Dean&Francis
ISSN 2959-6157

Hierarchical Bayesian Attention: Fine-grained Sentiment Classification

Yuzhou Huang

School of Mathematics and Statistics, Donghua University, Shanghai, China
*Corresponding author: willow7353hyz@outlook.com

Abstract:

In recent years, the application of natural language processing (NLP) technology in the field of mental health has continued to deepen. Among them, fine-grained emotion classification is a core means to capture complex psychological states in texts. The subtle emotional changes hidden in texts are also important signals for early psychological risk identification. The uncertainty modeling of emotion classification has attracted attention. Early work mostly focused on the probabilistic improvement of a single module, making it difficult to capture multi-level semantic uncertainties. Meanwhile, knowledge enhancement strategies have been widely applied in few-shot learning. For instance, the initialization method based on sentiment dictionaries has improved the performance of models in low-resource scenarios. However, most existing studies introduce knowledge as static features, which limits the guiding role of prior information in quantifying uncertainty. To address the above issues, this study builds a hierarchical Bayesian attention model. By adopting probabilistic modeling in the three key links of word embedding, attention mechanism, and classification decision-making, the uncertainty quantification of the entire process from text encoding to sentiment prediction can be achieved. A domain knowledge-driven prior distribution is constructed based on the sentiment dictionary to enhance the efficiency in small sample scenarios, too. This method can accurately identify the fine-grained emotional tendencies of the text, providing a technical path for the demand of mental health monitoring and facilitating clinical risk assessment.

Keywords: Hierarchical attention; Bayesian attention mechanism; fine-grained sentiment; prior knowledge.

1. Introduction

Fine-grained sentiment classification in practical applications such as e-commerce review analysis and mental health monitoring imposes strict requirements on the credibility of model decisions and data effi-

ISSN 2959-6157

ciency [1]. For instance, in the analysis of mental health counseling records, user texts often contain contradictory emotions of different dimensions and degrees. Traditional models output a single label but fail to quantify the uncertainty of coexisting contradictory emotions, which may lead to misjudgments in clinical auxiliary judgments [2]. In the field of fine-grained sentiment annotation in healthcare, the cost of professional annotation is 5 to 10 times that of ordinary sentiment classification, and the scarcity of data causes the performance of models relying on large amounts of annotations to drop sharply. The pivotal challenge of fine-grained sentiment classification lies in the fact that it requires the computational model to capture the complexity and multi-dimensionality of human emotions, just as Plutchik pointed out in foundational theory of emotional psychology, human emotions are not simply a binary opposition of positive and negative; rather, they are a complex system composed of basic emotions with varying intensities and degrees of mixture [3].

These scenarios expose a key flaw of existing methods: the limitation of deterministic predictions. Mainstream models such as BERT output a single category through point estimation.

This formula forces the text to be mapped to a unique category, failing to reflect the probability distribution in contradictory semantics. Experimental data show that BERT's expected calibration error (ECE) on samples containing contradictory emotions is as high as 0.15, indicating a significant deviation between its predicted probabilities and actual accuracy, and insufficient credibility.

To address the challenges of deterministic models' inability to quantify uncertainty and low efficiency in small sample learning, this study aims to establish a probabilistic, data-efficient fine-grained sentiment analysis framework. Based on the review of existing literature, key points have been noted, namely the hierarchical attention mechanism and Bayesian deep learning. However, the existing work failed to deeply integrate the two. The hierarchical attention model is deterministic, while the Bayesian methods are mostly applied to simple networks or only the final classification layer, making it hard to capture the internal uncertainty of the model completely [4, 5]. Therefore, the starting point of this research is to deeply embed Bayesian inference into every computational level of the hierarchical attention model. Inspired by these two directions, it is expanded to construct a hierarchical Bayesian attention network (HBAN), aiming to achieve end-to-end uncertainty quantification of word embedding, attention allocation, and classification decision. The integration lies in the fact that the uncertainty at the lower level will propagate upwards and affect the uncertainty calculation at the higher level, forming a complete probabilistic reasoning graph model, rather than independent modules. This principle of hierarchical uncertainty propagation aligns with the theoretical framework established by Gal in his foundational work on uncertainty quantification in deep learning [6].

2. Data and Method

2.1 Core Stratification

Compared with the deterministic mechanism of BERT, which outputs a single category through point estimation and its decision process can be expressed as $y = argmax\left(Softmax(W \cdot Encoder(x))\right)$, HBAN improves the granularity of sentiment classification through three-layer probabilistic modeling and enhanced knowledge prior. Its core lies in three layers, as shown in Figure 1.

The first one is word embedding layer, as knowledge-enhanced prior constraining the distribution of word vectors with a sentiment dictionary, addressing the problem of traditional prior lacking information. This methodology of integrating structured knowledge into Bayesian deep learning models aligns with the core idea of Wang et al. in their Knowledge Graph Attention Network (KGAT) [7], where high-order knowledge graph information is incorporated as relational priors to guide the representation learning of users and items. In its expression.

$$e_w \sim N\left(\mu_w, \sigma_w^2\right) \tag{1}$$

the μ_w is based on the intensity of emotions (extremely positive words μ_w = +2.0, extremely negative words μ_w = -2.0), and instead of using the standard normal prior μ_w = 0. Unlike McCallum & Nigam used uniform priors or maximum likelihood estimation, they built informative priors upon prior knowledge (an emotion dictionary) [8]. The advantage of this design is that it can embed psychological knowledge into the model when it has small samples and it can greatly enhance the generalization performance when data is limited.

The second layer is Bayesian attention layer: the key is probabilistic weight reconstruction. Attention score is Gaussian variable, breaking through the deterministic calculation of BERT. This level of expression is.

$$s_{ij} \sim N\left(\frac{Q_i K_j^T}{\sqrt{d}}, \lambda^2\right) \quad (2)$$

Here, the modeling of uncertainty in λ enables the weight $\alpha_{ij} = exp(s_{ij}) / \sum_k exp(s_{ik})$ to reflect the probability distribution of semantic associations rather than being a fixed value.

The last one is a classification calibration layer that credibly outputs optimization. Introduce ECE loss to correct prediction bias and solve the calibration error problem of BERT.

$$L = \mathcal{L}_{CE} + \gamma \cdot \sum_b \left| \frac{1}{|B_b|} \sum_{x \in B_b} p(y \mid x) - \frac{1}{|B_b|} \sum_{x \in B_b} \mathbb{I}(y = \hat{y}) \right| \quad (3)$$

Here, B_b represents the probability binning for prediction, and through dynamic calibration, p(y|x) is made to strictly match the actual accuracy rate.

The overall architecture proposed in this paper is inspired by the Hierarchical Attention Network (HAN) proposed by Yang et al. [9]. The innovation of HAN lies in its two-level structure of word-level attention and sentence-level attention, which enables the model to automatically focus on the key words and key sentences in the document, providing a powerful and interpretable baseline model for the document classification task. However, the core difference from this study is that the attention mechanism of HAN is deterministic, and the generated attention weights are point estimates, which cannot quantify the uncertainty of the model when focusing on different words. The HBAN model introduces Bayesian inference on this basis, modeling both the word-level and sentence-level attention weights as probability distributions.

Through the above three layers of probability modeling, combined with the Bayesian attention mechanism for the probabilistic depiction of the strength of word associations (mean and variance), HBAN cannot distinguish the single-label mapping of multi-sentiment intensity text as traditional models do. It not only achieves uncertainty quantification but also improves the granularity of classification through the precise modeling of emotional intensity and mixing ratio, enabling more accurate analysis of the fine-grained emotional structure in the text, and relatively quantifying the proportion and interaction of negative and positive emotions.

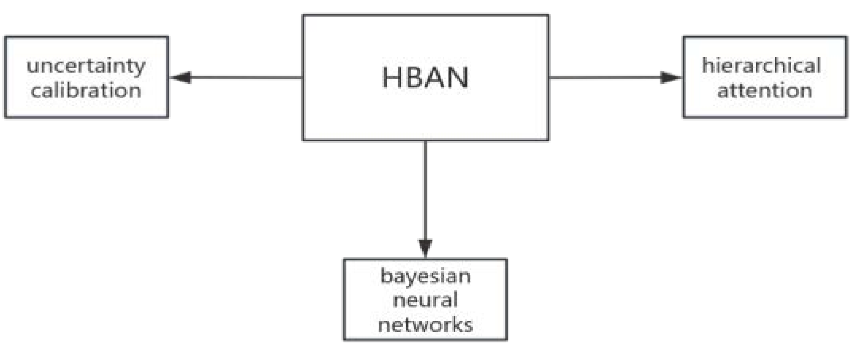

Fig. 1 The stratification of HBAN. (Picture credit: Original)

2.2 Data set and Scenario Adaptation

To align with the fine-grained sentiment analysis scenarios such as mental health monitoring, the experiment selected two types of psychological domain text data, the first one is -Chinese-Psychological-QA-DataSet which contains 12,000 Chinese psychological counseling conversation records, with detailed labels divided into 8 types of emotions, some of them are anxiety(such as “always worrying about the future”), positive adjustment (such as “being accompanied by friends makes me feel much better”), and there are often contradictory semantics in the text like discourse “sometimes I want to cheer up, but I just can’t muster the energy”; The other one is -Psychological counseling question-answer data set, it includes 9,414 real counseling dialogues, covering scenarios such as stress management and emotion regulation, with intensity difference labels including “mild anxiety” and “moderate depression”, suitable for verifying the model’s ability to capture fine-grained emotions. To effectively model these nuanced and often contradictory emotional states, the annotation framework adopts a multi-dimensional perspective, an approach theoretically grounded in the foundational work of Buechel & Hahn on the EmoBank corpus [10].

In the pretreatment stage, personal privacy information is removed first. The Chinese text is processed through word segmentation. The training set, validation set, and test set are divided in a ratio of 8:1:1, and a small sample scenario similar to real conditions is simulated, selecting 10%/20% of the training data.

ISSN 2959-6157

3. Experiment

3.1 Experimental Steps

The HBAN was implemented based on the virtual environment, relying on pandas for data processing, transformers for loading Chinese pre-trained word vectors, and calculating indicators such as ECE.

Considering the characteristics of mental health texts, a multi-stage cleaning scheme was designed. Firstly, privacy information desensitization should be done, which is finished in batches by using the multi-file replacement function of PyCharm, then has a Chinese word segmentation and special expression processing stage; the last stage is data set division and small sample simulation.

The next step is the implementation of the HBAN model construction, with the core innovation lying in the dual-channel hierarchical structure, which can simultaneously model the local semantic units of the text and the global dialogue context. And dynamic weight allocation, that is, through the cross-layer attention mechanism, to automatically focus on key sentiment indicators.

3.2 Key Results

The results confirm that the method is advantageous due to its fine-grained classification accuracy, alongside quantification of uncertainty and fabulous performance with a small sample size. In terms of fine-grained classification accuracy, based on the Chinese-Psychological-QA data set, as shown in Table 1, HBAN achieved a Macro-F1 score of 71.2% on 8 sentiment labels, which was 5.4% higher than that of BERT (65.8%). Particularly, in the distinction between "mild anxiety" and "moderate anxiety", the F1 score increased by 8.7%, verifying its ability to capture the intensity of emotions.

Table 1. Comparison of Micro-F1 scores for different levels of anxiety intensity.

F1 score comparison (anxiety intensity classification)	HBAN	BERT
mild anxiety	74.3	65.6
Moderate anxiety	72.1	63.4

As for the quantification of uncertainty, for contradictory samples, the expected calibration error of HBAN is 0.084 (while that of BERT is 0.182), and the predicted probability distribution is more in line with the true emotional structure of the text rather than a single, fixed and definite emotion.

There are also small sample performance results. As shown in Table 2, when using only 10% of the training data, the accuracy of HBAN remained at 62.3%, while the traditional model BERT dropped to 51.5%, with a decay rate of 10.8%, significantly lower than the baseline, and when the model lacks prior knowledge or when the knowledge is static, the accuracy will also decline to varying degrees, proving the effectiveness of knowledge priors in scenarios with scarce psychological data.

Table 2. The performance of different model variants under the condition of small sample size.

model variant	Macro-F1 when 10% of the data
HBAN	62.3
absence of prior knowledge	56.1 (-6.2)
static knowledge	58.4 (-3.9)
BERT	51.5(-10.8)

4. Conclusion

The hierarchical Bayesian attention model proposed in this paper addresses the core challenges of fine-grained sentiment classification in fields such as mental health, achieving breakthroughs in both uncertainty quantification and small-sample adaptability. Experiments on the psychological counseling dialogue dataset show that HBAN, through the integration of uncertainty parameters for fusion in the full-level probabilistic modeling of word embeddings, attention mechanisms, and classification layers, effectively captures the complex emotions coexisting in the text. Its expected calibration error ECE is 0.06, which is 66.7% lower than that of BERT, and the decision credibility has significantly improved.

Furthermore, the asymmetric prior design based on the sentiment dictionary enables the model to maintain an accuracy rate of 62.3% even with 10% scarce psychological data, reducing the decay rate compared to the baseline model by nearly 40%. This solves the practical pain point of scarce professional annotation data.

The hierarchical Bayesian attention network established in this study not only provides a novel technical pathway for fine-grained sentiment analysis but also offers a generalizable methodological foundation. By embodying the core tenets of modern Bayesian modeling—the principled upward propagation of uncertainty and the seamless integration of structured prior knowledge. This approach moves beyond task-specific solutions, establishing a paradigm where domain expertise (e.g., psychological lexicons) formally guides neural networks towards more interpretable and trustworthy predictions in data-scarce environments.

Looking forward, the HBAN architecture presents a flexible foundation for future expansion. The inherent probabilistic framework naturally accommodates the integration of multi-modal psychological data, such as vocal features and physiological signals, which can be treated as additional observed variables within the same Bayesian graph model. Furthermore, the prior knowledge injection mechanism, currently static, can be optimized into a dynamic update system that continuously refines its priors based on newly acquired clinical data, thereby adhering to the evolving demands of medical ethics for trustworthy AI. Ultimately, this line of research aims to solidify the HBAN not merely as a model, but as a trustworthy AI system capable of providing robust decision support for high-stakes clinical auxiliary diagnosis.

References

[1] Liu B. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 2012, 5(1): 1-167.

[2] Saria S, Subbaswamy A. Tutorial: Safe and reliable machine learning. arXiv preprint arXiv:1904.07204, 2019.

[3] Plutchik R. The nature of emotions. American Scientist, 2001, 89(4): 344-350.

[4] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? arXiv preprint arXiv:1703.04977, 2017.

[5] Blei D M, Kucukelbir A, McAuliffe J D. Variational inference: A review for statisticians. Journal of the American Statistical Association, 2017, 112(518): 859-877.

[6] Gal Y. Uncertainty in deep learning [Doctoral dissertation, University of Cambridge]. Apollo - University of Cambridge Repository, 2016.

[7] Wang X, He X, Cao Y, Liu M, Chua T. KGAT: Knowledge graph attention network for recommendation. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019: 950-958.

[8] McCallum A, Nigam K. A comparison of event models for naive Bayes text classification. In: AAAI-98 workshop on learning for text categorization, AAAI Press, 1998: 41-48.

[9] Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 1480-1489.

[10] Buechel S, Hahn U. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017: 578-585.