A Novel Perspective to Look At Attention: Bi-level Attention-based Explainable Topic Modeling for News Classification

Dairui Liu, Derek Greene and **Ruihai Dong**
Insight Centre for Data Analytics, Dublin
School of Computer Science, University College Dublin, Ireland
dairui.liu@ucdconnect.ie {derek.greene, ruihai.dong}@ucd.ie

Abstract

Many recent deep learning-based solutions have adopted the attention mechanism in various tasks in the field of NLP. However, the inherent characteristics of deep learning models and the flexibility of the attention mechanism increase the models' complexity, thus leading to challenges in model explainability. To address this challenge, we propose a novel practical framework by utilizing a two-tier attention architecture to decouple the complexity of explanation and the decision-making process. We apply it in the context of a news article classification task. The experiments on two large-scaled news corpora demonstrate that the proposed model can achieve competitive performance with many state-of-the-art alternatives and illustrate its appropriateness from an explainability perspective. We release the source code here[1].

1 Introduction

The attention mechanism is one of the most important components in recent deep learning-based architectures in natural language processing (NLP). In the early stages of its development, the encoder-decoder models (Bahdanau et al., 2015; Xu et al., 2015) often adopted an attention mechanism to improve the performance achieved by capturing different areas of the input sequence when generating an output in the decoding process to solve issues arising in encoding long-form inputs. Subsequently, researchers have applied the attention mechanism to large-scale corpora and developed a range of pre-trained language models (Kalyan et al., 2021), such as BERT (Devlin et al., 2019) and GPT-1 (Radford et al., 2018). This has yielded great progress across a range of NLP tasks, including sentiment analysis (Zhao et al., 2021) and news classification (Wu et al., 2021). However, the inherent characteristics of deep learning models

and the flexibility of the attention mechanism increase these models' complexity, thus leading to challenges in model explainability.

Today, there is still no consensus among researchers regarding whether attention-based models are explainable in theory. Some researchers believe that attention weights may reflect the importance of features during the decision-making process and thus can provide an explanation of their operation if we visualize features according to their weight distribution (Luong et al., 2015; Lu et al., 2018). However, other researchers have disagreed with this hypothesis. For example, Jain and Wallance's study demonstrated that learned attention weights are often uncorrelated with feature importance (Jain and Wallace, 2019). Some researchers have supported this viewpoint (Serrano and Smith, 2019), but treated with skepticism by others (Wiegreffe and Pinter, 2019).

In this paper, rather than validating the attention explainability theoretically, we propose a novel, practical explainable attention-based solution. Inspired by the idea of topic models (Blei et al., 2003), our proposed solution decouples the complexity of explanation and the decision-making process by adopting two attention layers to capture *topic-word* distribution and *document-topic* distribution, respectively. Specifically, the first layer contains multiple attentions, and each attention is expected to focus on specific words from a topic. The attention in the second layer is then used to judge the importance of topics from the perspective of the target document. In order to further improve the model's explainability, we add an entropy constraint for each attention in the first layer. To prove the effectiveness of our proposed solution, we apply it in the context of a news article classification task and conduct experiments on two large-scaled news article datasets. The results presented later in Section 4 show that our model can achieve competitive performance with many state-of-the-

[1] https://github.com/Ruixinhua/BATM

art transformer-based models and pre-trained language models, while also demonstrating its appropriateness from an explainability perspective.

2 Related Work

2.1 Attention Mechanism

The attention mechanism was first applied on machine translation tasks (Bahdanau et al., 2015) with the Seq2Seq model using RNN. To solve the dilemma in compressing long sequences by using an RNN-encoder, Bahdanau et al. (2015) introduced an attention mechanism by allowing RNN-decoder to assign attention weights to words in the input sequence. This strategy helps the decoder to effectively capture the relevant information between the hidden states of the encoder and the corresponding decoder's hidden state, which avoids information loss and makes the decoder focus on the relevant position of the input sequence. This attention mechanism is named *additive attention* or *Tanh attention* because it uses the Tanh activation function. In our work, we propose to use additive attention to discover the underlying mixture of topics within a document.

Furthermore, Vaswani et al. (2017) proposed a transformer architecture to replace RNNs entirely with multi-head self-attention. This approach makes it possible to compute hidden representation for all input and output positions in parallel. The advantage of parallelized training has led to the emergence of many large pre-trained language models, such as BERT (Devlin et al., 2019). The improvement of using the transformer-based language model for generating representations is significant compared with popular word embedding methods such as GloVe (Pennington et al., 2014). However, along with the considerable enhancement in performance, it makes the attention-based language models difficult to interpret. One potential solution is to use attention weights to provide insights into the model.

2.2 Attention as an Explanation

The visualization of attention weight alignment in (Luong et al., 2015; Vaswani et al., 2017) provides an intuitive explanation of the operation of additive attention and multi-head self-attention in machine translation tasks. But the faithfulness (i.e. accurately revealing the proper reasoning of the model) and plausibility (i.e. providing a convincing interpretation for humans) of using attention as an

explanation for some tasks are still in debate, and the questioning is mainly on faithfulness (Jacovi and Goldberg, 2020). This discussion is primarily focused on a simple model for specific tasks, such as text classification, using RNN models connecting an attention layer which is typically MLP-based (Bahdanau et al., 2015). A number of researchers have challenged the usefulness of attention as an explanation (Jain and Wallace, 2019; Serrano and Smith, 2019; Bastings and Filippova, 2020), concluding that saliency methods, such as gradient-based techniques, perform much better than using attention weights as interpretations in finding the most significant features of the input sequence that yield the predicted outcome. However, Wiegreffe and Pinter (2019) claimed that, despite the fact that explanations provided by attention mechanisms are not always faithful, in practice, this does not invalidate the plausibility of using attention as an explanation. We believe that the attention mechanism can provide a plausible explanation when applied correctly for an appropriate task.

2.3 Role of Attention Mechanism

Compared to simple additive attention, the Multi-Head Attention (MHA) mechanism, the core component of the big Transformer-based language model, is more complicated when attempting to interpret model behavior with complex weights distribution. Therefore, considerable work has attempted to understand the role played by the different attention heads (Rogers et al., 2020). For example, Voita et al. (2019) analyzed the patterns of attention heads by checking the survival of pruning, finding that the syntactic and positional heads are the final ones to be removed. Kovaleva et al. (2019) identified five attention patterns of MHA, while Pande et al. (2021) proposed a standardized approach for analyzing patterns of different attention heads in the context of the BERT model.

Instead of employing a complex transformer-like architecture with many MHA layers, we propose to start with a single MHA layer individually. Inspired by previous work, we focus on analyzing the role of attention heads in our architecture. We adopt a similar approach to (Lu et al., 2018) by modeling attention using topics. However, unlike the topic attention model (TAN), which uses a bag-of-words (BOW) model based on variational inference to align the topic space and word space with extracting meaningful topics (Panwar et al., 2021), we

assume that these multiple attention heads represent multiple topics in terms of their semantics.

3 Methodology

This section describes our proposed architecture Bi-level Attention-based Topical Model (BATM) as illustrated in Figure 1. It uses two attention layers to uncover a latent representation of the data and then makes use of attention weights as a form of topic distribution. We describe this architecture from the perspective of a news classification task. Our architecture consists of three components: an embedding layer, two attention layers, and a classification layer. After generating embedding vectors of words for the given news articles, we pass them to two attention layers to obtain the weight distribution of different words in each head (i.e. topic) and the weight distribution of different heads in the input articles. Then we generate the document representation vector based on these weights and finally classify the articles into different categories using a single linear layer. By analyzing the weight distribution of the attention layer on the entire news corpus, we find that some heads focus on the words related to the specific topics. These concentrated words help us understand the behavior of the attention mechanism.

3.1 Embedding Layer

There are two popular embedding methods: word-level embedding and contextual embedding, in general. Word-level embedding methods, such as GloVe, project different words into a word vector space and acquire a fixed-length word vector through a pre-trained embedding matrix. Contextual embedding models, such as BERT, generate different word vectors based on each word's context, so that the same word in different contexts can produce very different word vectors. For a given document x, suppose we have N tokens in total, we use an appropriate tokenizer to partition it into tokens $t_1, t_2, \ldots, t_N$ according to the embedding method. Then we can represent the document using its embedding vectors $e_1, e_2, \ldots, e_N$ as an input to the attention layer.

3.2 Multi-Head Attention Layer

We use a multi-head attention mechanism to allow the model to focus on different positions in the document from different representation subspaces through multiple attention heads. We compute the weight distribution g^k of the head vector h_k through a single-layer feed-forward network first:

$$\mathrm{g}_i^k = v_k \tanh\left(\mathrm{W}_k e_i + b_k\right) \tag{1}$$

We then use the *softmax* function to get the normalized weights distribution α^k among the document:

$$\alpha_i^k = \frac{e^{\mathrm{g}_i^k}}{\sum_j^N e^{\mathrm{g}_j^k}} \tag{2}$$

Finally, the head vector h_k is the weighted sum of word embedding vectors using the weights α^k, given by

$$h_k = \sum_i^N \alpha_i^k e_i \tag{3}$$

where trained parameters are $v_k \in \mathbb{R}^{D_k}$, $\mathrm{W}_k \in \mathbb{R}^{E \times D_k}$, and $b^k \in \mathbb{R}^{D_k}$. D_k is the projected dimension of each head in the middle, and E is the embedding dimension, while the dimension of head vector h_k is E which is the same as embedding vector e_i from Eqn. 3.

3.3 Additive Attention Layer

For a given number of attention heads K, we have a group of head vectors $H = \{h_1, h_2, \ldots, h_K\}$, which are fed into an additive attention network to generate the document-topic distribution.

$$\mu_k = c \tanh\left(\mathrm{W}_H h_k + b_H\right)$$
$$\beta_k = \frac{e^{\mu_k}}{\sum_i^K e^{\mu_i}} \tag{4}$$

Finally, the document representation d is the weighted sum of head vectors along with the weights distribution β:

$$d = \sum_i^K \beta_k h_k \tag{5}$$

where trained parameters are $c \in \mathbb{R}^{D_h}$, $\mathrm{W}_H \in \mathbb{R}^{E \times D_h}$, $b_H \in \mathbb{R}^{D_h}$, and the dimension of d is also E which is the same as h_k.

3.4 Classification Layer

Since the representation of each document d will be a dense vector containing a mixture of information about the document's content, we can use it as the feature vector for the final news classification task:

$$y = softmax\left(\mathrm{W}_C d + b_C\right) \tag{6}$$
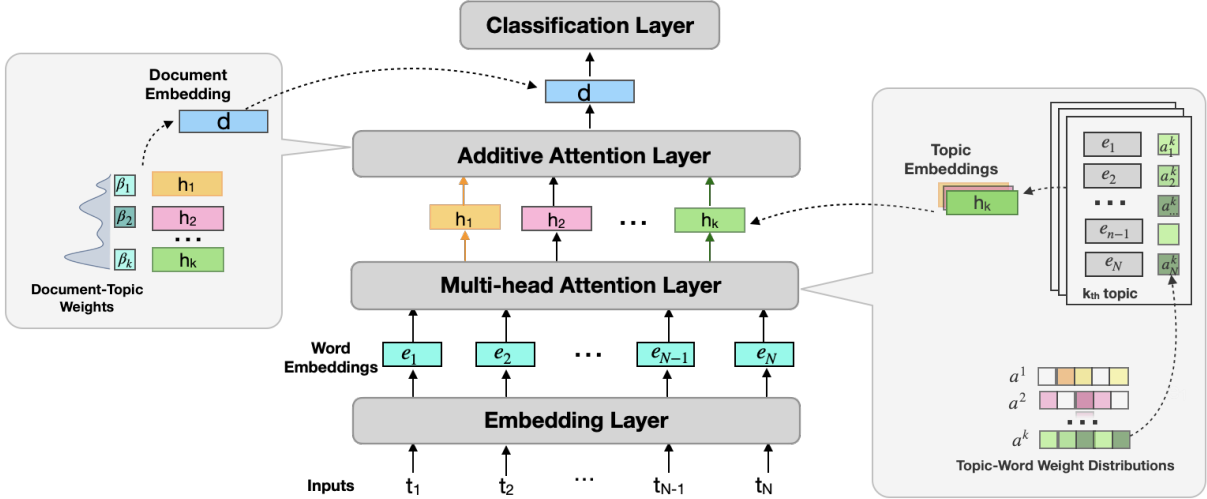
Figure 1: Structure of the proposed Bi-level Attention-based Topical Model (BATM).

3.5 Entropy Constraint

In order to further improve the explainability of our base model as described above, we now adjust the model so that each head only focuses on a specific set of words - i.e. we enforce topic-word weights distribution α^k not to spread over the document widely. We do this by computing the entropy of α^k as a part of the loss function. The entropy constraint penalizes the model when α^k has high entropy. Thus, the final loss with entropy constraint for the news classification task is:

$$\mathcal{L} = \mathcal{L}_{CE}(y, \hat{y}) + \lambda \frac{\sum_k^K \mathcal{E}_{\text{doc}}\left(\alpha^k\right)}{K} \qquad (7)$$

where $\mathcal{L}_{CE}(y, \hat{y})$ is the Cross-Entropy Loss between ground-truth class and predicted class, and λ is a hyper-parameter to scale the magnitude of average entropy calculated by α^k. The calculation for corresponding entropy $\mathcal{L}_{\text{entropy}}$ is by:

$$\mathcal{E}_{\text{doc}}\left(\alpha^k\right) = -\sum_i^N \alpha_i^k \log \alpha_i^k \qquad (8)$$

The entropy constraint applied on document-level in Eqn. 8 changes the distribution of topic-word weights α^k. However, our goal is to find more diverse topics, which means different topics should focus on different words. Therefore, it is necessary to know how entropy decreases at the token level (i.e. across the vocabulary as shown in Figure 2), which is defined by:

$$\mathcal{E}_{\text{token}}\left(M_i\right) = -\sum_k^K M_i^k \log M_i^k \qquad (9)$$
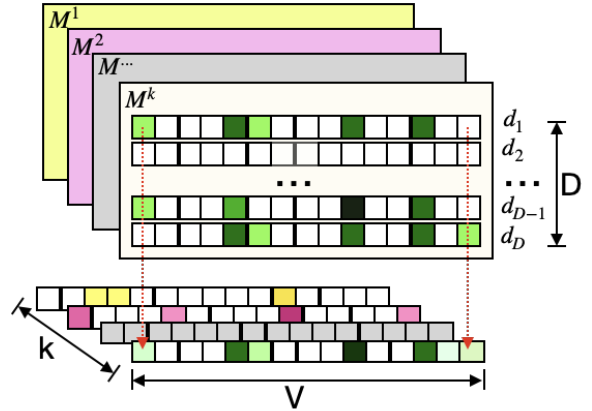

Figure 2: Structure of the topic-word weights α distribution among all documents.

To distinguish between the two variants of our model, we name the basic model as BATM-Base and use BATM-EC refer to the model with entropy constraints. From Eqn. 7, it is evident that if we set λ as 0, BATM-EC will be equivalent to the basic model.

3.6 Generating the Topic Distribution

After training our proposed BATM model, we analyze the attention weights generated from the first attention layer (MHA) over the corpus vocabulary to generate a global topic distribution. Let us assume that there are V words in the corpus and we have K heads corresponding to K topics. The resulting topic distribution takes the form of a $V \times K$ weight matrix, calculated from a trained MHA layer using embedded word vectors as inputs.

Moreover, to identify the most important words for each topic (which we can view as being the topic's *descriptor*), we extract the top-T words from the topic distribution, which can help us understand the heads and interpret them as topics. We examine the interpretations of these topic descriptors and display some examples in Section 4.5.

4 Experiments

We now evaluate the BATM model on two large-scale real-world datasets, and compare its performance with a number of state-of-the-art methods.

4.1 Datasets

We evaluate our proposed model on a news classification task and conduct extensive experiments on two public corpora. MIND (Wu et al., 2020) is a large-scale English dataset for news recommendation and categorization tasks. It contains information such as story title, abstract, and news category, but the public version does not include full article body content. We collected news articles from the Microsoft news website[2] to supplement it. There are 18 categories in the original MIND-large dataset, but three of them only have a small number of articles (< 10). Therefore, we exclude these categories from our experiment. The second one is the News Category Dataset[3], which contains approximately 200k news articles (each of them include a headline and a short news description) from 2012 to 2018 obtained from HuffPost. The original dataset has 41 categories, but some of these are duplicates. After merging the duplicated categories, there are 26 categories remain, which is denoted as News-26. We randomly split these two datasets into training/validation/test sets with a 80/10/10 split. Table 1 summarizes the divisions and the key statistics of the datasets.

4.2 Baseline Models

For the purpose of assessing classification performance, we first compare the effectiveness of our BATM base model relative to a number of attention-based and pre-trained language models:

- BERT (Devlin et al., 2019) composes of a bidirectional encoder of transformer and is pre-trained by using a combination of masked language modeling objective and next sentence prediction on a large corpus;

- DistilBERT (Sanh et al., 2019) is a small, fast, cheap, and light transformer model trained by distilling BERT base;

- XLNet (Yang et al., 2019) is an extension of the Transformer-XL (Dai et al., 2019) model, which utilizes an autoregressive method to learn bidirectional contexts by maximizing the expected likelihood over all permutations of input sequence factorization order;

- Roberta (Liu et al., 2019) is a robustly optimized BERT that modifies key hyperparameters, removing the next-sentence pre-training objective and training with much larger mini-batches and learning rates;

- Longformer (Beltagy et al., 2020) is based on RoBERTa (Liu et al., 2019) and uses sliding window attention and global attention to model local and global contexts;

- Fastformer (Wu et al., 2021) uses additive attention to perform multi-head attention, which is more efficient than a standard transformer.

The initial weights of these pre-trained language models (BERT, DistilBERT, XLNet, Roberta, and Longformer) are provided by Hugging Face Transformer (Wolf et al., 2020) library[4]. We use a linear classifier to receive the pooled output from previous transformer layers and then fine-tune these models to adapt them to the classification task. For the attention-based model, Fastformer, we initialize its embedding matrix using GloVe embedding and follow the hyper-parameter settings in (Wu et al., 2021).

4.3 Experimental Settings

In our experiments, we consider two ways to initialize our embedding matrix: *GloVe* embedding (Pennington et al., 2014) and context embeddings from a pre-trained language model *DistilBERT* (Sanh et al., 2019), where embedding weights are not fixed during the training procedure. We examine how different number of heads would influence the

---

Dataset	|Train|	|Validation|	|Test|	Avg. Len	#Class	|Vocabulary|
MIND-15	102,642	12,830	12,831	519.9	15	127,770
News-26	160,676	20,086	20,086	29.9	26	69,131

Table 1: Statistical information for the MIND-15 and News-26 corpora. Note the vocabulary size only refers to English words without any punctuation or numbers.
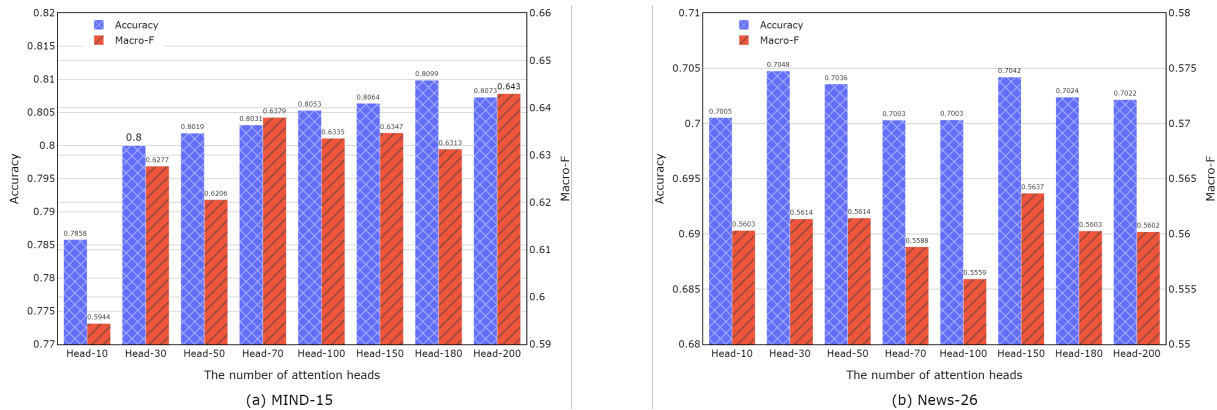

Figure 3: Performance of BATM-Base-GloVe with different number of attention heads on MIND-15 and News-26

performance of our proposed model on the validation set, the details is shown in Figure 3. Unsurprisingly, on the MIND data set, the model needs to set a relatively larger number of topics, because the average length of news articles in the MIND dataset and its vocabulary size are much larger than the News-26 dataset, as indicated in Table 1. We identify the number of topics for MIND-15 and News-26 as 180 and 30 for the rest of experiments, respectively. We use Adam (Kingma and Ba, 2015) for model optimization, and each epoch decays the learning rate by half.

4.4 Performance Comparison

The large pre-trained transformer variants perform better than the model with GloVe embedding, both for MIND-15 and News-26. Compared to Fastformer-GloVe, our BATM-Base-GloVe model achieves a similar result (variance in 0.3% of accuracy and 0.4% of Macro-F) for MIND-15 and a better result (variance in almost 0.4% of accuracy and 0.6% of Macro-F) for News-26. The differing results in MIND-15 and News-26 are due to the length of articles. As an efficient Fastformer can take a much longer sequence as input, it is advantageous to deal with long sequences which are unavailable in a short-length news dataset such as News-26. Using the pre-trained transformer-based embedding greatly improves the performance of our proposed BATM-Base model compared to the

GloVe embedding, although it adds to the difficulty of interpretation. The performance difference of the other pre-trained language models with the BATM-Base-DB model is less than 1% accuracy and approximately 2% Macro-F, both for MIND-15 and News-26. These experiments demonstrate the effectiveness of our proposed model in constructing document representations. Thus, the analysis of BATM's behavior using the topic-word distribution and document-topic distribution is essential to understanding the role of Bi-level attention layers.

4.5 Evaluation of Global Topic Representation

Besides the classification performance, we are also interested in whether each extracted topic descriptor as described in 3.6 has an intuitive meaning. We take the top-25 highest scoring terms from each topic and calculate topic coherence scores C_v (Röder et al., 2015). The average coherence scores of all topics of the *BATM-Base-GloVe* model are 0.58 and 0.56 on the MIND dataset and the news category datasets, respectively. Moreover, to more intuitively understand the meaning of topics mined by our model, we list a few topic examples whose coherence scores range from 0.3 to 0.8 along with a manually-assigned label in Table 3. The topics with coherence scores between 0.55 and 0.8 usually have precise meanings, such as the topic labeled as "Partisan" score of 0.76, where the vast major-

Models	MIND-15		News-26	
	Accuracy	Macro-F	Accuracy	Macro-F
BERT	82.12±0.31	67.09±0.47	75.01±0.31	62.08±0.36
DistilBERT	82.03±0.52	67.24±0.59	74.97±0.29	61.87±0.35
XLNet	82.37±0.18	67.75±0.43	73.99±0.29	60.8±0.44
Roberta	82.45±0.72	67.77±1.06	74.81±0.22	61.76±0.27
Longformer	82.71±0.16	68.09±0.4	74.87±0.29	61.79±0.35
Fastformer-GloVe	79.97±0.24	63.62±0.23	69.33±0.26	54.92±0.33
BATM-Base-GloVe	79.75±0.15	63.24±0.41	69.72±0.16	55.53±0.12
BATM-Base-DB	**82.82±0.15**	**68.79±0.26**	**75.74±0.17**	**63.01±0.23**

Table 2: Comparison of performance of models for the news classification task on MIND-15 and News-26 datasets. The best average scores are highlighted in bold.

Label	Topic Descriptor	C_v
Partisan	indictments voter votes fiscal impeachment petitions electorate partisanship repudiation treasonous repeal majorities dissent amendments judicial electoral repealing elections ratification partisan incompetence conviction impeach justification resignations	0.76
Household	cloth decorate towels embroidery basketballs suede bedding eggs fleece linen slippers cotton hooded porcelain bag plastic washed bowls clothes shirt flannel jacket jackets sweatshirt decorative	0.73
Unknown	serveware depositors mcadoo resold appliance cleats stockholders zoku horseshoes mailboxes frp hardwood holders multipacks disks unusable slugger noxzema laminate drawers tabletops ingvar costra memorabilia mailbox	0.61
Gender	bisexuals affectional transpeople asexuals genderqueer cisgender queerness cisgendered discimination heterosexism courtyards bisexuality cissexism ochre asexuality sexualities heterosexuality androgyny transphobia heterosexual butches trans slurs blacks heterosexuals	0.57
Diseases	triceps mumps soundproofed measles immunodeficiency listeria stepfamilies brees pronated workouts bestival talaq coronavirus stepfamily babyproofing salmonellosis obliterans varicella homestyle iguodala bomer griever botulism gbk cortisol	0.45
Schedule	said evening keynote annual month morning event scheduled weekend attended week according adjusted hosted inaugural host conferences conference attend telecast afternoon night will brightness sessions	0.38

Table 3: Examples of topics identified by our approach, in terms of extracted topic descriptors, topic coherence scores C_v, and manually-assigned labels.

ity of words are related to political activities and elections. However, some topics with a score in the range of $0.55 \sim 0.8$ are still tough to surmise the focus, as the unknown topic (labeled as "Unknown" with C_v value is 0.61) suggest, where the correlation of topic descriptors is non-intuitive. In contrast, some low-coherence topics may contain highly relevant words as well. For example, the topic "Schedule" a with a score of 0.38 (under 0.55) mainly includes words related to time and arrangement, which we can comprehend the central point of these words, but the automated metric unfairly evaluates it. Therefore, with the auxiliary of topic coherence measurement and manual verification, we are firmly convinced that topic descriptors extracted by the *BATM-Base-GloVe* model indeed

λ	MIND-15				News-26			
	Accuracy↑	Macro-F↑	Avg.$\mathcal{E}_{doc}$ ↓	Avg.$\mathcal{E}_{token}$ ↓	Accuracy↑	Macro-F↑	Avg.$\mathcal{E}_{doc}$ ↓	Avg.$\mathcal{E}_{token}$ ↓
0	80.50	63.40	3.171	8.542	70.03	55.88	2.175	9.022
1e-6	80.13	62.97	3.049	8.483	69.43	54.96	2.176	9.073
1e-5	80.16	64.07	3.076	8.599	69.55	55.12	2.129	8.995
1e-4	79.03	61.35	2.251	7.624	69.39	54.74	1.943	8.879
1e-3	72.86	50.58	0.041	5.947	58.16	38.74	0.080	7.071
1e-2	65.66	36.39	0.002	4.464	49.36	27.79	0.009	7.355

Table 4: Influence of λ of BATM-EC model on MIND-15 and News-26 datasets with 180 and 30 heads respectively.

have specific meanings.

5 Effect of Entropy Constraints

In the previous sections, the proposed BATM-Base-GloVe model demonstrates its competitive classification performance and excellent explainability. We now study the effect of adding an entropy constraint, as discussed in Section 3.5. In the extended model, referred to as BATM-EC, λ determines the degree of constraint that is imposed, so the BATM-Base-GloVe model is a special case when λ is zero.

This study assumes that a good topic (a first-level of attention) should only focus on specific words related to that topic. Its weight distribution on a news article should not be flat for the whole document, while its global weight distribution should also not be widely spread out across the entire vocabulary (i.e., it should have a relatively lower entropy). Therefore, we observe the dynamic of two entropy metrics $\mathcal{E}_{doc}$ and $\mathcal{E}_{token}$ (see calculation in Eqn. 8 and Eqn. 9) by setting different values of λ. We present the performance and entropy changes along with the values of λ in Table 4

The results meet our expectations. When λ reaches 1e-4, both entropy indicators decrease significantly with an acceptable trade-off in classification performance. When continually increasing the impact of entropy constraints, both entropy indicators and classification performance decrease dramatically. This is reasonable, as this experiment is conducted with a fixed number of heads. When attention focuses on a minimal number of topics, and the number of topics does not increase accordingly, information within article texts is likely to be lost, affecting the classification performance.

6 Discussion and Future Work

While the variant of our proposed model, BATM-base-DB, which is initialized by the contextual embeddings, can outperform all alternatives, the meaning of its topics is much worse than BATM-Base-GloVe. Each contextual embedding learned by pre-trained language models will merge the information from its surrounding words, which increases the difficulties of the proposed attention layer to capture the topics it focuses on, thus leading to more noise in their representations.

Another challenge we will address in the future is how to balance the computation cost, topic granularity, and classification performances. As discussed in the previous sections, it will affect the model's classification performance if we only introduce entropy constraints without incrementing the number of attention heads. However, increasing the number of attention heads will lead to the proportional increment of parameters, increasing the complexity of the model and resulting in a high computation cost. We will consider increasing the number of heads and the extending entropy constraint further, to improve classification performance while maintaining strong explainability.

7 Conclusion

In this paper, we presented a novel approach that harnesses a bi-level attention framework to decouple the text classification process as topic capturing, topic importance recognition and decision-making process to benefit explainability. We conducted the experiments on two large-scale text corpora. Compared with a number of state-of-the-art alternatives on a text classification task, our model can not only achieve a competitive performance, but also demonstrates a strong ability to capture intuitive meanings in the form of topical features, thus improving its explainability and transparency. In addition, by initializing it with contextual embeddings, our model outperforms all the baseline models.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yichao Lu, Ruihai Dong, and Barry Smyth. 2018. Coevolutionary recommendation model: Mutual learning between ratings and reviews. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 773–782, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Madhura Pande, Aakriti Budhraja, Preksha Nema, Pratyush Kumar, and Mitesh M. Khapra. 2021. The heads hypothesis: A unifying statistical approach towards understanding multi-headed attention in bert. In *AAAI*.

Madhur Panwar, Shashank Shailabh, Milan Aggarwal, and Balaji Krishnamurthy. 2021. TAN-NTM: Topic attention networks for neural topic modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3865–3880, Online. Association for Computational Linguistics.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Fastformer: Additive attention can be all you need. *ArXiv*, abs/2108.09084.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, Online. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

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

Lingyun Zhao, Lin Li, Xinhao Zheng, and Jianwei Zhang. 2021. A BERT based sentiment analysis and key entity detection approach for online financial texts. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1233–1238. IEEE.

A Appendix

A.1 Experimental Environment

Our experiments are conducted on the sonic system with Linux operating system. We use PyTorch 1.8.0 as the backend. The GPU type is Nvidia Tesla V100 and A100 with 32GB and 40GB GPU memory, respectively. We run each experiment 5 times with fixed random seeds by a single thread.

A.2 Preprocessing

We use the PyTorch default Tokenizer to preprocess texts. And we remove all the non-alphabetic characters when extracting the topic descriptors from the first attention layer.

A.3 Hyperparameter Settings

The dimension of the GloVe embedding and pretrained language model (PLM) is 300 and 768, respectively. The learning rate for the GloVe-based model and PLM model is 1e-3 and 5e-5, respectively. The maximum sequence length of all models is 512 on MIND-15 and 100 on News-26, except for Fastformer, which is 2048 on MIND-15. The batch size is 32 for all experiments, both on MIND-15 and News-26.

A.4 More Topic Examples

Label	Topic Descriptor	C_v
Gender	lgbtq divorce lgbt infertility divorced hiv transgenders stepparent hpv surrogacy divorcing honeymoons heterosexuals marriage honeymoon transgendered weddings prenuptial menopause premarital alimony stepfamily listeria queer prenups	0.71
Mood	disabling rebooting attacker accidently alerted viewer snapshots reset incriminating device disables inadvertently maliciously alerting securely jagged unintentionally sobering unsettling crashing gruesome wreckage jarring helpfully accidentally	0.70
Marriage	bridal wedding playdates preschooler brides toddlers bride gradeschool kindergarten mehndi kids toddler carolee weddings pacifier uighur preschoolers kyiv boomer udaipur design bridesmaid kid preschool kindergarden	0.67
Disease	epidemiology smashbox hilson dietetics nondairy deminers ijustine kimmel circadian vitamix presenteeism disinformers preparers disick keri fearless jwt integrative fassbender engelberg nutritionists swizz nivea juanes braff	0.67
Unknown	succinct republished talkbacks commenter peterman compiling errico excerpted newsfeeds reposted techdirt compiled dealnews compiles emailer tipsters editors crossposted postings downloaded collated tipster rnberg snarkiest khayr	0.53
Law	larceny forgery summonses unlawful misstatements offences felonies indictments wrongdoing audits contemplated misconduct misstatement breach burglary perjury incidents defendants tolerances irregularities misdemeanor fabricated misdemeanors comply statutory	0.5
Relationship	son playgroup aged daughter womb nieces playdate granddaughters mums playroom parents ladera swingset sons playdates picnicking tykes toddlers icmi eldest napped dad newborn children bedtimes	0.48
Unknown	workarounds reposting malicious voicemails emboldening excerpted harpersbazaar screenshots mischaracterizing defamatory incriminating formatted manipulates maliciously repost screenshot keystrokes enraging downloaded fallible poignant undeleted snapshots overwritten succinct	0.42
Sports	women bicycle home races bike boats racing run wheelchair floors wife walking Minnesota race rentals volleyball Tennessee girls couples basketball clubs flying cars beach golf	0.33
Unknown	bellefonte balcones ellijay intracoastal titusville asbury masterson kander riverhead hallandale whidbey bridgehampton hiawatha bedminster boylston rossville schertz bushnell chaska rayden riverdale boothbay simcoe deerfield millcreek	0.3

Table 5: Examples of topics identified by our approach, in terms of extracted topic descriptors, topic coherence scores C_v, and manually-assigned labels.