Classifying Scientific Publications with BERT - Is Self-attention a Feature Selection Method?

Andres Garcia-Silva(✉) and Jose Manuel Gomez-Perez

Research Lab, Expert.ai, Prof. Waskman 10, 28036 Madrid, Spain
{agarcia,jmgomez}@expert.ai
https://www.expert.ai

Abstract. We investigate the self-attention mechanism of BERT in a fine-tuning scenario for the classification of scientific articles over a taxonomy of research disciplines. We observe how self-attention focuses on words that are highly related to the domain of the article. Particularly, a small subset of vocabulary words tends to receive most of the attention. We compare and evaluate the subset of the most attended words with feature selection methods normally used for text classification in order to characterize self-attention as a possible feature selection approach. Using ConceptNet as ground truth, we also find that attended words are more related to the research fields of the articles. However, conventional feature selection methods are still a better option to learn classifiers from scratch. This result suggests that, while self-attention identifies domain-relevant terms, the discriminatory information in BERT is encoded in the contextualized outputs and the classification layer. It also raises the question whether injecting feature selection methods in the self-attention mechanism could further optimize single sequence classification using transformers.

Keywords: Neural language models · Text classification · Scholarly communications

1 Introduction

The annotation and classification of scientific literature is a crucial task to make scientific knowledge easily discoverable, accessible, and reusable, accelerating scientific breakthroughs by helping scholars locate and understand the right research, making connections, and overcoming information overload. Some examples of efforts to structure scientific literature include scientific search engines like Semantic Scholar [1] and Microsoft Academic [23]. Both rely on knowledge graphs to enable a structured representation of scientific knowledge that supports applications like topic-driven search and recommendation. Similarly, scientific publishers have released knowledge graphs such as SN SciGraph [7] in order to more effectively organize their publications and increase automation.

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12656, pp. 161–175, 2021.
https://doi.org/10.1007/978-3-030-72113-8_11

Other efforts like ORKG [8] rely on knowledge graphs to structure the actual contributions described in the publications, making research results on a specific topic comparable across the literature.

Publications are therefore being annotated with information about their content, which includes topics [1], fields of study [23], concepts [7], and research fields [8]. Such metadata is generally based on controlled vocabularies and arranged according to a taxonomy [7,8], thesaurus [1,23] or ontology [21]. In some cases, the annotation process can be fully automatic [1,23]. However, authors are often asked to manually classify their contribution in the right categories, which is tedious and error-prone. In other occasions, this task falls under the responsibility of a reduced number of senior expert editors, making the process expensive and slow [21].

In this paper, we focus on the task of classifying scientific publications against a taxonomy of scientific disciplines. A wide variety of approaches are suitable for this task, including machine learning classifiers that rely on high-dimensional sparse representations [10], deep learning classifiers using dense representations [11], and rule-based or heuristic methods [21]. Encouraged by the success of recent developments in natural language processing and understanding, where pre-trained transformer language models dominate the state of the art [27], herein we focus on BERT [5] and its different flavors specialized in the scientific domain: BioBERT [16] and SciBERT [2].

Our experiments confirm that using transformers to train scientific classifiers generally results in greater accuracies compared to linear classifiers that were until now regarded as strong baselines [11]. We also observe that fine-tuning pre-trained transformers on domain-specific corpora contributes to this goal. However, despite previous research focused on interpreting and understanding how transformers encode information [4,9,15,20,25], the actual mechanism by which fine-tuning impacts on our classification task is still unclear. In an effort to shed light on this matter, we focus on analyzing the self-attention mechanism inherent of the transformer architecture [26]. Our findings show that the last layer of BERT attends to words that are semantically relevant for the scientific fields associated with each publication. This observation suggests that self-attention actually performs some type of feature selection for the fine-tuned model.

We investigate the possible relation between self-attention and feature selection methods from different perspectives, including vocabulary overlap, ranking similarity, domain relevance, feature stability, and classification performance. Our results open a future research path to determine whether injecting feature selection methods in the self-attention mechanism could derive even better results for single sequence classification using transformer architectures.

Our main contributions in this paper are the following:

– We leverage the vertical pattern present in the transformer self-attention mechanism of BERT, SciBERT and BioBERT, where some words receive more attention on average than the rest of the words, and compare it against conventional feature selection methods used in text classification.

– We find that self-attention has interesting properties as a feature selection method. The most attended words are in general more relevant to the publication domain than those found using conventional approaches to feature selection. The stability of the features resulting from self-attention is in line with the results obtained through conventional approaches. However, when used to learn classifiers from scratch, methods like chi-square and information gain contribute to train better classifiers.
– We analyze from a semantic point of view the self-attention mechanism and quantify the amount of domain knowledge it encodes in the hidden states of the last layer. To this purpose, we rely on ConceptNet [24], a commonsense knowledge graph where attended words are mapped to concepts from which we derive their corresponding domains.

The remainder of the paper is structured as follows. Section 2 describes related work in the annotation of scientific publications, classification, transformer language models, and other work focused on the analysis of transformer self-attention. In Sect. 3, we present experimental results classifying research papers into a scientific taxonomy. In Sect. 4, we motivate the analysis of self-attention as feature selection with examples of attended words and scientific categories. In Sect. 5, we quantify the relation between self-attention and feature selection methods. Finally, Sect. 6 concludes the paper[1].

2 Related Work

Annotating research articles with entities such as research fields or topics is addressed in the literature using entity recognition and similarity measures between entity labels and their mentions [3]. In Microsoft Academic Graph [23] the candidate entities (field of study) are identified using string matching between the entity keywords and their paper mentions, then rules are applied to gather more candidates and to filter out the less relevant entities. Similarly, the CSO classifier [21], which assigns articles to concepts in the Computer Science Ontology[2], first identifies concepts explicitly mentioned in the text and then, in an effort to find entities not explicitly mentioned, it uses a similarity measure based on word embeddings. In the Semantic Scholar literature graph [1], an ensemble of tools is used to annotate entities: statistical models for entity span prediction and disambiguation, rules for string-based entity spotting, and off-the-shelf tools[3].

In addition, different models can be used for this task, including SVM [10] or softmax classifiers [14]. Mai et al. [17] proposed classifiers based on convolutional [13] and recurrent neural networks [30] to annotate research articles. However,

[1] Tables, datasets and notebooks to reproduce our experiments are available in https://github.com/expertailab/Is-BERT-self-attention-a-feature-selection-method.
[2] See http://cso.kmi.open.ac.uk/.
[3] https://sobigdata.d4science.org/web/tagme/tagme-help.

such deep learning classifiers need to be trained from scratch and depend on the network architecture. On the contrary, neural language models and particularly transformers like GPT-2 [19] or BERT [5] are pre-trained on a large corpus and then fine-tuned for classification by just adding a linear classifier to the model output. This approach has proven to successfully tackle several NLP tasks [27], including text classification. In the scientific domain, SciBERT [2] and BioBERT [16] have also reported state of the art results. Researchers are investigating the mechanics underlying BERT [20], analyzing its hidden states and outputs [9,25], as well as the self-attention mechanism [4,15]. Unlike previous approaches [4,15], we semantically analyze the words that are attended above average in the last hidden state, leveraging the commonsense knowledge represented in ConceptNet, and quantify the relation between attention and feature selection methods often used in text classification.

3 Fine-Tuning Language Models for Text Classification

We evaluate the use of language models on a text classification task where research articles are labeled with one or more knowledge fields. To this purpose, we choose: i) BERT and GPT-2, pre-trained on a general-purpose corpus, ii) SciBERT, pre-trained solely on scientific documents, and iii) BioBERT, pre-trained on a combination of general and scientific text. Table 1, provides relevant information about each language model, its pre-training and vocabulary. BioBERT uses the same tokenization method and vocabulary as BERT, while SciBERT adopts SentencePiece, based on WordPiece tokenization. The overlap between the vocabularies of BERT and SciBERT is 42%, which shows a substantial difference in the most frequently used words in the scientific domain and general-purpose documents. We choose the base version of BERT models

Table 1. Language models pre-training information.

Model	Tokenizer	Vocabulary	Corpus	Domains	steps/epochs
BERT	WordPIece	30K	BookCorpus (2.5B tokens) + Wikipedia (0.8B tokens)	General	1M steps
BioBERT 1.1	WordPiece	BERT	BERT corpus + PubMed abstracts (4.5B tokens)	General + Biomedic	1M steps
BioBERT 1.0	WordPiece	BERT	BERT Corpus + PubMed abstracts (4.5B tokens) + PMC full-text articles (13.5M tokens)	General + Biomedic	470K steps
SciBERT	SentencePiece	30K	Semantic Scholar (3.17B tokens) (1.14M full text papers)	18% Computer Science and 82% Biomedical	Not reported
GPT-2	Byte Pair Encoding (BPE)	50k	8 million web pages, except Wikipedia (40 GB of text)	General	Not reported

(12 layers, 768 hidden size, 12 attention heads per layer) and a comparable model for GPT-2.

To fine-tune BERT, BioBERT and SciBERT on our multilabel classification task, we follow the guidelines provided by Devlin et al. [5] for single-sentence classification. We take the last layer encoding of the classification token <CLS> and add an N-dimensional linear layer, with N the number of classification labels. We use a binary cross-entropy loss function to allow the model to assign independent probabilities to each label. For GPT-2 we also add a linear layer on top of the last hidden state for the classification token. We train the models for 4 epochs, with batch size 8 and 2e–5 learning rate.

As a baseline, we use an SVM with a linear kernel [6]. We follow a one-vs-all strategy to train a binary SVM classifier per category, with grid search for the regularization parameter. We use WordNet to lemmatize the words, whenever they exist in the WordNet lexicon, and remove stop words. In addition, we use fastText [11] to learn a hierarchical softmax classifier using n-gram embeddings. We learn binary classifiers for each category, with automatic hyperparameter optimization to fix learning rate, number of epochs, and n-gram length.

We gather our dataset of scientific articles from a broad range of knowledge fields in SciGraph [7], where articles are labelled following the ANZSRC[4] taxonomy. This taxonomy comprises 22 first level categories, such as *Economics*, *Law*, and *Computer Science*, each of them with their own subcategory tree. From SciGraph, we extract the titles and abstracts of articles published in 2011 and 2012, as well as their categories. In total, we gather 405K papers, 187K from 2011 and the rest from 2012. In average, each first level category has 20,164 articles with a standard deviation of 31,791, which shows how unevenly the different categories are covered. Some of them are well represented, like *Medical And Health Sciences*, with 138,728 articles, while others, like *Studies In Creative Arts And Writing*, have little over a hundred articles.

We fine-tune the language models to learn to classify papers on any of the 22 first level categories. We train on papers only from 2011 and evaluate using 5-fold cross validation. Table 2 shows that the transformers pre-trained on a scientific corpus generally achieve greater f-measure in this task. The exception is BioBERT-1.0, which scores under BERT. BioBERT-1.0 was pre-trained on a lower number of steps than the other transformers, which could be affecting its performance. GPT-2 is the model producing the lowest f-measure, which shows evidence of a potential mismatch between the vocabulary and quality of the scientific corpus and the Web corpus where it was pre-trained, which may be undermining its performance. Overall, transformers produce more accurate classifiers than the linear methods used as baselines.

To further explore the relation between the pre-training and fine-tuning corpora, we learn classifiers to label articles with second level categories in ANZSRC for some of the first level categories. For this experiment, we enlarge our dataset with articles published in 2012 and evaluate only the best language models, discarding BioBERT 1.0 and GPT-2. The results in Table 2 show that, in general,

[4] Australian and New Zealand Standard Research Classification.

Table 2. Evaluation results of the multilabel classifiers (f-measure) on first level categories (a), and on second level categories (b).

First level categories

Model	f-measure
SciBERT	**0.838**
BioBERT-1.1	0.825
BERT	0.819
BioBERT-1.0	0.818
GPT-2	0.808
SVM	0.807
fastText	0.790

Second level categories

Categories	Articles	Subcat.	Bert	BioBERT-1.1	SciBERT	SVM	fastText
Biological	65340	9	0.883	0.884	**0.887**	0.880	0.871
Medical and Health	58068	18	0.838	0.843	**0.854**	0.836	0.819
Chemical	40837	8	0.858	0.862	**0.865**	0.854	0.847
Mathematical	28723	5	0.886	0.883	**0.891**	0.884	0.878
Computer Sciences	20777	6	0.861	0.862	**0.864**	0.861	0.849
Language	2233	6	**0.911**	0.900	0.903	0.900	0.906
Hist. And Archelogy	2076	4	**0.955**	0.950	0.941	0.946	0.946
Built Environment	140	4	0.495	0.700	0.697	**0.808**	0.804
Creative Arts	132	4	0.639	0.788	0.781	**0.925**	0.828

scientific categories are dominated by SciBERT and BioBERT-1.1. However, for categories in humanities, e.g. *Language*, and *History and Archaeology*, BERT produces better classifiers, providing evidence that the general-purpose knowledge encoded in BERT is more relevant in those cases. Interestingly, when there are few examples, e.g., in categories *Built Environment* and *Creative Arts*, the general knowledge encoded in BERT is of little use for the classifiers, while the scientific knowledge in BioBERT-1.1 and SciBERT contributes to achieve higher f-measure. Linear classifiers outperform transformer-based models in such under-represented categories.

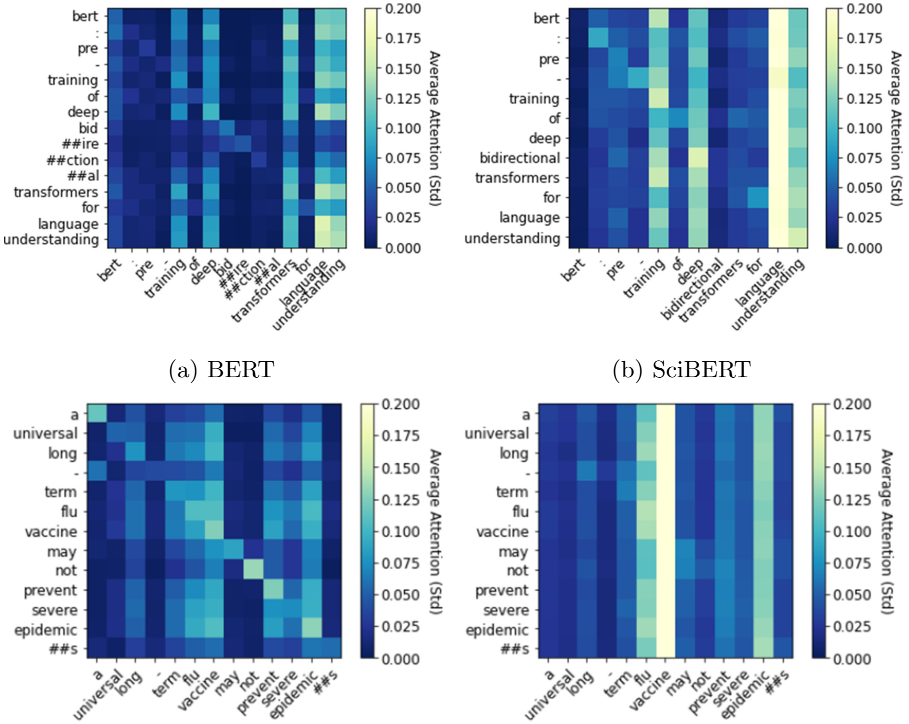
4 Exploring Self-attention Heads

Above we show that BERT-based models are able to produce high performance multilabel classifiers. However, we know little about what makes them good at this task. In this section, we inspect the self-attention mechanism underpinning such models as a key element to understanding this behavior.

According to Clark et al. [4], attention weights indicate how relevant a particular word is when computing the next representation for the current word. To illustrate this statement, Fig. 1, depicts the mean weights of the 12 self-attention heads in the last hidden state of the fine-tuned models for two papers titled *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"*, and *"A universal long-term flu vaccine may not prevent severe epidemics"*. The plots clearly show the so-called vertical pattern [15], where a few tokens receive most of the attention, such as *training, deep, transformer, language*, and *understanding* in the first sentence, and *flu, vaccine, prevent, severe* and *epidemic* in the second. Note how while the vocabulary captured by SciBERT includes the word *bidirectional*, BERT uses subwords to represent it.

We do not include special tokens <SEP> and <CLS> since the amount of attention received by these tokens makes the attention received by the other tokens barely noticeable. Clark et al. [4] speculate that the attention on <SEP> in one head could indicate that the attention heads function is not applicable, while Rogers et al. [20] interpret the attention on <CLS> as the attention on a pooled sentence-level representation.

From these two examples, we observe that the most attended words in the last hidden state are highly related to the research fields of the articles: *Computer science* and *Medical and Health Sciences*. So, we look into this relation and identify the words that receive most vertical attention in the last hidden state for a subset of our dataset where each first level category is represented with at most 500 papers. First, for each input sequence we calculate the mean weights for the 12 attention heads in the last hidden state. Next, we generate a new weight matrix grouping subwords into words by averaging the subword weights. Finally, we gather the words with a vertical mean attention above the mean attention in the weight matrix. This results in 8,840 attended words for BERT, 17,773 for BioBERT, and 12,265 for SciBERT, corresponding to 16%, 32%, and 22% of the vocabulary managed by each language model.



(a) BERT (b) SciBERT

(c) BioBERT-1.1 (d) SciBERT

Fig. 1. Average weights in the self-attention heads of the last hidden state.

Table 3 shows the top 20 most frequent attended words in three research fields: *Biology*, *Computer Science* and *History and Archaeology*. As can be noted, most of such words are highly related to the specific research field, appearing along a few punctuation marks and some stop words. While frequent attention to periods and commas was already reported in [4,15], the reason why this happens is not clear yet. Rogers et al. [20] suggest that it must be related to model overparameterization while Clark et al. [4] point at the high frequency of these tokens in the corpus. Stop words are also highly frequent words and the models could be learning to attend to them as in the case of punctuation marks.

Table 3. Most attended words above average attention in the fine-tuned models.

06 - Biological sciences			08 - Computer science			21 - History and archaeology		
BERT	BioBERT	SciBERT	BERT	BioBERT	SciBERT	BERT	BioBERT	SciBERT
,	of	.	,	the	.	,	the	.
Species	The	,	.	of	,	.	of	,
Gene	In	Gene	Data	Data	Data	History	In	History
Cell	.	Species	)	-	Information	)	History	Century
Cells	To	Cell	Image	Time	Algorithm	Historical	-	Historical
Protein	Species	The	network	Information	Network	Archaeological	To	Modern
.	And	Protein	Information	Model	Image	Cultural	Century	The
Genetic	For	Expression	Networks	System	Algorithms	The	.	Archaeological
Plants	-	Genes	Control	Algorithm	As	Social	and	Social
Plant	Gene	Genetic	Images	In	Networks	Archaeology	A	Cultural
Expression	A	Cells	Algorithms	Systems	Systems	Political	Historical	American
Growth	Cell	Growth	Software	Based	Model	Culture	Period	Human
Genes	Protein	Plants	Neural	A	Analysis	Women	Early	Literature
)	Genes	Plant	Optimization	Network	Software	Literary	Modern	Data
Molecular	Cells	Dna	Simulation	To	Time	Heritage	Archaeological	State
Dna	On	Proteins	Learning	Algorithms	Images	Precipitation	World	Women
Stress	With	Molecular	Search	Analysis	Control	Education	Social	Life
Populations	Expression	Populations	Web	Image	Simulation	Identity	On	Period
Population	Genetic	Population	A	Models	Problems	Literature	Years	Political
Genome	Plants	Water	Classification	User	Such	Past	American	Development

5 Feature Selection

In the previous section we show that fine-tuned BERT models concentrate their attention on a subset of the overall vocabulary that ranges between 16% to 32% of the words. Following this observation, we hypothesize that such attention on a selected fragment of the vocabulary is the transformer version of feature selection. However, rather than picking the most interesting features for a classifier, self-attention selects words that heavily influence the representation of the rest of the words in the same sequence. We investigate whether there is a relation between feature selection algorithms commonly used for text classification and the most attended words in the fine-tuned language models.

We center our analysis on four feature selection methods used for text classification [14,18,22]: Chi-square (chi), Information Gain (ig), Document Frequency (df), and Categorical Proportional Difference (pd). Chi-square measures the lack of independence between a word and a class; its value is zero if the word and the class are independent. Information Gain measures the entropy reduction of the dataset when it is split by a feature value. Thus, words with larger information gain discriminate the data ensuring a lower entropy. Document Frequency counts the number of documents where a term appears. Categorical Proportional Difference measures the degree to which a word contributes to differentiating a particular category from others.

We compare the most attended words with those selected by the above-mentioned feature selection methods, and measure how similar the rankings of words sorted by their average attention are to the rankings produced by each feature selection method. In Table 4, we report the vocabulary overlap of the most attended words and feature selection methods after filtering out the stop words. The number of features selected was limited to the top k words, where k is the number of words attended above average by each language model. Indeed, the results indicate a large overlap. Fine-tuned language models for text classification attend up to 64% of the common terms returned by dc, the most simple of our feature selection baselines, which itself performs similarly to ig and chi [29]. For all three models, their most attended words have the largest overlap with document frequency, followed by information gain, chi-square and, finally, proportional difference.

Table 4. Word overlap: most attended vs. feature selection.

LM	FS	%
BERT	dc	60%
	ig	54%
	chi	43%
	pd	12%
BioBERT-1.1	dc	64%
	ig	55%
	chi	44%
	pd	25%
SciBERT	dc	58%
	ig	49%
	chi	42%
	pd	20%

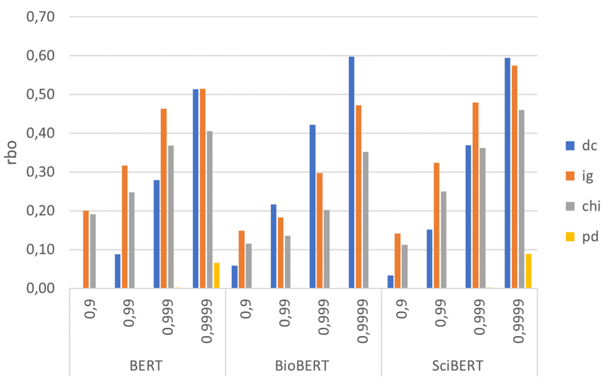


Fig. 2. Rank-biased overlap at different p values between most attended words and feature selection algorithms.

To measure the similarity between rankings we apply the Rank-Biased Overlap (RBO) [28] metric. RBO ranges between 0 to 1, from less to more similar, and was designed for non-conjoint rankings, i.e. both lists may have different items, may be incomplete and with different length. Through the p parameter, RBO models the probability to continue considering the overlap at the next rank, having examined the overlap at the previous rank. Figure 2 shows the RBO for the attention and feature selection rankings. We set p to 0.9, 0.99, 0.999, and 0.9999, indicating the model to assign the first 10, 100, 1,000, and 10,000 ranks respectively, approximately 85% to 86% of the weight of the evaluation.

While the BERT and SciBERT attended words rankings are more similar to the ranking of discriminative words (ig) for p values of 0.9 to 0.999, they finally converge with the ranking of common terms (dc), too. On the other hand, the BioBERT-1.1 ranking is clearly most similar to the common term rankings (dc). We think that the difference between the three models could be related to the subword vocabulary and pre-training corpus. Subword vocabularies are tightly related to the training corpus since they are generated to represent the whole corpus with the minimum number of word pieces. BERT trains its own subword vocabulary on a general corpus and during fine-tuning learns to attend more to discriminative words in the scientific domain. SciBERT also uses its own vocabulary trained on a limited scientific corpus, enabling the model to attend to discriminative words (like BERT) but also to common words due to the domain knowledge it encodes. BioBERT on the other hand reuses the BERT subword vocabulary and therefore many scientific terms are split in a suboptimal number of pieces. This has a negative impact on the ability of the self-attention mechanism to focus on discriminative words, and subsequently on the attention to common terms.

5.1 Domain Knowledge

We investigate the domain relevance of the words that are most attended by the language models and compare it with words produced by the feature selection methods. To this end, we search the words in ConceptNet and leverage the relation *HasContext* to identify the domains where they are commonly used. We manually map the 22 first level categories in ANZSRC to the corresponding concepts in ConceptNet. To deal with morphological variations like plurals and conjugations we use the *FormOf* relation, and to increase the coverage we traverse the *isA* type hierarchy one level up looking for the corresponding concept. For example, the word *networking* is a *FormOf* of the root word *network*, which in turn *HasContext Computer Science* and *Electronics*, and the concept *Electronics isA* type of *Physics*.

For each first level category, we gather the top 100 most attended words, as well as those with the highest scores according to each feature selection method. Then, for each word, we look for the corresponding context according to ConceptNet. Table 5 reports the domain relevance obtained for each category. In BERT and SciBERT, self-attention identifies more domain-relevant words than feature selection methods. However, this is not the case for BioBERT. Recall that in our sample dataset, the set of most attended words produced by BioBERT is the largest (32%) with respect to the vocabulary, which is a clear indication that the model spreads its attention more widely. Weighing the words by their term frequency (TF), attended words remain more domain-relevant than those obtained through feature selection. In fact, the domain relevance of the frequent attended

Table 5. Words per category matching the corresponding ConceptNet context.

	Mean Self-Att.			Feat. Sel.				Self-Att. (TF)			TF				TF/IDF			
Category	BERT	BioB.	SciB.	dc	ig	chi	pd	BERT	BioB.	SciB.	dc	ig	chi	pd	dc	ig	chi	pd
Mathematics	36	18	28	29	18	16	25	60	54	53	51	52	53	33	53	53	55	35
Physics	21	4	22	18	11	13	20	41	42	38	33	33	38	18	41	41	42	18
Chemistry	20	7	18	7	15	16	20	29	30	27	24	24	25	36	27	27	29	37
Biology	18	6	18	11	15	14	11	44	43	38	25	24	28	14	34	33	35	16
Agriculture	1	1	0	0	0	0	1	4	4	4	1	1	1	0	3	3	3	0
Comp. Science	6	4	7	11	5	5	4	20	17	18	14	14	16	11	15	15	16	12
Technology	5	0	3	1	1	1	0	3	2	2	1	1	1	1	1	1	2	0
Medicine	16	13	22	11	12	15	11	30	28	32	19	19	20	17	21	21	22	20
Education	2	1	1	1	1	1	3	4	8	6	4	4	4	1	5	5	5	4
Economics	2	4	1	1	2	2	0	8	10	9	8	8	7	0	9	9	9	0
Commerce	7	4	2	0	1	1	0	6	6	7	3	3	4	2	6	6	6	2
Psychology	2	2	0	4	1	0	2	8	7	5	6	6	7	7	9	9	9	7
Law	5	6	2	6	3	4	7	9	7	8	9	9	8	7	8	8	9	8
Literature	1	0	0	1	1	0	2	0	0	1	1	1	1	1	0	0	0	1
Language	1	0	0	0	2	2	0	2	1	1	2	2	2	0	1	1	1	0
History	10	9	12	23	11	9	11	11	21	21	26	25	26	16	26	25	26	15
Philosophy	16	0	7	15	6	6	10	17	15	18	19	19	18	10	20	20	22	10
Total	169	79	143	139	105	105	127	**296**	**295**	288	246	245	259	174	279	277	**291**	185

words is greater or on pair with those selected when TF/IDF is used to weigh the output of feature selection methods: self-attention takes into account not only the importance of words in the document (TF) but also their importance in the document collection (IDF).

5.2 Feature Evaluation

To evaluate the quality of the resulting features we measure their stability and their classification performance. Stability is the robustness of a feature subset generated from different training sets from the same distribution [12]. To measure stability we compute the mean Jaccard coefficient between the different subsets of words generated by each method. We apply 5-fold cross-validation and process each fold with the fine-tuned language models and the feature selection methods. Stability is reported on Table 6, where we can see that language models attend to the same words with stability values in line with those reported by document count. Attended words are more stable than the rest of the feature selection methods, including chi-square and information gain, which seems to be more volatile across folds.

Table 6. Stability of the features measured using Jackard similarity coefficient

SciBERT	BioBERT	BERT	dc	pd	ig	chi
0.87	0.84	0.83	0.86	0.77	0.65	0.58

In addition, we use the set of features to learn classifiers for the 22 first level categories using Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Neural Networks (NN), and SVM. The neural network comprises an embedding matrix of 100 dimensions and a fully connected layer using sigmoid as activation function. For the SVM the regularization parameter is tuned and for the remaining algorithms we use the recommended settings. We evaluate the classifiers using 5-fold cross validation on the subset of documents where each category was represented with up to 500 papers. The f-measure of the classifiers is shown in Fig. 3. In general, we observe that traditional feature selection methods like chi-square and information gain mainly help to learn more accurate classifiers than the set of most attended words by the language models. This observation clearly indicates that the success of BERT models in this task is not only driven by the self-attention mechanism but also by the contextualized outputs of the transformer, which are the input of the added classification layer.

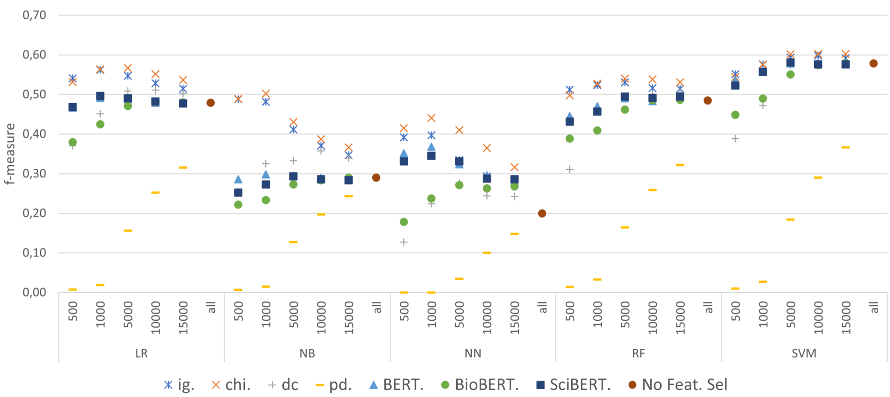


Fig. 3. Classifiers performance using distinct feature sets and number of features.

6 Conclusions

In this paper, we investigate the self-attention mechanism of BERT in a fine-tuning scenario for the classification of scientific articles over a taxonomy of research fields. We observe that attention in the fine-tuned model is focused on words that are highly relevant to the research field of each article. Furthermore, we notice that the most attended words represent just a fraction of the whole vocabulary: a hint that self-attention performs a sort of feature selection.

We systematically compare the most attended words against those resulting from feature selection methods normally used in text classification. We show that language models and feature selection methods like information gain and chi-square share between 42% to 55% of the selected words. We also observe that the attention-based word rankings produced by the transformers are more similar to those obtained using document frequency and information gain.

From our experiments we conclude that self-attention focuses more on words that are relevant to each research domain than the words produced through conventional feature selection. However, self-attention is not as good to learn classifiers from scratch, especially compared to chi-square and information gain. While self-attention identifies domain-relevant terms the discriminatory information in the fine-tuned model is encoded on the output representations and the additional classification layer. As future work, we plan to investigate the impact of integrating, perhaps as part of the loss function, optimal feature selection methods during fine-tuning of transformer for single sequence classification.

Acknowledgment. We gratefully acknowledge the EU Horizon 2020 research and innovation programme under grant agreement No. 825627 (ELG). We also thank Raul Ortega and Cristian Berrio for their contributions to the experimental evaluation.

References

1. Ammar, W., et al.: Construction of the literature graph in semantic scholar. In: NAACL-HLT (2018)
2. Beltagy, I., Lo, K., Cohan, A.: Scibert: pretrained language model for scientific text. In: EMNLP (2019)
3. Chernyak, E.: An approach to the problem of annotation of research publications. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, pp. 429–434. Association for Computing Machinery, New York (2015). https://doi.org/10.1145/2684822.2697032
4. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does BERT look at? an analysis of bert's attention. CoRR abs/1906.04341 (2019). http://arxiv.org/abs/1906.04341
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
7. Hammond, T., Pasin, M., Theodoridis, E.: Data integration and disintegration: Managing springer nature sciGraph with SHACL and OWL. In: Nikitina, N., Song, D., Fokoue, A., Haase, P. (eds.) International Semantic Web Conference (Posters, Demos and Industry Tracks). CEUR Workshop Proceedings, vol. 1963. CEUR-WS.org (2017). http://dblp.uni-trier.de/db/conf/semweb/iswc2017p.html#HammondPT17
8. Jaradeh, M.Y., et al.: Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, pp. 243–246. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3360901.3364435
9. Jawahar, G., Sagot, B., Seddah, D.: What does BERT learn about the structure of language? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3651–3657. Association for Computational Linguistics, Florence, July 2019. https://doi.org/10.18653/v1/P19-1356, https://www.aclweb.org/anthology/P19-1356

10. Joachims, T.: Training linear SVMS in linear time. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2006, pp. 217–226. Association for Computing Machinery, New York (2006). https://doi.org/10.1145/1150402.1150429
11. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
12. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. **12**(1), 95–116 (2007)
13. Kim, Y.: Convolutional neural networks for sentence classification. In: EMNLP (2014)
14. Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., Alsaadi, F.E.: Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. Appl. Soft Comput. **86**, 105836 (2020). https://doi.org/10.1016/j.asoc.2019.105836. http://www.sciencedirect.com/science/article/pii/S1568494619306179
15. Kovaleva, O., Romanov, A., Rogers, A., Rumshisky, A.: 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), pp. 4365–4374. Association for Computational Linguistics, Hong Kong, November 2019. https://doi.org/10.18653/v1/D19-1445, https://www.aclweb.org/anthology/D19-1445
16. Lee, J., et al.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
17. Mai, F., Galke, L., Scherp, A.: Using deep learning for title-based semantic subject indexing to reach competitive performance to full-text. CoRR abs/1801.06717 (2018). http://arxiv.org/abs/1801.06717
18. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, USA (2008)
19. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
20. Rogers, A., Kovaleva, O., Rumshisky, A.: A primer in bertology: what we know about how bert works. Trans. Assoc. Comput. Linguist. **8**, 842–866 (2020). https://doi.org/10.1162/tacl_a_00349
21. Salatino, A.A., Osborne, F., Birukou, A., Motta, E.: Improving editorial workflow and metadata quality at springer nature. In: Ghidini, C., et al. (eds.) ISWC 2019. LNCS, vol. 11779, pp. 507–525. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30796-7_31
22. Simeon, M., Hilderman, R.J.: Categorical proportional difference: a feature selection method for text categorization. In: AusDM (2008)
23. Sinha, A., et al.: An overview of microsoft academic service (MAS) and applications. In: Proceedings of the 24th International Conference on World Wide Web, WWW 2015 Companion, pp. 243–246. Association for Computing Machinery, New York (2015). https://doi.org/10.1145/2740908.2742839
24. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: an open multilingual graph of general knowledge. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI 2017, pp. 4444–4451. AAAI Press (2017)
25. Tenney, I., Das, D., Pavlick, E.: BERT rediscovers the classical NLP pipeline. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4593–4601. Association for Computational Linguistics, Florence, July 2019. https://doi.org/10.18653/v1/P19-1452, https://www.aclweb.org/anthology/P19-1452

26. Vaswani, A., et al.: Attention is all you need. CoRR abs/1706.03762 (2017). http://arxiv.org/abs/1706.03762
27. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: 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, pp. 353–355. Association for Computational Linguistics, Brussels, November 2018. https://doi.org/10.18653/v1/W18-5446, https://www.aclweb.org/anthology/W18-5446
28. Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. (TOIS) **28**(4), 1–38 (2010)
29. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning, ICML 1997, pp. 412–420. Morgan Kaufmann Publishers Inc., San Francisco (1997)
30. Yin, W., Kann, K., Yu, M., Schütze, H.: Comparative study of CNN and RNN for natural language processing. arXiv preprint arXiv:1702.01923 (2017)