From BERT's Point of View: Revealing the Prevailing Contextual Differences

Carolin M. Schuster
Technical University of Munich
carolin.schuster@tum.de

Simon Hegelich
Technical University of Munich

Abstract

Though successfully applied in research and industry large pretrained language models of the BERT family are not yet fully understood. While much research in the field of BERTology has tested whether specific knowledge can be extracted from layer activations, we invert the popular probing design to analyze the prevailing differences and clusters in BERT's high dimensional space. By extracting coarse features from masked token representations and predicting them by probing models with access to only partial information we can apprehend the variation from 'BERT's point of view'. By applying our new methodology to different datasets we show how much the differences can be described by syntax but further how they are to a great extent shaped by the most simple positional information.

1 Introduction

By taking on the perspective of BERT and presenting the methodology to explore this point of view we contribute a new approach to BERTology research, a field that has emerged for a number of good reasons.

Ever since the original BERT paper (Devlin et al., 2019) combined masked language modeling (MLM) with massive pretraining and the transformer architecture (Vaswani et al., 2017), models of the BERT family have achieved a variety of new Natural Language Processing (NLP) benchmarks. While their success is driven by the contextualization of words it is clear that these models do not yet have a real understanding of language (Bender and Koller, 2020). Still, researchers are struggling to find out what it is exactly that they learn and how they perform so well. Challenges are the high number of parameters over which the model knowledge is distributed, and the innumberable different patterns the models can potentially gather from text.

BERTology takes on this quest of understanding the inner workings of these large pretrained models to drive further improvements and identify the next steps towards general AI. Though the training of ever greater models with ever more data has been criticized because of the societal costs and risks these models bring with, including bias and discrimination (Bender et al., 2021). Nevertheless, because of their high performance they are already employed in research but also industry-applications, which exacerbates the need for their explainability.

The black box of Bidirectional Encoder Representations from Transformers (BERT) and its relatives commonly consists of 6-24 identical transformer encoder layers. Each layer comprises a multi-head-attention block followed by a fully connected block, with both being bypassed by residual connections (Vaswani et al., 2017). This stack of layers is primarily pretrained with MLM, the task of predicting a randomly masked (or replaced) word in an input text, and can afterwards be finetuned to specific tasks. Next to attention scores, layer activations are a popular choice for analysis, as they conflate the information from attention heads and skip-connections and represent the stages of the contextualization process.

BERT layer activations are most prominently scrutinized by the so-called edge probing design, in which they are treated as the fixed input to another neural network trained on specific NLP tasks (Tenney et al., 2019). Previous research has employed this method to test them for information on word senses or grammatical properties, but this does not reveal how much the information shapes the space.

By inverting the probing design we present a new way to analyze layer activations, specifically their prevailing patterns, complementing the existing methodology. In contrast to the regular process, we do not use them as input but as the output of

Findings of the Association for Computational Linguistics: ACL 2022, pages 1120 - 1138
May 22-27, 2022 ©2022 Association for Computational Linguistics

a model. We reduce their dimensionality by clustering and principal component extraction to capture the predominating differences within a dataset. By declaring these differences as our ground truth and explaining them in a second step, we can render visible the salient patterns and groupings from 'BERT's point of view'. We chose this term to signal the shift of the perspective, from any contextual information a human might deem important to the information that actually predominates the representative space of the models.

For different datasets we extract the layer activations of masked tokens, meaning that all analyzed tokens start out with identical representations. With this setup we can be sure that the differences we analyze are derived from context and are not caused by different pretrained embeddings. As masked language modeling continues to be a popular and successful pretraining objective, it is of particular interest which patterns are exploited when a model is determining the identity of a masked token.

For the prediction of the representative features we train three types of probing models that receive as input a simplified version of the token context: bag-of-words, ordered part-of-speech tags and simply the position of the token in a sentence. By the disentanglement of these information types, the probing results provide indications about their importance for shaping the representative space.

Social science shows that contrastive explanations are more relevant to humans than complete explanations (Miller, 2019), which affirms the necessity of methods that focus on the contrasts perceived by black box models. For a specific dataset of masked tokens our methodology reveals the most salient differences between their contexts.

Contributions With our methodology we offer a new perspective on the contextual representations inside masked language models: The contrasts within a dataset from BERT's point of view.

By its application we render visible how well syntax describes the coarse patterns of the space but further how much of this description is possible by mere simplistic positional information.

Finally we demonstrate the danger of misinterpreting the learned patterns of the models due to the correlational nature of separated information types which may also lead to an overestimation of the models' sophistication.

2 Related Work

In the field of BERTology (see Rogers et al., 2020 for a general overview) much research has focused on three components; the self-attention mechanism, a key component of the transformer architecture that provides intuitive explanations (see e.g. Kovaleva et al., 2019, Manning et al., 2020, Clark et al., 2019), individual neurons (e.g. Luo et al., 2021) and the layer activations that are scrutinized in our work. Frequently the edge probing design (Tenney et al., 2019) has been deployed to analyze the contents of these activations, in different settings such as after finetuning (Merchant et al., 2020) and with various modifications. Amnesic probing measures what information gets used in the probing tasks by removing selected properties (e.g. part-of-speech) from the activations (Elazar et al., 2021). Similarly O'Connor and Andreas (2021) measured usable information when increasing context size and ablating features of this additional context, e.g. by shuffling. In a parameter-free approach Wu et al. (2020) analyzed the output representation of a masked token by additionally masking other tokens in its proximity to determine their impact.

Another stream of research explores the geometrical space of layer activations. A common approach is the direct measurement of similarities, e.g. between instances of the same token and tokens of the same sentences (Ethayarajh, 2019; Peters et al., 2018) or between instances of homonyms and synonyms (Garcia, 2021). Further work analyzes the separability of predefined categories (e.g. word senses) by manifold analysis (Mamou et al., 2020), by measuring categorical cohesion with silhouette scores (Mickus et al., 2020) or a nearest-neighbor classifier (Coenen et al., 2019), or by searching for clustering solutions that correspond to the categories (Yenicelik et al., 2020). The similarities to our work are the focus on word level representations and the search for categories, though our clusters are not predefined by us but are the groupings inherent to our datasets from BERT's point of view.

Much of the described work concerns only representations of unmasked tokens, except for e.g. Wu et al. (2020) and Mamou et al. (2020), but as masked language modeling continues to be a popular training objective the study of contextual information of masked tokens is highly relevant.

3 Experimental Setup

Because of the various components of probing classifiers and their respective interactions, the design of such is non-trivial (Belinkov, 2021). This is also true for this new, inverted type of probing process that we present here.

3.1 Data

For this analysis of salient differences between large numbers of datapoints the composition of the dataset determines what can potentially be found. Differences may be related to semantics, syntax but also to artifacts that humans are unaware of. Which of the existing distinctions shape the representations in turn depends on if and how these patterns are utilized by the studied model.

The prevailing differences thus depend on:

- The availability of different patterns
- The frequency of available patterns
- BERT's attention to available patterns
- BERT's integration of available patterns

The choice of data is contingent on the objective; it thus can be a specific NLP dataset to understand a model's task performance or a new dataset that we wish to interpret. For an explorative analysis of BERT's view the data can be used as is, but for an analysis of the relevance of specific patterns it is necessary to control their availability and frequency within the dataset. Patterns that are the same across all examples do not influence the feature extraction process.

The tokens to be masked can be one or more frequent words, word senses or syntactical functions, e.g. Part-of-Speech, depending on the selected dataset and the contexts of interest. The diversity of contexts and contextual representations may differ much depending on the token, especially as contextual information not only gives clues about the word behind the mask but also about its interpretation - additional meaning that is attached to it. This is especially true for tokens that signify entities as they are subject to opinions, e.g. "person". This is also reflected by the great amount of sensible candidate words, e.g. named entities, professions or even insults, compared to a masked determinator token "the" or other stopwords.

We selected four datasets from two sources and with different masked tokens to demonstrate the varying patterns that are salient in different kinds of datasets. Data collection and preprocessing steps are listed in Appendix C.

SemCor&OMSTI noun-synsets Our first dataset stems from the combined word-sense annotated corpus (Raganato et al., 2017) of SemCor (Sense-tagged Semantic Corpus) (Miller et al., 1991) and OMSTI (One Million Sense-Tagged Instances) (Taghipour and Ng, 2015). We selected three frequent noun synsets for masking: person.n.01, manner.n.01 and line.n.16, and stratified according to synset, which resulted in a dataset of 6048 masked tokens. While these words are all nouns, they are still used in different syntactic settings.

SemCor&OMSTI person.n.01 From the same combined, sense-annotated corpus we masked all 7702 instances of the synset person.n.01. This includes instances of the word "person" but also named entities.

cctweets-random Our cctweets data consists of tweets about climate change activism that were collected during and after the UN Climate Change Conference in 2019. Ethical considerations of data privacy are elucidated in Appendix A. The discourse was highly polarized, containing diverging representations of the same issues, posing the question of what differences would be salient in the presence of such polarization. We masked random tokens for explorative analysis and as comparison to the cctweets-activist dataset. This dataset comprises 155952 instances.

cctweets-activist Our last dataset consists of climate change related tweets with 132710 masked mentions of a prominent climate change activist, as this person was the center of attention of the discourse. Therefore we could extract thousands of lexically identical instances with different depictions. This dataset represents the use case of searching for semantic groupings based on interpretations from context.

3.2 Feature Extraction

For the investigation of salient differences between the masked token representations we chose k-means clustering and principal component analysis to produce both categorical and continuous features. The appropriateness of either method depends on the properties of the data and we show the results for both, for all of our datasets.

input type	example tokenization	example tensor
Bag-of-Words	who is mask	[0 1 0 0 1 0 1 0 0]
Part-of-Speech	[PAD] [PAD] WP VBZ [MASK] [PAD] [PAD] [PAD] [PAD]	[0 0 4 6 1 0 0 0 0]
Position	[PAD] [PAD] [MASK] [MASK] [MASK] [PAD] [PAD] [PAD] [PAD]	[0 0 1 1 1 0 0 0 0]

Table 1: Input Format: Bag-of-Words, Part-of-Speech and position.

However it is achieved the dimensionality reduction helps humans to grasp the coarse patterns of the space, which is not possible with the raw distributions of meaning over 768 dimensions (bert-base-uncased). Categorizing and aligning datapoints along a single dimension, e.g. ranking them according to some quality, are furthermore tasks that humans not only understand but perform themselves on a daily basis, which underlines the importance of representing the data accordingly.

It has been shown that none of the layer representations of BERT are uniformly distributed with respect to direction (Ethayarajh, 2019) and it is thus important to note that the methods applied here are susceptible to this anisotropy. This does not contradict our design, as we want to describe distances as they are, also showing possible causes for anisotropy. The goal is not to tune our feature extraction methods but to understand how we may want to change the language models themselves.

K-Means Clustering The purpose of clustering is finding distinct groups of similar contexts and it is performed directly on the raw, extracted layer representations. We chose a robust, widely-used algorithm to capture obvious clusters, namely k-means, which we ran with the default configuration of the scikit-learn library. This means 10 runs with different centroid initializations, returning the best solution. We leave the experimentation with different clustering algorithms for future work, but it should be noted that feature extraction methods should remain simple, as complex features will take away from the explainability power of the method.

As we do not know the correct numbers of clusters we cluster for different values of k (2-30) and also utilize silhouette scores to identify the optimal value, thus showing what might be a useful granularity from BERT's point of view. Silhouette scores are a measure of how similar datapoints are to points within their cluster as opposed to points of neighboring clusters (Rousseeuw, 1987). We select common values 2 and 5 to perform the probing for better comparability between settings.

Principal Component Analysis By rotating our axes with principal component analysis (PCA) we obtain the uncorrelated dimensions along which there is the most variation. Thus they are continuous representations of the biggest divergences that are perceived by the BERT models. This is a useful, straightforward alternative to the categorization by clustering when the clusterability of the representative space is low. We choose to analyze the first two principal components with our probing method.

3.3 Pretrained Models

For the extraction of the masked representations we chose two models of the BERT family. First bert-base-uncased (Devlin et al., 2019), which is the standard sized original BERT model and second, deberta-base (He et al., 2020), a modification that has recently been a prominent name on NLP benachmark leaderboards, e.g. SuperGLUE (Wang et al., 2019). The models were retrieved from the Huggingface Transformers library (Wolf et al., 2020).

3.4 Probing

Our reversed probing methodology predicts the features we extracted from BERT representations and takes as inputs simpler features that we obtain from the texts. These inputs are chosen to provide different kinds of contextual information to our probing models. By optimizing these models we can then find out how well our coarse BERT features are described by this information.

To find out how much co-occurences — unordered meaning — and how much syntax shape our coarse BERT features, we disentangle these types of information from our context sentences by creating two input types. The first is a bag-of-words vector that considers all context words of a masked token, while the second input type is a part-of-speech embedding that retains the original order of the context tokens. Because a preliminary qualitative analysis of clusters showed that much of the performance of the syntax classifier may be due to the positional information it receives, we added a third position-only input type. This list of classifiers is not conclusive, but rather a starting point and may also be expanded depending on what additional information is available. Table 1 shows the overview of the selected information and input formats.

For better comparability and similar optimization, we chose to build the architecture of these classifiers identically except for the input layer. While the first layer of the BOW model is fully-connected and accepts a multi-hot vocabulary vector, the POS architecture requires an actual embedding layer. Position is retained simply by centering the masked token and padding on both sides until maximum sequence length. A linear layer aggregates the information over the sequence dimension, arriving at a fixed-length syntax embedding. The position classifier functions similarly, without the additional embedding dimension. For all probing models we append one hidden and one output layer with ReLU activations in-between.

Implementation The probing classifiers were implemented with the Huggingface Transformers Trainer Loop (Wolf et al., 2020) with AdamW optimizer (Loshchilov and Hutter, 2019) and a linear learning rate schedule. Hyperparameter search was realized with Optuna (Akiba et al., 2019) and is described in Appendix D. For the cluster prediction models the cross-entropy loss was calculated with balanced class weights and the best model was selected by Macro-F1 score, as we care equally about all identified clusters. The best models for the regression of principal components were determined by MSE-loss.

4 Results

For the investigation of prevailing differences discerned by the models, we are starting with a manual inspection of the PCA plots for bert-base-uncased in Figure 1 and deberta-base in Figure 3, observing that the contextual space for some combinations of datasets and layers exhibits quite distinct clusters. The presence of further clusters is indicated by their optimal number as determined by silhouette scores, shown in the lower right corner of the individual plots. The datapoints are colored according to positional information, here simply defined as the first character of the masked token divided by the number of characters in the sentence. From these visuals alone we can already learn that positional information greatly shapes the principal components and visible clusters. Some clusters are completely defined by a specific position while others are internally arranged by this feature.

The probing results for the test datasets are shown in Figure 2 for bert-base-uncased and Figure 4 for deberta-base (evaluation results can be found in Appendix E). For almost all studied representations the Part-of-Speech models perform best or are on par to the Bag-Of-Words models. The performance gap is more distinct for the explained variance of the principal components with an 0.21 average difference in R² but only 0.1 for the Macro-F1 scores of the cluster predictions. Notably, while the position models can never outperform the POS models, as they receive only the position-related subset of their contextual inputs, they achieve a large percentage of their performance for many settings, corroborating the finding of the visibly prevailing positional information. Here the performance gaps are 0.40 for R² and 0.26 for the Macro-F1 scores, averaged over all studied settings, showing how much the part-of-speech tags add to the explanations.

For some datasets the plots of BERT and deBERTa closely resemble one another, especially for the noun-synsets. Strikingly, though the data consists of three equally-sized groups of synsets, there are exactly two clusters visible in 2D. The cluster assignment plots for best values of k in Appendix F show that for some layers of BERT and deBERTa the k-means algorithm does manage to differentiate all three of the synsets. Since the probing results are similar as well, we can conclude that BERT's and DeBERTa's point of view do correspond for this dataset. Qualitative inspection found that the synsets person.n.01 and manner.n.01 adjoin while line.n.16 is spatially far removed. While positional information visibly permeates the clusters the distance between them is described almost perfectly by the POS models, thus by syntactic contextual differences. However, the almost equal performance of the BOW models shows the correlation of part-of-speech with bag-of-words patterns that can be exploited.

For the dataset of masked person.n.01 synsets the BERT and deBERTa 2D-projections appear less alike, especially for layer 6. In this setting the POS model's performance for the regression of BERT's principal components greatly exceeds all others with 88% of their variances explained. Further analysis showed that the cluster in the upper half of the plot contains only instances of masked to-

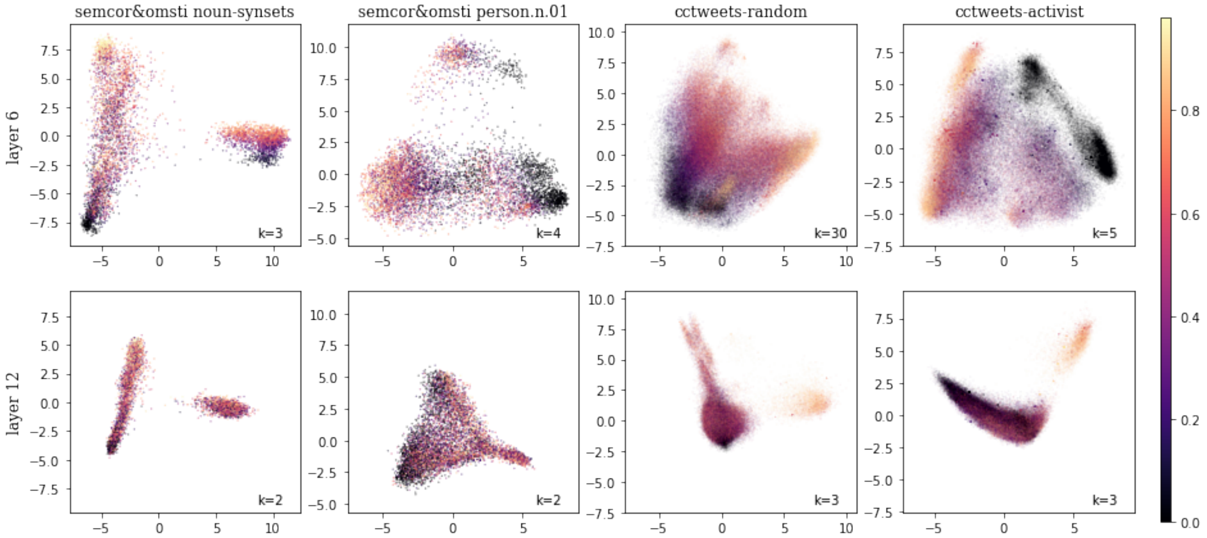

Figure 1: BERT-base-uncased 2D Principal Components. Datapoints are colored by positional information, calculated by the first character of the masked token divided by the number of characters of the sentence. K indicates the number of clusters with the best silhouette score. Extended plot with additional layers: Appendix 5.

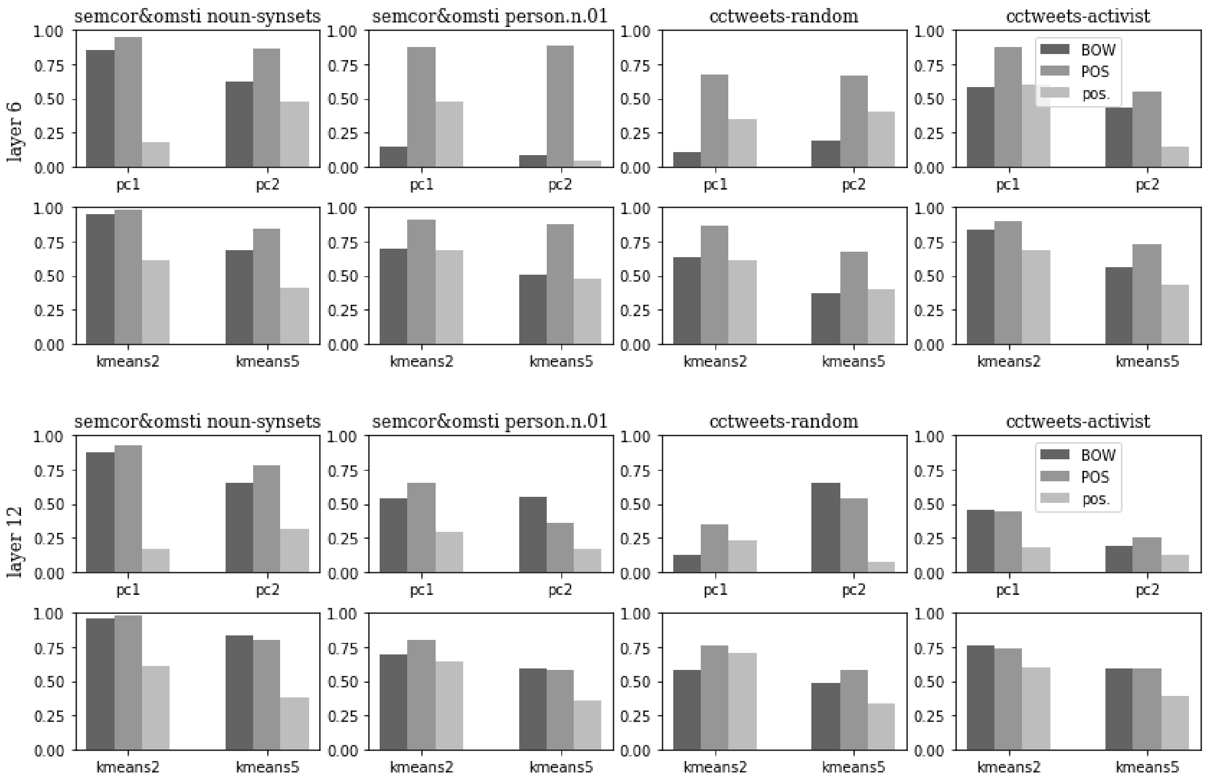

Figure 2: BERT-base-uncased Probing Test Results. Reported scores are Macro-F1 for k-means prediction and R² for the regression of principal components.

kens that were followed by an apostrophe, showing that his specific syntactic pattern is perceived as significantly different. Because of the very low performance of the BOW regression model, we can be sure that this difference is indeed caused by syntax and not by co-occurences. The diverging results for the k-means-2 model expose that the clustering algorithm found a different one than the visible grouping solution (see also plots with cluster assignments for k=2 in Appendix 6).

For both the twitter datasets of random masks and masked activist tokens the final layer of BERT

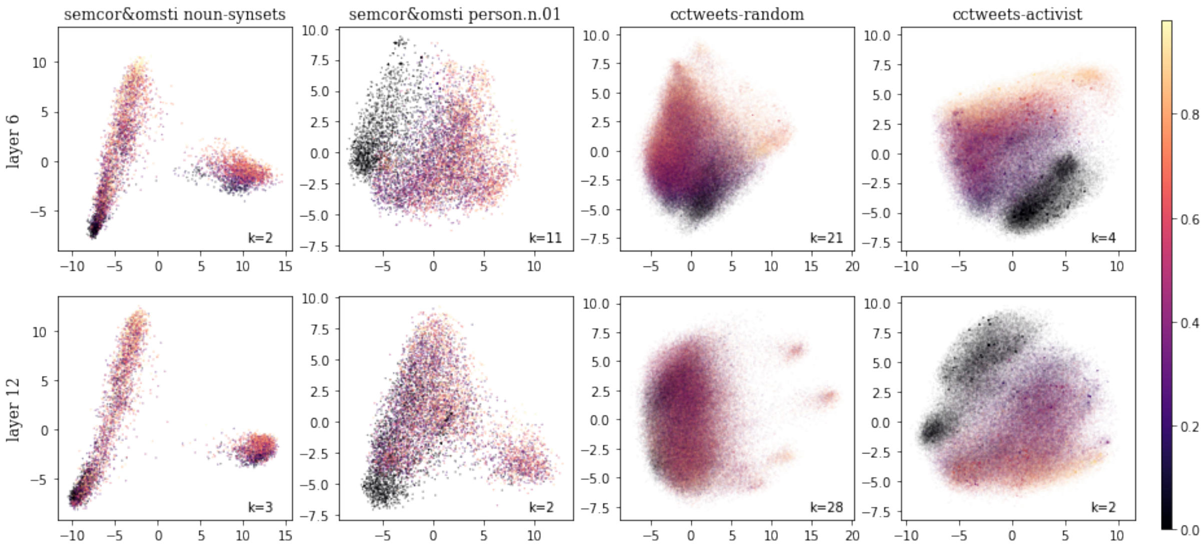

Figure 3: DeBERTa-base 2D Principal Components. Datapoints are colored by positional information, calculated by the first character of the masked token divided by the number of characters of the sentence. K indicates the number of clusters with the best silhouette score. Extended plot with additional layers: Appendix 8.

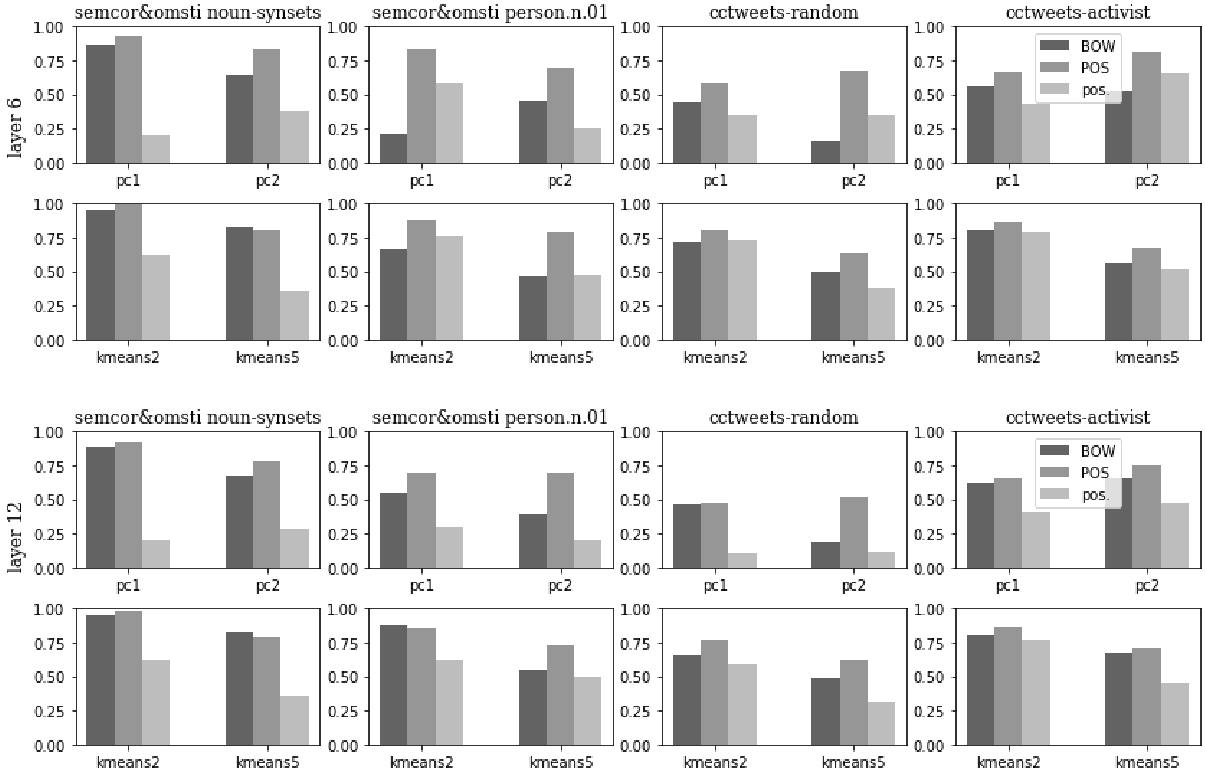

Figure 4: DeBERTa-base Probing Test Results. Reported scores are Macro-F1 for k-means prediction and R² for the regression of principal components.

singles out those that appeared at the end of the sentence. Qualitative investigation showed that this is the case regardless if the token was the final one or followed by a punctuation character. For DeBERTa the space of cctweets-activist is characterized by position to a greater extent than that of cctweets-random, as evident from the visuals and the performances of the position models. The numbers of clusters as suggested by silhouette scores are much higher for the random masks. While the BERT and DeBERTa perspectives on the dataset with equally sized groups of synsets seem quite identical, the cases of one synset or random masks reveal rather different perceived contextual differences.

5 Discussion & Conclusion

In contrast to much recent BERTology work of predicting specific syntactic and semantic information from layer activations, we invert the probing design to instead predict features of the representative space itself. From masked token representations we extract clusters and principal components of contextual information and explore their nature by probing models that receive as input detangled types of information. We thus paint a picture of the dissimilarities and groupings within a dataset from BERT's point of view, thereby expanding existing probing methodology by a crucial perspective.

Our analysis shows that the representative space of contextual information does exhibit clusters. Most clusters and principal components of our datasets are best described by the Part-of-Speech models, however, for many settings the positional probing models can achieve 50% or more of the POS performance. This shows how the representative space of both BERT and DeBERTa is greatly shaped by the most simple positional information, even though these models handle positional embeddings differently. As demonstrated by Geirhos et al. (2020), neural networks are prone to shortcut learning, and thus position may be one such shortcut. On the other hand, for the standard BERT it was shown that the representative space is anisotropic due to outlier neurons capturing positional information, which was attributed to Layer Normalization (Luo et al., 2021).

The usual probing classifier architecture that receives representations as input and predicts a specified linguistic property cannot clarify, if the representations are actually informed by the linguistic property of interest or by other, correlating properties of the training data (Belinkov, 2021). In our analysis the for some cases equal performance of detangled semantics (Bag-of-Words) and syntax (ordered Part-of-Speech tags) shows as well their correlational nature and the difficulty of pinpointing which features are actually utilized by large masked language models. When simplistic and meaningful features correlate this provides the danger of assuming that the models are much more sophisticated than they actually are.

We do not attempt to answer the question of what information should predominate the representational space, but it is likely that the optimum is not reached with features as simple as the position of a token in a sentence. We expect the best solutions to be defined by more sophisticated features that are not obtainable with simple string analysis, and which might even be utilizable for data analysis and hence other fields of research.

Concluding, our methodology delivers clues about the shortcomings of language models and the shortcuts that they are exploiting, to highlight directions of further adjustments of training objectives and processes in the future.

We hope that this work inspires more researchers to look at the world from BERT's point of view, to understand how it differs from ours. By recognizing the nature of their current primitivity we can generate new ideas on how to improve these large language models, gradually moving in the direction of a more general AI.

Limitations The prevailing contextual patterns that are revealed by this method are not universal but are always contingent on the analyzed datasets. Accordingly these have to be chosen and controlled depending on the research objective.

For the extraction of categories by clustering, selecting the appropriate number of clusters is non-trivial. Here the number of clusters was set to equal numbers to allow for a comparison between datasets and layers, but these may not reflect the inherent number of groupings.

Lastly this method, as with any method that analyzes individual parts of a network in isolation, does not explain how the identified prevailing information is utilized during task performance.

Future Work A promising extension of this work will be to enlarge the set of probing models to even better partition the different types of information, to better understand their contributions. Examples of further relevant input types are a windowed Bag-of-Words or Bag-of-Words filtered by word types. Furthermore it would be highly interesting to compare the POS model to other embedding models (e.g. simply word embeddings) with identical structures.

The settings that may be analyzed by this method are various, such as the finetuning process, to explore how the prevailing patterns shift when models adapt to particular tasks. A comparison between models of different languages may reveal different focuses and varying correlations of information types.

The described method is applicable also for the analysis of unmasked tokens though then the process of contextualization will differ from the very first layer depending on the token. The masked and unmasked contextualizations are moreover shaped by different objectives, predicting masked tokens and predicting potentially perturbed tokens, which may result in attention to different contextual patterns.

Finally it may be fruitful to utilize gradient-based attribution methods to pinpoint not just the relevance of input types but the relevance of specific inputs and positions from BERT's point of view.

Acknowledgements

This work was supported by the Heinrich Böll Foundation through a doctoral scholarship. We would like to thank the anonymous reviewers for their valuable feedback.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. *arXiv preprint arXiv:2102.12452*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

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.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and Measuring the Geometry of BERT. In *Advances in Neural Information Processing Systems*, volume 32, pages 8594–8603.

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.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. In *Transactions of the Association for Computational Linguistics*, volume 9, pages 160–175.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Marcos Garcia. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. 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 3625–3640, Online. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv:2006.03654 [cs]*.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Ziyang Luo, Artur Kulmizev, and Xiaoxi Mao. 2021. Positional artefacts propagate through masked language model embeddings. 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 5312–5327, Online. Association for Computational Linguistics.

Jonathan Mamou, Hang Le, Miguel Del Rio, Cory Stephenson, Hanlin Tang, Yoon Kim, and SueYeon Chung. 2020. Emergence of Separable Manifolds in Deep Language Representations. In *International Conference on Machine Learning*, pages 6713–6723.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, BERT? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290, New York, New York. Association for Computational Linguistics.

George Miller, R. Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1991. Introduction to WordNet: An On-line Lexical Database*. *International journal of lexicography*, 3(4):235–244.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? 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 851–864, Online. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Peter J. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Kaveh Taghipour and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 338–344, Beijing, China. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS 2017)*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Advances in Neural Information Processing Systems*, 32. ArXiv: 1905.00537.

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.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

David Yenicelik, Florian Schmidt, and Yannic Kilcher. 2020. How does BERT capture semantics? a closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 156–162, Online. Association for Computational Linguistics.

A Ethical Considerations

This work utilizes public discussions by private individuals on Twitter. The tweets were collected with the Twitter streaming API and all information of the tweeting users, including user names and ids, was discarded. However, sensitive information is also found within the analyzed texts, none of which are made public. The tweets are stored with restricted access and will be deleted upon research conclusion. Afterwards only the tweet ids will be available, which can be hydrated through the Twitter API only for tweets that still are public.

B Computing Infrastructure & Runtimes

A Nvidia GeForce RTX 2080 Ti graphics card was used for the training and evaluation of the probing models. The hyperparameter search with 50 runs lasted 28 minutes on average and the final models were optimized with an average of 6 minutes training time.

C Data Collection & Preprocessing

The unified sense-tagged corpus of SemCor and OMSTI was obtained from `http://lcl.uniroma1.it/wsdeval/training-data` (Raganato et al., 2017).

The Part-of-Speech tags for the POS probing models were obtained by the nltk python package. Specialized taggers for Twitter data are available but were not deemed necessary as most of the twitter-specific artefacts were removed during preprocessing.

For probing the datasets were split randomly into training, evaluation and test sets by the ratio 70:15:15.

Twitter datasets Tweets were collected through the Twitter Streaming API with keywords related to climate change and activism. The timeframe of collection was during and after the United Nations Climate Change Conference in 2019 (COP25): 2.-19.12.2019 As per Twitter policy only the ids of tweets are made available, which can be rehydrated with the Twitter API.

- Filters:
  - only English tweets
  - no replies
  - at least three words
  - only sentences / sentence-like phrases
  - duplicates removed
- Preprocessing:
  - removing URLs
  - removing hashtag and mention sequences if n > 1
  - pruning repeating characters and words if n > 3
  - random masking /masking first token that matches activist pattern
  - obtaining sentences / sentence-like phrases containing the mask token

D Hyperparameter Search

Hyperparameter search was performed for a sample of the analyzed probing settings: For each combination of the 4 datasets, 3 input types (bag-of-words, part-of-speech and position) and 2 output types (k-means, principal component), 3 settings were sampled and hyperparameter search was conducted with Optuna for 50 runs. The search results were then pooled for each combination.

The search space and pooling strategy are shown in Table 2. Preliminary experiments had shown that one hidden layer was a generally good choice for network depth, but network width was included as a search parameter. The determined values for the hyperparameters stayed within the search boundaries, except for two cases where n_hidden was equal to the maximum value. Thus additional trials were run to find out if representational capacity had to be increased further with the maximum value found to be 2060.

hyperparameter	search space	pooling
hidden_layer_size	128 - 2048	max
batch_size	8 - 64	mean
learning_rate	1e-5 - 1e-1	mean
n_steps	5000 - 50000	max + 5000

Table 2: Hyperparameter Search Space and Pooling Strategy.

For batch_size and learning_rate the values were aggregated by averaging, but to ensure a sufficient capacity of the network layer_size was set to the maximum. The number of training steps was set to the maximum plus additional 5000 steps to ascertain sufficient training for any configuration. As the best checkpoint is selected for testing, this does not hurt the performance of faster converging models.

dataset	input_type	output_type	n_hidden	batch_size	learning_rate	max_steps
cctweets-activist	BOW	kmeans	1882	14	0.00075	53988
cctweets-activist	BOW	pc	1617	33	0.00076	31841
cctweets-activist	pos.	kmeans	2060	11	0.00026	45684
cctweets-activist	pos.	pc	1715	34	0.00057	36678
cctweets-activist	POS	kmeans	1654	30	0.00317	38124
cctweets-activist	POS	pc	1970	47	0.00247	24824
cctweets-random	BOW	kmeans	1308	26	0.00079	37955
cctweets-random	BOW	pc	1721	23	0.00112	48627
cctweets-random	pos.	kmeans	1187	12	0.00035	47055
cctweets-random	pos.	pc	1674	47	0.00041	54746
cctweets-random	POS	kmeans	1482	25	0.00412	47735
cctweets-random	POS	pc	2048	48	0.00252	38101
S&O noun-synsets	BOW	kmeans	1726	21	5e-05	14053
S&O noun-synsets	BOW	pc	698	16	0.02717	26494
S&O noun-synsets	pos.	kmeans	1098	11	0.00029	34690
S&O noun-synsets	pos.	pc	597	10	0.00332	15622
S&O noun-synsets	POS	kmeans	736	17	0.04023	45072
S&O noun-synsets	POS	pc	299	24	0.00947	13580
S&O person.n.01	BOW	kmeans	1983	21	0.00022	32521
S&O person.n.01	BOW	pc	518	29	0.00786	40901
S&O person.n.01	pos.	kmeans	945	20	0.00222	37165
S&O person.n.01	pos.	pc	1047	19	0.00229	17568
S&O person.n.01	POS	kmeans	923	18	0.00102	24021
S&O person.n.01	POS	pc	1075	19	0.02934	18444

Table 3: Hyperparameter Settings.

The resulting hyperparameter settings are listed in Table 3.

E Extended Results

F Extended Plots

		noun-synsets				person.n.01				cctweets-random				cctweets-activist			
		k2	k5	pc1	pc2	k2	k5	pc1	pc2	k2	k5	pc1	pc2	k2	k5	pc1	pc2
BOW	eval	0.96	0.69	0.84	0.59	0.7	0.52	0.19	0.2	0.64	0.37	0.11	0.19	0.83	0.57	0.56	0.45
BOW	test	0.95	0.69	0.86	0.62	0.7	0.51	0.15	0.08	0.63	0.37	0.1	0.19	0.83	0.56	0.58	0.43
POS	eval	0.98	0.83	0.93	0.85	0.92	0.9	0.89	0.9	0.88	0.69	0.69	0.67	0.9	0.74	0.88	0.56
POS	test	0.98	0.84	0.95	0.87	0.91	0.88	0.88	0.89	0.87	0.68	0.68	0.67	0.9	0.73	0.88	0.55
pos.	eval	0.62	0.44	0.17	0.45	0.73	0.48	0.54	0.03	0.61	0.41	0.34	0.41	0.69	0.43	0.6	0.16
pos.	test	0.61	0.41	0.18	0.47	0.69	0.47	0.48	0.04	0.61	0.4	0.35	0.4	0.69	0.43	0.6	0.15

Table 4: BERT-base-uncased Layer 6 Probing Evaluation and Test Results.

		noun-synsets				person.n.01				cctweets-random				cctweets-activist			
		k2	k5	pc1	pc2	k2	k5	pc1	pc2	k2	k5	pc1	pc2	k2	k5	pc1	pc2
BOW	eval	0.96	0.77	0.87	0.68	0.75	0.59	0.61	0.54	0.58	0.48	0.12	0.64	0.76	0.62	0.45	0.21
BOW	test	0.96	0.83	0.88	0.66	0.7	0.59	0.54	0.55	0.58	0.48	0.13	0.66	0.76	0.59	0.45	0.19
POS	eval	0.98	0.81	0.9	0.74	0.81	0.6	0.66	0.36	0.77	0.59	0.37	0.53	0.74	0.6	0.44	0.28
POS	test	0.98	0.8	0.93	0.78	0.8	0.58	0.66	0.36	0.76	0.58	0.35	0.54	0.74	0.59	0.44	0.25
pos.	eval	0.63	0.37	0.16	0.29	0.66	0.38	0.32	0.14	0.73	0.35	0.26	0.07	0.6	0.4	0.19	0.14
pos.	test	0.61	0.38	0.17	0.32	0.64	0.36	0.29	0.17	0.71	0.34	0.23	0.07	0.6	0.39	0.18	0.13

Table 5: BERT-base-uncased Layer 12 Probing Evaluation and Test Results.

		noun-synsets				person.n.01				cctweets-random				cctweets-activist			
		k2	k5	pc1	pc2	k2	k5	pc1	pc2	k2	k5	pc1	pc2	k2	k5	pc1	pc2
BOW	eval	0.96	0.78	0.86	0.63	0.67	0.49	0.29	0.43	0.73	0.5	0.43	0.15	0.8	0.57	0.57	0.53
BOW	test	0.95	0.82	0.87	0.64	0.66	0.46	0.21	0.45	0.72	0.5	0.44	0.16	0.8	0.56	0.56	0.53
POS	eval	0.98	0.82	0.91	0.82	0.9	0.81	0.82	0.7	0.8	0.63	0.59	0.68	0.86	0.68	0.66	0.81
POS	test	0.99	0.8	0.93	0.83	0.88	0.79	0.83	0.7	0.8	0.63	0.58	0.68	0.86	0.68	0.67	0.81
pos.	eval	0.63	0.37	0.18	0.37	0.8	0.49	0.61	0.24	0.73	0.39	0.36	0.35	0.78	0.51	0.42	0.64
pos.	test	0.62	0.36	0.2	0.38	0.76	0.47	0.58	0.25	0.73	0.38	0.35	0.35	0.79	0.52	0.43	0.65

Table 6: DeBERTa-base Layer 6 Probing Evaluation and Test Results.

		noun-synsets				person.n.01				cctweets-random				cctweets-activist			
		k2	k5	pc1	pc2	k2	k5	pc1	pc2	k2	k5	pc1	pc2	k2	k5	pc1	pc2
BOW	eval	0.96	0.81	0.88	0.68	0.91	0.57	0.62	0.4	0.65	0.48	0.47	0.19	0.8	0.69	0.62	0.65
BOW	test	0.95	0.82	0.89	0.68	0.88	0.55	0.55	0.39	0.65	0.48	0.46	0.19	0.8	0.68	0.62	0.66
POS	eval	0.98	0.81	0.91	0.76	0.88	0.76	0.68	0.69	0.78	0.63	0.47	0.54	0.86	0.71	0.65	0.75
POS	test	0.98	0.79	0.92	0.78	0.85	0.73	0.7	0.7	0.77	0.62	0.47	0.52	0.86	0.71	0.65	0.75
pos.	eval	0.63	0.37	0.18	0.26	0.67	0.52	0.33	0.17	0.59	0.33	0.1	0.11	0.77	0.45	0.4	0.47
pos.	test	0.62	0.36	0.2	0.28	0.62	0.49	0.29	0.2	0.59	0.32	0.1	0.11	0.77	0.45	0.41	0.48

Table 7: DeBERTa-base Layer 12 Probing Evaluation and Test Results.

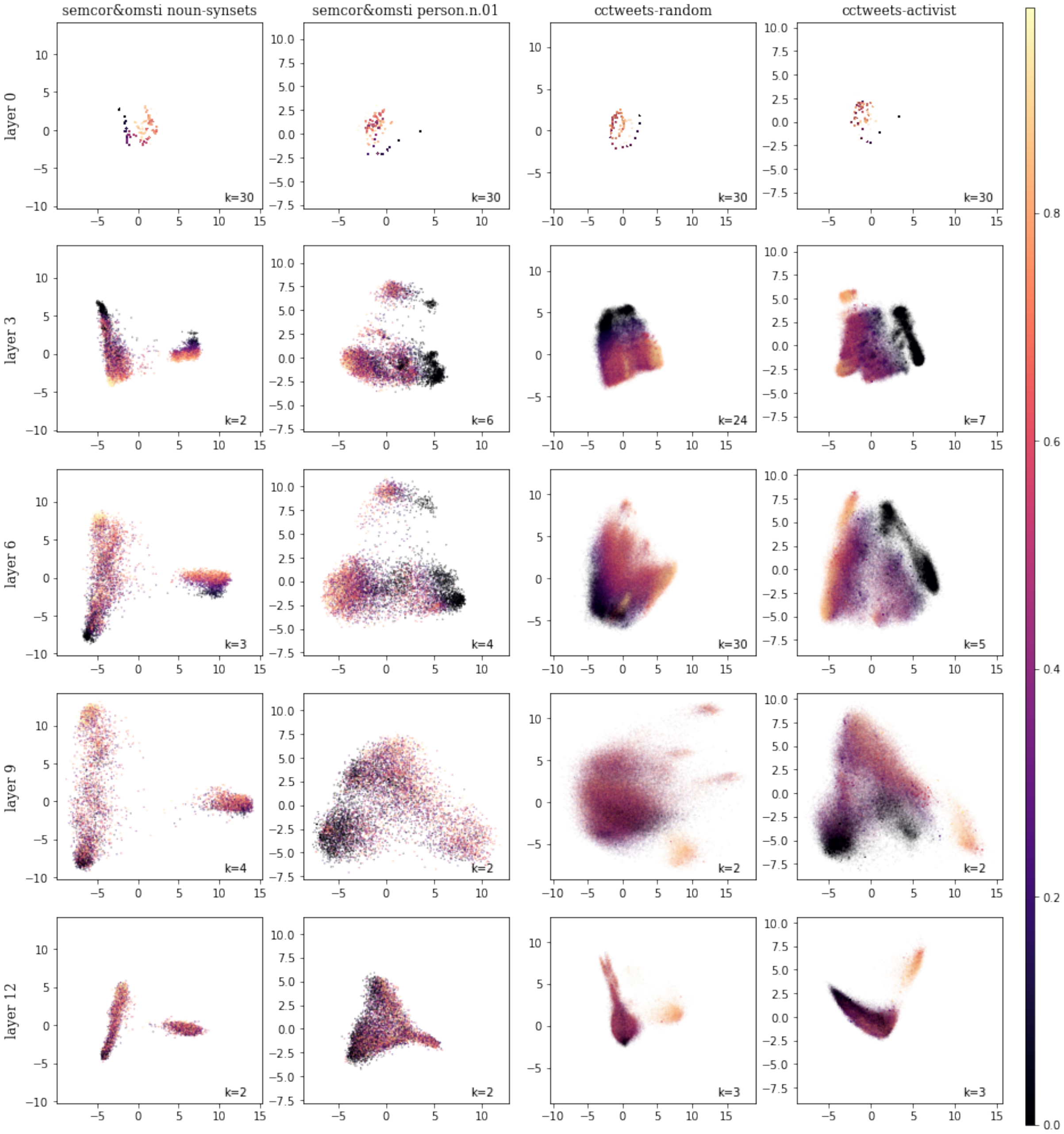

Figure 5: BERT-base-uncased 2D Principal Components. Datapoints are colored by positional information, calculated by the first character of the masked token divided by the number of characters of the sentence. K indicates the number of clusters with the best silhouette score.

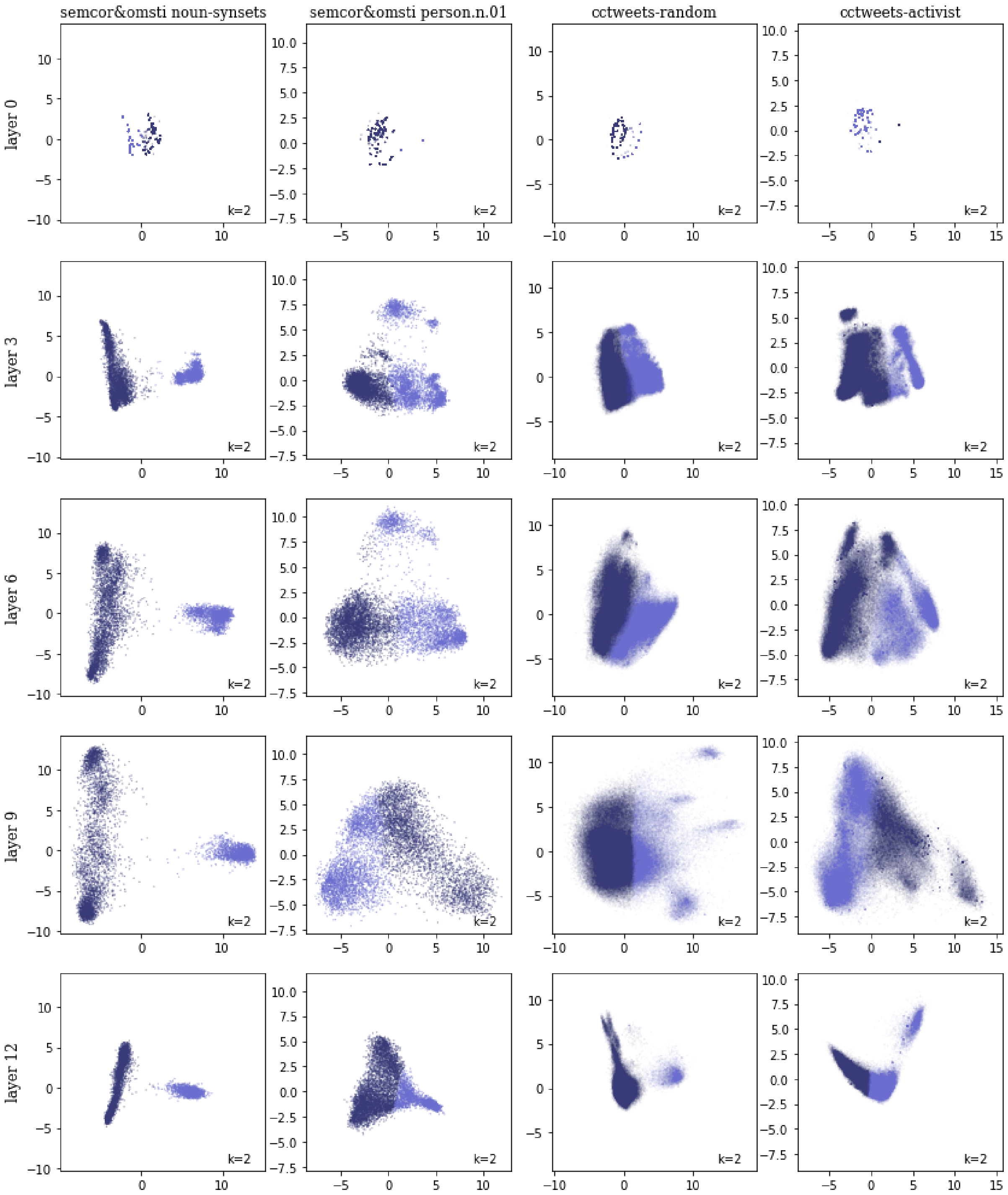

Figure 6: BERT-base-uncased 2D Principal Components and Cluster Assignments for k=2. K (lower right corners) indicates the number of clusters with the best silhouette score.

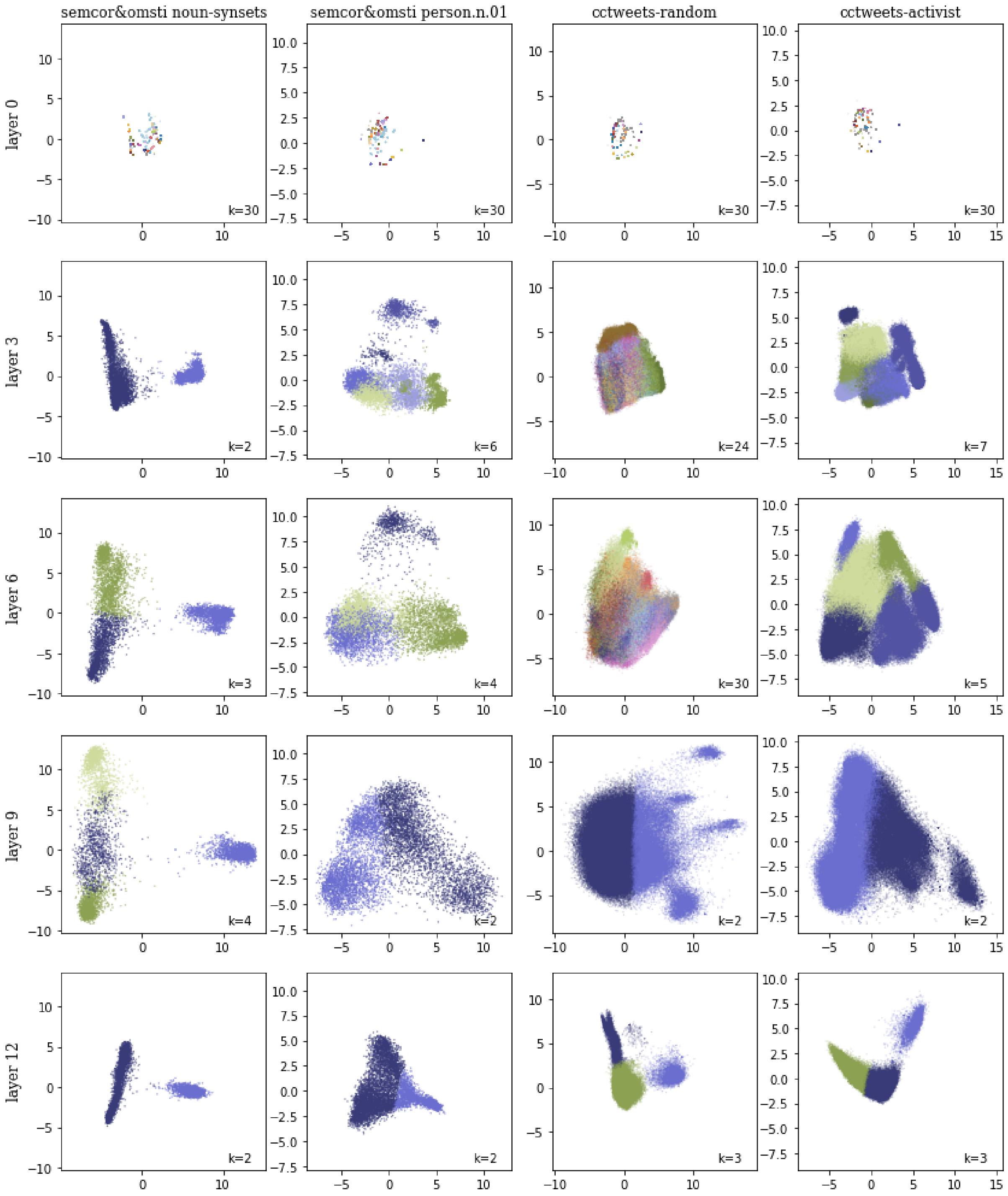

Figure 7: BERT-base-uncased 2D Principal Components and Cluster Assignments for best Values of K as determined by Silhouette Scores.

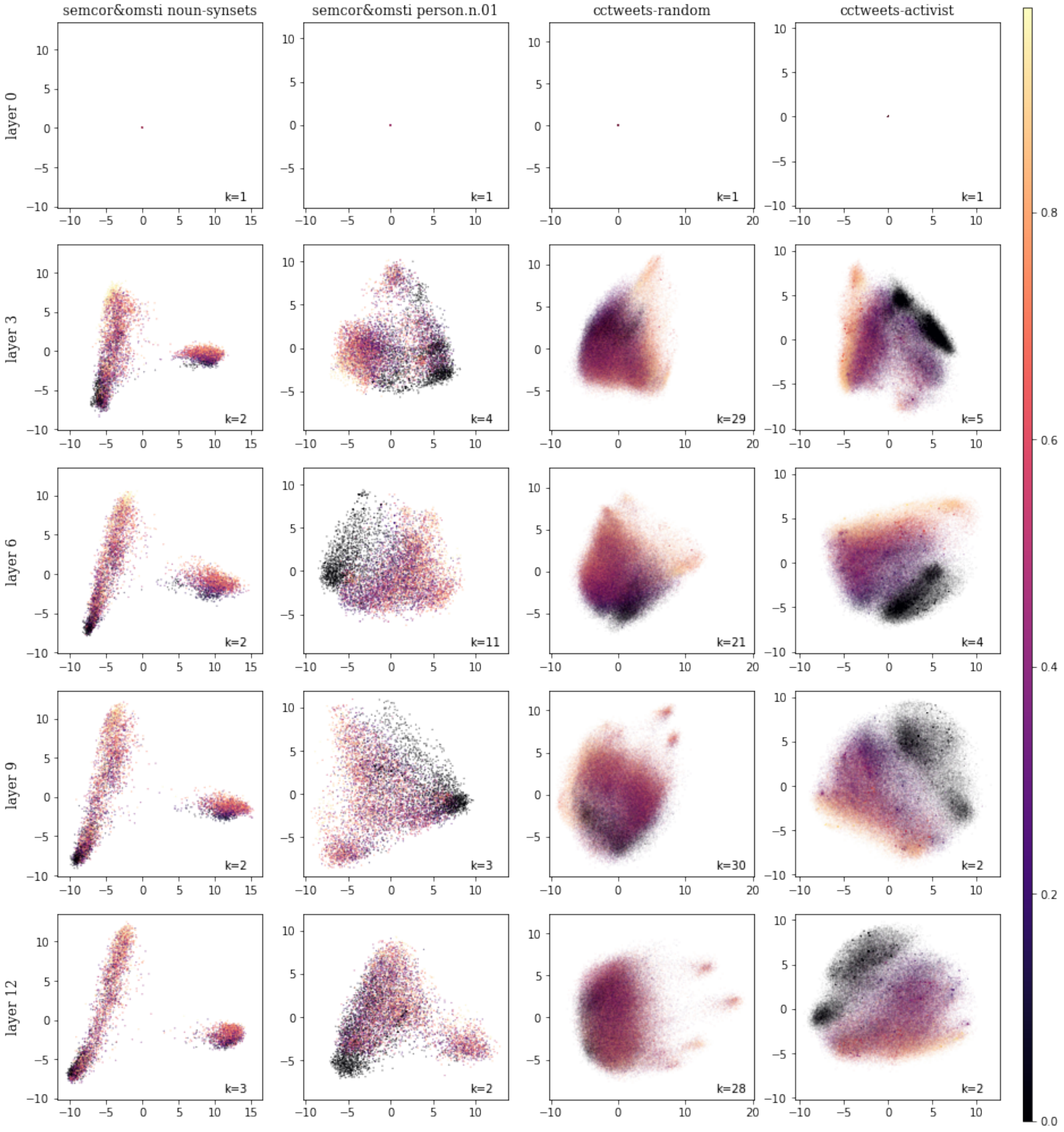

Figure 8: DeBERTa-base 2D Principal Components. Datapoints are colored by positional information, calculated by the first character of the masked token divided by the number of characters of the sentence. K indicates the number of clusters with the best silhouette score.

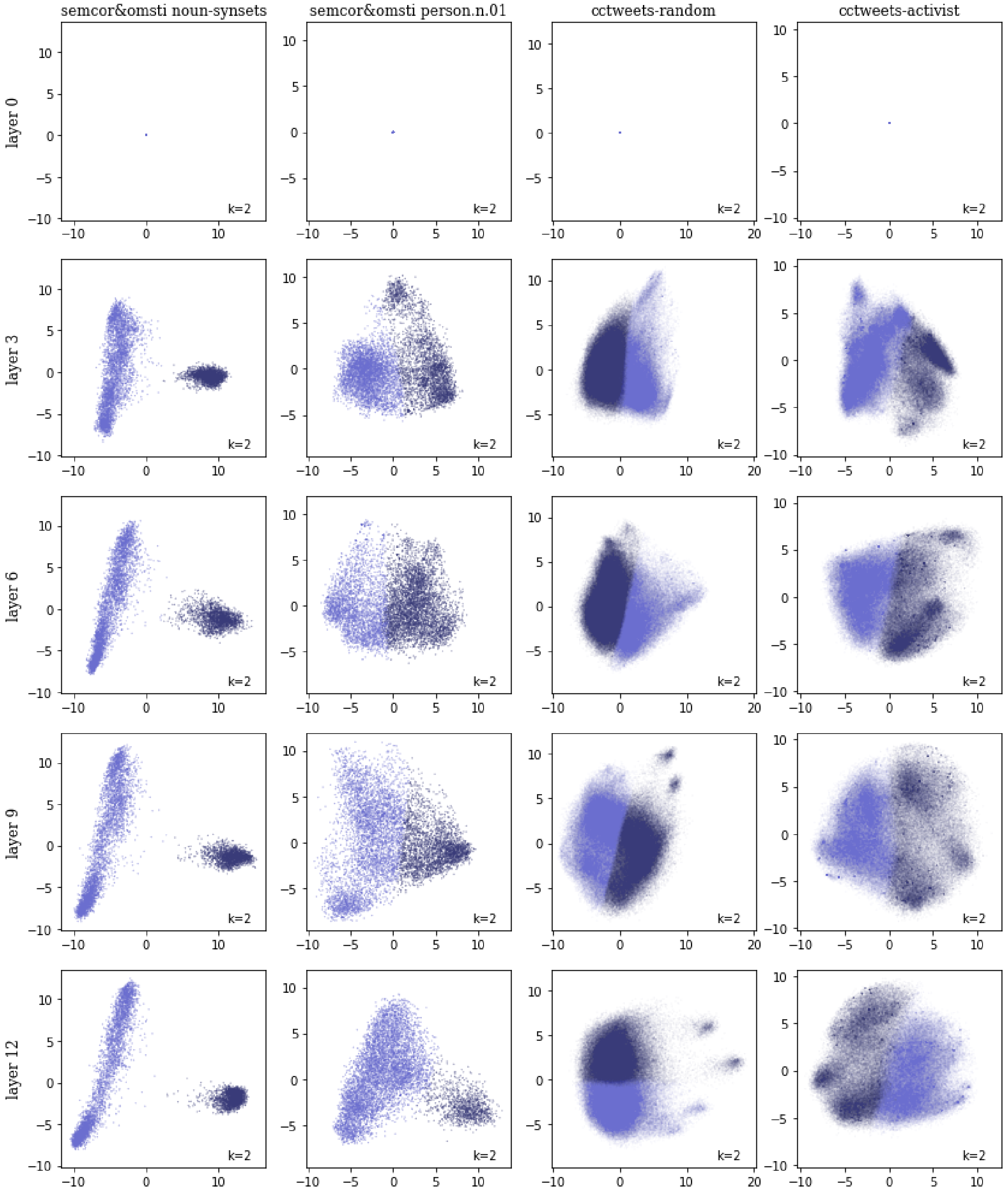

Figure 9: DeBERTa-base 2D Principal Components and Cluster Assignments for k=2. K (lower right corners) indicates the number of clusters with the best silhouette score.

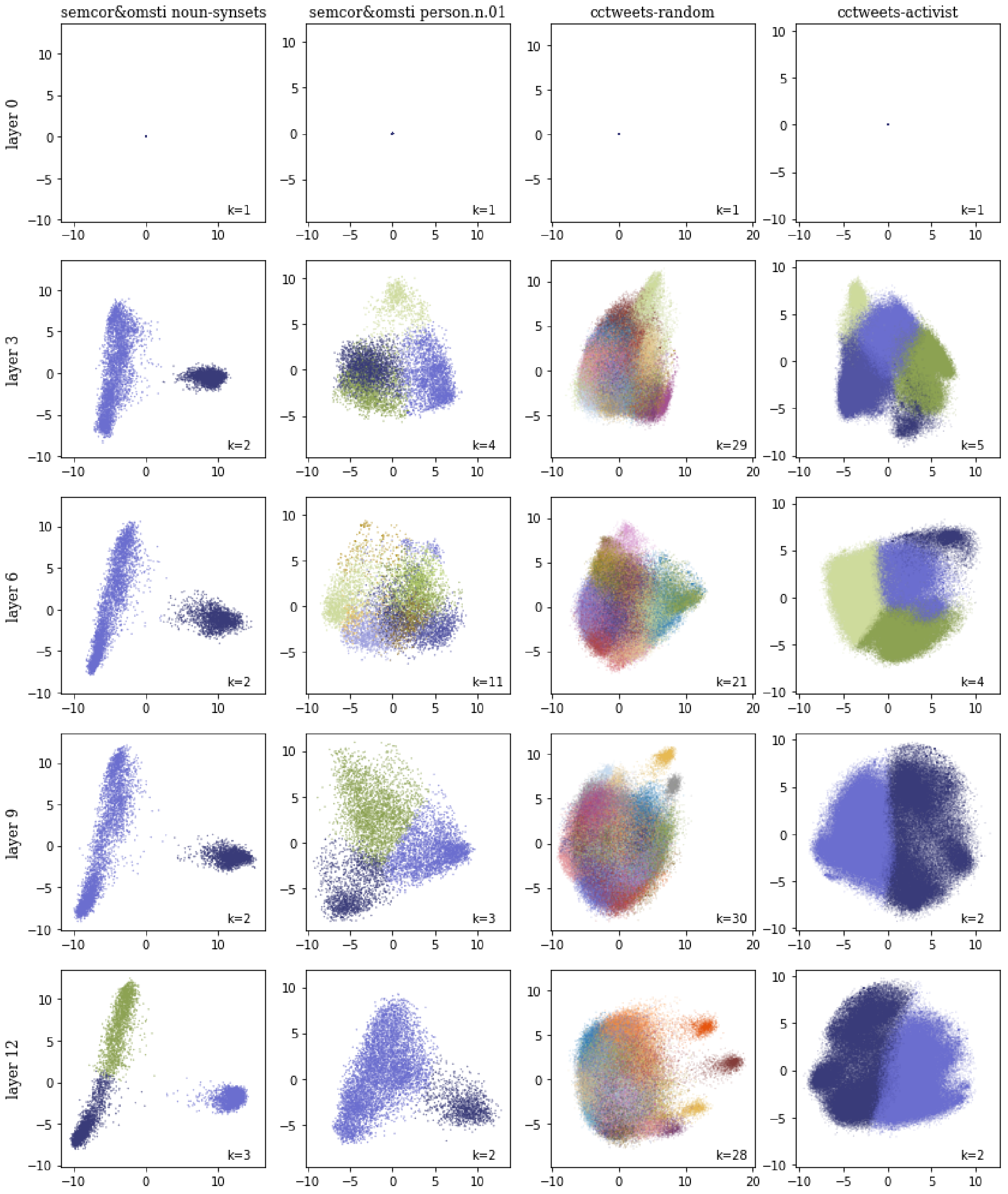

Figure 10: DeBERTa-base-uncased 2D Principal Components and Cluster Assignments for best Values of K as determined by Silhouette Scores.