Sentiment Analysis with Neural Models for Hungarian

László János Laki, Zijian Győző Yang

Hungarian Research Centre for Linguistics
Benczúr u. 33, H-1068 Budapest, Hungary
laki.laszlo@nytud.hu; yang.zijian.gyozo@nytud.hu

MTA-PPKE Hungarian Language Technology Research Group
Práter u. 50/A, H-1083 Budapest, Hungary
laki.laszlo@itk.ppke.hu; yang.zijian.gyozo@itk.ppke.hu

Abstract: Sentiment analysis is a powerful tool to gain insight into the emotional polarity of opinionated texts. Computerized applications can contribute to the establishment of next-generation models that can provide us with data of unprecedented quantity and quality. However, these models often require substantial amount of resources in order to meet the desired performance expectations. Therefore, numerous research efforts are targeted to achieve high-quality results while lowering the resource needs by improving the structure and function of the models used. From a cognitive perspective, it is important to understand the mental state of users when they engage in activities that potentially reflect their feelings and emotions. With the emergence of the widespread use of digital solutions, users post opinionated texts on social media, which can be used as a valuable source to detect their underlying sentiments. Therefore, these platforms offer an unparalleled opportunity to perform sentiment analysis. In recent years, natural language processing tasks, like sentiment analysis, can be solved with high performance, if a pre-trained language model is fine-tuned. Herein we present the first neural transformer-based sentiment analysis model for Hungarian, which achieved state-of-the-art performance. Several limitation factors can occur during fine-tuning, such as the lack of training corpora with appropriate size or the complete absence of usable training material. In our experiment, we use data augmentation methods, specifically machine translation and cross-lingual transfer, to increase the size of our training corpora. Here, we demonstrate our experimentation with 9 different language models. Our work provides evidence for the increased efficiency of the trained models if translation text is added to the training corpora. Furthermore, using the augmentation technique, we could further increase the performance of our models. Consequently, our findings represent an important milestone in the advancement of sentence-level and aspect-based sentiment analysis in the Hungarian language.

Keywords: sentence-level sentiment analysis; aspect-based sentiment analysis; data augmentation; transformer models; BERT

1 Introduction

Natural languages provide an important platform for thought and communication, which are considered as pivotal human cognitive characteristics. Therefore, natural language processing can provide valuable insights into human cognitive processes [1]. Communication between a human and an artificially cognitive system is called inter-cognitive communication (information transfer occurs between two cognitive beings with different cognitive capabilities) [2] [3]. Machine Learning methods are vital elements of modern cognitive infocommunications systems because they can be used in various ways such as behavior modeling or sentiment analysis [4].

Sentiment analysis is the automatized identification of sentiments in a text and the categorization of these sentiments into categories like negative, neutral or positive. With the increased number of social media users, vast amount of text information is already present on the internet, which can be especially useful in identifying the underlying emotions of the authors who wrote the text. Since it offers an exceptional insight with potential applicability in multiple ways (e.g. analysis of the popularity of politicians, customer feedback analysis, social media monitoring, emotion detection in psychology), both academic and industrial stakeholders become more and more interested in the extraction of sentiment data from texts [5].

The development of neural language modeling (LM) has resulted in a breakthrough for most Natural Language Processing (NLP) tasks. The language models differ not only in different training data, but also in the internal structure of neural networks and the used training methods. Consequently, a specific NLP task could be solved by a properly chosen LM. The state-of-the-art technique to solve an NLP task is to fine-tune a pre-trained language model with a smaller task-specific data. The quality of these systems not only depends on the pre-trained models but the size of the tuning set. NLP tasks like Hungarian sentiment analysis have a great interest in the industry, but it has a limited amount of freely available data and models. Further on we have not found any previously published solution for Hungarian neural sentiment analyser, which means our solution can be considered a pioneer in this application area.

During our work a machine translation (MT) system was used to translate an English sentiment analysis dataset to Hungarian. A translated corpus was generated using our internally trained machine translation tool, which was later integrated into our systems. It is important to point out that our work offers a novel approach for applying machine translation and data augmentation procedures to expand the available repertoire of corpora in Hungarian.

In Section 2 the previous solutions, in Section 3 the used corpora, in Section 4 the used neural models, in Section 5 the baseline experiments and results, in Section 6 the data disparity handling experiments and results, in Section 7, the data augmentation experiments and results, and in Section 8 the aspect term extraction experiments and evaluations were described.

2 Related Work

Sentiment analysis is a very complex natural language processing task and has a wide range of application areas, for instance social media monitoring [6], supporting the decision making process of investors by the analysis of semantic textual content in financial news [7], digital marketing [8], assessment of psychological state [9], and many more fields where the application of such advanced text-mining approaches can be especially beneficial.

Currently, multiple approaches are being developed for sentiment analysis. Initial attempts have been made to classify documents and texts based on the overall polarity (negative, positive or neutral) [10]. Another main direction is the aspect-based method, which is more fine-grained and its main focus is to identify the aspects of an object or an entity that is responsible for the elucidation of the sentiment [11]. An alternative strategy is called the sentence-level sentiment analysis, which chooses the sentence as the investigational entity, thus the goal is to determine how a given sentence in the document is opinionated [12].

The emergence of artificial intelligence-driven solutions in sentiment analysis is clearly in line with the research trends that can be observed in the cognitive infocommunications field. It is of high importance to better understand the theoretical framework behind text-mining, which can facilitate the development of novel applications towards an unequivocal prediction of user sentiment or crowd opinions [2]. We are convinced that these efforts will synergistically enhance the progress of cognitive sciences and the related interdisciplinary domains.

For Hungarian, only a few sentiment analysis corpora and tools exist. OpinHuBank [13] is a human-annotated corpus to aid the research of opinion mining and sentiment analysis in Hungarian. It consists of 10,000 sentences containing person names from major Hungarian news sites and blogs. Each entity occurrence was tagged by 5 human annotators for sentiment polarity in its sentence (neutral, positive or negative). Using the OpinHuBank, Hangya et al. [14] trained different supervised machine learning models to detect the sentiment of the entities. In their research, MALLET tool [15], polarity lexicon and dependency parser (magyarlanc [16]) were used.

HuSent [17] is a deeply annotated Hungarian sentiment corpus. It is composed of Hungarian opinion texts written about different types of products, published on the homepage [18]. The corpus contains 154 opinion texts, and comprises ~17,000 sentences and ~251,000 tokens. Steinberger et. al. in their research [19], aspect-based sentiment corpus was created with multilingual parallel corpora that contained a Hungarian subcorpus.

In this research, we have done the first neural network-based sentiment analysis research for Hungarian, in both sentence-level and aspect-based sentiment analyses. All of our models and scripts can be found on our Github site [20]. You can try out our sentiment analysis application on our demo site [21].

3 Corpora

For training sentence-level sentiment analysis the Hungarian Twitter Sentiment Corpus (HTS) was used [22] that was created by Precognox Kft. In the case of aspect-based sentiment analysis, we used the OpinHuBank [13] (OHB). We found these corpora as the only freely available annotated corpora for Hungarian sentiment analysis. In the case of HTS, we created a binary subcorpus (HTS2) to allow the binary sentiment analysis experiments. In this paper, we refer to the original HTS corpus (with five classes) as HTS5. In Table 1, the characteristics of the HTS and OHB corpora can be seen. The labels of the corpora are the following:

- **HTS5**: 1-very negative, 2-negative, 3-neutral, 4-positive, 5-very positive.
- **HTS2**: We have converted the 1 and 2 scores as negative, 4 and 5 scores as positive ones: 0-negative, 1-positive. We did not consider the score 3 to avoid the ambiguities.
- **OHB3**: -1-negative, 0-neutral, 1-positive. For convenience, we have converted them: 1-negative, 2-neutral, 3-positive.

Table 1
Characteristics of corpora

	SST2	SST5	ACL14	HTS2	HTS5	OHB3
Sentence	70,045	11,855	6,940	2,737	4,000	10,005
Token	671,257	213,812	150,904	50,036	71,235	308,407
Type	15,665	20,555	16,405	13,679	18,394	47,492
Avg word #	9.54	19.15	17.64	12.15	11.67	26.36
Labels	0;1	1;2;3;4;5	1;2;3	0;1	1;2;3;4;5	1;2;3
Training set	67,350	8,544	2,468	3,600	6,248	9,005
Test set	873	1,101	269	400	692	1,000

For the transfer and translation experiments, we have used the SST2 and SST5 corpora from GLUE benchmark [23] for the sentence-level research, over and above the acl-14-short-data [24] (ACL14) corpus for the aspect-based research. All of these corpora contained English sentences. With machine translation we used these corpora as additional data (SST2_hu, SST5_hu, ACL14_hu).

In our research, in the case of HTS5 and OHB3, we have split the corpus into 90%-10% training and test corpora. The first 10% of the corpora are our test corpora. In the case of HTS2, documents with score 3 are omitted. In the case of OHB, five sentiment scores (from five annotators) were assigned to each sentence. In Table 4, the inter-annotator agreement scores of the sentiment labels are presented. Since the agreement scores are not considerably high, we used the most common label for each sentence. The low agreement value can be attributed to the difficulties in determining the difference between neutral and negative/positive sentiment values in many cases, due to the limited context of the sentences. This could have a

negative effect on the performance of the models. However, this is the only available aspect-based sentiment corpus for Hungarian.

Table 4
Evaluation of inter-annotator agreement of OHB

	OpinHuBank
Fleiss's Kappa	0.6551
Krippendorff's Alpha Coefficient	0.6548
Scott's pi	0.6548
Average Pairwise Cohen's Kappa	0.6549

Table 3 shows the distributions of labels in the different corpora.

Table 3
Distribution of labels in corpora

	train	test	train	test	train	test	train	test
label	SST2		HTS2					
0	29,755	428	1,021	108				
1	37,539	444	1,448	162				
	SST5		HTS5		ACL14		OHB3	
1	1,089	139	93	12	1,560	173	2,253	136
2	2,200	289	936	88	3,127	346	5,205	565
3	1,594	229	1,111	150	1,561	173	1,548	299
4	2,259	279	1,349	141				
5	1,266	165	111	9				

4 Neural Models

In our experiments we have used 8 different types of monolingual Hungarian contextual language models, 2 types of multilingual contextual language models and a classical word embedding model. The short description of the used models will be seen in the second part of this section.

huBERT [25]: A Hungarian BERT base language model trained on the Webcorpus 2.0, which is composed of the Common Crawl web archive and the Hungarian Wikipedia. BERT (Bidirectional Encoder Representations from Transformer) is defined as a multi-level, bidirectional Transformer encoder [26] architecture. The BERT model is pre-trained on two language modeling tasks: word masking and next sentence prediction. Importantly, the fine-tuned huBERT model is considered the current state-of-the-art in several NLP tasks for Hungarian. The huBERT model was not fine-tuned for Hungarian sentiment analysis task before.

HILBERT [27]: A BERT large model for Hungarian, HILBERT offers high performance in Hungarian language processing tasks. HILBERT was trained on 4BN NYTK-BERT corpus. The model achieves remarkable results in various tasks, such as Name Entity Recognition (NER) and summary generation [28]. The advantage of this model against the huBERT is that it contains much more parameters, but on the other hand it was trained on less data.

HIL-RoBERTa [29]: One of the key challenges in language model optimization is encountered in the course of pre-training. Since pre-training is an especially resource-intensive process, it is important to research and develop new ways that can provide significant improvements. RoBERTa is a Robustly optimized BERT pre-training approach, which achieves state-of-the-art results on tasks like GLUE [23], RACE [30] and SQuAD [31], while using less resources due to its optimized pre-training paradigm. HIL-RoBERTa is a cased RoBERTa [32] small model, which is trained on Hungarian Wikipedia.

HIL-ALBERT [33]: Multiple efforts have been made to increase language model performance on end-tasks while optimizing the resource needs during pre-training. A Lite BERT (abbreviated as ALBERT) attempts this by incorporating parameter-reduction techniques [34]. In order to apply this paradigm to the Hungarian language, two pre-trained, uncased ALBERT models were created: one was trained on Hungarian Wikipedia (part of the Webcorpus 2.0 dataset), the other was on a part of the NYTK-BERT corpus. In our research, HIL-ALBERT NYTK was used.

HIL-ELECTRA [29]: Approaches designated as the Efficiently Learning an Encoder that Classifies Token Representation Accurately (abbreviated as ELECTRA) represent a successful alternative to masked language modeling (MLM) by the application of replaced token detection, which is a self-supervised pre-training task used to train the model to distinguish between real input and synthetically created reinstatements. The ELECTRA models are established upon the application of the Generative Adversarial Network method. The experimental evidence supports that this alternative is efficient and high-performing compared with other methods [35]. As for the Hungarian language implementation of ELECTRA, two models were created, the ELECTRA wiki and the ELECTRA NYTK-BERT, trained on Hungarian Wikipedia and NYTK-BERT v1 corpus (contains Hungarian Wikipedia as well), respectively. In our research, HIL-ELECTRA NYTK was used.

HIL-ELECTRIC [29]: Electric offers an implementation to the cloze task using an energy-based model [36]. The Electric model is an efficient solution to determine the distribution of possible tokens at a certain position by assigning energy scores to the token positions. As for the noise distribution, Electric applies a two-tower cloze model, which includes two Transformers operating in opposite directions and uses the context to both sides of the tokens. Electric has the capability of calculating likelihood scores simultaneously for all input tokens and not only for the masked ones. As for the Hungarian language implementation of ELECTRIC, two models were created, the ELECTRIC nytk and the ELECTRA nytk 10%, trained on one

tenth of Hungarian NYTK-BERT v1 corpus (contains Hungarian Wikipedia as well) respectively.

HILBART [37]: Models based on the combination of Bidirectional and Auto-regressive Transformers (abbreviated as BART) represents a powerful tool in sequence-to-sequence model pre-training. BART is especially potent when applied for text generation tasks, but it can achieve remarkable performance on discriminative and summarization tasks as well [38]. BART outperforms all previously established models in summarization tasks. Accumulating evidence suggests that BART performs the best when applied for Natural Language Generation (NLG), but achieves remarkable results in translation and comprehension tasks as well. BART was applied to Hungarian language as well resulting in HILBART models. These are HILBART base web and HILBART base wiki, trained on 1% of Webcorpus 2.0, 10% of Webcorpus 2.0 and on Hungarian Wikipedia, respectively. In our research, HILBART base web was used.

NYTK-GPT-2 [39]: GPT models are decoder-only transformer models. Generative Pre-Training (GPT) designates the concept of pre-training a language model on large datasets, which is followed by fine-tuning for a downstream task. The application of the GPT paradigm can foster significant advancements in NLP, especially in the area of classification, question-answering and investigation of semantic similarity. GPT models use a Transformer Decoder architecture [40]. GPT-2 achieved significant performance in several tasks already in a zero-shot setting [41]. NYTK-GPT-2 is an experimental GPT-2 model that was trained on Hungarian Wikipedia.

mBERT [26]: Multilingual BERT (abbreviated as mBERT) is a model that is established on the architectural principles of BERT, it also uses the same training paradigm with the key difference that the pre-training is performed on a concatenated dataset of Wikipedia texts of 104 different languages. The application of mBERT is especially advantageous in the case of low-resource languages, e.g. when only a relatively small number of annotated sentences is available for a language or a set of given languages. Cross-language pre-training models including mBERT have been applied for the Name Entity Recognition (NER) task in Hungarian and Uyghur languages [42]. In our research, BERT multilingual base model (cased) was used.

XLM-RoBERTa [43]: Cross-Language Understanding (XLU) is a key challenge and serves as an accelerator to the development of multilingual models. In 2020, the Facebook AI team published an article presenting XLM-RoBERTa (abbreviated as XLM-R as well), which is a transformer-based multilingual masked language model. The pre-training was performed on the CC-100 corpus, which contains texts from 100 different languages including Hungarian (number of Hungarian tokens: 7807 M; size of the Hungarian corpus: 58.4 GiB). The authors reported that XLM-RoBERTa achieved competitive results on several benchmarks in comparison with monolingual models, such as RoBERTa.

fastText [44] [45]: fastText is a solution developed by Facebook AI, which aims to facilitate text classification and representation learning. The paradigm is based on the incorporation of character n-grams into the skipgram model resulting in a fast and efficient method without the need for any preprocessing or supervision [46] As for text classification, fastText is comparable with other deep learning-based classifiers in accuracy and it is a much faster option than those for training and evaluation. The platform offers word vectors for English and 157 other languages. Therefore, it represents a powerful tool in multilingual language processing.

5 Neural Sentiment Analysis Baseline Experiments and Results

For the better comparison, we have used the same hyperparameters for almost all models. The hyperparameters are the following: learning rate: 2e-5, batch size: 32 per device (4 x GPU), epoch 4, max seq length: 128. In the case of HILBERT we used batch size as 8 per device in order to avoid the CUDA out of memory error. In the case of ELECTRA, the models used only one single GPU. Finally, fastText did not use GPU at all, it has used only CPU and the batch size was 1. For all experiments we used 4 x GeForce RTX 2080 Ti type video cards and 40 x Intel(R) Xeon(R) Silver 4114 CPU-s. For fine-tuning, we have used the code provided by huggingface transformers text classification library [47], google electra library [48] and fastText tool [49].

All models have been fine-tuned with their training set (described in Table 3) and evaluated on the test sets of the datasets. The results are shown in Table 5. First of all, the results of HTS2 dataset are presented. Two quality categories can be distinguished. The winners are huBERT, HILBERT and XLM-RoBERTa with around 83-84% F1-score, while the next cluster contains all other systems (71-79%). The absolute winner is the huBERT just like for most of the NLP benchmark tests. It is worth mentioning how well the multilingual models performed. Their performance is comparable with the monolingual ones. The second experiment was the HTS5 dataset, where the same quality clusters could be defined. In this case, HILBERT could outperform huBERT. In this case a third cluster could be seen which contains the systems between 51-55. Finally, the abstractive sentiment analysis task was evaluated. huBERT gained the whole task with statistically significant quality gain (~82% F-score). The second cluster is between 73-80%, the third cluster is between 64-69% and the final one contains the systems less than 63%. The last section of the table describes the performance of the English systems. An interesting phenomenon is that while the quality of the classification is excellent on the binary data set, the performance on the five labelled dataset is only average.

Table 5
Sentiment analysis baseline results

	Sentence-level		Aspect-based
	HTS2	HTS5	OHB3
huBERT	**84.07**	66.00	**81.99**
HILBERT	83.33	**68.00**	57.80
HIL-RoBERTa	75.92	59.15	68.50
HIL-ALBERT	75.56	55.49	63.99
HIL-ELECTRA	78.89	59.11	65.37
HIL-ELECTRIC	76.67	58.00	63.66
HILBART	71.11	51.25	60.39
NYTK-GPT-2	77.40	57.49	73.69
mBERT	78.51	57.74	75.49
XLM-RoBERTa	83.33	63.49	79.69
fastText	71.9	53.2	59.5
	SST2	SST5	ACL14
mBERT	90.02	49.97	73.69
XLM-RoBERTa	92.77	53.96	73.26

6 Data Equalization Experiments and Results

The deeper analysis of the results explained in Section 5 has shown us that the main issue of the classification (primarily in the case of multi-level ones) is that the training data of the different labels are not uniform. The edge categories contain only 3-3% of the data, which lead the systems not to use these categories as a prediction. For example, huBERT and HILBERT systems did not produce any sentences with the very negative or the very positive labels at all.

A possible solution for this problem is to balance the data. The perfect solution would be adding more training data. Unfortunately, in this research we do not have sufficient resources to create new data. In our first experiment we use the same amount of data as the lowest label has (called minus or "-"). Secondly, we fulfilled the smaller corpus with the duplication of the data (called plus or "+").

Table 6
Results of sentiment analysis data disparity handling

		Sentence-level		Aspect-base
		HTS2	HTS5	OHB3
huBERT	original	84.07	66.00	**81.99**
	+	85.92	**67.50**	81.00
	-	**86.49**	39.75	77.99

HILBERT	original	83.33	**68.00**	**57.80**
	+	**86.66**	50.49	57.40
	-	86.29	37.74	56.99
HIL-RoBERTa	original	75.92	**59.15**	**68.50**
	+	**78.14**	57.99	68.00
	-	77.03	38.74	66.29
HIL-ALBERT	original	75.56	55.49	63.99
	+	**78.51**	**56.49**	**65.10**
	-	74.07	38.99	60.90
HIL-ELECTRA	original	**78.89**	**59.11**	65.37
	+	75.09	40.10	**69.07**
	-	71.38	30.58	67.87
HIL-ELECTRIC	original	76.67	**58.00**	63.66
	+	**77.78**	35.75	**65.47**
	-	75.19	34.25	64.26
HILBART	original	71.11	51.25	**60.39**
	+	**78.51**	**52.24**	60.29
	-	75.55	31.74	51.49
NYTK-GPT-2	original	77.40	57.49	**73.69**
	+	**79.25**	**58.24**	72.89
	-	77.40	31.49	70.80
XLM-RoBERTa	original	83.33	**63.49**	**79.69**
	+	**87.03**	61.00	78.50
	-	86.29	36.50	75.40
mBERT	original	78.51	**57.74**	**75.49**
	+	**78.88**	55.25	75.19
	-	78.88	36.75	74.09
fastText	original	71.9	53.2	59.5
	+	**72.2**	**53.7**	**62.0**
	-	70.0	29.7	60.2

Based on Table 6 one can observe that the duplication technique could increase the quality of the binary classification, while it does not have the significant benefit in the other data sets. This technique is facilitated by the systems that started to use the edged labels. As expected, the size reduction of the training data significantly decreased the quality significantly. On the other hand, the precision of the classification of edged labels became better, while other ways of classification resulted in a setback.

8 Data Augmentation Experiments and Results

8.1 Machine Translation and Cross-Lingual Transfer

As it was described above, the size of the training data is crucial for training neural models. Unfortunately, we are in the absence of good quality data and it is a really expensive task to create it manually. In our research, machine translation and cross-lingual transfer methods were used for increasing our training dataset [50].

Our idea was to use already existing English corpora and use its translation as an auxiliary training set. The idea comes from machine translation (MT), where back translated corpora have been used to increase the quality of the translation of a low resourced language pair [51]. During our work MarianNMT [52] was used, which is a freely available software package written in C++. It is an easy to install, memory- and resource-optimal implementation, which makes it the most commonly used tool by academic users and developers [53]. A transformer-based encoder-decoder architecture was used with SentencePiece tokenization [54]. The tokenizer used common vocabulary for both languages and the vocabulary size was set to 32,000. We used the default parameters of the framework for the size of hidden layers and for the optimization metric. For training data the English-Hungarian language pairs of the ParaCrawl [55] corpora were used. The total training data contains ~45.5M segments and ~573M English tokens. The system reached 35.54% BLEU word level score on the test set. We had achieved the state of the art performance in English-Hungarian language pair [56]. Using our machine translation system, ACL14 and SST corpora were translated into Hungarian.

There are two ways to use translated corpora. First of all, cross-lingual data transfer could be used, where an English corpus could be used as a first stage fine-tune dataset before the use of the in-domain high quality one (we will call it as *translate+finetune*). Secondly, the auxiliary corpora could be concatenated with the in-domain one (we will refer to it as *mix*). The first fine-tuning stage has been done with the concatenated data (*mix*) followed by a second fine-tuning with the in-domain one (*mix+finetune*).

8.2 Experiments and Results

In our research, 7 different experiments were carried out:

- **original**: all pre-trained models were fine-tuned on the original HTS corpora. This will be our baseline method.
- **zeroshot**: multilingual models are able to predict for Hungarian NLP task. In this case the English corpora were used for fine-tuning and the system was used to predict for Hungarian sentences.

- **transfer**: multilingual models were fine-tuned on SST corpora, followed by further fine-tuning on HTS train corpora.
- **translate**: all pre-trained models were fine-tuned on translated SST corpora (SST_hu).
- **translate+finetune**: all pre-trained models were fine-tuned on SST_hu corpora, then fine-tuned with HTS train corpora.
- **mixed**: all pre-trained models were fine-tuned on the concatenation of SST_hu and HTS train corpora, then tested on HTS test corpora.
- **mixed+finetune**: all pre-trained models were fine-tuned on the concatenation of SST_hu and HTS train corpora, then further fine-tuned on HTS train corpora.

All experimental results were evaluated on HTS and OHB test corpora.

In Table 7, the results of our experiments are presented. Adding translation text to the training corpora could enhance the performance of sentiment analysis classification in most cases. For all the applied models, we can state that one of our translation methods could gain higher results than the baseline method.

We can define three distinct quality clusters based on the performance of the used models. The weakest systems were produced by HILBART and fastText. These are expected results, because the HILBART is created primarily for that text generation tasks, while the fastText is an obsolete static non-contextual word representation method which underperformes compared to contextual language models. On the other hand, we should take into account that fastText model requires much less resources to train the system and for online prediction it uses only CPU-s.

The second group contains the systems between 77-80% accuracy score of the binary classification and 58-63% accuracy score of the 5-class task. Finally, there are three systems in the top cluster (huBERT, XLM-RoBERTa and HILBERT) with about 85.5% accuracy of binary classification and about 66-69% accuracy score of the 5-class task. There is an interesting result that the XLM-RoBERTa multilingual model could achieve higher performance in HTS2 task, than the Hungarian language-specific huBERT model, which is the state-of-the-art LM model for most of the NLP tasks. Furthermore, the HILBERT model also could outperform the huBERT in HTS2 task, which was expected as well, because even though it was trained on less data, it is a large model and it operates with more parameters.

Table 7
Data augmentation results

		Sentence-level		Aspect-based
		HTS2	HTS5	OHB3
huBERT	original	84.07	66.00	81.99
	translate	73.33	29.25	63.70
	translate+finetune	85.55	66.50	81.69

	mixed	**85.55**	**68.99**	**82.30**
	mixed+finetune	84.81	68.00	81.49
HILBERT	original	83.33	68.00	57.80
	translate	74.07	34.75	52.10
	translate+finetune	82.59	67.75	54.40
	mixed	82.22	**68.50**	51.49
	mixed+finetune	**85.56**	68.00	**58.60**
HIL-RoBERTa	original	75.92	59.15	68.50
	translate	48.89	29.75	54.90
	translate+finetune	**79.63**	56.75	69.49
	mixed	76.66	**59.25**	**69.80**
	mixed+finetune	77.78	57.99	66.69
HIL-ALBERT	original	75.56	55.49	63.99
	translate	52.59	28.75	49.30
	translate+finetune	77.03	56.75	61.40
	mixed	72.22	60.50	**64.60**
	mixed+finetune	**77.41**	**60.75**	64.09
HIL-ELECTRA	original	78.89	59.11	65.37
	translate	55.02	37.34	56.86
	translate+finetune	**79.93**	61.15	67.97
	mixed	76.58	60.90	68.17
	mixed+finetune	79.18	**62.66**	**70.57**
HIL-ELECTRIC	original	76.67	58.00	63.66
	translate	52.79	37.34	54.65
	translate+finetune	78.52	56.75	66.57
	mixed	75.46	56.39	63.46
	mixed+finetune	**80.37**	**59.75**	**67.27**
HILBART	original	71.11	51.25	60.39
	translate	47.77	31.00	41.99
	translate+finetune	74.07	53.25	**62.09**
	mixed	71.48	52.50	59.70
	mixed+finetune	**76.66**	**54.75**	61.19
NYTK-GPT-2	original	77.40	57.49	73.69
	translate	60.37	31.74	59.79
	translate+finetune	78.51	**57.99**	73.79
	mixed	79.62	51.49	**74.80**
	mixed+finetune	**82.59**	57.49	73.90
mBERT	original	78.51	57.74	75.49
	zeroshot	47.41	30.50	61.19
	transfer	78.51	57.99	75.70
	translate	48.88	28.75	41.60
	translate+finetune	**79.25**	56.75	61.40

	mixed	77.77	56.99	75.49
	mixed+finetune	78.89	**59.75**	**76.39**
XLM-RoBERTa	original	83.33	63.49	79.69
	zeroshot	68.88	40.99	66.79
	transfer	84.81	66.25	79.79
	translate	68.51	35.25	63.89
	translate+finetune	85.18	66.00	**80.59**
	mixed	85.18	66.25	79.69
	mixed+finetune	**85.56**	**66.50**	77.70
fastText	original	71.9	53.2	59.5
	translate	62.2	32.0	55.5
	translate+finetune	73.3	**56.2**	59.5
	mixed	74.1	51.7	59.5
	mixed+finetune	**75.6**	53.5	**60.2**

In Figure 1, we have compared the performances and F1 results of 5 different types of sentence-level models. The state-of-the-art Hungarian language model huBERT, the HILBERT large model, the non-contextual fastText and the two multilingual models were compared. The sole significant result is that only the huBERT and HILBERT have predicted score 1 and they have predicted more score 5 than the multilingual models or fastText. The fastText did not predict neither score 1 and 5. It means the score 1 nor 5 occur infrequently in the training corpus (see Table 3). The huBERT and HILBERT are Hungarian models and huBERT trained on the corpus that contains 9 billion tokens, the HILBERT is a large model with 340 million parameters, thus they could learn more sophisticated details.

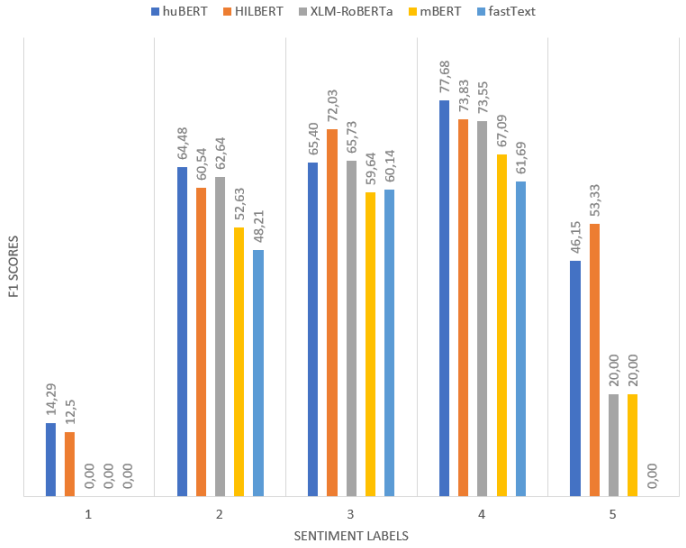

Figure 1
F1 score comparison of HTS5 task

Conclusions

Our study proposes new approaches in sentence-level and aspect-based sentiment analysis for Hungarian language. We have constructed the first neural sentiment analysis models for Hungarian, which achieved state-of-the-art performance and can be considered a new artificial cognitive capability in this field. We conclude that the addition of translation texts to the corpora generally increases the performance of our models, which is an important implication with reference to the optimization of sentiment analysis pipelines. Remarkably, our data augmented models could outperform the our state-of-the-art models in multiple tasks, which offers promising new ways to apply the results presented in this paper to facilitate the progression of the areas that are based on the proceedings of sentiment analysis. Our findings are especially relevant for the development of novel strategies that can contribute to the efficient collaboration of interdisciplinary teams working in different domains connected to cognitive infocommunication.

References

[1] C. Vogel and A. Esposito, "Linguistic and Behaviour Interaction Analysis within Cognitive Infocommunications," in *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2019

[2] P. Baranyi and Á. Csapó, "Definition and Synergies of Cognitive Infocommunications," *Acta Polytechnica Hungarica,* Vol. 9, pp. 67-83, 2012

[3] P. Baranyi, A. Csapo and G. Sallai, Cognitive Infocommunications (CogInfoCom), Springer International Publishing, 2015

[4] B. Bogdándy, Á. Kovács and Z. Tóth, "Case Study of an On-premise Data Warehouse Configuration," in *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2020

[5] M. Hoang, O. A. Bihorac and J. Rouces, "Aspect-Based Sentiment Analysis using BERT," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Turku, 2019

[6] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros and T. By, "Sentiment Analysis on Social Media," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012

[7] J. R. Saura, P. Palos-Sanchez and A. Grilo, "Detecting Indicators for Startup Business Success: Sentiment Analysis Using Text Data Mining," *Sustainability,* vol. 11, p. 917, February 2019

[8] S. A. Kinholkar and P. K. C. Waghmare, "Enhance Digital Marketing Using Sentiment Analysis and End User Behavior," in *International Research Journal of Engineering and Technology (IRJET)*, 2016

[9] H. Jo, S.-M. Kim and J. Ryu, *What we really want to find by Sentiment Analysis: The Relationship between Computational Models and Psychological State,* 2018

[10] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 2002

[11] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos and S. Manandhar, "SemEval-2014 Task 4: Aspect Based Sentiment Analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, 2014

[12] R. Feldman, "Techniques and Applications for Sentiment Analysis," *Commun. ACM,* Vol. 56, pp. 82-89, April 2013

[13] M. Miháltz, "OpinHuBank: szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez (OpinHuBank: open source annotated corpora for Hungarian sentiment analysis)," in *IX. Conference on Hungarian Computational Linguistics*, Szeged, Hungary, 2013

[14] V. Hangya, R. Farkas and G. Berend, "Entitásorientált véleménydetekció webes híranyagokból (Aspect-based sentiment detection in online news data)," in *XI. Conference on Hungarian Computational Linguistics*, Szeged, Hungary, 2015

[15] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," 2002

[16] J. Zsibrita, V. Vincze and R. Farkas, "magyarlanc: A Tool for Morphological and Dependency Parsing of Hungarian," in *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, Hissar, 2013

[17] M. K. Szabó, V. Vincze, K. I. Simkó, V. Varga and V. Hangya, "A Hungarian Sentiment Corpus Manually Annotated at Aspect Level," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, 2016

[18] Dívány, "Dívány," 2022 [Online] Available: https://divany.hu [Accessed 09 02 2022]

[19] J. Steinberger, P. Lenkova, M. Kabadjov, R. Steinberger and E. van der Goot, "Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora," in *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, Hissar, 2011

[20] Hungarian Research Centre for Linguistics, "Github - nytud/sentiment-analysis - Sentiment Analysis," 2022 [Online] Available: https://github.com/nytud/sentiment-analysis [Accessed 09 02 2022]

[21] Hungarian Research Centre for Linguistics, "Demo site," 2022 [Online] Available: https://juniper.nytud.hu/demo/sentana. [Accessed 09 02 2022]

[22] Precognox, "opendata.hu - Hungarian Twitter Sentiment Corpus," 2022 [Online] Available: http://opendata.hu/dataset/hungarian-twitter-sentiment-corpus. [Accessed 09 02 2022]

[23] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy and S. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, 2018

[24] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou and K. Xu, "Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification," in *The 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014

[25] D. M. Nemeskey, "Introducing huBERT," in *XVII. Conference on Hungarian Computational Linguistics*, Szeged, Hungary, 2021

[26] 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, 2019

[27] Á. Feldmann, R. Hajdu, B. Indig, B. Sass, M. Makrai, I. Mittelholcz, D. Halász, Z. G. Yang and T. Váradi, "HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben (HILBERT, Training a Hungarian BERT-large Model in a Cloud Environment)," in *XVII. Conference on Hungarian Computational Linguistics*, Szeged, Hungary, 2021

[28] Z. G. Yang, Á. Agócs, G. Kusper and T. Váradi, "Abstractive text summarization for Hungarian," *Annales Mathematicae et Informaticae,* VOL. 53, pp. 299-316, 2021

[29] Z. G. Yang, Á. F. Feldmann and T. Váradi, "A kis HIL-ELECTRA, HIL-ELECTRIC és HIL-RoBERTa - Magyar kísérleti nyelvi modellek tanítása kevés erőforrással (The small HIL-ELECTRA, HIL-ELECTRIC and HIL-RoBERTa - Training Hungarian experimental models with low resources)," in *XVIII. Conference on Hungarian Computational Linguistics*, Szeged, Hungary, 2022

[30] G. Lai, Q. Xie, H. Liu, Y. Yang and E. Hovy, "RACE: Large-scale ReAding Comprehension Dataset From Examinations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, 2017

[31] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, 2016

[32] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach,* 2019

[33] C. Hungarian Intelligent Language Applications, "HILANCO - HIL-ALBERT," 2022 [Online] Available: https://hilanco.github.io/models/albert.html [Accessed 09 02 2022]

[34] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," in *Proceedings of the Eighth International Conference on Learning Representations*, 2020

[35] K. Clark, M.-T. Luong, Q. V. Le and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," in *ICLR*, 2020

[36] K. Clark, M.-T. Luong, Q. V. Le and C. D. Manning, "Pre-Training Transformers as Energy-Based Cloze Models," in *EMNLP*, 2020

[37] C. Hungarian Intelligent Language Applications, "HILANCO - HILBART," 2022 [Online] Available: https://hilanco.github.io/models/hilbart.html. [Accessed 09 02 2022]

[38] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020

[39] Z. G. Yang, ""Az invazív medvék nem tolerálják a suzukis agressziót" - Magyar GPT-2 kísérleti modell ("Invasive bears do not tolerate Suzuki aggression" - Hungarian GPT-2 experimental model)," in *XVIII. Conference on Hungarian Computational Linguistics*, Szeged, Hungary, 2022

[40] A. Radford and K. Narasimhan, "Improving Language Understanding by Generative Pre-Training," 2018

[41] A. Radford, J. Wu, R. Child, D. Luan, D. Amodeia and I. Sutskever, "Language models are unsupervised multitask learners," 2019

[42] S. Chen, Y. Pei, Z. Ke and W. Silamu, "Low-Resource Named Entity Recognition via the Pre-Training Model," *Symmetry,* Vol. 13, 2021

[43] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer and V. Stoyanov, *Unsupervised Cross-lingual Representation Learning at Scale,* 2020

[44] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, "Bag of Tricks for Efficient Text Classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, 2017

[45] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou and T. Mikolov, "FastText.zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651,* 2016

[46] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics,* Vol. 5, pp. 135-146, 2017

[47] Hugging Face, "Github - huggingface/transformers - Text classification examples," 2022 [Online] Available: https://github.com/huggingface/transformers/tree/master/examples/pytorch/text-classification. [Accessed 09 02 2022]

[48] Google, "Github - google-research/electra," 2022 [Online] Available: https://github.com/google-research/electra. [Accessed 09 02]

[49] Facebook Inc., "fastText," 2020 [Online] Available: https://fasttext.cc. [Accessed 09 02 2022]

[50] M. Fadaee, A. Bisazza and C. Monz, "Data Augmentation for Low-Resource Neural Machine Translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, 2017

[51] A. Poncelas, D. S. Shterionov, A. Way, G. M. de Buy Wenniger and P. Passban, "Investigating Backtranslation in Neural Machine Translation," *CoRR,* Vol. abs/1804.06189, 2018

[52] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins and A. Birch, "Marian: Fast Neural Machine Translation in C++," in *Proceedings of ACL 2018, System Demonstrations*, Melbourne, 2018

[53] L. Barrault, O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. MĂźller, S. Pal, M. Post and M. Zampieri, "Findings of the 2019 Conference on Machine Translation (WMT19)," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, 2019

[54] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, 2018

[55] C. f. E. L. Broader/Continued Web-Scale Provision of Parallel, "ParaCrawl," 2022 [Online] Available: https://paracrawl.eu/index.php [Accessed 09 02 2022]

[56] L. J. Laki and Z. G. Yang, "Neural machine translation for Hungarian," *Acta Linguistica Academica,* Vol. 69, No. 4, pp. 501-520, 2022