataset was used. Despite this dataset is quite old, it is commonly used and gives a good base for comparison with other studies. This dataset is rich in named entities and makes a good fit for this research. Table 1 contains information about the dataset. The purpose of this paper is to shed light on how data augmentations impact on performance of transformer-based models with a limited amount of labeled data extended with synthetic augmented data. To accomplish this goal, Table 2 demonstrates 6 different subsets of the initial train dataset, which was used. Validation and test parts of the datasets were used without any changes.

Table 1. Information about used datasets.

CoNLL 2003 dataset, English						
	Sentences	Tokens	LOC	MISC	ORG	PER
Training set	14,041	203,621	7,140	3,438	6,321	6,600
Validation set	3,250	51,362	1,837	922	1,341	1,842
Test set	3,453	46,435	1,668	702	1,661	1,617

Table 2. CoNLL-based train datasets were used during the research.
Validation and test parts are unchanged.

Abbreviation	**Sentences**	**Description**
S100	14,041	Contains all sentences from the original dataset without changes.
S20	2,808	Contains 20% of the initial train dataset records: the first 10% of sentences and the last 10%.
S10	1,404	Contains 10% of the initial train dataset records: the first 5% of sentences and the last 5%.
S5	702	Contains only the first 5% of the dataset.
R10	1,404	Contains 10% of initial dataset, chosen by random.
R5	702	Contains 5% of initial dataset, chosen by random.

The primary focus of this research was on word-level embeddings. This allows to preservation of sentence structure and named entities labeling in the sentence. The next data augmentation approaches were selected:

- Antonym augmentation – apply substitution to some percent of the words in a sentence for its antonym.
- Synonym augmentation – apply substitution to some percent of the words in a sentence for its synonym.
- Word embedding augmentation – apply substitution to some percent of the words in a sentence with its “similar” word based on the word2vec model [19].

At the same time, BERT, RoBERTa, DistilBERT, and ALBERT models were chosen for the experiment:

- BERT model – one of the first implementations of the transformer architecture and leveraged state-of-the-art in multiple NLP tasks.
- ALBERT model – purpose to optimize BERT model architecture and train process and achieve similar results. This model has 18x fewer parameters than BERT and almost 2x faster training time.
- DistilBERT model – has the same purpose as the ALBERT model to optimize the BERT model through optimization of the size and size reduction.
- RoBERTa model – while having the same goal to optimize the initial BERT model, is more comprehensive, and uses a significantly bigger initial training set, dynamic masking, etc.

To build and perform the experiment, the HuggingFace platform [20] was used. It has multiple useful tools to work with datasets, models save and load, and perform model fine-tuning and evaluation. Also, the portal contains multiple basic models with an easy way to configure

required pipelines for experiments. Also, for data augmentation library nlpag was used [21]. Experiments were executed on the Google Colab platform utilizing T4 GPU High RAM runtime.

The experiment was built with the next structure:

- Dataset S100 was used without augmentations to train all four models as a comparison basis.
- For generation of the new train data was used 8 approaches, as described in Table 3. For each sentence, additional sentences were added. Also, during the generation of the synthetic data, all original named entities were preserved.
- Each transformer model was trained on dataset variations from Table 2, except S100 and synthetic data generation based on Table 3. In total, 40 trained and estimated models. Also, models utilized the same datasets for the same configuration.
- The evaluation was accomplished using the F1 score using seqeval library [22].
- For fine-tuning, the next pre-trained models from HuggingFace portal database: bert-base-uncased, roberta-base, albert-base-v2 and distilbert-base-uncased.
- All models were fine-tuned with the same initial training parameters: number epochs 3, learning rate $5e^{-5}$, weight decay 0.01, and batch size 16.

Table 3. Train synthetic data generation approaches.

Scenario	**Count**	**Description**
Antonyms	2	Two sentences with antonyms
Synonyms	2	Two sentences with synonyms
Word embeddings	2	Two sentences with word embeddings
Antonyms + Synonyms	1+1	For each augmentation approach, add one sentence
Antonyms + Word embeddings	1+1	For each augmentation approach, add one sentence
Synonyms + Word embeddings	1+1	For each augmentation approach, add one sentence
Synonyms + Antonyms + Word embeddings	1+1+1	For each augmentation approach, add one sentence

Measurement system.

Approaches to measure the effectiveness of the model for token labeling tasks can depend on the expected result. For example, in [23] for OpenAI GPT models' estimation was used measurement system to consider only extracted named entities without any reference to their position in the text. In this paper, for estimation of the performance the F1 score was used, label was recognized properly only in cases, when all its parts were recognized properly. It was described in [18] as a measurement system for the CoNLL dataset.

Results and analysis.

Overall, the result of this research makes sense to split into three blocks:

- Preparation of the datasets – overview of dataset preparation time for different augmentation approaches (Table 4).
- Training process – shed light on how different configuration of the datasets impacts on training process (Table 5 and Figure 1).

- Models' effectiveness based on F1 score – review how different dataset sizes and data augmentations impact model performance (Table 6).

Based on data in Table 4, the 2 different groups were identified: antonyms and synonyms augmentations are extremely fast and word embeddings are a hundred(s) times slower. In the rest of this research, the term "simple augmentations" would be used for any augmentations, based on antonyms, synonyms, and any their combination. On the other side, "complex augmentations" – for any augmentation, which utilizes the usage of word embeddings.

Moving forward, significant differences in dataset preparation between "simple augmentations" and "complex augmentations" were expected results since the operation of determining a synonym or antonym for a certain word was reduced to finding the word in a dictionary and randomly selecting one of the associated words. On the other hand, in the case of word embeddings, when searching for a similar word, a large semantic graph is analyzed and requires significant computational resources. Since the difference in time required to construct synthetic data using these augmentation methods is significant, simple augmentation methods must be chosen if greater speed is needed.

In Table 5, the training times were demonstrated. Based on this information, all four chosen models showed similar behavior: training time almost linearly depends on the size or number of sentences in the initial dataset. For example, the RoBERTa model for dataset S100 contains around 14000 sentences and has a learning time of 983 seconds, for dataset S20 and synonym + antonym applied augmentation – around 8500 sentences and a learning time is 627 seconds. That is 60% of the initial dataset size and training took only 63% of the time, compared with the S100 dataset. This factor is obvious as a key factor, that impacts training time in several sentences, but applied data augmentations don't change significantly the length of the sentences, the difference is usually up to 3 tokens (up to 10%). Regarding the difference in token quantity per sentence, it doesn't make a significant impact due to the models' architecture.

Table 4. Time in seconds for dataset augmentation.

	S20	**S10**	**R10**	**S5**	**R5**
Antonyms	13	7	6	2	3
Synonyms	12	6	6	2	3
Antonyms + Synonyms	12	6	6	2	3
Word embeddings	5,520	2,789	2,598	1,201	1,392
Antonyms + Word embeddings	2,921	1,334	1,346	607	706
Synonyms + Word embeddings	2,899	1,326	1,397	606	706
Synonyms + Antonyms + Word embeddings	2,775	1,350	1,383	607	694

Table 5. Time in seconds for model training based on dataset type and augmentation approach.

	S100	**S20**	**S10**	**R10**	**S5**	**R5**
BERT model						
Original train dataset	965	189	95	96	42	43
Extended with 2 additional sentences (all augmentations, except the synonyms + antonyms + word embeddings)	--	618	304	301	141	147
Synonyms + Antonyms + Word embeddings	--	847	418	417	198	207
ALBERT model						
Original train dataset	1072	209	103	105	44	47
Extended with 2 additional sentences (all augmentations, except the synonyms + antonyms + word embeddings)	--	702	338	343	158	165
Synonyms + Antonyms + Word embeddings	--	975	475	478	225	235
DistilBERT model						
Original train dataset	536	104	53	52	23	24
Extended with 2 additional sentences (all augmentations, except the synonyms + antonyms + word embeddings)	--	340	165	165	78	80
Synonyms + Antonyms + Word embeddings	--	463	228	229	108	112
RoBERTa model						
Original train dataset	983	193	89	97	43	43
Extended with 2 additional sentences (all augmentations, except the synonyms + antonyms + word embeddings)	--	627	302	307	143	146
Synonyms + Antonyms + Word embeddings	--	856	419	420	199	211

On the other hand, as shown in Fig. 1, models demonstrated fast learning approached the asymptote of the loss function in 0.5-1 epochs. Nevertheless, all models demonstrated a slower learning rate for datasets S20, S10, and R10 without data augmentations. In this case, the amount of data is 3-4 times less than for augmented datasets and a few steps respectively. Also, training loss at the end of the training process is bigger. As a fact for these sizes of datasets, models were adopted better for augmented datasets, diversity of the data with the same NER labels allows the model to fit better initial dataset.

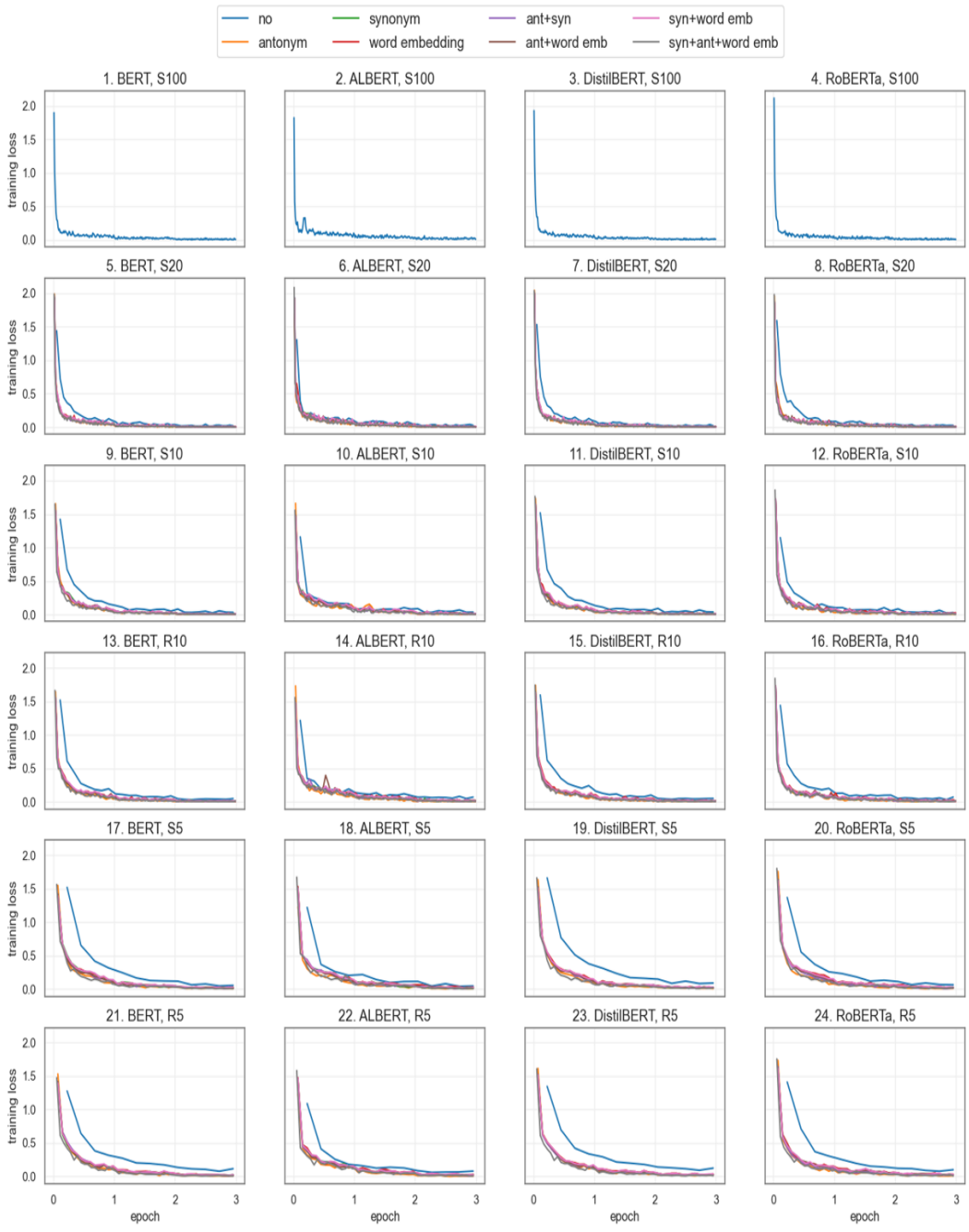

Fig. 1. Training loss change during fine-tuning process for all 164 models based on different datasets and augmentation approaches.

Table 6. F1 scores for models, datasets, and augmentations approaches

	S100	**S20**	**S10**	**R10**	**S5**	**R5**
BERT model						
Without augmentations	**94.38**	90.03	85.51	88.86	77.16	77.59
Antonyms	--	90.95	**86.25**	90.03	83.01	**88.59**
Synonyms	--	90.23	86.08	90.06	81.04	86.73
Word embeddings	--	90.78	84.29	89.72	81.75	86.52
Antonyms + Synonyms	--	90.75	85.15	90.23	**83.35**	88.1
Antonyms + Word embeddings	--	**91.07**	84.66	89.83	83.19	87.22
Synonyms + Word embeddings	--	90.22	85.07	89.62	82.65	87.8
Synonyms + Antonyms + Word embeddings	--	90.88	85.21	**90.26**	81.13	87.34
ALBERT model						
Without augmentations	**93.4**	**89.65**	**85.69**	87.46	81.89	83.9
Antonyms	--	89.45	84.4	**89.56**	**83.63**	86.03
Synonyms	--	88.14	83.11	87.65	82.2	84.14
Word embeddings	--	87.87	83.53	88.16	82.67	84.19
Antonyms + Synonyms	--	89.07	83.26	88.7	82.72	84.94
Antonyms + Word embeddings	--	89.22	83.26	89.28	82.07	85.72
Synonyms + Word embeddings	--	88.13	82.91	88.3	82.64	84.2
Synonyms + Antonyms + Word embeddings	--	88.12	82.44	88.5	81.98	**86.33**
DistilBERT model						
Without augmentations	**94.16**	89.93	83.86	87.68	73.34	77.66
Antonyms	--	**90.25**	**85.17**	88.77	81.51	85.84
Synonyms	--	89.39	83.73	88.62	80.25	85.4
Word embeddings	--	89.51	83.71	89.19	80.56	**85.99**
Antonyms + Synonyms	--	89.95	83.97	89.04	81.65	85.81
Antonyms + Word embeddings	--	89.4	84.72	**89.25**	**82.44**	85.88
Synonyms + Word embeddings	--	89.47	84.0	89.24	81.7	85.9
Synonyms + Antonyms + Word embeddings	--	89.79	83.86	89.28	81.55	86.58
RoBERTa model						
Without augmentations	**95.77**	91.61	87.46	90.22	82.92	83.16
Antonyms	--	92.76	88.74	92.26	**87.12**	89.87
Synonyms	--	**93.03**	88.75	91.66	85.08	89.38
Word embeddings	--	92.49	88.05	91.94	85.23	89.53
Antonyms + Synonyms	--	92.77	**89.05**	91.45	86.86	88.73
Antonyms + Word embeddings	--	92.2	88.69	92.08	86.29	89.3
Synonyms + Word embeddings	--	92.76	88.44	91.95	86.05	88.72
Synonyms + Antonyms + Word embeddings	--	92.73	88.36	**92.63**	85.91	**90.01**

On the other side, an interesting finding is that for extremely low datasets like S5 and R5, additional augmented data significantly improves model learning possibilities. Based on this information, extremely small datasets augmentation even with simple augmenters is reasonable.

Table 6 demonstrates the measurement of the models with all dataset variations and augmentation approaches. In contradistinction to the train part of the dataset, all evaluations during the fine-tuning of the models and evaluation process were performed with full original validation and test parts of the CoNLL dataset. In brief, all models demonstrated the best results on the full dataset S100. This is an expected result as the generation of augmented data doesn't create new named entities and the training dataset in S20, S10, S5, R10, and R5 is limited to available named entities inside it, while S100 should contain a bigger diversity of the named entities. On the other side, data augmentations extend the context for available named entities in which they are in use.

Results, demonstrated in Table 6 possible to split into 3 different groups:

- Original dataset S100 without augmentations: models demonstrated the best performance with 3-4% better results compared to the same model on smaller datasets.
- Datasets S20, S10, R10: models demonstrated controversial results, no obvious favorite, results spread 1-2%. Even though small difference in the results, RoBERTa demonstrates a positive impact of the augmentations with an improvement of 1-2%, DistilBERT, also, demonstrates a positive impact on performance in most cases, while ALBERT and BERT show slightly less influence of the data augmentations and could demonstrate even worse results.
- Datasets S5, and R5: models demonstrated great results with an increase of up to 11% (BERT model with Antonyms augmentation) with an average score of 5-8%. This fact demonstrates, that for very low initial datasets, simple word-level augmentations could significantly increase the diversity of the dataset and its performance on NER recognition. Models on these datasets with data augmentations demonstrated very close results to 2-4x bigger datasets like S20, S10, and R10.

Also, research demonstrated, that choosing different context and meaning sentences from the initial dataset demonstrates better results with data augmentations. For example, models demonstrated on average 3-6% better results on augmented datasets with randomly selected records compared to sequential data. Randomly selected data contains more different named entities and synthetic data allows the model to better fit the domain.

Conclusion.

This paper investigated the influence of the word level of data augmentations on named entity recognition. CoNLL dataset was utilized with six different variations: full dataset, 20% of initial train dataset, two types of 10% of initial train dataset – one with straight part of the data and one with randomly chosen and two types with 5% of initial train dataset – the same as for 10%. For data augmentation, word-level data augmentations with synonyms, antonyms, and word embeddings and their combinations. Word-level data augmentations were chosen to preserve initial labeled named entities in the training dataset. The research was applied to four transformer-based models: BERT, RoBERTa, ALBERT, and DistilBERT.

Despite on fact, that with data augmentations dataset sizes were increased 3-4 times, models demonstrated poor performance improvement on 0.5-2% compared with non-augmented ones for 20% and 10% of initial size datasets. Even though on small improvement, synthetic data doesn't require significant resources to produce it, but could produce additionally

recognized records. This could be useful when data labeling is a significantly complex task, for example, for low-resource domains. Overall, this is the expected result because data augmentation with our approach doesn't produce new named entities in the train dataset, but increases contexts, in which available named entities were used.

A key finding is that for the very low size of the initial high-quality labeled dataset, applying data augmentations could provide significant improvement in model performance. BERT model demonstrated improvement for up to 11% from 77.59% up to 88.59% with applied antonym augmentation. Average improvement has been on level 5-8%, which is significant. For the very low size of the training dataset, increasing the context, in which different named entities could be used is important. Additionally, a significant impact on performance for 3-6% showed the diversity of the initial dataset. In other words, better to build a dataset with diverse information compared to a monotonous one.

To summarize further research directions, promising area of the research is to have the possibility to produce augmented information, which will contain some set of named entities, that we want to recognize. To achieve this goal could be useful approaches with text summarization, applying Large Language Models like OpenAI GPT series or open-source models like LLaMa 2/3, MT5, etc.

References

[1] *Li J., Sun A., Han J., Li C.* (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1), 50–70.

[2] *Roy, A.* (2021) "Recent trends in named entity recognition (NER)." *arXiv preprint* arXiv:2101.11420v1

[3] *Yadav V., Bethard S.* (2019) "A survey on recent advances in named entity recognition from deep learning models." *arXiv preprint* arXiv:1910.11470v1.

[4] *Shen Y., Yun H., Lipton Z. C., Kronrod Y., Anandkumar A.* (2017). "Deep active learning for named entity recognition." *arXiv preprint* arXiv:1707.05928v3.

[5] *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Polosukhin I.* (2017). "Attention is all you need." *Advances in neural information processing systems* 30

[6] *Devlin, J., Chang MW., Lee K., Toutanova K.* (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint* arXiv:1810.04805v2.

[7] *Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Stoyanov V.* (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint* arXiv:1907.11692v1.

[8] *Sanh V., Debut L., Chaumond J., Wolf T.* (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper, and lighter. *arXiv preprint* arXiv:1910.01108v4.

[9] *Lan Z., Chen M., Goodman S., Gimpel, K, Sharma P., Soricut R.* (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint* arXiv:1909.11942v6.

[10] *Pavlyshenko B. M.* (2023). Analysis of Disinformation and Fake News Detection Using Fine-Tuned Large Language Model. *arXiv preprint* arXiv:2309.04704v1.

[11] *Pavlyshenko B. M.* (2023). Financial News Analytics Using Fine-Tuned Llama 2 GPT Model. *arXiv preprint* arXiv:2308.13032v2.

[12] *Chen J., Tam D., Raffel C., Bansal M., Yang D.* (2021). An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. *arXiv preprint* arXiv:2106.07499v1.

[13] *Feng S., Gangal V., Wei J., Chandar S., Vosoughi S., Mitamura T., Hovy E.* (2021). A survey of data augmentation approaches for NLP. *arXiv preprint* arXiv:2105.03075v5.

[14] *Chen S., Aguilar G., Neves L., Solorio T.* (2021). Data Augmentation for Cross-Domain Named Entity Recognition. *arXiv preprint* arXiv:2109.01758v1.

[15] *Dai X., Adel H.* (2020). An Analysis of Simple Data Augmentation for Named Entity Recognition. *arXiv preprint* arXiv:2010.11683v1.

[16] *Pavlyshenko B., Stasiuk M.* (2023). Augmentation in a binary text classification task. *2023 IEEE 13th International Conference on Electronics and Information Technologies (ELIT)*, Lviv, Ukraine, pp. 177–180.

[17] *Pavlyshenko B., Stasiuk M.* (2024). Data augmentation in text classification with multiple categories. *Electronics and information technologies*, Issue 25. – P. 67-80. DOI: http://dx.doi.org/10.30970/eli.25.6

[18] *Sang, E. F., De Meulder F.* (2003). "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition." *arXiv preprint arXiv:cs/0306050v1*.

[19] Mikolov T., Le Q., Sutskever I. (2013). Exploiting Similarities among Languages for Machine Translation *arXiv preprint arXiv:1309.4168v1*.

[20] HuggingFace [Electronic resource]. Access mode: https://huggingface.co/

[21] NlpAUG library repository [Electronic resource]. Access mode: https://github.com/makcedward/nlpaug

[22] Seqeval library repository [Electronic resource]. Access mode: https://github.com/chakki-works/seqeval

[23] *Pavlyshenko B, Drozdov I.* (2023). Named entity recognition using OpenAI GPT series models. *Electronics and information technologies*, Issue 23. – P. 46-58. DOI: http://dx.doi.org/10.30970/eli.23.5

ВПЛИВ АУГМЕНТАЦІЇ ДАНИХ НА РОЗПІЗНАВАННЯ ІМЕНОВАНИХ СУТНОСТЕЙ ЗА ДОПОМОГОЮ МОДЕЛЕЙ НА БАЗІ ТРАНСФОРМЕРІВ

Б. Павлишенко, І. Дроздов

Львівський національний університет імені Івана Франка,
вул. Драгоманова 50, 79005 Львів, Україна
bohdan.pavlyshenko@lnu.edu.ua, ihor.drozdov@lnu.edu.ua

Моделі на основі трансформерів показали свою ефективність для задач обробки природної мови. Тренування цих моделей потребують дуже великих обсягів текстових даних. Створення якісного набору даних потребує великих ресурсів для збору, обробки та розмітки даних. Також, підготовка достатнього набору даних може бути проблематичним для рідковживаних мов або доменів, оскільки доступної інформації може бути. Аугментація

даних є одним з підходів до створення штучної інформації, що дозволяє збільшити початковий набір даних та результативність моделі.

Головна мета цієї статті – дослідити можливості застосування аугментації даних для покращення можливостей розпізнавання іменованих сутностей популярних моделей, основаних на трансформерах: BERT, ALBERT, DistilBERT та RoBERTa. Для дослідження використано один з найбільш популярних датасетів для дослідження розпізнавання іменованих сутностей - CoNLL 2003. Під час експериментів було створено зменшені варіанти початкового датасету до 20%, 10% та 5% з різними підходами до відбору речень в ці датасети. Для аугментації даних було використано аугментатори на рівні слів: аугментація антонімами, аугментація синонімами та вставка контекстних слів та їхні комбінації без аугментації іменованих сутностей. Експерименти було виконано на однаковому обладнанні для отримання порівнюваних результатів. Оцінка результатів базується на оцінці F1. Результати продемонстрували ефективність застосування аугментації даних для невеликих наборів даних. В цих випадках вдалося досягнути суттєвих покращень результатів. При більших наборах даних, вплив аугментації знижується, але дозволяє при наявних ресурсах досягти незначного покращення результатів.

Ключові слова: розпізнавання іменованих сутностей, обробка природної мови, аугментація, BERT, ALBERT, DistilBERT, RoBERTa.

The article was received by the editorial office on 18.06.2024.
Accepted for publication on 01.07.2024.