# The Dark Side of the Language: Pre-trained Transformers in the DarkNet

**Anonymous ACL submission**

## Abstract

Pre-trained Transformers are challenging human performances in many natural language processing tasks. The gigantic datasets used for pre-training seem to be the key for their success on existing tasks. In this paper, we explore how a range of pre-trained natural language understanding models perform on truly novel and unexplored data, provided by classification tasks over a DarkNet corpus. Surprisingly, results show that syntactic and lexical neural networks largely outperform pre-trained Transformers. This seems to suggest that pre-trained Transformers have serious difficulties in adapting to radically novel texts.

## 1 Introduction

Pre-trained Transformers (Peters et al., 2018; Devlin et al., 2019; Zhang et al., 2019; Radford and Narasimhan, 2018) are outperforming humans in many natural language processing tasks (Wang et al., 2018, 2020) and, thus, are wiping out all other methods for natural language understanding. Pre-training seems to give Transformers crystal clear models of target languages. BERT is pre-trained on an English corpus of 3,300M words consisting of books (Zhu et al., 2015a) and Wikipedia. The English version of the last ERNIE (Sun et al., 2021) is trained on an even bigger corpus, and its Chinese version is trained on 14TB corpus. MEGATRON-LM (Shoeybi et al., 2019) is trained on an incredible corpus of 174 GB. The race is always towards training over bigger corpora.

The gigantic datasets used for pre-training seem to be the key to the success of Transformers. It may seem that Transformers have success in downstream tasks because they have seen large parts of possible sentences. Sometimes, this possible shortcoming is taken into consideration when a novel Tranformer is introduced (Radford et al., 2019; Shoeybi et al., 2019). Radford et al. (2019) have excluded Wikipedia pages for pre-training as it is a common data source for other datasets. Yet, when using off-the-shelf pre-trained models, this effect is generally disregarded. For example, the discovering ongoing conversation (DOC) task was found challenging for humans but BERT baseline model achieved the astonishing 88.4 F1 score (Wang et al., 2020). DOC consists of determining if two utterances are contiguous in classical theatrical plays. These plays may be included in the book dataset (Zhu et al., 2015a) used for pre-training BERT.

Corpora and related tasks derived from the DeepWeb and DarkWeb (Avarikioti et al., 2018; Choshen et al., 2019) offer a tremendous opportunity to study the effect of overfitting for different natural language understanding models. Indeed, it is extremely rare that texts extracted from these sources are included in pre-training corpora. Moreover, language on the DarkNet may have very different characteristics with respect to the one accessible from the surface web (Choshen et al., 2019).

In this paper, we aim to explore how pre-trained natural language understanding models behave on really unseen data or really unexplored linguistic registers and styles. This unseen data is given by the DarkNet corpus along with a classification task. We experimented with: Stylistic Classifiers based on the bleaching text model (van der Goot et al., 2018), with Lexical Neural Netowrks based on GloVe (Pennington et al., 2014) and word2vec(Mikolov et al., 2013), with Syntatic-based neural networks based on KERMIT (Zanzotto et al., 2020), and with holistic Transformers such as BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), ERNIE (Zhang et al., 2019) and Electra (Clark et al., 2020). Results show that syntactic and lexical neural networks surprisingly outperform pre-trained Transformers. This seems to suggest that pre-trained Transformers have serious difficulty in adapting to really unseen texts.

The rest of the paper is organized in: Material and Methods; Results and Discussion; and, Conclusions.

## 2 Material and Methods

### 2.1 Material: A Dark Web Dataset

Corpora scraped from DarkWeb to fight illegal actions are good testbeds for studying large pre-trained models on totally new texts, as these are not covered by the corpora used for pre-training.

Nabki et al. (2019), following Choshen et al. (2019)'s instructions, sampled "Darknet Usage Text Addresses" (DUTA-10k) from the DarkWeb. This dataset proposes the task of classifying legal and illegal activities on the domain of forums and drug markets. To compare with the data from surface web, Nabki et al. (2019) have extracted item descriptions from eBay as well. The descriptions were selected by searching the keywords (marijuana, weed, grass, and drug); these were divided by paragraphs and filtered, producing a corpus without repetition. The texts of the corpus were extracted from links provided by Choshen et al. (2019) [1] and pre-processed by removing: HTML tags, non-linguistic content such as buttons, encryption keys, metadata, and common words such as "Show more results".

The corpus DUTA-10k contains data collected and divided into five different subsets: (1) eBay items, (2) legal drugs, (3) illegal drugs,(4) forums discussing legal activities and (5) forums discussing illegal topics. The number of samples of each dataset and their corresponding categories is presented in table1. Since the aim is to classify legal vs illegal activities (Choshen et al., 2019), the subsets are used for four different experiments: (1) eBay vs. legal drugs, (2) legal vs illegal drugs, (3) legal vs illegal forums and finally, (4) legal and illegal drugs training data vs the test set of legal and illegal forums.

| dataset | # tokens | # samples | # samples in class | |
|---|---|---|---|---|
| Ebay vs legal drugs - train | 24,795 | 924 | Ebay 456 | legal drugs 468 |
| - dev | 2,623 | 103 | 53 | 50 |
| - test | 2,802 | 115 | 62 | 53 |
| Onion forums - train | 15,409 | 924 | illegal 468 | legal 456 |
| -dev | 1,478 | 103 | 50 | 53 |
| -test | 1,640 | 115 | 53 | 62 |
| Onion drugs - train | 25,582 | 924 | illegal 468 | legal 456 |
| -dev | 2,416 | 103 | 50 | 53 |
| -test | 2,995 | 115 | 53 | 62 |

Table 1: Distribution of examples and classes

| *Corpus* | *Size* |
|---|---|
| BooksCorpus (Zhu et al., 2015b) | 800M words |
| 2010-and-2014-English Wikipedia dump | 2,500M words |
| Giga5 (Parker et al., 2011) | 16GB |
| Common Crawl (Crawl, 2019) | 110GB |
| ClueWeb (Callan et al., 2009) | 19GB |
| Penn Treebank (Marcus et al., 1993) | 1M words |

Table 2: Pre-traning corpora with their size. All corpora are derived from the surface web.

### 2.2 Methods: Classification Models

This section introduces the models which we used to investigate the role of pre-training in transformers when applied to truly uncovered texts.

**Stylistic Classifier** Legal and illegal activities may be described with different styles of language: a formal language vs a more informal style of writing. For this reason, we tested an SVM classifier that uses some stylistic characteristics captured from the surface properties of the tokens. This classifier is used to determine if analyzed tasks are purely stylistic.

*Bleaching text* (van der Goot et al., 2018) is a model proposed to capture the style of writing at the word level. Originally, it has been applied for cross-lingual author's gender prediction. To capture the style, this model converts sequences of tokens, e.g., '1x Pcs Mobile Case!? US$65', into abstract sequences according to the following rules presented with the effect on the example: (1) each token is replaced by its length (effect: '02 03 06 06 05'); (2) alphanumeric characters are merged into one single letter and other characters are kept (effect: 'W W W W!? W$W'); (3) punctuation marks are transformed into a unified character (effect: 'W W W WPP W'; (4) upper case letters are replaced with 'u', lower case letters with 'l', digits with 'd', and the rest to 'x' (effect: 'dl ull ull ullxx uuxdd'); (5) consonants are replaced with 'c', vowels to 'v' and the rest to 'o' (effect: 'oc ccc cvcvcv cvcvoo vcooo'). Finally, a sample is represented by the concatenation of all the above transformations. For classification, we use a linear SVM classifier with a binary bag of word representation.

**Lexical-based Neural Networks** To investigate the role of pre-trained word embeddings, we used a classifier based on a vanilla feed-forward neu-

[1] data and code are available in Choshen et al. (2019) GitHub repository `https://github.com/huji-nlp/cyber`

ral networks (FFN) over a bag-of-word-embedding (BoE) representation of sentences. In BoE, sentence representations are computed as the sum of word embeddings representing their words.

We experimented with two versions of the classifier: BoE(GloVe) and BoE(re-train). BoE(GloVe) uses GloVe word embeddings (Pennington et al., 2014) trained on 2014 Wikipedia dumps and Giga5 (see Table 1). BoE(re-train) uses word embeddings learnt on the novel corpus using a CBOW model of word2vec (Mikolov et al., 2013). This latter model is trained with 300 dimensions for 5 epochs.

The supporting FFNs of BoE(GloVe) and BoE(re-train) are slightly different. In BoE(GloVe), the FFN consists of an input layer of dimension 300, 2 hidden layers of 150 and 50 dimensions with the $ReLU$ activation function. In the BoE(re-train), the FFN consists of two layers of 150 neurons. $tanh$ activation function is used for each layer.

**Syntactic-based Neural Networks** To evaluate the role of "pre-trained" universal syntactic models, we used the Kernel-inspired Encoder with Recursive Mechanism for Interpretable Trees (KERMIT) (Zanzotto et al., 2020). This model positively exploits parse trees in neural networks as it increases performances of pre-trained Transformers when it is used in combined models.

The version used in the experiments encodes parse trees in vectors of 4,000 dimensions. The rest of the feed-forward network is composed of 2 hidden layers of dimension 4,000 and 2,000 respectively, finally the output layer of dimension 2. Between each layer the $ReLU$ activation function and a dropout of 0.1 is used to avoid overfitting on the train data.

Even in this case, the model is somehow 'pre-trained'. In fact, KERMIT exploits parse trees produced by a traditional parser. In our experiments, we used the English constituency-based parser in CoreNLP (Zhu et al., 2013). The parser is trained on the standard WSJ Penn Treebank (Marcus et al., 1993), which contains only around 1M words.

**Holistic Transformers** We tested the following Transformers to cover the majority of cases of pre-training size (see Table 2) and models:

- $BERT_{base}$ (Devlin et al., 2019), the architecture Bidirectional Encoder Representations from Transformers, trained on the BooksCorpus (Zhu et al., 2015b) and English Wikipedia and the Multi-lingual $BERT_{multi}$ (Pires et al., 2019) trained on a Wikipedia dump of 100 languages. Both implementations are from the Huggingface's Transformers library (Wolf et al., 2019);

- XLNet (Yang et al., 2019), which is based on a generalised autoregressive pre-training technique that allows the learning of bidirectional contexts by maximising the expected likelihood over all permutations of the factorization order and to its autoregressive formulation. XLNet is trained on 32.89 billion tokens, taken from datasets gathered from the surface web or publicly available datasets, such as Wikipedia, Bookcorpus, Giga5, Clueweb and Common Crawl.

- ERNIE (Sun et al., 2021) introduced a language model representation that addresses the inadequacy of BERT and utilises external knowledge graph for named entities. ERNIE is pre-trained on Wikipedia corpus and Wikidata knowledge base.

- ELECTRA (Clark et al., 2020) Compared to BERT, instead of masking an input token, they "corrupt" it by replacing it with a token that potentially fits the place. Training procedure is a classification of each token on if it is a corrupted input or not. To make its performance comparable to BERT, they have trained the model on the same dataset that BERT was trained on.

## 3 Results and Discussion

We explored the performance of all the pre-trained models on the dataset and the tasks described in section 2.1. Results reported in Table 3 show unexpected behavior of these models.

The proposed tasks cannot be solved using only stylistic features. Stylistic models are performing worse with respect to lexical, syntactic and combined models in three tasks out of four. The task where stylistic models are performing better is the one where models are trained on legal/illegal Drugs and tested on legal/illegal Forums. In this case, lexicon only cannot help in drawing decisions and stylistic features are useful discriminating factors.

General lexical knowledge is basically important when dealing with completely novel texts. Indeed, pre-trained lexical models have generally

| | eBay/Legal Drugs | Drugs | Forums | Drugs/Forums |
|---|---|---|---|---|
| *NB (POS)* (Choshen et al., 2019) | 91.4 | 77.6 | 74.1 | **78.4** |
| *SVM (POS)* (Choshen et al., 2019) | 63.8 | 63.8 | **85.3** | 62.1 |
| ***Holistic Transformers*** | | | | |
| $BERT_{base}$ | 65.30(±2.6) | 64.63(±3.4) | 52.60(±0.7) | 47.40(±3.93) |
| $BERT_{multi}$ | 49.50(±2.3) | 51.30(±2.93) | 51.32(±2.42) | 48.29(±3.85) |
| *Electra* | 70.20(3.8) | 58.60(±4.36) | 52.70(±2.84) | 49.39(±4.62) |
| *XLNet* | 57.30(±3.6) | 54.30(±2.77) | 51.60(±1.93) | 50.83(±2.68) |
| *Ernie* | 67.65(±4.73) | 56.87(±4.29) | 50.61(±3.8) | 48.25(±2.53) |
| ***Lexical Models*** | | | | |
| *BoE(GloVe)* | 91.50(±0.5) | 81.60(±1.4) | 54.60(±1.4) | 53.50(±1.5) |
| *BoE(re-trained)* | 87.13(±0.01) | 74.08(±0.01) | 57.22(±0.01) | 50.26(±0.02) |
| ***Syntactic Models:** KERMIT* | 90.50(±1.0) | 79.00(±1.0) | 66.60(±1.4) | 58.37(±1.26) |
| ***Stylistic models:** Bleaching text* | 81.73 | 79.13 | 55.65 | 54.78 |
| ***Lexical and Syntactic Models*** | | | | |
| *BoE(GloVe) + KERMIT* | **93.54**(±1.46) | **83.10**(±1.4) | 66.20(±1.4) | 54.30(±2.34) |
| *BoE(re-trained)+KERMIT* | 88.69(±1.23) | 80.03(±0.97) | 58.50(±1.4) | 52.34(±2.3) |

Table 3: Accuracy of the different pre-trained models on the Legal vs. Illegal Classification Task on the DarkWeb Corpus (Choshen et al., 2019). The first two lines are results provided in (Choshen et al., 2019). Experiments with neural networks are obtained over 5 runs with different seeds.

higher results with respect to re-trained lexical models: BoE(Glove) outperforms BoE(re-trained) on three out of the four tasks (see Table 3). Hence, re-training word embeddings with a small corpus seem to be useless. In fact, re-training adds information in only one sub-task: dealing with legal vs. illegal forums (57.22 vs. 54.60).

Surprisingly, holistic Transformers have poor performance on this totally uncovered corpus and on the defined tasks. $BERT_{base}$, $BERT_{multi}$, $Electra$, $XLNet$ and $Ernie$ have worse performances with respect to all the other models. Considering that there is an overlap between the data used for training the *BoE(GloVe)* model and the transformer-based models, their poor performance is unexpected.

However, neural network models based on syntax have extremely interesting performances on this dataset. KERMIT (Zanzotto et al., 2020) behaves better than holistic Transformers, showing that these tasks are sensitive with respect to syntactic information. The major difference is that KERMIT uses a parser (Manning et al., 2014), which is pre-trained on a definitely smaller training set.

Moreover, the combined "pre-trained" lexical and syntactic model, that is, BoE(GloVe) + KERMIT, outperforms previous state-of-the-art on two subtasks out of four. This shows that the two combined models can exploit their pre-training on totally new, unseen language and tasks.

In conclusion, selected tasks are on a completely novel dataset and are sensitive with respect to lexical and syntactic information. Yet, pre-trained Transformers seem not to be able to solve these tasks, although these Transformers are able to deal with lexical and syntactic information (Jawahar et al., 2019; Hewitt and Manning, 2019; Hu et al., 2020). This contradiction seems to be a possible evidence of the fact that large pre-training may force Transformers to overfit on seen data. This overfitting possibly happens at the sentence level so they cannot capture stylistic and syntactic differences.

## 4 Conclusion

Transformers are successful on many downstream tasks, and it stems from the huge corpora that they are trained on. As a result, investigation of their strengths and weaknesses is important. In this paper, we aimed to explore how pre-trained natural language understanding models perform in totally unknown and unprecedented contexts, such as the DarkNet. We conducted extensive experiments to investigate the performance of stylistic, lexical style, syntactic, and holistic approaches. The results show that syntactic and lexical neural networks surprisingly outperform pre-trained Transformers, which indicates that Transformers have difficulty adapting to unknown texts.

## References


Georgia Avarikioti, Roman Brunner, Aggelos Kiayias, Roger Wattenhofer, and Dionysis Zindros. 2018. Structure and content of the visible darknet.

Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. Clueweb09 data set.

Leshem Choshen, Dan J. Eldad, Daniel Hershcovich, Elior Sulem, and Omri Abend. 2019. The language of legal and illegal activity on the darknet. In *ACL*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Common Crawl. 2019. Common crawl. URL: http://commoncrawl.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mhd Wesam Al Nabki, Eduardo FIDALGO, Enrique Alegre, and Laura Fernández-Robles. 2019. Torank: Identifying the most influential suspicious domains in the tor network. *Expert Syst. Appl.*, 123:212–226.

R Parker, D Graff, J Kong, K Chen, and K Maeda. 2011. English gigaword fifth edition ldc2011t07 (tech. rep.). Technical report, Technical Report. Linguistic Data Consortium, Philadelphia.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert?

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Mohammad Shoeybi, Mostofa Ali Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-LM: Training multi-billion parameter language models using model parallelism. *ArXiv*, abs/1909.08053.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *ArXiv*, abs/2107.02137.

Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Melbourne, Australia. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.0.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. 2020. KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 256–267, Online. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities.

Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo Zhu. 2013. Fast and accurate shift-reduce constituent parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 434–443, Sofia, Bulgaria. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015a. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015b. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.