Cross-task Interaction Mechanism for Humour Prejudice Detection

Minna Peng[1], Nankai Lin[2,*]

[1]School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, Guangdong, PR China

[2]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, Guangdong, PR China

Abstract

The expression of prejudice is the most common strategy used to hurt people of minority groups. Nowadays, humour becomes a space in which these prejudiced attitudes are maintained. The IberLEF 2023 shared task, titled "HUrtful HUmour," encompasses three distinct subtasks for the humour prejudice detection [1]. From the perspective of multi-task learning, this paper aims to construct a model to deal with three tasks. This paper proposes a cross-task interaction mechanism to increase the interaction between different tasks. The experimental results show the effectiveness of our method. In the final testing phase, our method achieved second place in subtask 1 and fifth place in subtask 2A.

Keywords

Humour prejudice detection, Multi-task learning, Cross-task interaction mechanism

1. Introduction

The expression of prejudice is the most common strategy used to hurt people of minority groups. Prejudice is defined as the negative pre-judgment of members of a race or religion or of any other socially significant group, regardless of the facts that contradict it. The expression of prejudice is an issue directly related to stereotyping. Stereotypes are beliefs about the characteristics of a social group that are originated in a pre-judgment, i.e. a prejudice that regards a certain group as "different". In the present era, characterized by the widespread use of social media platforms, novel avenues have emerged for the propagation of prejudiced views. Often these messages make use of humour to avoid the moral judgment that penalizes discrimination. In fact, when a society begins to overcome its prejudices towards certain social groups, we can observe that humour becomes a space in which these prejudiced attitudes are maintained.

The IberLEF 2023 shared task, titled "HUrtful HUmour," encompasses three distinct subtasks [1]. Firstly, it aims to discern whether a tweet containing prejudice is intended to be humourous. Secondly, it involves identifying the targeted groups within each tweet, which can be considered as a multi-label classification task. Lastly, the task evaluates the degree of prejudice present in the messages, specifically focusing on the average impact on minority groups.

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

2388304434@qq.com (M. Peng); neakail@outlook.com (N. Lin)

0000-0003-2838-8273 (N. Lin)

From the perspective of multi-task learning, this paper aims to construct a model to deal with three tasks. This paper proposes a cross-task interaction mechanism to increase the interaction among different tasks and ensure that the information between tasks can be transferred to each other, so as to improve the performance of each task. The experimental results show the effectiveness of our method. In the final testing phase, our method achieved second place in subtask 1 and fifth place in subtask 2A.

2. Related Work

2.1. Detection of Humour Spreading Prejudice

Recently, humour becomes a space in which these prejudiced attitudes are maintained. Individuals may employ humor as a means of conveying their prejudiced and discriminatory views occasionally. Such humor might appear benign, but in reality it can intensify and perpetuate prejudice and discrimination towards specific groups. Mpofu [2] demonstrated the pernicious utilization of disparagement humour in perpetuating racist tendencies and body-shaming practices, leading to consequential outcomes. Off-colour humour represents a genre of comedy widely criticized for its perceived lack of decorum and excessive vulgarity. This type of humour typically encompasses content featuring derogatory remarks targeting specific ethnic groups or genders, depictions of violence, domestic abuse, sexually explicit acts, and the use of excessive swearing or profanity. Among the various manifestations of off-colour humour, notable subcategories include blue humour, dark humour, and insult humour. While insult humour is explicitly designed to provoke offense, both blue and dark humour often suffer from misclassification due to the presence of insulting and harmful language. Addressing the pressing need to distinguish between dark and blue humour and offensive humour, Ahuja et al. [3] presented an innovative approach and a novel dataset comprising nearly 15,000 instances. Their contribution to resolving this challenge was crucial for preserving unrestricted freedom of speech on the internet.

The field of natural language processing(NLP) has proposed many tasks related to humour detection. The HAHA 2018 [4] and HAHA 2019 [5] competition provided a corpus consisting of Spanish tweets, where each tweet contained two attributes of whether it was humourous and a funniness score. The challenger needed to use this corpus to complete the automatic detection and automatic rating of humour in Spanish tweets, that is, to decide whether a tweet was humourous or not, and to predict a funniness score value of the tweet. Based on the tasks of HAHA 2018 and 2019, the HAHA 2021 competition added two new tasks: humourous logic mechanism classification and humour target classification [6]. However, the above tasks only aim at detecting whether the text is humourous, not at detecting offensive humour. Greenwood and Gautam [7] highlighted the power of social media as a vehicle for disparaging humour to activate, reinforce, and reproduce bias by examining whether gender, anti-obesity attitudes, and sexism influence joke perception, as well as moderate perception of joke-related targets, and highlighted the importance of taking jokes seriously in online Settings. Therefore, it is important to detect and identify such prejudiced humour.

To delve into the impact of hurtful and emotive language on sarcasm detection, Frenda et al. [8] proposed an innovative transducer-based system named AlBERToIS. This approach effectively combined the pre-trained AlBERTo model with linguistic features, leading to superior performance in both sarcasm detection and sarcasm identification tasks, particularly on the IronITA dataset. Building upon this research, Merlo et al. [9] explored the representation of humorous text by leveraging statistically significant differences in various features using data from the HaHackaton task. Through the application of a reduction test, the most relevant features were identified to distinguish non-offensive jokes from highly offensive ones. Merlo [10] employed computational linguistics to discern the distinctive features indicative of the offensive level present in humorous texts. The SemEval 2021 task 7, HaHackathon, introduced a groundbreaking shared task by amalgamating the previously separated domains of humor detection and offense detection. This task involved the manual annotation of 10,000 texts collected from Twitter and Kaggle short joke datasets, with assessments for humor and offense. Contestants were required to predict humor ratings, offense ratings, and determine if the variance in humor ratings surpassed a specific threshold [11].

2.2. Multi-task Learning

In natural language processing (NLP) tasks, multi-task learning models have shown great success. Multi-task learning(MTL) model is a machine learning algorithm that can handle multiple related tasks at the same time and share the learned knowledge between different tasks to improve the overall performance. Most researchers use multi-task learning to improve the performance of NLP related tasks. In multi-task text classification, Xiao et al. [12] innovatively integrated a gate mechanism into a multi-task convolutional neural network (CNN), thereby introducing a novel gated sharing unit. This proposed unit effectively filtered the flow of features across tasks, resulting in a substantial reduction in interference. Liu et al. [13] proposed two architectures for multi-task learning with neural sequence models, which can dynamically learn the relationship between different tasks. At the same time, a general framework of graph multi-task learning was proposed, so that different tasks can effectively communicate with each other. In deep-learning-based facial expression recognition task, Zhao et al. [14] proposed a selective feature sharing method and built a multi-task network for facial expression recognition and facial expression synthesis, which can effectively transfer beneficial features between different tasks and filter out those useless and harmful information. Previous research has established that multi-task learning has achieved good results in processing NLP tasks.

3. Our Method

The architecture of our model is depicted in Figure 1, illustrating the sequential steps involved. Initially, sentences are encoded using the BERT model, enabling the extraction of their semantic representations. Subsequently, the original semantic representations are further expanded through a linear layer, yielding distinct semantic representations corresponding to the three tasks. To foster enhanced interplay among the tasks and facilitate the transfer of information between them, we propose a cross-task interaction mechanism. This mechanism ensures that relevant information is effectively shared across tasks, thereby bolstering the performance of each individual task. The updated task representations, obtained through the cross-task

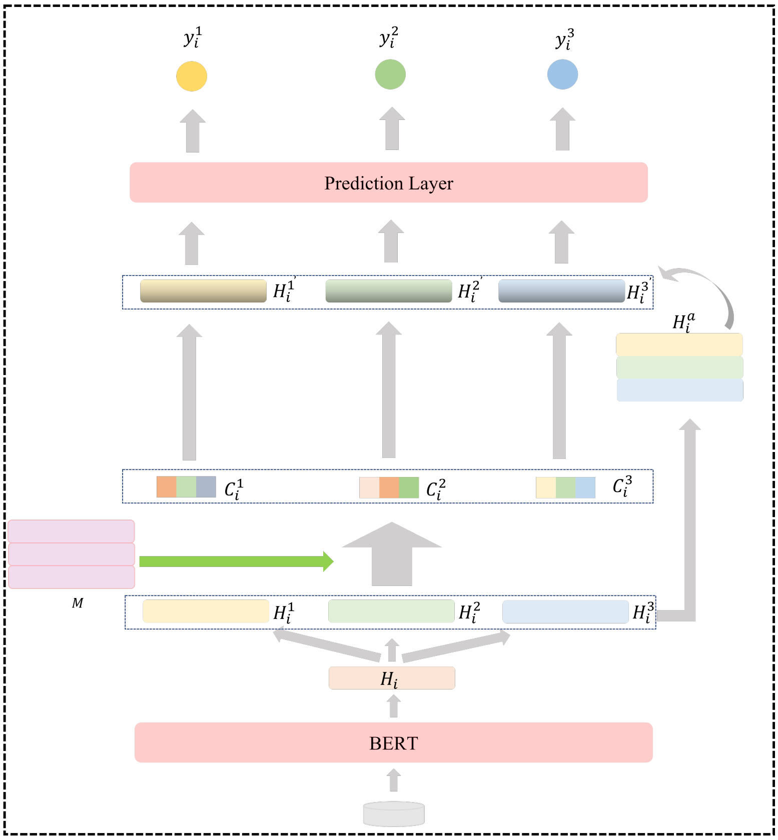

Figure 1: Model Framework Diagram.

interaction mechanism, are then employed for prediction and output generation pertaining to the respective tasks.

3.1. Text Representation

In the text representation module, sentences are encoded using a pre-trained model that performs well in the semantic representation of text. In the pre-training stage, non-autoregressive language models learn general language representations from massive corpora by unsupervised training, and learn a large amount of prior linguistic, syntactic and lexical information for downstream tasks.

The BERT model is a language model based on multi-layer bidirectional Transformer, which internally uses Transformer as the encoding structure. The input to the BERT model is a sum of 3 vectors. For each input token, the representation consists of three parts: token embeddings, segment embeddings, and position embeddings. The token embeddings represent the representation of the current token. The segment embeddings represent the position encoding of the sentence in which the current token is located, and the position embeddings represent the position encoding of the current word. In the classification task, the input sentence uses the unique tokens [CLS] and [SEP] as opening and ending markers. The fully connected layers are connected at the [CLS] position of the last encoder layer, and finally, the softmax layer completes the classification of sentences or sentence pairs. We chose the Spanish version of BERT (bertbasespanishwwmuncased) as the pre-trained model.

The semantic feature H_i is obtained by encoding the input sequence S_i and the calculation process is as follows:

$$H_i = BERT(a_i, b_i, c_i) \tag{1}$$

where a_i, b_i, c_i are the token embeddings, segment embeddings, and position embeddings of the input sequence S_i.

3.2. Cross-task Interaction Mechanism

To enable the model to handle different tasks, we construct three linear layers to project the semantic representation H_i into three different representations:

$$H_i^1 = W_1^T H_i + b_1 \tag{2}$$

$$H_i^2 = W_2^T H_i + b_2 \tag{3}$$

$$H_i^3 = W_3^T H_i + b_3 \tag{4}$$

where, W_1, W_2, W_3, b_1, b_2 and b_3 are the learnable parameters of the three linear layers respectively to learn the information of different tasks.

In order to further increase the interaction between different tasks, that is, information between tasks can be transmitted to each other to improve the performance of each task, we construct a learnable matrix $M \in R^{t \times e}$, where each behavior of the matrix represents a corresponding representation of a task, where t is the number of tasks handled by the model and e is the dimension of semantic representation. The matrix can be backpropagated to learn the representation of each task. By using matrix M to multiply the semantic representation of the three tasks and softmax, we can get the degree of information correlation between each pair of tasks:

$$C_i^1 = softmax(M^T H_i^1) \tag{5}$$

$$C_i^2 = softmax(M^T H_i^2) \tag{6}$$

$$C_i^3 = softmax(M^T H_i^3) \tag{7}$$

For task j, its correlation degree C_i^j represents how much information should be transferred to task j by each task. Based on the information correlation degree, the semantic representation of three tasks is updated and the new semantic representations across tasks are obtained:

$$H_i^{1'} = C_i^1 H_i^a \tag{8}$$

$$H_i^{2'} = C_i^2 H_i^a \tag{9}$$

$$H_i^{3'} = C_i^3 H_i^a \tag{10}$$

where $H_i^a = [H_i^1, H_i^2, H_i^3]$. In essence, the cross-task interaction mechanism can be regarded as a special gate mechanism.

3.3. Text Prediction

We further use the new semantic representations through the cross-task interaction mechanism to predict different tasks.

For subtask 1 humor recognition, we treat it as a binary classification task, and input the semantic representation of task 1 into a linear classifier with the softmax function, which is formulated as follows:

$$y_i^1 = softmax(W_4^T H_i^{1'} + b_4) \tag{11}$$

where W_4 and b_4 are learnable parameters, and y_i^1 is the predicted probability.

For subtask 2A target group identification, we treat it as a multi-label classification task. We predict the targeted groups probability distribution for each sentence S. Its corresponding feature vector $H_i^{2'}$ is put into a linear classifier with the softmax function, which is formulated as follows, where W_5 and b_5 are trainable parameters:

$$y_i^2 = softmax(W_5 \cdot H_i^{2'} + b_5) \tag{12}$$

where $y_i^2 \in R^l$ and l is the number of labels. Since the model assigns one or more labels to each sentence in the subtask 2A, we set a probability threshold μ to assign the labels exceeding the threshold to the corresponding emotions of the sentence:

$$y_{ij}^{2'} = \begin{cases} 1 & y_{ij}^2 > \mu \\ 0 & p_{ij}^2 < \mu \end{cases} \tag{13}$$

where $j \in (1, l)$. $y_{ij}^{2'}$ denotes the assignment of the label j to the corresponding label of the sentence.

For subtask 2B prejudice degree measure, we build a linear classification layer for regressing a continuous value to represent the judgment degree:

$$y_i^3 = W_5^T H_i^{3'} + b_5 \tag{14}$$

where y_i^3 is the predicted prejudice value, and its value range is $\{y | y \geq 0\}$.

The three sub-tasks choose binary cross-entropy, multi-label classification cross-entropy and Huber loss as the loss functions, and the total loss of the model is the sum of the losses of the three sub-tasks.

4. Eeperiments

4.1. Experimental Setup

All experimental procedures are conducted utilizing the NVIDIA 3060 6-GB GPU. The feed-forward layer is initialized using weights drawn from a truncated normal distribution with a standard deviation of 2e−2, while the bias is initialized to zero. A fixed initial learning rate of 2e−5 is consistently applied across all experiments. The maximum sequence length is set to 128, representing the prescribed constraint on the number of tokens within a sentence. To optimize training, a warmup proportion of 1e-3 is implemented. The training episodes span 20 epochs, utilizing a batch size of 8.

In order to ensure a comprehensive evaluation of the effectiveness of our strategies, we employ a 5-fold cross-validation methodology. This approach involves dividing the datasets into five distinct subsets, enabling the construction of an ensemble model that exhibits enhanced generalization capabilities. Among these subsets, four are assigned for training purposes, while the remaining subset is utilized for verification. The evaluation results pertaining to the effectiveness of our strategies are derived by averaging the outcomes obtained from the five cross models.

4.2. Experimental results

Table 1
Main results.

Five-fold Cross Validation	Subtask	Individual Training	Merge Trainging	Our Model
Dev	1	0.7523	0.7523	0.7586
	2A	0.9274	0.9217	0.9244
	2B	0.8056	0.8047	0.8039
Test	1	-	0.7980	0.7990
	2A	-	0.7540	0.7580
	2B	-	0.9370	0.9410

As presented in Table 1, we conducted a comparative analysis of three distinct models trained individually, the integration of the three tasks during training, and the incorporation of a cross-task interaction mechanism within the combined training approach. The experimental findings highlight the effective performance of our proposed method, particularly evident in achieving the best results for subtask 1 and subtask 2B during five-fold cross-validation.

During the final evaluation phase, our method exhibited exemplary performance by achieving the highest scores in subtask 1 (0.7990) and subtask 2A (0.7580). Notably, our approach secured the second position in subtask 1 and the fifth position in subtask 2A on the leaderboard, further underscoring its competitive performance within the task benchmarks.

5. Conclusion

Within the context of multi-task learning, the primary objective of this study is to devise a model capable of effectively addressing three distinct tasks. To foster increased interplay between these tasks, a cross-task interaction mechanism is proposed. The experimental findings corroborate the efficacy of our proposed approach. Notably, during the final testing phase, our method attained a commendable position, securing second place in subtask 1 and fifth place in subtask 2A.

Acknowledgments

This work was supported by the Guangdong Philosophy and Social Science Foundation (No. GD20CWY10), the National Social Science Fund of China (No. 22BTQ045), and the Science and Technology Program of Guangzhou (No.202002030227).

References

[1] R. Labadie-Tamayo, B. Chulvi, P. Rosso, Everybody hurts, sometimes. overview of hurtful humour at iberlef 2023: Detection of humour spreading prejudice in twitter, in: Procesamiento del Lenguaje Natural (SEPLN), volume 71, 2023.

[2] S. Mpofu, 'If Ever I Offended You I Am Sorry': Disparagement Humour, Black Twitectives and the Dream Deferred, Springer International Publishing, Cham, 2021, pp. 215–231. URL: https://doi.org/10.1007/978-3-030-81969-9_11. doi:10.1007/978-3-030-81969-9_11.

[3] V. Ahuja, R. Mamidi, N. Singh, From humour to hatred: A computational analysis of off-colour humour, in: M. Zhang, V. Ng, D. Zhao, S. Li, H. Zan (Eds.), Natural Language Processing and Chinese Computing, Springer International Publishing, Cham, 2018, pp. 144–153.

[4] S. Castro, L. Chiruzzo, A. Rosá, Overview of the haha task: Humor analysis based on human annotation at ibereval 2018, in: IberEval@SEPLN, 2018.

[5] L. Chiruzzo, S. Castro, M. Etcheverry, D. Garat, J. J. Prada, A. Rosá, Overview of haha at iberlef 2019: Humor analysis based on human annotation, in: IberLEF@SEPLN, 2019.

[6] L. Chiruzzo, S. Castro, S. Góngora, A. Rosá, J. A. Meaney, R. Mihalcea, Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish, Proces. del Leng. Natural 67 (2021) 257–268.

[7] D. Greenwood, R. Gautam, What's in a tweet? gender and sexism moderate reactions to antifat sexist humor on twitter, HUMOR 33 (2020) 265–290. URL: https://doi.org/10.1515/humor-2019-0026. doi:doi:10.1515/humor-2019-0026.

[8] S. Frenda, A. T. Cignarella, V. Basile, C. Bosco, V. Patti, P. Rosso, The unbearable hurtfulness of sarcasm, Expert Systems with Applications 193 (2022) 116398. URL: https://www.sciencedirect.com/science/article/pii/S0957417421016870. doi:https://doi.org/10.1016/j.eswa.2021.116398.

[9] L. I. Merlo, B. Chulvi, R. Ortega-Bueno, P. Rosso, When humour hurts: linguistic features to foster explainability, Proces. del Leng. Natural 70 (2023) 85–98.

[10] L. I. Merlo, When humour Hurts: A Computational Linguistic Approach, Ph.D. thesis, Universitat Politècnica de València, 2022.

[11] J. A. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, W. Magdy, SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 105–119. URL: https://aclanthology.org/2021.semeval-1.9. doi:10.18653/v1/2021.semeval-1.9.

[12] L. Xiao, H. Zhang, W. Chen, Gated multi-task network for text classification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 726–731. URL: https://aclanthology.org/N18-2114. doi:10.18653/v1/N18-2114.

[13] P. Liu, J. Fu, Y. Dong, X. Qiu, J. C. Kit Cheung, Learning multi-task communication with message passing for sequence learning, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 4360–4367. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4346. doi:10.1609/aaai.v33i01.33014360.

[14] H. Zheng, R. Wang, W. Ji, M. Zong, W. K. Wong, Z. Lai, H. Lv, Discriminative deep multi-task learning for facial expression recognition, Information Sciences 533 (2020) 60–71. URL: https://www.sciencedirect.com/science/article/pii/S0020025520303601. doi:https://doi.org/10.1016/j.ins.2020.04.041.