PLM-AS: Pre-trained Language Models Augmented with Scanpaths for Sentiment Classification

Duo Yang* and Nora Hollenstein

Center for Language Technology, University of Copenhagen

Abstract

Recent research demonstrated that deep neural networks could generate meaningful feature representations from both eye-tracking data and sentences without designing handcrafted features, which achieved competitive performance across cognitive NLP tasks, such as sentiment classification over gaze datasets, but the previous works mainly encode the text and gaze data separately without considering the interaction between these two modalities or applying large-scaled pre-trained models. To address these challenges, we introduce PLM-AS, a novel framework to take full advantage of textual and eye-tracking features by sequence modeling in a highly interactive way for multimodal fusion. It is also the first attempt to combine large-scaled pre-trained models with eye-tracking features in the cognitive reading task. We show that PLM-AS captures cognitive signals from eye-tracking data and shows improved performance in sentiment classification within and across three datasets of different domains.

1 Introduction

Recent research studies have shown that eye-tracking features reflect cognitive information and lead to stable improvement in natural language tasks [1, 12, 16, 24, 28], such as sentiment classification [14, 17], sarcasm detection [18], named entity recognition [8], coreference resolution [3]. It can be mainly explained in the following aspects: 1) Entities, lengthy and complex words could catch attention more easily than common words in terms of lexical level. 2) While implicit expressions may also lead to a longer duration of fixation with a second check in terms of semantic level when reading the content. 3) Sentiment judgment is an auxiliary task to content comprehension [2], the participants will face difficulty in comprehension and will review the whole sentence several times due to the complex phrasal structure in terms of syntactic level [11]. In all these aspects, human gaze data can provide a wealth of cognitive information for content comprehension and support sentiment classification at the same time.

There have been breakthroughs in many fields by the advances in deep neural architectures in the recent decade, research on how to model text and human gaze data with deep neural networks was also conducted. A convolutional neural network was first applied to learn feature representations from both text and human gaze data [19], and the gaze component in their model handled with two fundamental eye-tracking attributes, including fixation and saccade. Another multi-task deep neural framework based on recurrent neural network LSTM also achieved competitive performance with gaze features [20]. The sentence-level attention corresponding to fixated words and adjacent words could also be applied to sentiment classification [2]. These works have limited capabilities from two aspects: (a) The text and gaze representations were learned by these neural networks without any interactions between two modalities, and the models just concatenated the final outputs for multimodal fusion (b) It is difficult to apply large-scaled transformer architecture directly within such a two-tower framework.

In this paper, instead of encoding two different modalities separately with neural networks, we propose a novel neural network structure that allows us to encode text modality first and then per-

*Corresponding Author: yd0301@outlook.com

https://doi.org/10.7557/18.6797

form sequence modeling leveraging the fixation order of words from gaze scanpaths intuitively with the support of pre-trained language models. We use the abbreviation **PLM-AS** for our proposed model **P**re-trained **L**anguage **M**odels **A**ugmented with **S**canpaths in the paper. We conduct various experiments for evaluation based on different gaze datasets, ETSA-I Dataset, ETSA-II Dataset and ZuCo Dataset released by [9, 11, 18], respectively. In summary, our contributions in this work are:
(1) We introduce a novel framework PLM-AS, combing the text representation with eye-tracking data in a more intuitive way than previous early-fusion models, specifically by leveraging the processing information encoded in the fixation order.
(2) It is also the first attempt to combine contextualized embeddings from pre-trained language models with eye-tracking data over gaze datasets on sentiment classification.
(3) We conduct a series of controlled studies by organizing the outputs from pre-trained language models to investigate the impact of eye fixations towards the framework, e.g., fixation words, fixation order.
(4) We also test the cross-corpus capabilities of this PLM-AS framework and analyze the results in the aspect of generalization.

2 Motivation

The concept of scanpath is first proposed by [22], which refers to the trajectories (paths) of the eyes when scanning the visual field and viewing and analyzing any kind of visual information. When it comes to human reading [25], the scanpath mainly demonstrates the sequence of eye fixations (50–1500 ms pause of viewing a fragment of text), revealing the saccades (a quick movement between two or more phases of fixation in the same direction) and the regression (backward saccade to a previously visited fragment).

The inspiration for our proposed model is that human reading is not a linear process in only one direction strictly, but the trajectory of the eye-movements could still be organized as the time series of eye fixations, and the composition of this new sequence is highly overlapped with the text itself, which means that we could represent the fixation sequence using different fragments of text, and it should work naturally in recurrent neural networks [26], e.g. GRU architecture [5], due to its sequential nature. Since the annotations are evaluated by the subjects, this means that the cognitive information would be automatically included in the scanpaths when they read the sentences. Applying the eye-tracking scanpaths into the deep neural networks directly is equivalent to combining the cognitive features with the corresponding text features in a more intuitive way.

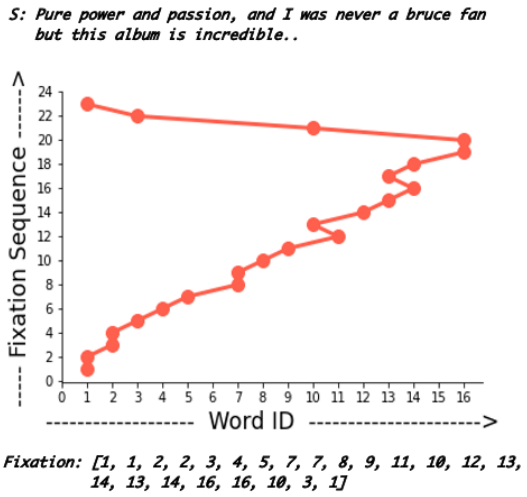

Figure 1: The scanpath of reading the sentence S from ETSA-II Dataset [18]. The fixation sequence records the positions of fixation words in the sentence, following the time series of eye fixations.

Our framework would follow this way: 1) Firstly, contextualized word representation is generated by transformer architecture over the reading sentences, and we take full advantage of the final layer from BERT [7] as text representation; 2) By retrieving the features of the corresponding positions step by step from BERT according to the index sequences of fixated words (scanpaths), we would have the text feature sequences in the fixation order; 3) Since recurrent neural network is designed for sequence modeling in deep learning, the new generated feature sequences are regarded as the input of the scanpath encoder, the GRU architecture [4], for the final multimodal fusion; 4) According to the actual length of the fixation sequence, the output of scanpath encoder in the final step is picked up for sentiment polarity prediction.

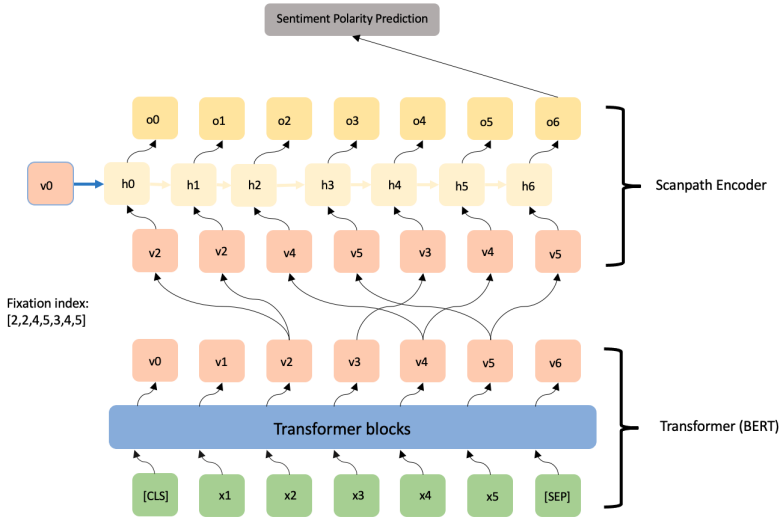

Figure 2: PLM-AS: Pre-trained Language Models Augmented with Scanpaths

3 Proposed Model

In this section, we introduce the PLM-AS framework with contextualized representation and the scanpath encoder, and give the details on loss functions over different label settings for sentiment classification.

Text Representation: Given a sentence S, each word would be cut into subword level by wordpiece tokenizer before the BERT architecture. Then we have subword sequence x_0 with two special tokens [CLS] and [SEP] inserted at the beginning and the end of the sequence, respectively.

$$x_0 = [[CLS], w_1, ..., w_{S-1}, [SEP]] \tag{1}$$

Each token could be represented by the concatenation of word embedding, position embedding, and segment embedding, and undergoes bidirectional multi-head self-attention across multiple transformer blocks:

$$x_l^{'} = MSA(LN(x_{l-1})) + x_{l-1}, l = 1...L \tag{2}$$

$$x_l = MLP(LN(x_l^{'})) + x_l^{'}, l = 1...L \tag{3}$$

where

$$MSA(X) = W_{\text{att}}[Att_1(X), ..., Att_m(X)]^\top \tag{4}$$

$$Att_i(X) = \text{softmax}\frac{((W_{Q_i}X)^\top W_{K_i}X)}{\sqrt{D/m}}(W_{V_i}X)^\top \tag{5}$$

Finally, we have the output sequence v from the final layer of BERT as a text representation (P refers to the dimension of hidden layer in BERT):

$$v = x_L \tag{6}$$

Scanpath Encoder: Given a subword-based fixation index sequence f, we retrieve the features of the corresponding position step by step from the output sequence v according to the index sequence of fixated words, then generate the new scanpath feature sequence s (N refers to the set of the fixation word index)

$$f = [f_1, f_2, f_3, ..., f_m], \ f_i \in N \tag{7}$$

$$s_i = v_{f_i}, \ i \in M \tag{8}$$

$$s = [s_1, s_2, s_3, ..., s_m], \ s_i \in \mathbb{R}^P \tag{9}$$

The scanpath feature sequence s is then passed to the GRU architecture and we have the output sequence o from the scanpath encoder: (Q refers to the output dimension of the scanpath encoder)

$$o_i = GRU_{\text{scanpath}}(s_i), \ i \in M \tag{10}$$

$$o = [o_1, o_2, o_3, ..., o_m], \ o_i \in \mathbb{R}^Q \tag{11}$$

Finally, we use the output in the last time step for training and evaluation (t refers to the actual length of the fixation index sequence).

$$o_{\text{final}} = o_t \tag{12}$$

Sentiment Polarity Classification: For the final classification, we take the outputs of GRU in the last step as the final features, according to the actual length of fixation index sequence.

1) Binary classification: The feature vector is then passed to the linear layer with a sigmoid activation function to predict the sentiment label {0,1}, positive or negative.

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(\hat{y}) + (1 - y_i) \cdot log(1 - \hat{y}) \quad (13)$$

We optimize the model with binary cross-entropy loss between the true labels and the predicted values during the training stage.

2) Multi-label classification The feature vector is passed to the linear layer with a softmax function, we pick up the index with the highest probability as the sentiment label {0,1,2}, positive, negative, or neutral.

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{3} y_{ij} \cdot log(\hat{y}_j) \quad (14)$$

We optimize the model with softmax cross-entropy loss between the true labels and the predicted values during the training stage.

4 Experiment Setup

4.1 Datasets

All the experiments followed the principle: a pair of one single scanpath and one sentence was treated as an example, instead of multiple reading scanpaths with one sentence, so we reconstructed the examples in this way over three cognitive datasets. Apart from that, we removed some examples with senseless annotation results from readers.

ETSA-I: We also worked on another cognitive reading dataset, Eye-Tracking and Sentiment Analysis-I, which have been used by [11] for sentiment classification. The dataset contains 1059 sentences in total from movie reviews and tweets, and the annotations come from five subjects, including eye-tracking data recorded by a remote eye-tracker Tobii TX 300 with sentiment labels (positive, negative, and neutral) for each sentence.

ETSA-II: We first applied our proposed framework based on the cognitive reading dataset released by [18]. The original dataset, Eye-Tracking and Sentiment Analysis-II Dataset, mainly supplemented with advanced eye-movement information over NLP dataset, contains fixation sequence data with 383 positive and 611 negative sentences, including sarcastic quotes, short movie reviews, and tweets. Eye-tracking data from 7 subjects are all included for each sentence, recorded by an SR-Research Eyelink-1000 eye-tracker during the reading.

ZuCo: Experiments were also carried out for cross-domain learning based on a cognitive dataset, the Zurich Cognitive Language Processing Corpus released by [9], combining EEG and eye-tracking recordings from subjects reading natural sentences as a resource for the investigation of the human reading process in adult English native speakers. This dataset includes simultaneous EEG and eye-tracking signals collected from 12 subjects during natural text reading, but in this case, we just extracted the textual features and the gaze features. The gaze data was recorded by an SR-Research Eyelink-1000 Plus eye-tracker. The corpus contains 400 sentences in total, of which 140 are positive, 123 are negative, and 137 are neutral, including movie reviews and biographical sentences.

4.2 Parameter Settings

As for ETSA-I and ETSA-II datasets, we simply split the dataset into two subsets, 90% of the dataset are treated as training samples, while 10% of them are used for validation. We follow the instruction in [21] to perform 25 runs for each model setting with the different random initialization, using the same data split and the same hyper-parameter settings, and the final results are averaged over these runs. The training is performed for 20 epochs with the batch size of 16, we adopt the AdamW optimizer by [15] with a learning rate of 0.0002 to minimize the loss and the default settings in PyTorch framework are kept unchanged, the learning rate is linearly increased for the first 10% of steps and linearly decayed to zero afterward.

All these settings are applied to BERT and the scanpath encoder equally. The scanpath encoder is designed as a single-direction GRU with one recurrent layer, the hidden size of GRU is set to 768, and the dropout with 0.1 are applied to the recurrent layer in GRU. We initialize the hidden state of scanpath encoder by using the special token [CLS] outputs from the final layer of BERT. Our implementation uses the PyTorch framework, and pre-trained models are loaded from HuggingFace Transform-

	Configuration	ETSA-II			ETSA-I		
		P	**R**	**F**	**P**	**R**	**F**
Traditional systems based on textual features*	Naïve Bayes	63.0	59.4	61.14	50.7	50.1	50.39
	Multi-layered Perceptron	69.0	69.2	69.2	66.8	66.8	66.8
	SVM (Linear Kernel)	72.8	73.2	72.6	70.3	70.3	70.3
CNN architectures* [19]	Text only	72.17	70.91	71.53	60.51	59.66	60.08
	Gaze only	65.2	60.35	62.68	52.52	51.49	52.0
	Text and Gaze	79.89	74.86	77.3	63.93	60.13	62.0
BERT	Text only	89.81	91.67	89.74	83.61	82.75	82.95
PLM-AS	Text and Gaze	**90.09**	**91.75**	**90.48**	**84.34**	**83.6**	**83.82**

Table 1: Performance evaluation over cognitive reading datasets [11, 18]. Except for removing a few noisy samples, we applied the same way to split the datasets as the previous work (*) did in [19]. We report macro-averaged precision (P), recall (R), and F1 score (F).

ers [27], an open source machine learning library in Python.

5 Performance Evaluation

Similar to previous cognitive studies in [19], we evaluate the PLM-AS over two cognitive reading datasets for sentiment classification task. The goal of our experiments is to investigate if the proposed model could take full advantage of textual and eye-tracking features for multimodal fusion over sentiment classification task and analyze where the improvement comes from by controlled baselines. Table 1 presents the performance of the previous works and our proposed model. In addition, we also evaluate our proposed model in cross-domain learning over three different datasets, shown in Table 3.

Single modality vs. Multimodality: The previous works in [19] show that CNN architectures learned from both text and eye-tracking data outperform those settings with single modality only. However, applying large-scaled pre-trained models has become the mainstream approach across different natural language tasks. The results show that the BERT model become another strong baseline on this task, even with text input only, but our proposed framework, PLM-AS, could perform multimodal fusion and beat the new baseline over both these datasets by taking advantages of large-scaled pre-trained models and gaze features at the same time. It would always be good to replace the BERT with other advanced pre-trained models for text representation, e.g. RoBERTa in [13] to achieve more gains over all these related settings, but it is not our main research purpose here.

Effect of fixation words (a): We consider the fixation words are selected subconsciously by the human cognitive process during reading, contributing to sentiment judgments after the content comprehension [2]. We also question if our proposed model could work smoothly with random word choices instead of this kind of certain word choices from human. To further investigate this question in PLM-AS, we carried out our first controlled baseline by randomly shuffling the BERT outputs before feeding them into the scanpath encoder, the results, Table 2, show that the performance of PLM-AS is better than the setting (a) over both datasets, to some extent, all these word choices selected during the natural reading by human share the common ground in cognition and support the sentiment judgments within our proposed model.

Effect of fixation order (b): The core idea of our proposed model is to capture the eye-tracking features by the fixation sequences, which provide cognitive information about the word choices and fixation order. To better understand the impact of the fixation order in PLM-AS, we try to shuffle the fixation order before feeding them into the scanpath encoder but with the word choices remained. Unsurprisingly, PLM-AS is better than the shuffled setting (b) from Table 2, which means that the fixation words could not contribute to the overall performance individually without the order information in PLM-AS, at least not in such a RNN sequential model setting [26] of the scanpath en-

	Configuration	ETSA-II			ETSA-I		
		P	**R**	**F**	**P**	**R**	**F**
BERT	Text only	89.81	91.67	89.74	83.61	82.75	82.95
PLM-AS	Text and Gaze	**90.09**	**91.75**	**90.48**	**84.34**	**83.6**	**83.82**
Other baselines	Shuffle BERT outputs (a)	89	91.1	89.26	83.56	82.3	82.73
	Shuffle fixation sequence (b)	89.38	91.45	89.81	83.4	82.51	82.81
	Replace fix. sequence with natural text (c)	89.35	91.5	89.78	84.05	83.13	83.38

Table 2: Performance evaluation based on controlled baselines. Noted that the text inputs of pre-trained language model stay the same without any shuffle or replacements in order to provide the text representation across all these settings, but in the next stage of encoding with GRU: (a) we create a subword-based index sequence with random words from the sentence to replace the fixation sequence; (b) we create another index sequence by shuffling the fixation sequence; (c) we replace the fixation sequence with natural text, the same order as text inputs for the pre-trained language model.

coder. RNN architecture might not be the only option for modeling the fixation feature sequences, especially in capturing the order information, but we left it to future research.

Effect of encoder architecture (c): We also question that the improvement might come from the scanpath encoder itself rather than the eye-tracking features, so we carried out our third controlled baseline by replacing the fixation sequences with word sequences of natural text, the only difference between this setting and BERT is by adding an extra GRU architecture, and it becomes a text-only setting. The results in Table 2 show that the performance of this setting (c) is close to the BERT baseline but lower than the performance of PLM-AS, which indicates the GRU architecture itself without any cognitive features might not contribute a lot to the overall performance of PLM-AS.

Cross-domain evaluation: Apart from these controlled baselines, we also perform a cross-domain evaluation based on the ZuCo dataset. The results in Table 3 show that our PLM-AS framework can achieve more competitive cross-domain performance to the BERT baseline while the models are trained on ETSA-II Dataset rather than ETSA-I Dataset. Noted that the reading texts in the ETSA-II Dataset are collected from two popular sarcastic quote websites, Tweet and the Amazon Movie Corpus, [23], with a higher level of complexity and diversity than the ETSA-I Dataset and the ZuCo dataset. Nearly half of the reading texts in the ETSA-II Dataset are sarcastic, it could be assumed that the eye-tracking data (scanpaths) in ETSA-II Dataset would be more abundant and diverse, which improve the overall performance. In addition, human scanpaths might vary not only from the text domains but also from person to person due to reading behaviors. Instead of providing eye-tracking features on average across all the subjects at a time, our PLM-AS framework might learn the reading patterns from a certain group of subjects and face challenges in generalizing these learned reading patterns to other subjects in such a subject-based sample construction. When it comes to the testing stage of cross-domain measurements, to some extent, the inconsistency between datasets' subjects should also be considered for the undesirable results when the models are trained on the ETSA-I Dataset.

Train	Test	Models	P	R	F
ETSA-I	ZuCo	BERT	**87.97**	**87.47**	**86.9**
		PLM-AS	87.66	86.5	85.77
ETSA-II	ZuCo	BERT	67.11	67.67	62.66
		PLM-AS	**68.7**	**68.53**	**64**

Table 3: Cross-domain evaluation over three datasets.

6 Conclusion

In this paper, we propose a novel framework to fully combine text representations with eye-tracking features by scanpath modeling and carry out experiments to evaluate our model (PLM-AS) for sentiment classification. The results show that PLM-AS captures cognitive signals from the eye-tracking data and shows improved performance on senti-

ment classification within and across three datasets of different domains. This indicates that the order of fixation during text reading carries linguistic information that is useful for NLP tasks.

Since all the experiments are carried out on small datasets with limited texts and unstable results are observed when applying large-scale language models, we decide to follow the evaluation strategy in [21], to obtain convincing results. Since it is not always practical to obtain related eye-tracking data for augmentation at the test time, many research studies have been proposed for gaze feature prediction on text [10] and image [6] in recent years. However, scanpath prediction on text has not yet been explored sufficiently, which could be investigated as an auxiliary task over different NLP challenges during training, to be free of this limitation.

References

[1] M. Barrett and N. Hollenstein. Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. *Language and Linguistics Compass*, 14(11):1–16, sep 2020. doi: 10.1111/lnc3.12396.

[2] X. Chen, J. Mao, Y. Liu, M. Zhang, and S. Ma. Investigating human reading behavior during sentiment judgment. *International Journal of Machine Learning and Cybernetics*, 13(8):2283–2296, mar 2022. doi: 10.1007/s13042-022-01523-9.

[3] J. Cheri, A. Mishra, and P. Bhattacharyya. Leveraging annotators' gaze behaviour for coreference resolution. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 22–26, Berlin, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1904.

[4] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics, 2014. doi: 10.3115/v1/w14-4012.

[5] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. doi: 10.3115/v1/d14-1179.

[6] R. A. J. de Belen, T. Bednarz, and A. Sowmya. ScanpathNet: A recurrent mixture density network for scanpath prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, jun 2022. doi: 10.1109/cvprw56347.2022.00549.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

[8] N. Hollenstein and C. Zhang. Entity recognition at first sight:. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1001.

[9] N. Hollenstein, J. Rotsztejn, M. Troendle, A. Pedroni, C. Zhang, and N. Langer. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5(1), dec 2018. doi: 10.1038/sdata.2018.291.

[10] N. Hollenstein, E. Chersoni, C. L. Jacobs, Y. Oseki, L. Prévot, and E. Santus. CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.cmcl-1.7.

[11] A. Joshi, A. Mishra, N. Senthamilselvan, and P. Bhattacharyya. Measuring sentiment annotation complexity of text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2014. doi: 10.3115/v1/p14-2007.

[12] S. Klerke, Y. Goldberg, and A. Søgaard. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2016. doi: 10.18653/v1/n16-1179.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2020. doi: 10.48550/arXiv.1907.11692.

[14] Y. Long, L. Qin, R. Xiang, M. Li, and C.-R. Huang. A cognition based attention model for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017. doi: 10.18653/v1/d17-1048.

[15] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. doi: 10.48550/arXiv.1711.05101.

[16] S. Mathias, D. Kanojia, A. Mishra, and P. Bhattacharya. A survey on using gaze behaviour for natural language processing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4907–4913. International Joint Conferences on Artificial Intelligence Organization, 2020. doi: 10.24963/ijcai.2020/683.

[17] E. S. McGuire and N. Tomuro. Sentiment analysis with cognitive attention supervision. *Proceedings of the Canadian Conference on Artificial Intelligence*, jun 2021. doi: 10.21428/594757db.90170c50.

[18] A. Mishra and P. Bhattacharyya. Predicting readers' sarcasm understandability by modeling gaze behavior. In *Cognitively Inspired Natural Language Processing*, pages 99–115. Springer Singapore, 2018. doi: 10.1007/978-981-13-1516-9_5.

[19] A. Mishra, K. Dey, and P. Bhattacharyya. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017. doi: 10.18653/v1/p17-1035.

[20] A. Mishra, S. Tamilselvam, R. Dasgupta, S. Nagar, and K. Dey. Cognition-cognizant sentiment analysis with multitask subjectivity summarization based on annotators' gaze behavior. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.12068.

[21] M. Mosbach, M. Andriushchenko, and D. Klakow. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*, 2021. doi: 10.48550/arXiv.2006.04884.

[22] D. Noton and L. Stark. Scanpaths in eye movements during pattern perception. *Science*, 171(3968):308–311, jan 1971. doi: 10.1126/science.171.3968.308.

[23] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219855.

[24] B. Plank. Keystroke dynamics as signal for shallow syntactic parsing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 609–619, Osaka, Japan, Dec.

2016. The COLING 2016 Organizing Committee. doi: 10.48550/arXiv.1610.03321.

[25] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998. doi: 10.1037/0033-2909.124.3.372.

[26] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, oct 1986. doi: 10.1038/323533a0.

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6.

[28] R. Zhang, A. Saran, B. Liu, Y. Zhu, S. Guo, S. Niekum, D. Ballard, and M. Hayhoe. Human gaze assisted artificial intelligence: A review. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, jul 2020. doi: 10.24963/ijcai.2020/689.