Can BERT eat RuCoLA? Topological Data Analysis to Explain

Irina Proskurina[1], Irina Piontkovskaya[2], Ekaterina Artemova[3]
[1]Université de Lyon, Lyon 2, ERIC UR 3083, France
[2]Huawei Noah's Ark lab
[3]Center for Information and Language Processing (CIS), LMU Munich, Germany
Correspondence: Irina.Proskurina@univ-lyon2.fr

Abstract

This paper investigates how Transformer language models (LMs) fine-tuned for acceptability classification capture linguistic features. Our approach uses the best practices of topological data analysis (TDA) in NLP: we construct directed attention graphs from attention matrices, derive topological features from them, and feed them to linear classifiers. We introduce two novel features, chordality, and the matching number, and show that TDA-based classifiers outperform fine-tuning baselines. We experiment with two datasets, CoLA and RuCoLA,[1] in English and Russian, typologically different languages.

On top of that, we propose several black-box introspection techniques aimed at detecting changes in the attention mode of the LMs during fine-tuning, defining the LM's prediction confidences, and associating individual heads with fine-grained grammar phenomena.

Our results contribute to understanding the behavior of monolingual LMs in the acceptability classification task, provide insights into the functional roles of attention heads, and highlight the advantages of TDA-based approaches for analyzing LMs. We release the code and the experimental results for further uptake.[2]

1 Introduction

Language modelling with Transformer (Vaswani et al., 2017) has become a standard approach to acceptability judgements, providing results on par with the human baseline (Warstadt et al., 2019). The pre-trained encoders and BERT, in particular, were proven to have an advantage over other models, especially when judging the acceptability of sentences with long-distance dependencies (Warstadt and Bowman, 2019). Research examining linguistic knowledge of BERT-based language models (LMs) revealed that: (1) individual attention heads can store syntax, semantics or both kinds of linguistic information (Jo and Myaeng, 2020; Clark et al., 2019), (2) vertical, diagonal and block attention patterns could frequently repeat across the layers (Kovaleva et al., 2019), and (3) fine-tuning affects the linguistic features encoding tending to lose some of the pre-trained model knowledge (Miaschi et al., 2020). However, less attention has been paid to examining the grammatical knowledge of LMs in languages other than English. The existing work devoted to the cross-lingual probing showed that grammatical knowledge of Transformer LMs is adapted to the downstream language; in the case of Russian, the interpretation of results cannot be easily explained (Ravishankar et al., 2019). However, LMs are more insensitive towards granular perturbations when processing texts in languages with free word order, such as Russian (Taktasheva et al., 2021).

In this paper, we probe the linguistic features captured by the Transformer LMs, fine-tuned for acceptability classification in Russian. Following recent advances in acceptability classification, we use the Russian corpus of linguistic acceptability (RuCoLA) (Mikhailov et al., 2022), covering tense and word order violations, errors in the construction of subordinate clauses and indefinite pronoun usage, and other related grammatical phenomena. We provide an example of an unacceptable sentence from RuCoLA with a morphological violation in the pronoun usage: a possessive reflexive pronoun 'svoj' (oneself's/own) instead of the 3rd person pronoun.

(1) * Eto byl pervyj chempionat mira v **svoej** kar'ere. ("It was the first world championship in **own** career.")

Following the recently proposed Topological Data Analysis (TDA) based approach to the linguistic acceptability (LA) task (Cherniavskii et al., 2022),

[1]Arugula or rocket salad in English
[2]https://github.com/upunaprosk/la-tda

Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023), pages 123–137

we construct directed attention graphs from attention matrices and then refer to the characteristics of the graphs as to the linguistic features learnt by the model. We extend the existing research on the acceptability classification task to the Russian language and show the advantages of the TDA-based approach to the task. Our main contributions are the following: *(i)* we investigate the monolingual behaviour of LMs in acceptability classification tasks in the Russian and English languages, using a TDA-based approach, *(ii)* we introduce new topological features and outperform previously established baselines, *(iii)* we suggest a new TDA-based approach for measuring the distance between pre-trained and fine-tuned LMs with large and base configurations. *(iv)* We determine the roles of attention heads in the context of LA tasks in Russian and English.

Our initial hypothesis is that there is a difference in the structure of attention graphs between the languages, especially for the sentences with morphological, syntactic, and semantic violations. We analyze the relationship between models by comparing the features of the attention graphs. To the best of our knowledge, our research is one of the first attempts to analyse the differences in monolingual LMs fine-tuned on acceptability classification corpora in Russian and English, using the TDA-based approach.

2 Related Work

Acceptability Classification. First studies performed acceptability classification with statistical machine learning methods, rule-based systems, and context-free grammars (Cherry and Quirk, 2008; Wagner et al., 2009; Post, 2011). Alternative approaches use threshold scoring functions to estimate the likelihood of a sentence (Lau et al., 2020). Recent research has been centered on the ability of omnipresent Transformer LMs to judge acceptability (Wang et al., 2018), to probe for their grammar acquisition (Zhang et al., 2021), and evaluate semantic correctness in language generation (Batra et al., 2021). In this project, we develop acceptability classification methods and apply them to datasets in two different languages, English and Russian.

Topological Data Analysis (TDA) in NLP. Recent work uses TDA to explore the inner workings of LMs. Kushnareva et al. (2021) derive TDA features from attention maps to build artificial text detection. Colombo et al. (2021) introduce BARYSCORE, an automatic evaluation metric for text generation that relies on Wasserstein distance and barycenters. Chauhan and Kaul (2022) develop a scoring function which captures the homology of the high-dimensional hidden representations, and is aimed at test accuracy prediction. We extend the set of persistent features proposed by Cherniavskii et al. (2022) for acceptability classification and conduct an extensive analysis of how the persistent features contribute to the classifier's performance.

How do LMs change via fine-tuning? There have been two streams of studies of how fine-tuning affects the inner working of LM's: (i) what do sub-word representation capture and (ii) what are the functional roles of attention heads? The experimental techniques include similarity analysis between the weights of source and fine-tuned checkpoints (Clark et al., 2019), training probing classifiers (Durrani et al., 2021), computing feature importance scores (Atanasova et al., 2020), the dimensionality reduction of sub-word representations (Alammar, 2021). Findings help to improve fine-tuning procedures by modifying loss functions (Elazar et al., 2021) and provide techniques for explaining LMs' predictions (Danilevsky et al., 2020). Our approach reveals the linguistic competence of attention heads by associating head-specific persistent features with fine-grained linguistic phenomena.

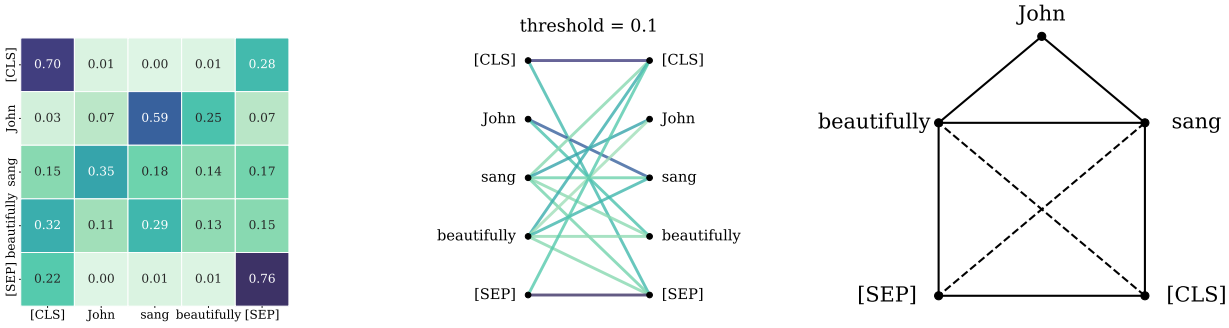
3 Methodology

We follow Warstadt et al., 2019 and treat the LA task as a supervised classification problem. We fine-tune Transformer LMs to approximate the function that maps an input sentence to a target class: acceptable or unacceptable.

3.1 Extracted Features

Given an input text, we extract output attention matrices from Transformer LMs and follow Kushnareva et al., 2021 to compute three types of persistent features over them.

Topological features are properties of attention graphs. We provide an example of an attention graph constructed upon the attention matrix in Figure 1. An adjacency matrix of attention graph $A = (a_{ij})_{n \times n}$ is obtained from the attention matrix



(a) (b) (c)

Figure 1: An example of an attention map (a) and the corresponding bipartite (b) and attention (c) graphs for the CoLA sentence *"John sang beautifully"*. The graphs are constructed with a threshold equal to 0.1.

$W = (w_{ij})_{n \times n}$, using a pre-defined threshold thr:

$$a_{ij} = \begin{cases} 1 & \text{if } w_{ij} \geq thr \\ 0 & \text{otherwise,} \end{cases}$$

where w_{ij} is an attention weight between tokens i and j and n is the number of tokens in the input sequence. Each token corresponds to a graph node. Features of directed attention graphs include the number of strongly connected components, edges, simple cycles and average vertex degree. The properties of undirected graphs include the first two Betti numbers: the number of connected components and the number of simple cycles. We propose two new features of the undirected attention graphs: the matching number and the chordality. The matching number is the maximum matching size in the graph, i.e. the largest possible set of edges with no common nodes.

Consider an attention matrix depicted in Figure 1a and a simple undirected attention graph (Figure 1c) constructed based on the bipartite graph (Figure 1b) with a threshold of 0.1. The matching number of that attention graph is equal to two. One example of a maximum matching in that graph is a set of edges: {(*John* - *sang*), ([SEP] - [CLS])}. That matching is maximum because there are no more edges that are not incident to the already matched 4 nodes (tokens). The chordality is a binary feature showing whether the attention graph is chordal; that is, whether the attention graph does not contain induced cycles of a length greater than 3. For example, the plotted graph in Figure 1c is chordal because it does not contain induced cycles with more than 3 edges. If there were no dotted edges (chords) in the graph, there would be a cycle [SEP]-*beautifully*-*sang*-[CLS]-[SEP] of length 4, meaning that the graph is not chordal.

We expect these novel features to express syntax phenomena of the input text. The chordality feature could carry information about subject-verb-object triplets. The maximum matching can correspond to matching sentence segments (subordinate clauses, adverbials, participles, introductory phrases, etc.).

Features derived from barcodes include descriptive characteristics of $0/1$-dimensional barcodes and reflect the survival (death and birth) of connected components and edges throughout the filtration.

Distance-to-pattern features measure the distance between attention matrices and identity matrices of pre-defined attention patterns, such as attention to the first token [CLS] and to the last [SEP] of the sequence, attention to previous and next token and to punctuation marks (Clark et al., 2019). We use a publicly available implementation to compute features.[3]

3.2 Experimental Framework

Data We use two publicly available LA benchmarks in two typologically different languages: Russian (RuCoLA; Mikhailov et al., 2022) and English (CoLA; Warstadt et al., 2019). Both selected corpora consist of in- and out-of-domain data and contain sentences collected from linguistics publications; each is marked as acceptable or unacceptable. Unacceptable sentences are annotated with syntactic, morphological and semantic phenomena violated in them. RuCoLA, in addition, covers synthetically generated data by generative LMs. We provide examples of acceptable sentences from observed corpora (2a, 3a) along

[3] https://github.com/danchern97/tda4atd

with sentences with semantic violations (2b, 3b).

(2) a. The dog bit the cat.
b. * The **soundly and furry** cat slept.

(3) a. Koshki byli svyashchennymi zhivotnymi v Drevnem Egipte. ("Cats were sacred animals in ancient Egypt.")
b. * **Bliz** kresla na nebol'shom kovrike lezhala koshka. ("**Outside of** an armchair on a small rug a cat was lying.")

Table 4 (Appendix A) reports statistics of the used corpora. For per-category evaluation, we use RuCoLA error annotations, and for CoLA, we use minor grammatical phenomena annotations to group erroneous sentences. We provide more details in Table 5 (Appendix A).

Models Our baseline model architectures, fine-tuning and evaluation scripts are taken from the Transformers library (Wolf et al., 2020). We use the following case-sensitive monolingual Transformer LMs for the experiments: (1) base size En-BERT[4] (Devlin et al., 2019) and Ru-BERT,[5] (2) large size En-RoBERTa[6] (Liu et al., 2019) and Ru-RoBERTa.[7] To estimate the effect of fine-tuning, we compare two types of models: pre-trained LMs with frozen weights (frozen) and fine-tuned LMs on the training sets. Transformer LMs are fine-tuned for 5 epochs on in-domain training data, with a batch size of 32 and an optimal set of hyper-parameters determined by the authors of the datasets. To mitigate class imbalance, we use weighted cross-entropy loss. We provide fine-tuning details in Table 6 (Appendix A).

TDA Classifiers We extract a range of persistent (TDA) features listed in Section 3.1 from Transformer LMs and refer to them as training features fed to a linear classifier. We reduce the feature space dimensionality with principal component analysis (PCA). Next, we train Logistic Regression classifiers with adjusted class weights on the reduced feature space. We iterate over a range of inverse regularization parameter values $C \in \{10^{-3}, 10^{-2}, 0.1\}$ and the number of principal components $\#PC \in [10, 20 \dots 200]$. We choose the value 200 as the upper bound of the PC grid to ensure that the number of latent features is at least two times less than the size of the in-domain development (IDD) or out-of-domain development (OODD) sets. We tune hyper-parameters to maximize the classifier performance on the IDD set. We compare the performance of two feature sets, by reporting results of classifiers trained on (i) basic TDA features by Kushnareva et al., 2021 (dubbed as TDA) and (ii) TDA features with two novel features added (dubbed as TDA_{ext}).

3.3 Evaluation

Performance Metrics Following Warstadt et al., 2019, we measure performance with Accuracy (Acc.) and Matthews Correlation Coefficient (MCC). MCC is used as the main performance metric for finding hyperparameters, evaluating trained models, and adjusting the decision threshold.

Fine-tuning Effect We estimate changes in attention weights between pre-trained and fine-tuned LMs with two methods. First, we follow Hao et al., 2020 and employ Jensen-Shannon (JS) divergence:

$$D_{JS}(M_t||M_0) = \frac{1}{N}\frac{1}{H}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{W}\sum_{i=1}^{K} D_{JS}(W_t^h(token_i)||W_0^h(token_i))$$

where M_t and M_0 are fine-tuned and frozen models respectively, N is number of sentences, H is a number of attention heads ($H = 12$ for base-configuration LMs, $H = 24$ for large LMs), K is the number of tokens in the sentence n, and $W_t^h(token_i)$ is an attention weight of attention head h at token i in model M_t.

Second, we estimate the difference between attention graphs as an average correlation distance between the TDA_{ext} features across attention heads:

$$D_{TDA}(M_t, M_0) = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{F}\sum_{f=1}^{F} D_{corr}(V_{tf}^h, V_{0f}^h)$$

where F is the number of features, V_{tf}^h are values of the feature f, computed over attention matrix W_t^h, extracted from the model M_t.

4 Results

4.1 Acceptability Classification

Table 1 reports LA classification results. Linear classifiers trained on the TDA features boost Transformer LMs performance; that trend is consistent across all models, with the MCC score gain of

[4] hf.co/bert-base-cased
[5] hf.co/sberbank-ai/ruBert-base
[6] hf.co/roberta-large
[7] hf.co/sberbank-ai/ruRoberta-Large

Model	Fine-tuned LMs				Frozen LMs			
	IDD		OODD		IDD		OODD	
	Acc.	MCC	Acc.	MCC	Acc.	MCC	Acc.	MCC
RuCoLA								
Ru-BERT	80.3	0.420	75.1	0.438	62.4	0.079	54.7	0.112
+ TDA	80.1	0.440	75.1	0.447	76.5	0.314	62.3	0.253
+ TDA$_{ext}$	80.1	0.478	73.2	0.440	76.7	0.331	62.6	0.270
Ru-RoBERTa	83.5	0.530	79.3	0.530	72.8	0.313	58.1	0.241
+ TDA	85.0	0.581	**81.0**	**0.584**	77.0	0.374	**64.7**	0.343
+ TDA$_{ext}$	**85.7**	**0.594**	80.1	0.558	**77.2**	**0.391**	64.2	**0.358**
CoLA								
En-BERT	85.0	0.634	82.0	0.561	62.6	0.039	64.3	0.124
+ TDA	85.6	0.649	81.4	0.548	77.0	0.484	68.4	0.335
+ TDA$_{ext}$	**88.2**	**0.726**	81.0	0.556	81.4	0.543	73.1	0.369
En-RoBERTa	87.3	0.692	**84.9**	**0.637**	74.0	0.317	75.0	0.362
+ TDA	86.3	0.680	83.5	0.620	81.2	0.543	**78.5**	0.464
+ TDA$_{ext}$	87.3	0.695	83.1	0.604	**83.1**	**0.604**	77.3	**0.476**

Table 1: Acceptability classification results of monolingual LMs and linear classifiers trained on the sets of features by the benchmark. **IDD**=in domain development set. **OODD**=out of domain development set. TDA$_{ext}$=TDA features+chordality and the matching number. The best score is in bold, and the second-best one is underlined.

+0.252 at most for the Russian LMs and a more substantial +0.504 increase falling on En-BERT. Proposed chordality and matching number features are beneficial and help improve performance, proving that they capture linguistic information.

Unlike base LMs, large frozen LMs exhibit grammatical knowledge even before fine-tuning. Base LMs' MCC scores fluctuate around zero, while large LMs achieve at least 0.3 MCC.

That observation aligns with the recent works showing that pre-trained large En-RoBERTa can achieve competitive scores without further fine-tuning in tasks such as lexical complexity prediction (Rao et al., 2021).

At the same time, TDA classifiers outperform fine-tuned models by a minor margin enhancing scores by at best +0.064 MCC for Russian and +0.092 MCC for English. We believe that fine-tuning may cause the LM to lose general grammatical skills and forget language phenomena that are not present in the fine-tuning set (Miaschi et al., 2020). Thus, the features extracted from the fine-tuned models may require a thorough feature selection with non-linear models to mitigate feature redundancy issues. TDA classifies for RuCoLA achieve scores on par with the baseline LMs. However, for CoLA, the TDA$_{ext}$ classifier coupled with En-RoBERTa outperforms the baseline. We report classification results on OOD test data in Table 7 and Table 8, Appendix B.1.

4.2 Sensitivity to Violation Categories

Next, we analyze gains in recall by TDA classifiers with respect to violation category. Table 2 reports scores of Ru-BERT and En-BERT baselines and TDA classifiers averaged between IDD and OODD sets with respect to 5 grammatical violations. TDA classifiers outperform LMs in unacceptable sentences; that uptrend holds for both languages, while there is a drop for acceptable sentences.

In contrast to English, the TDA$_{ext}$ classifier trained on Ru-BERT features is more sensitive to syntactic violations reaching the overall 76.6 recall; that is, the increase in the score is around 20 recall points, compared to fine-tuned Ru-BERT. As for the rest grammar categories, the TDA$_{ext}$ classifier outperforms the fine-tuned Ru-BERT by a large margin, especially in sentences with word-level morphological violations, where the recall of Ru-BERT is more than doubled.

Next, we manually analyze the errors of the fine-tuned Ru-BERT and our classifier TDA$_{ext}$ in OODD sentences in Russian. First, we compare the unacceptable sentences, which are misclassified by Ru-BERT but correctly classified by the TDA$_{ext}$ classifier. We find that the error span in OODD sentences is relatively short, with at most three tokens. In particular, in these sentences, such violations as non-existing words are most often encountered, the misuse of which is quite common among native speakers (4a, word formation error 'ekhaj'), local inverse word order (4b), or nonsense (4c). Common false predictions of both models include long sentences that mix grammatical phenomena, contain long-distance agreement violations and complex errors in punctuation.

(4) a. * A ty **ekhaj** pryamo k direktoru teatrov. ("**You should gotta to** the director of theatres.")

b. * V etom lesu **vodyatsya volki**. ("There are **in this forest wolves**.")

c. * Oni chitali moi zhaloby **na sebya**. ("They read my complaints **onto themselves**.")

The domain shift from ID to OOD introduces new types of unacceptable phenomena are not present in ID data. Overall, the scores for OOD data are lower than for ID data (Table 2, Table 9, Appendix B.1). Hence LMs do not generalize well to unseen unacceptable phenomena and have little knowledge about the unseen linguistic properties.

Model	Acceptable	Hallucination	Morphology	Semantics	Syntax
Ru-BERT	92.1	53.9	20.0	25.0	55.7
+TDA$_{ext}$	80.6	73.9	53.9	46.6	76.6
En-BERT	94.3	68.5	69.4	63.0	55.6
+TDA$_{ext}$	84.5	78.8	82.5	76.3	73.0

Table 2: Overall per-category recall by the benchmark.

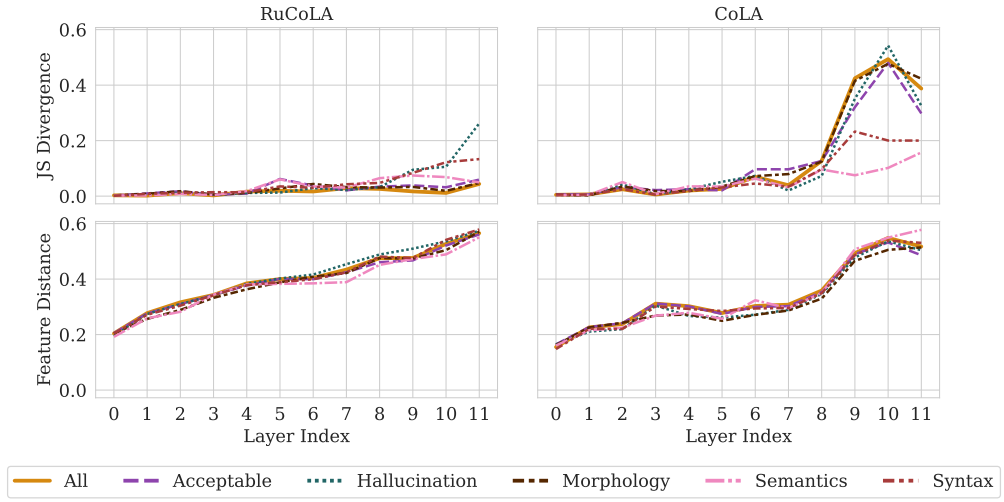


Figure 2: Per-layer feature distance and JS divergence of attention scores between the frozen and fine-tuned Ru-BERT and En-BERT.

4.3 Fine-tuning Effect

We investigate the dynamics of LM fine-tuning and measure per layer distance between TDA$_{ext}$ features extracted from frozen and fine-tuned LMs on OODD subsets (§3.3). Figure 2 illustrates layer-wise feature distance and JS divergence for Ru-BERT and En-BERT (Figure 3, Appendix B.2 for large models). Overall, we find that the distance between features rises steadily from the bottom to higher layers, whilst for English LMs, the most noticeable changes occur only in the last four layers. That observation implies that there is a noticeable difference in fine-tuning dynamics between En-BERT and Ru-BERT.

For both languages, the feature distance trend differs from JS divergence, especially in the first six layers. This indicates that the TDA$_{ext}$ features can be used to detect minor changes in the lower layers that are poorly expressed when using the JS divergence. For example, TDA-based distance is sensitive to small changes in the attention weights at lower predefined thresholds where large attention weights remain unchanged. JS divergence is not capable of capturing such cases.

The distance between features is uniform with respect to the violation category. The trends for acceptable and unacceptable sentences almost coincide, albeit there are noticeable differences in JS divergence. For Russian models, JS divergence in sentences with syntactic violations and hallucinations is more evident in higher layers compared to other categories. In turn, the JS divergence for English shows that the attention mode is more consistent with the frozen En-BERT on the sentences with semantic and syntactic violations; for acceptable and other sentences, the peak is reached at the penultimate layer. Similar to LMs with the base configuration, there is a steady increase in feature dissimilarity across all the layers, while for English, the main changes appear in higher layers.

4.4 Head Importance

We probe linguistic phenomena with the help of persistent features: we exploit the learnt feature weights in the linear classifiers (Appendix B.3). The higher the weight of the feature, the more it contributes to the final prediction. We aggregate

Error type	Sentence	Feature	Head
Morphology	Recept **chipy** s syrom, **maniokom** i yajcami. ("Recipe of **cheps** with cheese, **maniokom** and eggs.")	$c_{thr0.25}$	(9,5)
Syntax	Bylo nachato **stroit' novyj rajon**. ("**Of new district building** was started.")	$c_{thr0.1}$	(9,5)
Semantics	Vchera v dva chasa magazin **zakryt**. ("The store **closing** at two o'clock yesterday.")	[CLS]	(11,0)

Table 3: Examples of the most important Ru-BERT TDA$_{ext}$ features for judging RuCoLA unacceptable sentences by error type. c = the number of simple cycles in a graph, thr = threshold used for constructing attention graph, [CLS] = distance-to-[CLS]-token.

features derived from each head: the importance of the head is derived as a number of important features. We define two types of heads: (1) heads that contribute the most to true positive and true negative predictions (i.e. correct predictions), dubbed as agreeing heads, and (2) heads that contribute the most to false negative and false positive predictions (i.e. classifier's errors), dubbed as disagreeing heads. First, we explore the importance of each individual head. Figure 4, Appendix B.4 shows how important the head is for the final prediction. En-BERT and Ru-BERT have similar patterns for the heads of type (1) as the most useful features for Ru-BERT are housed in middle to higher layers. For En-BERT, these tend to be localized mostly in the last two layers.

Next, we compute the feature importance with respect to the violation category. Heads of middle layers contribute more to detecting syntactic and morphology violations in English and Russian. Heads of type (2) do not overlap with the heads of type (1) with a few exceptions, which are head 10 and head 0 from the last layer of Ru-BERT and En-BERT, respectively. Judging by the number of type (2) heads Ru-BERT struggles the most to distinguish sentences with hallucinations from acceptable sentences. This might be due to multiple reasons: (i) hallucinated sentences are not seen during training, (ii) hallucinated sentences are mainly well-formed but semantically incorrect, so there are no surface or syntactical clues to rely on.

Next, we determine the set of sentences that are the most challenging for the TDA classifier and, thus, the corresponding LM since TDA features are extracted from its attention map. To do so, we define the LM's confidence as the sum of absolute feature weights for predicting acceptable and unacceptable classes. The lower the score, the more confused the LM is and the more attention heads tend to disagree with the desired prediction. We consider those sentences challenging that obtain the lowest confidence scores. The most challenging sentences are long, consist of multiple clauses and contain terms or named entities, see the unacceptable sentence in 5 for example. For the sake of completeness, we conduct the same analysis for CoLA sentences and provide an example of the most confusing sentence for TDA$_{ext}$ classifier (6). The results align well. The most challenging sentences contain long-distance dependencies and named entities.

(5) * Eta gruppa obnaruzhila **(nepravil'no) chto severnyj predel** Merrimak byl bliz togo, chto teper' izvestno kak ozero **Vinnipesuki** v **N'yu-Gempshire**.
("This group found **(poorly), that the northern watershed** of the Merrimack was near what is now known as Lake **Vinnipesaukee** in **New Gampshire**.")

(6) * Gould's performance of Bach on the piano doesn't please me anywhere as much as **Ross's on the harpsichord**.

Finally, we explore the feature contribution on the sentence level. Our TDA-based approach allows explaining predictions for every single sentence. To this end, the contribution (=importance) of each feature is the feature value multiplied by the learnt weight of the linear classifier. We observe the following patterns across unacceptable sentences in Russian and Ru-BERT:

1. Distance-to-pattern features appear to be useful for classifying unacceptable sentences

with word-level violations, including spelling, punctuation, and agreement errors;

2. Topological features and features derived from barcodes contribute equally to more complicated grammatical phenomena.

Table 3 provides examples of unacceptable sentences along with the feature importance values. Chordality, the matching number, the number of simple cycles, and the average vertex degree derived at thresholds 0.1 or 0.25 frequently become important to predict unacceptable sentences in Russian. Similarly, the average number of vertex degrees has the most discriminative power for English and En-BERT. Important features are housed across different layers in the LMs. For English, the most important features are extracted from the last layer, while for Russian, they appear at the earliest at layer 6.

However, when it comes to the discrepancy in attention graphs between acceptable and unacceptable sentences, we find the following common for both languages. The number of connected components in attention graphs for unacceptable sentences is larger at the lowest and the highest thresholds. At the highest threshold, these components consist of one token; at the lowest one, they consist of a few ones. It means that the values of attention maps in unacceptable sentences do not deviate much from each other. On the contrary, for acceptable sentences, there is a tendency to put the most attention weight on a single token, which is usually the sentence's head verb. In terms of the TDA feature values, this effect can be seen as the sign of the correlation coefficient between the feature value and the target class correlation. Thus, there is an obvious shift towards positive correlation at a threshold of 0.5 for average vertex degree features (Figure 5).

To sum up, such an analysis helps better explain the classifiers' prediction. Since persistent features are attributed to individual heads, we can trace the role and importance of each head. A fine-grained annotation of language phenomena allows us to associate specific linguistic skills with individual heads.

5 Conclusion

In this paper, we adopt and improve methods for acceptability classification by using best practices from topological data analysis (TDA). We showcase the developed methods in two typologically different languages by using the datasets in English and Russian, CoLA and RuCoLA, respectively. In particular, we introduce two novel features, chordality and the matching number, and compare the performance of TDA-based classifiers to fine-tuning. TDA-based classifiers boost the performance of pre-trained language models.

TDA-based classifiers have advantages over LM fine-tuning because they are more interpretable and help to introspect the inner workings of LMs. To this end, we introduce a TDA feature-based distance measure to detect changes in the attention mode of LMs during fine-tuning. This distance measure is sensitive even to small changes occurring at the bottom layers of LMs that are not detected by the widespread Jensen-Shannon divergence. What is more important, we show how TDA features reveal the functional roles of attention heads. We compare heads that contribute to making correct and incorrect predictions based on their importance. This way we discover heads that store information about word order, word derivation, and complex semantic phenomena in unacceptable sentences and heads that attend to acceptable sentences.

Given the sentence, we evaluate the prediction confidence based on the contribution of the features. We determine the set of sentences in which LMs are less confident and find that those sentences usually consist of multiple clauses and frequently include named entities. Finally, we find a distinct pattern that is frequently present in the attention maps of unacceptable sentences in English and Russian.

We hope that our results shed light on the performance of LMs in Russian and English and help understanding their fine-tuning dynamics and the functional roles of attention heads. We are excited to see the adoption of TDA by NLP practitioners to other languages and downstream problems.

Limitations

Acceptability judgments datasets Acceptability judgments datasets use linguistic literature as source of unacceptable sentences. Such approach is subject to criticism on two counts: (i) the reliability and reproducibility of acceptability judgments (Gibson and Fedorenko, 2013; Culicover and Jackendoff, 2010; Sprouse and Almeida, 2013; Linzen and Oseki, 2018), (ii) representativeness, as linguists' judgments may not reflect the errors that

speakers tend to produce (Dąbrowska, 2010).

Computational complexity The computation complexity of the proposed features is linear. For chordality features, we rely on the implementation of linear $O(|E| + |V|)$ time algorithm (Tarjan and Yannakakis, 1984), where $|E|$ and $|V|$ are the numbers of edges and nodes, respectively. We use a greedy algorithm with linear complexity $O(|E|)$ to find the maximum matching. When calculating simple cycles with the exponential-time algorithm (in the worst case), we use a constraint equal to 500 to do an early stopping. We suggest that simple cycles features are less informative when that value is exceeded. Kushnareva et al., 2021 discuss the time complexity of the rest features.

Acknowledgements

We thank Laida Kushanareva, Daniil Cherniavskii, Vladislav Mikhailov, Serguei Barannikov, Alexander Bernstein, and Dmitri Piontkovski for their comments at its early stages, and we thank Max Ryabinin for providing scripts to process the RuCoLA dataset.

References

J Alammar. 2021. Ecco: An open source library for the explainability of transformer language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 249–257, Online. Association for Computational Linguistics.

Sam Bowman Alex Warstadt, Amanpreet Singh. 2018. Cola out-of-domain open evaluation.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Soumya Batra, Shashank Jain, Peyman Heidari, Ankit Arun, Catharine Youngs, Xintong Li, Pinar Donmez, Shawn Mei, Shiunzu Kuo, Vikas Bhardwaj, Anuj Kumar, and Michael White. 2021. Building adaptive acceptability classifiers for neural NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 682–697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jatin Chauhan and Manohar Kaul. 2022. BERTops: Studying BERT Representations under a Topological Lens. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2022. Acceptability judgements via examining the topology of attention maps. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 88–107, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Colin Cherry and Chris Quirk. 2008. Discriminative, syntactic language modeling through latent SVMs. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Research Papers*, pages 65–74, Waikiki, USA. Association for Machine Translation in the Americas.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021. Automatic text evaluation through the lens of Wasserstein barycenters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10466, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter W Culicover and Ray Jackendoff. 2010. Quantitative methods alone are not enough: Response to gibson and fedorenko. *Trends in Cognitive Sciences*, 6(14):234–235.

Ewa Dąbrowska. 2010. Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep NLP models? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4947–4957, Online. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Edward Gibson and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1-2):88–124.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2020. Investigating learning dynamics of BERT fine-tuning. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 87–92, Suzhou, China. Association for Computational Linguistics.

Jae-young Jo and Sung-Hyon Myaeng. 2020. Roles and utilization of attention heads in transformer-based neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404–3417, Online. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2021. Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 635–649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Tal Linzen and Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics*, 3(1).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matt Post. 2011. Judging grammaticality with tree substitution grammar derivations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 217–222, Portland, Oregon, USA. Association for Computational Linguistics.

Gang Rao, Maochang Li, Xiaolong Hou, Lianxin Jiang, Yang Mo, and Jianping Shen. 2021. RG PA at SemEval-2021 task 1: A contextual attention-based model with RoBERTa for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 623–626, Online. Association for Computational Linguistics.

Vinit Ravishankar, Memduh Gökırmak, Lilja Øvrelid, and Erik Velldal. 2019. Multilingual probing of deep pre-trained contextual encoders. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 37–47, Turku, Finland. Linköping University Electronic Press.

Jon Sprouse and Diogo Almeida. 2013. The empirical status of data in syntax: A reply to gibson and fedorenko. *Language and Cognitive Processes*, 28(3):222–228.

Ekaterina Taktasheva, Vladislav Mikhailov, and Ekaterina Artemova. 2021. Shaking syntactic trees on the sesame street: Multilingual probing with controllable perturbations. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 191–210, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert E Tarjan and Mihalis Yannakakis. 1984. Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM Journal on computing*, 13(3):566–579.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Joachim Wagner, Jennifer Foster, and Josef van Genabith. 2009. Judging Grammaticality: Experiments in Sentence Classification. *Calico Journal*, 26(3):474–490.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. 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*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt and Samuel R. Bowman. 2019. Linguistic analysis of pretrained sentence encoders with acceptability judgments. *arXiv: Computation and Language*.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

A Experiment Setup

	CoLA	RuCoLA
Language	English	Russian
Data type	Real	Real, Synthetic
α	0.86	0.89
# Train sent.	8551	7869
# Dev sent.	527	983
# Test sent.	516	1804
%	70.5	71.8

Table 4: Statistics of language acceptability corpora. α = Average annotator agreement rate. % = Percentage of acceptable sentences.

Grammatical feature	Error type
Extra/Missing Word	Hallucination
Semantic Violations	Semantics
Infl/Agr Violations	Morphology
Other	Syntax

Table 5: CoLA features aggregated by error type. Infl/Agr =Inflection and Agreement. Other=the rest of grammar violation phenomena present in CoLA annotation for unacceptable sentences.

Model	Learning rate	Weight decay
En-BERT	$3 \cdot 10^{-5}$	0.01
En-RoBERTa	$2 \cdot 10^{-5}$	10^{-4}
Ru-BERT	$3 \cdot 10^{-5}$	0.1
Ru-RoBERTa	10^{-5}	10^{-4}

Table 6: Hyperparameter values used for finetuning transformers.

B Experiment Results

B.1 Linguistic Acceptability Classification

Model	Expert		Machine	
	Acc.	MCC	Acc.	MCC
Ru-BERT	77	0.37	75	0.44
+ TDA$_{ext}$	75	0.39	72	0.42
Ru-RoBERTa	84	0.55	80	0.56
+ TDA$_{ext}$	83	0.53	80	0.56

Table 7: Linguistic acceptability classification results of monolingual LMs and linear classifiers on RuCoLA out of domain test set by source.[8]

Model	MCC
En-BERT	0.509
+ TDA$_{ext}$	0.469
En-RoBERTa	0.608
+ TDA$_{ext}$	0.616

Table 8: Acceptability classification results of monolingual LMs and linear classifiers on CoLA out of domain test set (Alex Warstadt, 2018).

Model	Acceptable	Hallucination	Morphology	Semantics	Syntax
RuCoLA IDD					
Ru-BERT	93.9	-	12.5	24.0	56.0
+TDA$_{ext}$	86.2	-	56.2	45.0	75.4
Ru-RoBERTa	95.9	-	50.0	37.0	70.9
+TDA$_{ext}$	96.3	-	31.2	34.0	72.4
RuCoLA OODD					
Ru-BERT	90.3	53.9	26.6	25.9	55.4
+TDA$_{ext}$	75.0	73.9	51.6	48.1	77.7
Ru-RoBERTa	90.9	64.3	54.7	42.0	75.5
+TDA$_{ext}$	89.9	63.9	53.1	39.5	71.4
CoLA IDD					
En-BERT	94.8	65.0	69.0	72.2	61.2
+TDA$_{ext}$	87.9	77.5	86.2	83.3	82.4
En-RoBERTa	94.8	72.5	88.9	75.9	64.7
+TDA$_{ext}$	87.3	75.0	79.3	88.9	70.6
CoLA OODD					
En-BERT	93.8	72.0	69.7	53.8	50.0
+TDA$_{ext}$	81.0	80.0	78.8	69.2	63.5
En-RoBERTa	93.5	76.0	87.9	76.9	56.2
+TDA$_{ext}$	83.1	80.0	81.8	92.3	63.5

Table 9: Per-category recall on the IDD and OODD sets by benchmark.

[8] https://rucola-benchmark.com

B.2 Fine-tuning effect

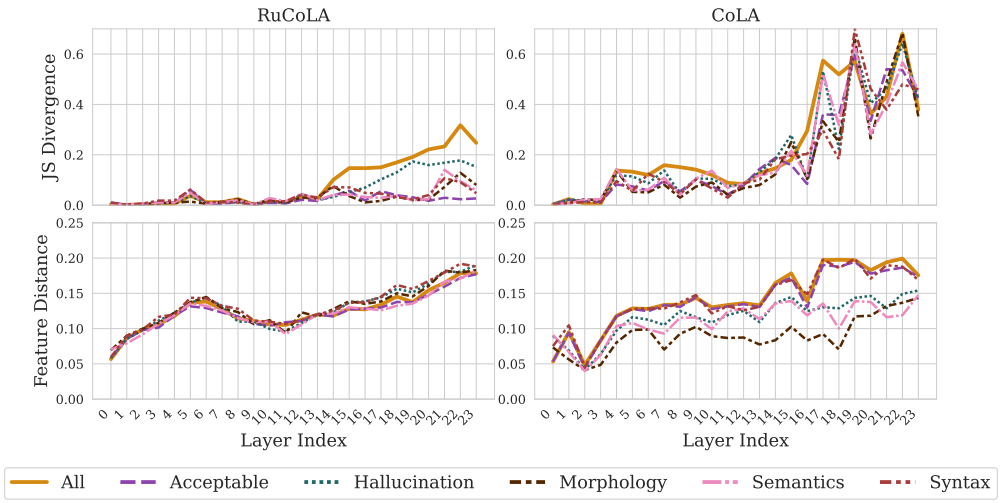


Figure 3: Per-layer feature distance and Jensen-Shannon divergence of attention scores between the frozen and fine-tuned Ru-RoBERTa and En-RoBERTa.

B.3 Feature Importance

Consider a linear classifier with L1 regularization, then the output probability for the sentence i is:

$$p_i \sim \exp(X_{0i}^T C^T w + c),$$

where X_{0i} are the input TDA features, C is the principal component matrix, w^T is a vector of PCs coefficients in the decision function, and c is the added bias. $C^T w$ is the feature contribution to prediction.

B.4 The Roles of Attention Heads

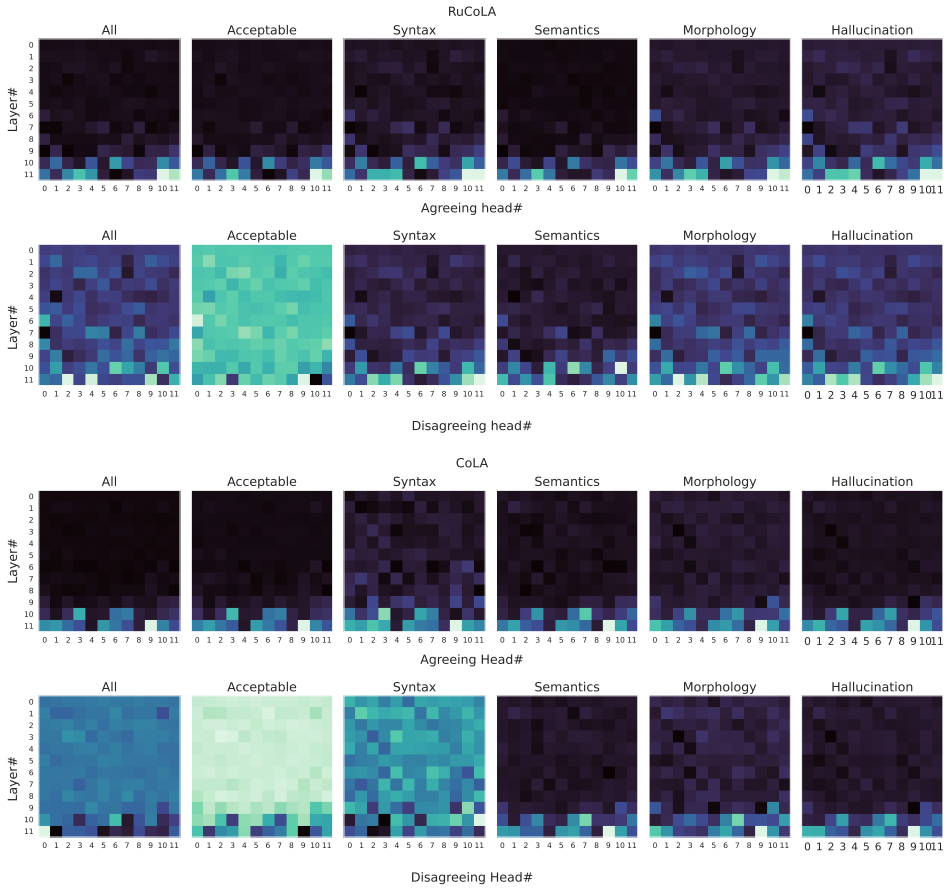


Figure 4: Mean feature weights in TDA_{ext} classifiers with respect to the dataset. TDA_{ext} are extracted from fine-tuned Ru-BERT and En-BERT, respectively. Features of an *agreeing head* contribute to correct prediction. Features of an *disagreeing head* contribute to incorrect prediction. Brighter colors stand for higher mean feature weights.

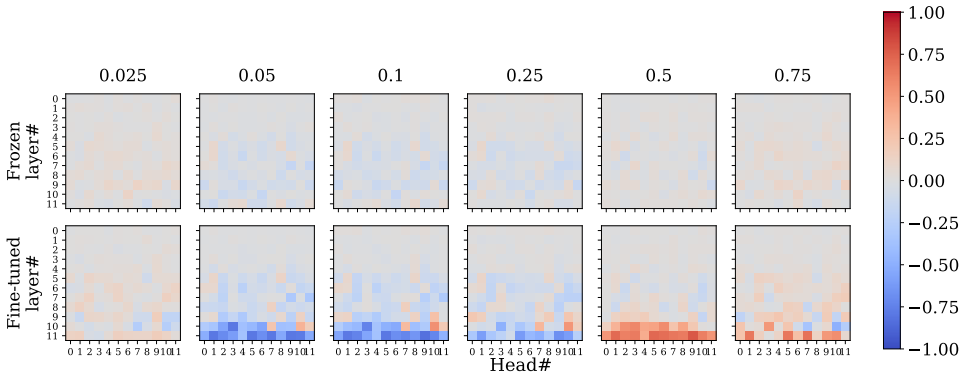


Figure 5: Correlation coefficients between average vertex degree features and target labels for frozen and fine-tuned Ru-BERT by the threshold used to construct attention graph.