Attention Mechanism with BERT for Content Annotation and Categorization of Pregnancy-Related Questions on a Community Q&A Site

Xiao Luo
Department of CIT
IUPUI
Indianapolis, USA
luo25@iupui.edu

Haoran Ding
Department of ECE
IUPUI
Indianapolis, USA
hd10@iu.edu

Matthew Tang
Department of CS
University of Illinois at Urbana-Champaign
Champaign, USA
mt13@illinois.edu

Priyanka Gandhi
Department of CIS
IUPUI
Indianapolis, USA
prgandh@iu.edu

Zhan Zhang
Department of CSIS
Pace University
New York City, USA
zzhang@pace.edu

Zhe He
School of Information
Florida State University
Tallahassee, USA
zhe@fsu.edu

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313379

***Abstract*—In recent years, the social web has been increasingly used for health information seeking, sharing, and subsequent health-related research. Women often use the Internet or social networking sites to seek information related to pregnancy in different stages. They may ask questions about birth control, trying to conceive, labor, or taking care of a newborn or baby. Classifying different types of questions about pregnancy information (e.g., before, during, and after pregnancy) can inform the design of social media and professional websites for pregnancy education and support. This research aims to investigate the attention mechanism built-in or added on top of the BERT model in classifying and annotating the pregnancy-related questions posted on a community Q&A site. We evaluated two BERT-based models and compared them against the traditional machine learning models for question classification. Most importantly, we investigated two attention mechanisms: the built-in self-attention mechanism of BERT and the additional attention layer on top of BERT for relevant term annotation. The classification performance showed that the BERT-based models worked better than the traditional models, and BERT with an additional attention layer can achieve higher overall precision than the basic BERT model. The results also showed that both attention mechanisms work differently on annotating relevant content, and they could serve as feature selection methods for text mining in general.**

***Index Terms*—AI Interpretation, Content Annotation, Consumer's Question Classification, NLP**

I. Introduction

Social media has become a popular tool that enables users' creation and exchange of information. Social media allows users to form groups or online communities to provide information and emotional support to peers. In recent years, the social web has been increasingly used for health information seeking, sharing, and health-related research [1]. Research has shown that women often use social media platforms to seek pregnancy-related information, share experiences, and communicate with each other [2] [3] [4]. The stages discussed on these platforms range from before, during, and after pregnancy and may include information related to fertility, symptoms, diet/nutrition, fetal growth and development, labor, breastfeeding, and newborns. Categorizing the questions to a relevant category and annotating the important words would help people find the key information in a large amount of content efficiently and inform the design of question retrieving and routing tools to meet information needs. The extant research mostly focuses on qualitative analysis of the information obtained from a social media website [4] or statistical analysis on the characteristics of specific population subgroups [5]. However, computational methods to facilitate pregnancy-related information retrieval and information seeking are scarce.

In this study, we collect pregnancy-related questions from the Yahoo! Answers [6] and explore different attention mechanisms with BERT to annotate the important words in the questions. Since 2018, the BERT model [7] has been used effectively in many natural language processing (NLP) tasks including question classification [8]. However, like typical deep neural networks, it is still a "black box", posing challenges to interpret the classification decisions. In this research, we propose to use the attention mechanism to interpret the classification. We investigate two attention mechanisms with BERT: (1) built-in self-attention of BERT and (2) a newly proposed BERT-Attention model, which applies an attention layer to the output of base BERT to improve the interpretability of the classification by identifying the salient terms that are important to the corresponding class. Both attention mechanisms can be used to annotate the relevant terms in the context of pregnancy-related questions. However, the built-in self-attentions do not explicitly provide the interpretation

This project was partically supported by NSF REU award 1852105, and UF-FSU CTSA Hub under award number UL1TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

(salient words) to the classification decisions. In contrast, the additional attention layer on top of BERT explicitly links to the classification layer that drives the classification outcomes.

We evaluate the performances of the BERT models and the baseline traditional machine learning models on eight categories of the collected pregnancy-related questions. The classification results show that the BERT models performed better than the traditional models. We then select correctly and incorrectly classified questions with highlighted terms identified by two different attention mechanisms to evaluate the content annotation. The case study demonstrates that both attention mechanisms can track the words relevant to the main topics within the questions. However, different attention mechanisms highlight some different terms. These analyses yield insights into the models' interpretation of the question classification and enable us to extract terms relevant to the corresponding category automatically.

The main contribution of this paper is evaluation of two attention mechanisms, including the BERT-attention model proposed in this research, for pregnancy-related question categorization and relevant term annotation to interpret the decisions.

II. System Design and Models

In this section, we describe the two attention mechanisms for content annotation using the BERT model. Both models are capable of classifying the questions while annotating the important words within the questions.

A. Content Annotation using the Self-Attention Extraction from the BERT

Devlin et al. [7] introduced the BERT Transformer based on bidirectional self-attention. Unlike other language embedding generating architecture such as Word2Vec [9], the BERT model inputs are not vectors that represent words. Instead, the input includes the token, segment, and position embeddings. The BERT model can be fine-tuned for text classification by merely adding a softmax classification layer on top of the BERT model to predict the class of a given text sequence. The input of the softmax layer is the output of the last hidden layer of the first token that represents the original text sequence.

Previously, some researchers [10] [11] claimed that the bidirectional encoder structure of BERT can understand the context. However, few research investigated how to make use of the encoder structure for key information extraction and annotation. In this research, we investigate the feasibility of the information annotation using the multi-head self-attention mechanism of the BERT.

The attention mechanism of BERT works as Query (Q), Key (K), and Value (V) that start a linear transformation to "dynamically" generate weights for different connections, and then feed them into the scaling dot product. In the definition of self-attention, Q is K itself. d_k is the dimension of Q and K. Scaling the dot product prevents the dot product from growing too fast; if not addressed properly, it may cause the gradient of softmax function (shown as Equation 1) being too small [12]. The multi-head attention calculates N self-attention in parallel. Each self-attention is called a 'head.' To make the self-attention more flexible, each $head_i$, $i \in N$, is not calculated on the original Q, K, and V, but assigned a group of random parameter matrices on Query (${W_i}^Q$), Key (${W_i}^K$) and Value (${W_i}^V$), so each $head_i$ (shown as Equation 2) trains its attention map [12].

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

$$Head_i = Attention(QW_i{}^Q, KW_i{}^K, VW_i{}^V) \quad (2)$$

BERT is unsupervised, and the multi-head attention mechanism shows diverse and independent attention [13]. Clark et al. [14] investigated the attention mapping and extraction from BERT architecture. They analyzed different aspects of the attentions at each layer based on the input content patterns and found that none of the layers is dedicated to a specific NLP task [14]. Based on their finding, it is hard to tell which layer(s) should be used to extract attention associated with the importance of the words within the context. On the other hand, there are no trade-offs between the heads in each layer since they are calculated in parallel, and the heads also work well on co-reference resolution [14]. Therefore, keeping all the heads is conducive to improve the performance of the transformers [13]. Figure 1 shows the self-attentions between the tokens of two heads extracted from the last layer of the BERT model. The darker the line is, the higher the attention between the two tokens is. In this research, we utilize the self-attentions received by each token in the heads of all layers within BERT for word annotation.

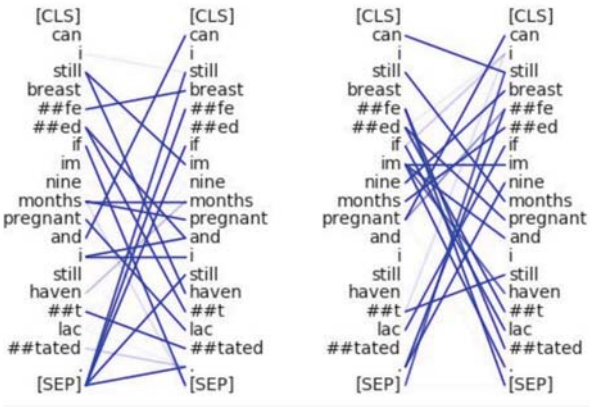

Fig. 1: Visualization of Self-Attentions within BERT

The sum of attention received by each independent participle in the layers and heads (shown as Equation 3) is used to present each token's attention. We assume that the more attention the token receives, the more important the token is towards context understanding and final decision.

$$attentionW_{token} = \sum_{layers} \sum_{heads} attention_i \quad (3)$$

Since BERT works with tokens, often, a word is broken down into multiple tokens. To calculate each word's attention, we sum up the attentions of the tokens of a word, as shown in Equation 4. The special characters, the punctuation, and the beginning and end sign tokens of the sentence are kept as is.

$$attentionW_{word} = \sum_{token \in word} attentionW_{token} \quad (4)$$

To annotate a chunk of text (pregnancy-related question in our case), we first need to fine-tune a pretrained BERT model for the pregnancy-related question corpus, then load the fine-tuned BERT model trained for the question classification. Since the BERT is trained on a large number of sentence pairs, to utilize a trained BERT model, the input question needs to be converted into sentence pairs. Our strategy is to construct self-pairs for each question. The built-in tokenizer of BERT then further transforms self-question pairs into tokens to feed into the BERT. Then, we extract the word attentions from the BERT model for annotation. In this research, we extract the top m word with high attention values and annotate them as important words.

B. Content Annotation by Adding an External Attention Layer to BERT (BERT-Attention Model)

Although the fine-tuned base BERT model can be used for text classification, the salient terms that drive the model's decision process are not explicitly linked to the classification layer. As such, the classification cannot be interpreted. Inspired by the BiLSTM with attention architecture, in this research, we propose to add an attention layer on top of the base BERT to capture the attention of the neural network on each token to first interpret the importance of the tokens for classification, as represented in the middle part of Figure 2. The maximum tokens in the text collection define the number of neurons in the attention layer. The attention layer's output connects to the classification layer by applying the relu activation function to capture the relationship between tokens and the output. The attention weight of each token $attentionW_i$ is defined by softmax of the output of the attention layer o, shown as Equation 5, where K is the number of tokens. It is used to identify the importance of tokens.

$$attentionW_i = \frac{exp(o_i)}{\sum_K exp(o_j)} \quad (5)$$

The architecture of the BERT-Attention model generates tokens with attention weights. To identify importance of tokens, we develop an algorithm. Given a sequence of attention weights $A = \{attentionW_1, ..., attentionW_n\}$ obtained from input tokens, we calculate the highest attention weight $attentionW_{max}$. Then, we calculate the difference between the attentions of the tokens to $attentionW_{max}$. The value of the n^{th} percentile of the difference can be used as a threshold

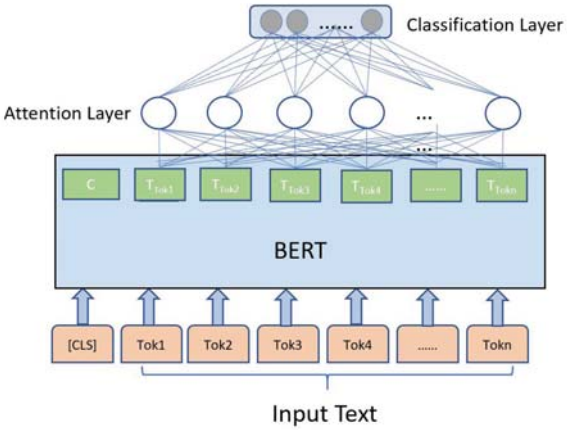

Fig. 2: BERT-Attention model for Question Classification and Relevant Content Annotation

to find the important words or tokens for the input document. The tokens are then combined into terms (single words or multi-word phrases). Since BERT tokenizer breaks the words into tokens and not all the tokens of a word have the same attention weights, if one token has an attention weight over the threshold (θ which is given in Equation 6), the whole word is extracted as an important word for annotation.

$$\theta = p((attentionW_{max} - attentionW_i), n) \quad (6)$$

III. Experiment

A. Data Set

In this work, we used questions posted in the Pregnancy and Parenting section of Yahoo! Answers from 2009 and 2014 [6]. Two annotators read through 662 posted question titles and content, and then grouped them into three primary pregnancy stages and eight categories. The numbers of questions included in each category are shown in Table I.

- "Pre-pregnancy" stage contains questions before pregnant. For this stage, we organized questions into the following categories: "birth control", "miscarriage", "whether pregnant or not" and "trying to conceive".
- "Pregnancy" stage contains questions during pregnancy and related to health concerns, general inquiries about pregnancy, and labor. We categorized the questions in this stage into two categories: "labor" and "pregnancy symptoms".
- "Postpartum" stage contains questions after the pregnancy is over. The questions in this stage are related to breastfeeding, baby and newborn, and nurturing newborns; the posts are classified into two categories: "baby and newborn" and "breastfeeding".

B. Experiment Setup

1) Baselines: The traditional machine learning methods, such as logistic Regression (LR), Random Forest (RF), SVM, Deep Neural Networks (DNN), and k-Nearest-Neighbors

(kNN) were utilized for question classification in the literature. Hence, we compared the BERT models against them for classification performance.

2) Data Preprocessing and Representation: For the baseline methods, the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme was used to construct the feature vectors to feed the models. For the BERT models, we fine-tuned the pre-trained BERT model (BERT-Base, Uncased) for the task for question categorization and content annotation.

3) Experimental Parameters: We split our dataset into 75% for training and 25% for testing. For the baseline algorithms, we used the built-in functions in the sklearn package, and tuned the parameters to achieve the best results. For DNN, there is one hidden layer with 512 neurons. For the BERT-Attention mode, the size of the attention layer is 512 given that the max length of the questions is 361 based on the number of words. For both BERT models, the training epoch was set to 5, the batch size was 16, and the learning rate was 2e-5.

C. Question Categorization Evaluation

Table I shows the classification performance of all the models on the test dataset. The results showed that both BERT based models work better than the baseline traditional learning models. Based on the Micro-Average and Macro-Average values, the BERT-Attention model worked slightly better than the base BERT model. BERT-Attention model gained higher precision on a few categories, such as 'Birth Control', 'Breast Feeding', and 'Labor'. The main reason is that the BERT-attention model takes advantage of the attention layer to extract the relevant words for classification, which is also demonstrated in the following content annotation using the attention mechanisms. BERT-Attention model did not perform well on the category "Pregnancy Symptoms". After investigation, we found that a few instances were misclassified as "Trying to Conceive", because for those cases, many words that received greater attention occur in both categories. LR cannot recognize the questions in the test dataset and belong to two small categories - "Labor" and "Breastfeeding".

D. Contention Annotation using Attentions

The main objective of this research is to explore whether the attention mechanisms can be used to annotate the relevant words that drive the classification. In this section, we demonstrate a few questions with terms that have high attention weights. Dark green and light green were used to highlight the words identified by self-attention of BERT and BERT-attention, respectively. Red was used to highlight the words identified by base BERT or BERT-attention if they are misclassified. Color grey was used to highlight the words that are identified by both attention mechanisms.

Figure 3 shows the relevant terms of two sentences that were correctly classified into the 'Baby and Newborn' category. For question in Figure 3, the relevant terms identified by both attention mechanisms are: 'baby' and 'diaper change'. These terms indicate that this question was about a baby. Both BERT and BERT-Attention models classified it correctly. Both models also identified other words that are relevant to the topic. The self-attention identified 'rest', 'day,' and 'average', whereas the BERT-attention model identified 'waking up' and 'how'. From this question, the main concern is 'baby waking up', BERT-Attention works better on capturing the key point.

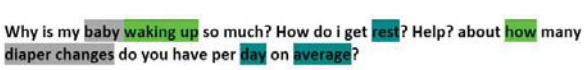

Fig. 3: Post Pregnancy - Baby and Newborn

Figure 4 is a question about 'Whether pregnant or not' and asked by a woman who might be pregnant. Both attention mechanisms captured the terms 'pregnancy', 'pregnancy test', and 'symptoms' with high attention weights. Each attention mechanism captured more special terms, such as 'negative', 'pies', 'abdomen', 'odd feeling' which can provide more detailed context of the question.

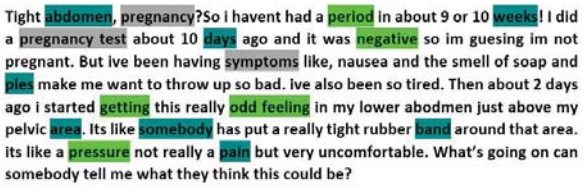

Fig. 4: Pre-Pregnancy - Whether Pregnant or not

The category 'Labor' contains the posts often asked by women concerning the health status before labor or the process of labor. Figure 5 presents a question where a woman is concerned with delivering her child. Both attention mechanisms identified the terms 'child', 'edema', 'high BP', 'pregnancy', 'due date', 'son' and 'labor', which are important for the model to correctly classify the questions.

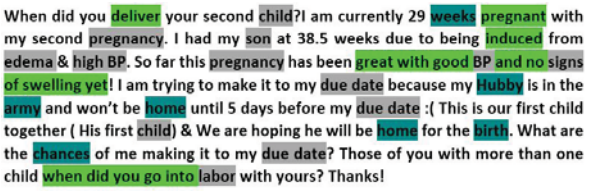

Fig. 5: Pregnant - Labor

The question in Figure 6 shows the questions asked by a woman who is facing problems in conceiving. BERT-Attention model correctly classified it to the category 'Trying to Conceive', whereas the base BERT model classified it to 'Whether Pregnant or not'. The words highlighted by both attention mechanisms are not closely relevant to the 'Trying to Conceive' category, and they seem to be more relevant to the category 'Whether Pregnant or not'. However, the BERT-Attention model also identified the words 'conceived' and 'ovulating', which often co-occur in the questions asked by a woman trying to pregnant. Hence, it was correctly classified this question to 'Trying to Conceive' category by the BERT-Attention model.

TABLE I: Comparison of the Question Classification of Different Categories.

Category (# of Questions)	Results																				
	BERT-Attention			BERT			SVM			DNN			RF			LR			kNN		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baby & Newborn (178)	**0.92**	**1.00**	**0.96**	**0.92**	0.98	0.95	0.90	0.96	0.92	0.93	0.91	0.92	0.67	0.98	0.79	0.69	1.00	0.82	0.79	0.91	0.85
Birth Control (32)	**1.00**	0.62	0.77	0.86	**0.75**	**0.80**	0.67	0.50	0.57	0.75	0.38	0.50	1.00	0.25	0.40	1.00	0.25	0.40	0.31	0.62	0.42
Breastfeeding (38)	**1.00**	**0.78**	**0.88**	0.71	0.56	0.63	0.67	0.67	0.67	0.50	0.67	0.57	1.00	0.22	0.36	0.0	0.0	0.0	0.56	0.56	0.56
Labor (24)	**1.00**	0.33	0.50	0.80	**0.67**	**.73**	**1.00**	0.33	0.50	0.80	**0.67**	**0.73**	0.50	0.40	0.44	0.0	0.0	0.0	0.50	0.33	0.40
Miscarriage (19)	**1.00**	0.40	0.57	0.50	0.40	0.44	0.50	0.20	0.29	0.50	0.20	0.29	**1.00**	0.60	0.75	**1.00**	0.20	0.33	0.10	0.20	0.13
Whether Pregnant or Not (172)	0.74	**0.91**	**0.81**	**0.77**	0.79	0.78	0.68	0.74	0.71	0.66	0.67	0.67	0.64	0.86	0.73	0.66	0.86	0.75	0.61	0.63	0.62
Pregnancy Symptoms (54)	0.42	0.57	0.48	0.63	**0.86**	**0.73**	**0.70**	0.50	0.58	0.58	0.50	0.54	0.67	0.14	0.24	0.33	0.07	0.12	0.50	0.21	0.30
Trying to Conceive (145)	**0.97**	0.78	**0.86**	0.84	0.75	0.79	0.71	**0.83**	0.77	0.70	**0.83**	0.76	0.81	0.72	0.76	0.77	**0.83**	0.80	0.76	0.53	0.62
Micro Average	**0.85**	**0.82**	**0.82**	0.81	0.81	0.80	0.75	0.75	0.74	0.73	0.73	0.72	0.71	0.70	0.65	0.63	0.70	0.63	0.65	0.62	0.62
Macro Average	**0.88**	0.67	**0.73**	0.75	**0.72**	**0.73**	0.73	0.59	0.63	0.68	0.60	0.62	0.72	0.47	0.50	0.56	0.40	0.40	0.52	0.50	0.49

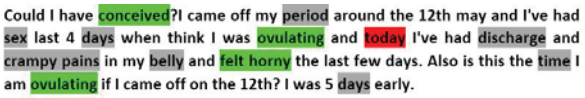

Fig. 6: Pre-Pregnancy - Try to conceive

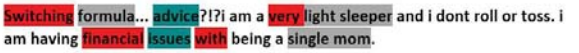

Fig. 7: Post Pregnancy - Baby and Newborn

We also investigated the posts that were misclassified by BERT-Attention but correctly classified by base BERT model. Figure 7 presents a question misclassified by the BERT-Attention model as 'Breastfeeding.' It was found that the BERT-Attention model captured the words 'switching', 'financial', and 'single mom', which also occurred in questions about breastfeeding. Hence, it led to misclassification of this case. Whereas the self-attention of the BERT model captured words 'formula', 'advice', 'issue', which are more relevant to the category 'Baby and Newborn'.

In summary, these cases demonstrate that the attention mechanisms can capture content words that drive the classification decisions. We found that the words captured by two attention mechanisms inform the main topics of the questions that were correctly classified.

IV. Conclusion and Future Work

In this paper, we investigated two BERT models for pregnancy-related question classification and content annotation. The BERT-Attention model is proposed to compare with the self-attentions within BERT for content annotation. The evaluation results showed that both base BERT and BERT-Attention model work better than the traditional learning models on question classification. Both attention mechanisms can be used to highlight the salient terms that drive the classification decision. Through the case studies, we showed that when the salient terms are identified correctly by the attention mechanism, the classification results are also improved. The quantitative analysis of the effectiveness of the attention mechanism for annotation and the evaluation of the BERT-Attention model on leveraging the results of the inherent attention mechanism of BERT will also be done in our future work. We also plan to improve the BERT-Attention model by modifying the attention layer to cross-reference the semantic similarity of the tokens Due to the transformative nature of the models, we also plan to apply it to other QA tasks for critical health issues such as COVID-19.

References

[1] J. Bian, Y. Guo, Z. He, and X. Hu, *Social Web and Health Research*. Springer, 2019.

[2] T. Harpel, "Pregnant women sharing pregnancy-related information on facebook: Web-based survey study," *J Med Internet Res*, vol. 20, no. 3, p. e115, Mar 2018. [Online]. Available: http://www.jmir.org/2018/3/e115/

[3] M. Larsson, "A descriptive study of the use of the internet by women seeking pregnancy-related information," *Midwifery*, vol. 25, no. 1, pp. 14–20, 2009.

[4] C. Zhu, R. Zeng, W. Zhang, R. Evans, and R. He, "Pregnancy-related information seeking and sharing in the social media era among expectant mothers: Qualitative study," *J Med Internet Res*, vol. 21, no. 12, p. e13694, Dec 2019. [Online]. Available: https://www.jmir.org/2019/12/e13694

[5] J. Radin, S. Steinhubl, A. Su, H. Bhargava, B. Greenberg, B. Bot, M. Doerr, and E. Topol, "The healthy pregnancy research program: transforming pregnancy research through a researchkit app," *npj Digital Medicine*, vol. 1, 12 2018.

[6] Y. QA, "Yahoo! health question and answering," n.d., https://answers.yahoo.com/dir/index?sid=396545018.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] Z. Lv, D. Liu, H. Sun, X. Liang, T. Lei, Z. Shi, F. Zhu, and L. Yang, "Autohome-orca at semeval-2019 task 8: Application of bert for fact-checking in community forums," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 870–876.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[10] X. Wu, S. Lv, L. Zang, J. Han, and S. Hu, "Conditional bert contextual augmentation," in *International Conference on Computational Science*. Springer, 2019, pp. 84–95.

[11] G. Jawahar, B. Sagot, and D. Seddah, "What does bert learn about the structure of language?" 2019.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[13] J. Li, Z. Tu, B. Yang, M. R. Lyu, and T. Zhang, "Multi-head attention with disagreement regularization," *arXiv preprint arXiv:1810.10183*, 2018.

[14] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does bert look at? an analysis of bert's attention," *arXiv preprint arXiv:1906.04341*, 2019.