Linguistic Knowledge Application to Neuro-Symbolic Transformers in Sentiment Analysis

1st Joanna Baran
Department of Artificial Intelligence
Wrocław University of Science and Technology
Wrocław, Poland
joanna.baran@pwr.edu.pl

2nd Jan Kocoń
Department of Artificial Intelligence
Wrocław University of Science and Technology
Wrocław, Poland
jan.kocon@pwr.edu.pl

***Abstract*—Neuro-symbolic approaches explore ways to combine neural networks with traditional symbolic knowledge. These methods are gaining attention due to their efficiency and the requirement of fewer data compared to currently used deep models. This work investigated several neuro-symbolic models for sentiment analysis focusing on a variety of ways to add linguistic knowledge to the transformer-based architecture. English and Polish WordNets were used as a knowledge source with their polarity extensions (SentiWordNet, plWordNet Emo). The neuro-symbolic methods using knowledge during fine-tuning were not better or worse than the baseline model. However, a statistically significant gain of about three percentage points in the F1-macro was obtained for the SentiLARE model that applied domain data — word sentiment labels — already at the pretraining stage. It was the most visible for medium-sized training sets. Therefore, developing an effective neuro-symbolic model is not trivial. The conclusions drawn from this work indicate a further need for a detailed study of these approaches, especially in natural language processing. In the context of sentiment classification, it could help design more efficient AI systems that can be deployed in business or marketing.**

***Index Terms*—neuro-symbolic, sentiment analysis, WordNet, linguistic knowledge**

I. Introduction

Deep learning models have been widely used in many artificial intelligence (AI) tasks. However, despite their performance, neural networks suffer from a lack of interpretability and the need for large amounts of data for learning. In contrast, traditional *symbolic* methods that use human-prepared information are easier to explain and can work with less data. Unfortunately, they depend heavily on the quality of external knowledge and do not capture complex correlations. The shortfalls of these two techniques have led researchers to explore a novel area of AI called ***neuro-symbolic***, a combination of deep learning architectures with symbolic reasoning. This approach has already outperformed state-of-the-art (SOTA) deep learning models, in domains such as image and video reasoning with significantly less training data [1], [2].

A good understanding of the language semantics is very important for **text classification** problems when the model has to learn to predict the appropriate label from the input text. This work focuses primarily **sentiment analysis** which aims to decide whether the presented text has a positive, negative, or neutral polarity. In a more advanced scenario, handling the sentiment ambivalence can also be considered [3]. Standard machine learning techniques have been applied to this task, such as decision trees, SVM, Naive Bayes or random forest. From the classical symbolic approach, a text is evaluated using external knowledge bases, e.g., sentiment dictionaries, due to which words from the text are linked to proper polarization derived from such dictionaries. The final sentiment is an aggregation over all words [4], [5]. Currently, the most common methods for sentiment analysis are deep learning architectures based on transformer models [6].

Many lexicon resources for various languages have been developed in the past decade. Princeton WordNet (PWN) [7] is a major one for English, but similar knowledge bases have also been created for other languages. Some contain emotive annotations for specific word meanings assigned by people (e.g. SentiWordNet [8], plWordNet Emo [9]). In addition, natural language processing (NLP) tools were created to analyze data like human understanding, such as part-of-speech (PoS) tagger or word sense disambiguation (WSD) systems. Given the difficulty of the sentiment analysis task, which connects NLP with psychology and cognitive science, intuitively, it makes a good candidate for using neuro-symbolic methods via incorporating external knowledge prepared by linguists. It could improve the results of standard deep learning models, as already proven for other NLP problems [10], [11].

The main objective of this article is to answer the following Research Questions: **(RQ1)**: How to efficiently provide linguistic knowledge to the transformer-based model and improve the model's classification results in relation to the baseline? **(RQ2)**: Will the model performance change when knowledge is incorporated during fine-tuning versus the pretraining stage? **(RQ3)**: Does the neuro-symbolic approach deliver on its promises regarding the requirements for fewer data while performing as well as the baseline on full dataset?

2375-9259/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDMW58026.2022.00059

II. Related Work

Sentiment analysis has recently become a very active research area in AI that has led to many proposals of traditional machine learning techniques and those based on neural network architectures [6]. Some use human knowledge in graphs, lexicons, or other symbolic representations. One of the main NLP resources is a knowledge graph called **WordNet**. It is a large lexical database where nouns, verbs, adjectives, and adverbs are grouped into cognitive synonyms (*synsets*), each expressing a distinct concept. The first WordNet was created for English – **Princeton WordNet** (PWN) [7]. The Polish version **plWordNet** (*pl. Słowosieć*) [12] was created soon after and has been continuously extended until today. The most useful for sentiment classification are extensions of wordnets focusing on opinion mining. For English, it is **SentiWordNet** [8] which assigns to each synset of PWN three sentiment scores (float values): positivity, negativity, and objectivity. In case of Polish, **plWordNet Emo** [13] provides 80k units tagged for emotional overtones - positive, negative, ambiguous, or neutral. This annotation was performed on the level of lexical units and their use examples [14]. Furthermore, the labeling of units was expanded to include eight basic emotions and valuations [15].

Another source of polarity knowledge is various kinds of **lexicons**. Many of them differ mainly in the way they were labeled and in the source of the data provided for the annotation process. The vast majority are obviously for English. One of the most known is the **NRC Word-Emotion Association Lexicon**, commonly called *EmoLex* [16]. The annotations of ≈25k English senses were made here by crowdsourcing and consisted of eight basic emotion labels and two sentiments (negative and positive). We can also distinguish automatically generated lexicons, such as the NRC Hashtag Sentiment Lexicon [17] or the Sentiment140 Lexicon [18]. The last interesting collection is a database **SenticNet** [19], which provides a set of semantics, sentics, and polarity associated with 400k natural language concepts.

Symbolic methods to classify sentiment use linguistic knowledge, rules, or lexicons. They are mainly based on statistics, algebra, feature engineering, and traditional machine learning models like SVMs or decision trees [20]. A standard mathematical approach to predict sentiment is to **sum up the polarity values of all words in the document** that exist in a sentiment lexicon. This idea was used in [21] with Princeton WordNet as a knowledge source. Another popular method in the literature is **feature extraction**, which mainly involves counting words with a given polarity. Extensive research on this topic on Twitter data was done by [22] who obtained the prior polarity of words from Whissel's DAL [23] dictionary and prepared emotion lexicons. In-depth feature analysis showed that those of Twitter-specific type (emoticons, hashtags, etc.) add value to the classifier only marginally. Authors highlighted the **urgent need to develop sentiment lexicons** even when they are created on a different domain of data than the target task. Polarity feature engineering was also reviewed in [18] where two classifiers were developed for SemEval-2013 Task 2 [24]. The solution presented there uses several surface-form, semantic, and sentiment features, e.g., count of negated words, emoticons polarity, number of hashtags, word n-grams presence, etc. Each text was represented as a feature vector, an input to the SVM model. Again, **sentiment lexicon features led to the most gain in performance**.

The main advantage of neural networks is their ability to capture context-dependent semantic composition effects over sentences. The first neuro-symbolic approaches have evolved closely around already known classical ones. A simple one was proposed in [25] where the total sentiment score of a given sentence was treated as a **weighted sum of prior polarity scores** of negation and sentiment words, where the weights are learned by the neural model. The LSTM architecture was explored in work [26] combined with **regularization techniques**. Three types of linguistic resources were addressed in that paper: sentiment lexicon as a source of the prior polarity of a word, negation words -– typical sentiment shifters, and intensity words that change the valence degree of the current polarity score. The key idea was to regularize the difference between the predicted sentiment distribution of the current position and that of the previous or next positions. Another group of incorporating external knowledge into neural-based models is focused on adding **lexicon-aware input vectors** to the usual context embeddings. This idea was highly exploited in research. [27] utilized lexical features into transformer-based BERT model [28] for abusive or offensive remarks detection in social media posts. Authors introduced two architectures – *HurtBERT Encoding* and *HurtBERT Embedding*. Features of a vector or matrix were given in a separate model branch and then concatenated with the BERT output. The work proved that proposed methods improved the results over the BERT baseline in most of the provided in-domain and cross-domain experiments.

MOre recent ideas exploit the potential of deep network architectures. SentiLARE model introduced in [29] injects knowledge into the RoBERTa on the **pretrain stage** by modifying the standard masked language model (MLM) training task. This way, authors enriched the input sequence with the external information, including part-of-speech tag and sentiment polarity, to capture the relationship between sentence-level language representation and word-level linguistic knowledge. A different idea to infuse factual knowledge into the pretrained model was described in [30]. Proposed method *Knowledge Embedding and Pre-trained Language Representation* (KEPLER) takes advantage of knowledge graphs delivering the information about entities. An example of such a resource type may be the WordNet, where entities or senses are connected with different types of relations.

III. Data preparation

This section describes the datasets used and the preprocessing steps taken to efficiently provide the external knowledge of the polarity scores and emotion annotations from the WordNets – Princeton PWN and Polish plWN.

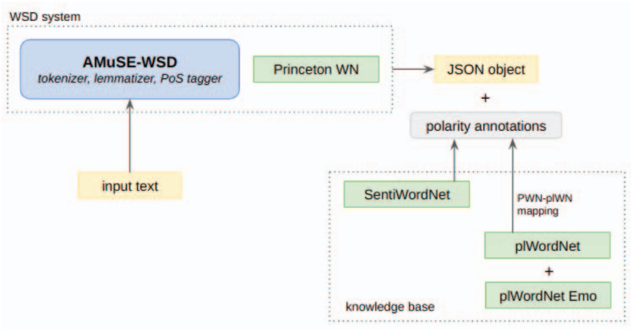


Fig. 1: Data preprocessing steps

A. Datasets

The focus was on open-source data for the English and Polish languages only.

- **IMDB** [31] - a collection of English reviews from the Internet Movie Database (IMDb) website labelled as positive or negative. Two class setup was derived from a 10-point rating scale. In this dataset, only highly polarizing reviews are considered -– a *negative* text has a score ≤ 4 out of 10, and a *positive* review has a score ≥ 7 out of 10.
- **Movie Reviews** [32] - a dataset of English movie review sentences labeled for their overall sentiment polarity or subjective rating scale. Binary classification.
- **Multiemo/PolEmo2.0** [33] - a benchmark dataset for the multilingual sentiment analysis task. The collection was created on the basis of the Polish written dataset PolEmo2.0 [34] which was translated automatically into ten other languages. The following classes were assigned: minus, zero, ambiguous, and plus.
- **GoEmotions** [35] - a human-annotated English dataset of carefully curated Reddit comments labelled for 27 fine-grained emotion categories or neutral. These data are intended for the *multi-label* classification of emotions. The distribution of labels is uneven; the neutral tag strongly dominates the other emotions, making this set the most difficult to get good classification results.
- **Allegro Reviews** [36] - a sentiment analysis dataset consisting of product reviews written in Polish and extracted from the Allegro website – a popular e-commerce marketplace. Each review contains at least 50 words and has a rating on a 5-point scale.

B. Preprocessing

All texts from the datasets were preprocessed by the AMuSE-WSD tool introduced in [37] which is an end-to-end system with a SOTA multilingual model for the Word Sense Disambiguation (WSD) task. External domain knowledge (polarity scores) was added in the next steps. The whole pipeline is shown in Figure 1.

Two sources of sentiment information were chosen for this work - SentiWordNet and plWordNet Emo. The output from AMuSE-WSD returned the Princeton WordNet sense offsets, thus the easiest way to get polarity scores was from SentiWordNet. Adding emotive tags and sentiment annotations from plWordNet Emo was more complex. Thanks to the partial integration of plWN with PWN, the PWN-plWN synset IDs mapping was obtained through synonymy, hyperonymy, and hyponymy relations. As a result, separate files containing word-level annotations were created for each dataset.

TABLE I: Token annotation coverage in preprocessed datasets (*pos* - positive, *neg* - negative)

Dataset	SentiWordNet (%)			plWordNet Sent (%)			plWordNet Emo (%)	
	all	pos	neg	all	pos	neg	emotions	valuations
IMDB	44.9	12.0	5.4	27.5	3.9	3.5	23.0	23.4
Movie Reviews	52.3	14.4	7.2	32.1	5.6	5.0	26.5	26.8
Polemo2.0	48.1	8.3	5.9	31.5	3.8	2.9	24.5	25.1
Multiemo	45.0	10.5	5.0	28.6	3.3	2.3	22.8	23.2
Allegro Reviews	47.9	8.3	6.0	28.2	3.7	2.8	22.4	22.7
GoEmotions	45.8	12.5	5.9	25.9	3.8	3.5	24.9	25.2

Each preprocessed dataset was analyzed in terms of token annotation attribute coverage. A summarization is presented in Table I. All collections for $\approx 50\%$ of their words have a sentiment value assigned from SentiWordNet. There is also a significant predominance of neutral words over positive and negative ones, with negative words being the minority.

IV. Models

The work aimed to explore methods from simple external knowledge delivery to advanced integration into model weights. Four different neuro-symbolic architectures using the transformer were chosen based on these conditions. All of them were taken from sources in the literature by re-implementation or adaptation to the research conditions.

A. HurtBERT

This method was proposed in [27] for the abusive language detection task. It extracts features from the HurtLex lexicon, which are then processed by a separate branch and concatenated with contextual embeddings before the classification layer. Lexical information can be used in two ways:

- **HurtBERT Encoding** - using a frequency count for the lexicon categories;
- **HurtBERT Embedding** - obtained with an embedding layer mapping a given word to a one-hot representation, and LSTM network (Figure 2).

The second method is more expressive, as it considers token order. As the amount of categories in plWordNet differs from the ones used in the original paper (17 specific to the HurtLex), the dimensionality of the sentiment embedding layer had to be modified to 24. This number includes all fine-grained sentiment scores and possible emotion tags from plWordNet Emo.

B. Tailored KEPLER

An adaptation of KEPLER [30] which infuses descriptive information from a knowledge graph into a pretrained language model (PLM). It uses factual entries about entities

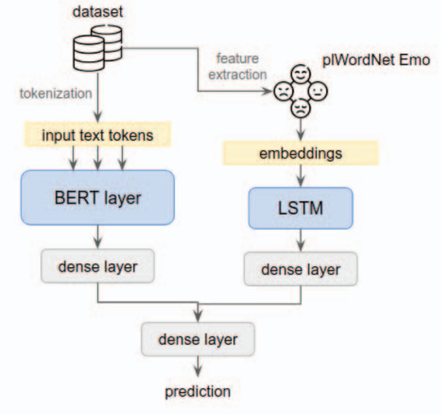


Fig. 2: HurtBERT Embedding architecture

contained in resources such as Princeton WordNet or Wikidata [38]. Graph structure is represented as TransE [39] triplets (h, r, t) where h, t emblems the head and the tail entities; r is a relation type. To harness knowledge, its entities' representations are obtained by encoding their textual descriptions with PLM. Thus, PLM can be learned with the Knowledge Embedding objective and a task objective.

The presented approach differs from the initial implementation, where extra knowledge is used during the pretraining stage via unsupervised masked language modeling. *Tailored KEPLER* is modified to a single task — it incorporates entity data during fine-tuning (Figure 3). As a graph knowledge source, the plWordNet is used, from which the relations between lexical units and synsets are extracted along with their definitions and usage text samples. Since the data processing step was based on synsets, the considered triplets were narrowed down to the relationships like synonymy, hypernymy, and hyponymy. The relation types are encoded by a randomly initialized, learnable embedding table.

$$\mathcal{L}_{\mathrm{KE}} = -\log\sigma(\gamma - d_r(\mathbf{h}, \mathbf{t})) - \sum_{i=1}^{n}\frac{1}{n}\log\sigma(d_r(\mathbf{h}'_{\mathbf{i}}, \mathbf{t}'_{\mathbf{i}}) - \gamma) \tag{1}$$

$$d_r(\mathbf{h}, \mathbf{t}) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\| \tag{2}$$

To fine-tune the pretrained model, multitask loss $\mathcal{L} = \mathcal{L}_{\mathrm{KE}} + \mathcal{L}_{\mathrm{NLP}}$ is applied, where $\mathcal{L}_{\mathrm{NLP}}$ is responsible for a downstream NLP task. Only those triplets that include synsets visible in the training dataset split are given during learning.

C. SentiLARE

The last model is taken directly from [29] where knowledge is injected into transformer-based RoBERTa architecture by modification of a standard MLM task, creating a new *label-aware* masked language model (LA-MLM). Besides textual sequence, the SentiLARE input requires external information – part-of-speech tags and sentiment values. In the original work, PoS for each word is acquired via Stanford Log-Linear tagging tool [40] and then used to match the polarity scores from SentiWordNet. One lemma can have several meanings, therefore, the sentiment values are obtained by the unsupervised WSD method [25].

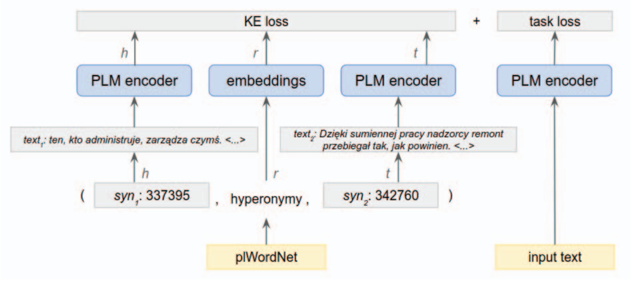

KE loss
+
task loss
h
r
t
PLM encoder
embeddings
PLM encoder
PLM encoder
text$_1$: ten, kto administruje, zarządza czymś. <...>
text$_2$: Dzięki sumiennej pracy nadzorcy remont przebiegał tak, jak powinien. <...>
h
r
t
(syn$_1$: 337395 , hyperonymy , syn$_2$: 342760)
plWordNet
input text

Fig. 3: KEPLER architecture

TABLE II: Sample sizes in dataset splits

Dataset	#Train	#Validation	#Test
IMDB	22,500	2,500	25,000
Movie Reviews	8,534	1,078	1,050
Polemo2.0/Multiemo	6,573	823	820
Allegro Reviews	9,577	1,002	1,006
GoEmotions	43,410	5,426	5,427

The model is pretrained in the LA-MLM setup, consisting of two subtasks – *early fusion* and *late supervision* — combined in one loss function. In early fusion, SentiLARE is required in addition to predicting the word in masked positions, also predicting its PoS tag and word-level polarity conditioned by the whole sentence sentiment label. The latter subtask aims to predict the sentence-level label and the word information. As a result, this approach delivers **knowledge-aware representation vectors** as a direct output from RoBERTa.

V. Training Setup

The selected datasets already had a publicly available ready division into three files: training, validation, and testing set. Only IMDB and Movie Reviews data do not have validation sets, so the final three-split form was used from the repository [41]. The overview of split sizes is shown in Table II.

In an attempt to answer question **RQ3**, the size of the training set had to be manipulated during the model learning. A list of train sizes was defined for each type of dataset, starting with 250 examples and ending with the full set. Subsequent values were selected proportionally to the size of each training split so that at least approximately 5, 10, 15, 30, and 60% of the complete data were tested. In the case of the GoEmotions dataset, experiments were started with a training set size of 1000, as it was found that smaller sizes were not sufficient for any model — either with baseline or neuro-symbolic ones — to learn the classification task. This is most likely caused by the difficulty of the multi-label prediction task configuration. To maintain a distribution of classes in training subsets similar to the full version, a stratified splitting was used.

As a baseline model, SOTA transformer-based architectures with linear classification heads were fine-tuned for each task.

HerBERT [42] was used for Polish datasets and BERT for English. To properly compare the performance of the SentiLARE model, the baseline was changed to RoBERTa. The training took maximally 50 epochs with an early stopping patience parameter equal to 20 epochs. The batch size was set to 32 and the model was optimized using AdamW with a learning rate of $1\epsilon^{-5}$ along with a warmup linear scheduler.

VI. Experiments and results

The experiments were divided into two parts according to the stage of knowledge transfer to the models — during fine-tuning (**Finetune-Exp**) or pretraining (**Pretrain-Exp**). The first group contains HurtBERT and Tailored KEPLER architectures and the second – SentiLARE approach with modified MLM task. The authors of the SentiLARE paper have shared the checkpoint of the trained model [41] which was used for the purpose of this study. Due to the nature of RoBERTa, only the English datasets could be considered for fine-tuning using the published model. Each experiment was repeated 10 times and averaged F1-macro score was reported.

The polarity source for HurtBERT data was a combination of SentiWordNet and plWordNet Emo where plWN was taken first; the emotion information was only provided by the latter resource. In case of SentiLARE, datasets were enriched in two ways - using only SentiWN polarity scores, and using two knowledge graphs together. Thus, two versions of neuro-symbolic trials were distinguished accordingly: *SentiLARE (SentiWN)* and *SentiLARE (plWN)*.

The obtained results presented in Fig. 4a show that the HurtBERT and Tailored KEPLER models performed the same or worse than baseline BERT. Architectures using a simple way of providing knowledge at the fine-tuning stage failed to deliver expectations of the neuro-symbolic model. On the contrary, results achieved for SentiLARE (Fig. 4b) proved the superiority of the neuro-symbolic approach over the baseline model. It was possible to outperform the RoBERTa for most of the datasets; even a 2-4 percentage point metric increase was noted for the Multiemo and IMDB collections. The advantage was the highest for small to medium-sized training sets. However, no improvement in classification quality for the GoEmotions dataset was also noted. These insights were supported by Student's t test [43] between the SentiLARE and its baseline. The assumption of distribution normality was checked using the Shapiro-Wilk [44] test. No statistical significance was obtained only for GoEmotion. Performance comparison between versions of the model, *SentiLARE (SentiWN)* and *SentiLARE (plWN)*, also resulted in no significant difference. The addition of partial sentiment knowledge from Polish WordNet did not improve but also did not worsen the results, which may prove that plWN is well integrated with PWN.

To further investigate how the neuro-symbolic SentiLARE model improve learning process, Differential Data Cartography was plotted for Multiemo and IMDB dataset. The focus was on a medium training size, here equal to 1000 samples. Visualization was inspired by work [45] where three different training dynamics measures were presented.

- Confidence - captures how confidently the model assigned a *true* label (y_i^*) to the sample, calculated as a mean probability across epochs:

$$\hat{\mu} = \frac{1}{E}\sum_{e=1}^{E} p_e(y_i^*|x_i), \tag{3}$$

where p_e is a model probability at the end of epoch e.

- Variability - measures how model was indecisive about sample label during training, low value means stable prediction of one label, and high value - often change of assigned label; calculated as a standard deviation:

$$\hat{\sigma} = \sqrt{\frac{\sum_{e=1}^{E}(p_e(y_i^*|x_i) - \hat{\mu_i})^2}{E}}. \tag{4}$$

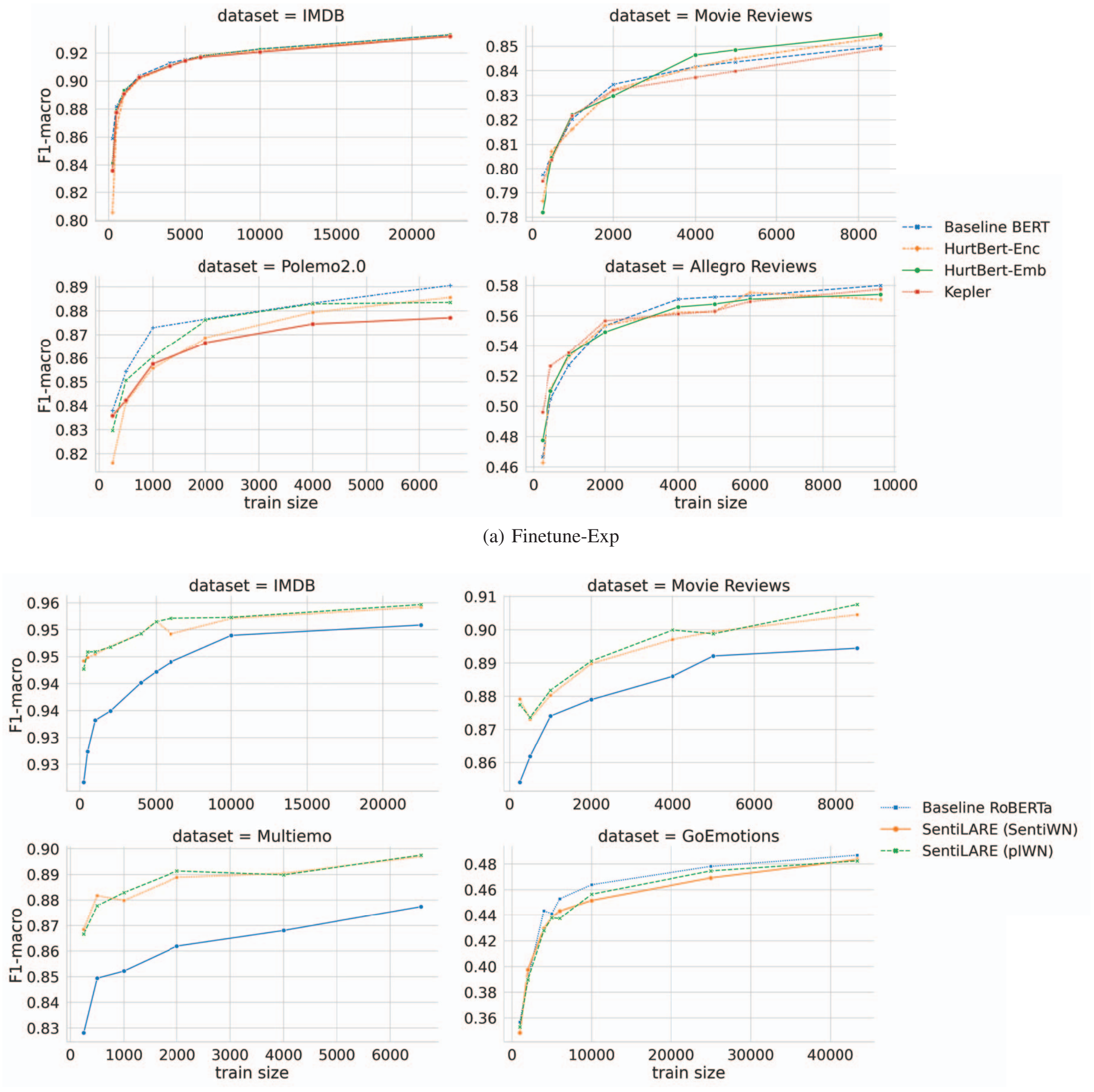
- Correctness - fraction of correctly predicted labels across training epochs.

Data maps in Fig. 5 and 6 show the difference in above measures between SentiLARE and baseline architectures. The values for the data points were obtained by subtracting the baseline's training dynamics metrics from the SentiLARE model's. Each data point symbolize one training sample. The most important part of the figure is the 4th quadrant of the coordinate system where majority of points are located. It denotes the increase in the model confidence and a reduction in its variability. Such changes are the most desirable and proves that the addition of prior sentiment knowledge at the pretraining stage improved the quality of the model.

VII. Discussion

RQ1: How to efficiently provide linguistic knowledge to the model and improve the model's classification results in relation to the baseline?

From the review conducted, the key to success was definitely **providing domain data at the pretraining stage** and injecting it into the *knowledge* of the model (its weights). SentiLARE was the only considered neuro-symbolic approach to the sentiment classification task that yielded a significant increase in prediction measures. However, in the context of the obtained results, the question arises of why the other methods did not work. For the Finetune-Exp group, all models failed to outperform the SOTA baselines. A possible explanation for HurtBERT-Encoding and HurtBERT-Embedding methods is that high-level embedded text does not cooperate well concated with simpler feature encoding before the classification layer. Overall, the values from both vectors had a different distribution, which apparently could interfere with the correct training of the network — it could come down to ignoring the values coming from the sentiment database. In case of Tailored KEPLER approach, the fault may have been in the knowledge graph representation way. The TransE method does not adequately express transitive relations, which are very common in WordNets [46]. Moreover, added knowledge



(a) Finetune-Exp

(b) Pretrain-Exp

Fig. 4: F1-macro scores for different train sizes

was of a more general type (definitions and usage examples) instead of sentiment-specific.

On the other hand, the Pretrain-Exp group's results showed that for the GoEmotions dataset the SentiLARE method was insufficient to beat the baseline. First, it could be explained by the fact that SentiLARE was created with the idea of classifying sentiment rather than emotion. Therefore, the model was heavily biased towards sentiment knowledge type.

In summary, the effectiveness of the neuro-symbolic model largely depends on how the knowledge is incorporated. Adding it to the deep-neural network weights at the pretraining level seems to be one of the most reasonable choices for sentiment classification.

RQ2: Will the model performance change when knowledge is incorporated during fine-tuning versus at the pretrain stage?

When infusing knowledge during pretraining, the SentiLARE had brought the desired increase in prediction quality, contrary to models with knowledge added at the fine-tuning stage. However, it cannot be rejected that there is or will be another way to infuse knowledge at the fine-tuning level of transformer-based architectures, which was not considered in this research. It could be also dependent on the type of task being solved and the type of knowledge, e.g., Tailored KEPLER did not work with transitive WordNet relations, but when the original KEPLER was trained on the Wikipedia with non-transitive relationships, it gave a boost in model quality measures in relation classification or entity typing tasks [30].

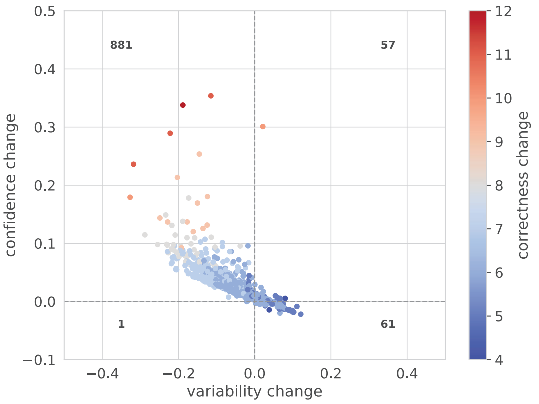


Fig. 5: Differential Data Cartography for SentiLARE and RoBERTa models on IMDB

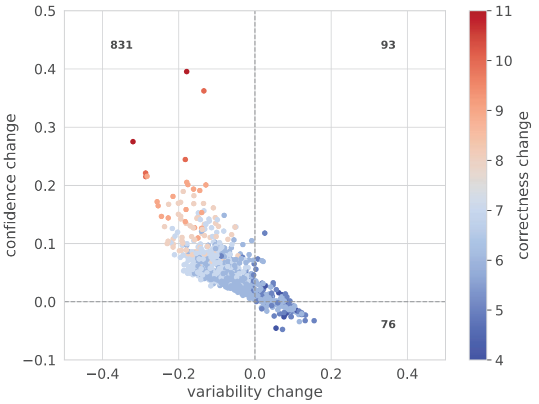


Fig. 6: Differential Data Cartography for SentiLARE and RoBERTa models on Multiemo

RQ3: Does the neuro-symbolic approach deliver on its promises regarding requirements for fewer data while performing as well as the baseline on full dataset?

Meeting these requirements is possible with a properly designed neuro-symbolic architecture which can access external symbolic knowledge effectively. In the scope of this study, only SentiLARE met this promise. It was especially reported for three datasets: IMDB, Movie Reviews, and Multiemo, where the gain was the greatest for small and medium-sized training sets. This issue requires further research.

VIII. Conclusions and Future Works

This work presented the review of several carefully selected *neuro-symbolic* models for text classification problems related to sentiment or emotion prediction. It is a novel, future-proof area in the machine learning field, and reports very promising results in various deep learning tasks.

An extensive research was done to analyze and validate neuro-symbolic solutions. They were conducted on five different datasets for English and Polish using four transformer-based approaches. A few ways of infusing linguistic knowledge into the models have been tested, from the simplest one in the form of a separate feature vector during fine-tuning, to injecting knowledge during pretraining of the language encoder. Then, the quality of the models was measured and compared.

The results obtained revealed the difficulty of designing the proper knowledge transfer method so that it could be compatible with transformers that operate at a high level of abstraction. Most of the neuro-symbolic architectures tested were unsuccessful; they did not improve F1-macro scores relative to the baseline. The only successful way to add external knowledge turned out to be by adding it directly into the model weights while pretraining the transformer blocks. This approach (SentiLARE) led to an approximately 3 pp. increase for some collections. It was an excellent result compared to the SOTA models.

The most important conclusion is that there are many methods of feeding human-prepared knowledge into neural networks, and how to do it effectively is still an open question. The conducted experiments may be a prelude to further research on the topic of neuro-symbolic approaches in the sentiment analysis. There is certainly a need to better explore the SentiLARE architecture and its full potential for the Polish language.

The effective usage of human knowledge in deep neural models will definitely require the development of new architectures in the near future. They will probably be of a completely different nature from the current ones, to better absorb external information. This potential breakthrough will also open the door to other areas of NLP, such as making inferences from textual data or its general understanding. Until now, neural networks have not been able to learn such tasks that require logical thinking. With the help of symbolic knowledge, this can change.

IX. Acknowledgements

This work was financed by (1) the National Science Centre, Poland, project no. 2019/33/B/HS2/02814 and 2021/41/B/ST6/04471; (2) the Polish Ministry of Education and Science, CLARIN-PL; (3) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19; (4) the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology.

References

[1] A. d. Garcez and L. C. Lamb, "Neurosymbolic ai: The 3rd wave," 2020.

[2] Z. Susskind, B. Arden, L. K. John, P. Stockton, and E. B. John, "Neuro-symbolic ai: An emerging class of ai workloads and their characterization," 2021.

[3] Z. Wang, S.-B. Ho, and E. Cambria, "Multi-level fine-scaled sentiment sensing with ambivalence handling," *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, vol. 28, pp. 683–697, 2020.

[4] Ł. Augustyniak, T. Kajdanowicz, P. Kazienko, M. Kulisiewicz, and W. Tuligłowicz, "An approach to sentiment analysis of movie reviews: Lexicon based vs. classification," in *HAIS'14*. Springer, 2014.

[5] A. Kotelnikova, D. Paschenko, K. Bochenina, and E. Kotelnikov, “Lexicon-based methods vs. bert for text sentiment analysis,” 2021.

[6] J. Joseph, S. Vineetha, and N. Sobhana, “A survey on deep learning based sentiment analysis,” *Materials Today: Proceedings*, vol. 58, pp. 456–460, 2022, international Conference on Artificial Intelligence& Energy Systems.

[7] G. A. Miller, “Wordnet: A lexical database for english,” *Commun. ACM*, vol. 38, no. 11, p. 39–41, nov 1995.

[8] S. Baccianella, A. Esuli, and F. Sebastiani, “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” in *LREC*, N. Calzolari et al., Ed. European Language Resources Association, 2010.

[9] A. Janz, J. Kocoń, M. Piasecki, and M. Zaśko-Zielińska, “plwordnet as a basis for large emotive lexicons of polish,” 11 2017.

[10] K. Hamilton, A. Nayak, B. Božić, and L. Longo, “Is neuro-symbolic ai meeting its promise in natural language processing? a structured review,” 2022.

[11] Z. Liu, Z. Wang, Y. Lin, and H. Li, “A neural-symbolic approach to natural language understanding,” 2022.

[12] M. Piasecki, S. Szpakowicz, and B. Broda, *A Wordnet from the Ground Up*. Wroclaw: Oficyna Wydawnicza Politechniki Wroclawskiej, 2009.

[13] A. Janz, J. Kocoń, M. Piasecki, and M. Zaśko-Zielińska, “plwordnet as a basis for large emotive lexicons of polish,” 11 2017.

[14] M. Zaśko-Zielińska and M. Piasecki, “Towards emotive annotation in plWordNet 4.0,” in *Proceedings of the 9th Global Wordnet Conference*. Nanyang Technological University (NTU), Singapore: Global Wordnet Association, Jan. 2018, pp. 153–162.

[15] P. Ekman, “An argument for basic emotions,” *Cognition & Emotion*, vol. 6, pp. 169–200, 1992.

[16] S. M. Mohammad and P. D. Turney, “Crowdsourcing a word-emotion association lexicon,” *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[17] S. Kiritchenko, X. Zhu, and S. M. Mohammad, “Sentiment analysis of short informal texts,” vol. 50, no. 1, p. 723–762, may 2014.

[18] S. M. Mohammad, S. Kiritchenko, and X. Zhu, “Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets,” in *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June 2013.

[19] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, “SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 3829–3839. [Online]. Available: https://aclanthology.org/2022.lrec-1.408

[20] P. D. Turney, “Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews,” p. 417, 2001.

[21] M. Hu and B. Liu, “Mining and summarizing customer reviews,” *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177, 2004.

[22] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, “Sentiment analysis of Twitter data,” in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. Portland, Oregon: Association for Computational Linguistics, Jun. 2011, pp. 30–38.

[23] C. M. Whissell, “Chapter 5 - the dictionary of affect in language,” in *The Measurement of Emotions*, R. Plutchik and H. Kellerman, Eds. Academic Press, 1989, pp. 113–131. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780125587044500116

[24] P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson, “Semeval-2013 task 2: Sentiment analysis in twitter,” 2019.

[25] Z. Teng, D.-T. Vo, and Y. Zhang, “Context-sensitive lexicon features for neural sentiment analysis,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1629–1638.

[26] Q. Qian, M. Huang, J. Lei, and X. Zhu, “Linguistically regularized LSTM for sentiment classification,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1679–1689.

[27] A. Koufakou, E. W. Pamungkas, V. Basile, and V. Patti, “HurtBERT: Incorporating lexical features with BERT for the detection of abusive language,” in *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, Nov. 2020, pp. 34–43.

[28] 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.

[29] P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang, “SentiLARE: Sentiment-aware language representation learning with linguistic knowledge,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 6975–6988.

[30] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang, “Kepler: A unified model for knowledge embedding and pre-trained language representation,” 2019.

[31] 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: Association for Computational Linguistics, Jun. 2011, pp. 142–150.

[32] B. Pang and L. Lee, “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales,” in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 115–124.

[33] J. Kocoń, P. Miłkowski, and K. Kanclerz, “MultiEmo: Multilingual, multilevel, multidomain sentiment analysis corpus of consumer reviews,” 2021, CLARIN-PL digital repository.

[34] J. Kocoń, M. Zaśko-Zielińska, and P. Milkowski, “Polemo 2.0 sentiment analysis dataset for conll,” 2019.

[35] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, “GoEmotions: A dataset of fine-grained emotions,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4040–4054.

[36] P. Rybak, R. Mroczkowski, J. Tracz, and I. Gawlik, “Klej: Comprehensive benchmark for polish language understanding,” 2020.

[37] R. Orlando, S. Conia, F. Brignone, F. Cecconi, and R. Navigli, “AMuSE-WSD: An all-in-one multilingual system for easy Word Sense Disambiguation,” in *Proceedings of EMNLP*, Punta Cana, Dominican Republic, nov 2021.

[38] D. Vrandečić and M. Krötzsch, “Wikidata: A free collaborative knowledgebase,” *Commun. ACM*, vol. 57, no. 10, p. 78–85, sep 2014.

[39] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, “Translating embeddings for modeling multi-relational data,” in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013.

[40] K. Toutanova, D. Klein, C. Manning, and Y. Singer, “Feature-rich part-of-speech tagging with a cyclic dependency network,” *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology—NAACL ’03*, vol. 1, 03 2004.

[41] P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang, “Sentilare,” https://github.com/thu-coai/SentiLARE, 2021.

[42] R. Mroczkowski, P. Rybak, A. Wróblewska, and I. Gawlik, “Herbert: Efficiently pretrained transformer-based language model for polish,” 2021.

[43] Student, “The probable error of a mean,” *Biometrika*, pp. 1–25, 1908.

[44] S. S. SHAPIRO and M. B. WILK, “An analysis of variance test for normality (complete samples),” *Biometrika*, vol. 52, no. 3-4, pp. 591–611, dec 1965.

[45] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi, “Dataset cartography: Mapping and diagnosing datasets with training dynamics,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 9275–9293.

[46] T. Song, J. Luo, and L. Huang, “Rot-pro: Modeling transitivity by projection in knowledge graph embedding,” 2021.