Prompt Based Framework for Violent Event Recognition in Spanish

Guanqiu Qin[1], Junheng He[1], Qifeng Bai[1], Nankai Lin[1], Jigang Wang[1], Kang Zhou[1], Dong Zhou[1] and Aimin Yang[1,*]

[1]School of Computer, Guangdong University of Technology, Guangzhou 510006, China

Abstract

Violent incident detection is a new task in recent years. Violent content on social networking platforms harms people's psychology but also helps people track and prevent incidents that are happening in time. The DAVINCIS@IberLEF 2022 shared task is a natural language problem that aims to detect violent incidents on Spanish tweets. In this task, we participated in subtask 2, namely, violent event category recognition via multi-class multi-label classification. We built a framework based on prompt learning and AsyLoss to solve the problem of data imbalance in the dataset. In order to enhance the performance of classification, we built an extra feature extraction module to enhance the feature extraction capability of the framework, and then we fused those different features in the multi-label classification module In the final phase of subtask 2, our approach achieves the best results.

Keywords

Prompt learning, CNN, Bi-LSTM, Multi-label classification, Data imbalance

1. Introduction

Violence has a significant negative impact on those who witness or experience violence which includes a high incidence rate of depression, anxiety, post-traumatic stress disorder [1]. The government is responsible for ensuring the safety of the people, but violent incidents will directly and seriously endanger the safety of the people. For ongoing violence, timely violence detection can provide the public sector with the conditions for rapid response. Therefore, timely violence detection and monitoring are of great significance to society, the government and the people. Twitter has gradually become one of the most popular social networking platforms in the world. Every day, a large number of tweets are published in real-time, including news, personal views and experience sharing, which brings urgent needs to the task of violence detection and monitoring, and also can help to achieve timeliness. However, traditional event detection is usually accomplished after event trigger extraction, therefore, the short text of tweets presents a challenge to traditional event detection.

IberLEF 2022 organized a Spanish twitter violence detection and monitoring task called DAVINCIS [1]. The task includes two subtasks: (i) violent event identification, which determines whether a given tweet is related to a violent event according to the content of the tweet; (ii)

IberLEF 2022, September 2022, A Coruña, Spain.

*Corresponding author.

violent event category identification, which determines whether the tweet belongs to an accident, murder, non-violent event, robbery, and kidnapping according to the content of the tweet. In particular, different from the traditional event detection methods that need trigger words, subtask 2 is set as a text multi-label multi-classification task according to the characteristics of tweets. Based on prompt learning, our model uses Bidirectional Long Short Term Memory (Bi-LSTM) and Convolutional Neural Network (CNN) to learn context features, and label-text further attention. In addition, we deployed AsyLoss as the model objective function for the case of data imbalance. Our method works best. In this report, we will review the solutions to this task. An example of the numbered list is as follows.

2. Related Work

Violence detection is a new task emerging with the rise of social networks. Nowadays, many countries in the world require that if the content published by users on social networking platforms contains bad speech information about violence, the content should be restricted or filtered in time to avoid adverse effects. At the same time, timely detection of accidents in the community can effectively prevent further development of the situation. This task is usually set as text or image classification, Yuvaraj et al. [2] used Artificial Neural Network (ANN) and Deep Reinforcement Learning (DRL) method to classify Cyberbullying in social network text. Won et al. [3] used ResNet to detect and perceive violence of pictures on social networks. In addition, Lam et al. [4] proved that using both images and text as input is better than a single of them. Spanish is one of the three languages most used on social media. In recent years, there have been many studies on violent content detection in Spanish, Plaza-Del-Arco et al. [5] based on a dictionary, detect misogyny and xenophobic violence in Spanish tweets via supervised deep learning. Schick et al. [6] used vocabulary resources to detect aggression in Mexican Spanish tweets.

Multi-label Classification is an extension of Single-label Classification. Most studies are based on Pre-training language models such as Bert [7]. For multi-label classification, the simplest solution is to set a binary classifier for each label. Xiao et al. [8] proposed that acquire labels' representation by embedding, and then fusing them with word embedding as the input of the classifier can achieve better results. For the potential relevance of labels in multi-label tasks, Xiao et al. [8] and Pal et al. [9] used the label-aware attention method to improve the prediction accuracy of combinations between labels, Zhang et al. [10] used multi-task learning to find the relationship of labels.

Fine-tuning with pre-trained language model (PLM) is now a common method in NLP tasks. However, in the case of some low resources and little information contained in the data, fine-tuning cannot perform well in the downstream tasks. So, a new training mode – Prompt learning came into being. Instead of adapting the PLM to downstream tasks, prompt learning reformats downstream tasks, makes the PLM more like solving tasks with the help of text prompts. Simply put, Huerta-Velasco et al. [11] redefines the input examples as cloze style phrases to help the LM understand a given task. These phrases are then used to assign soft labels to a large number of unlabeled corpus, and finally perform standard supervised training. This method can manipulate the behavior of the model so that the LM itself can be used to predict the desired output. The

advantage of cue learning is that given appropriate cues, the LM solves a large number of tasks through pre-training. Recent research on Prompt-learning is like Automatic-prompt [12]: Methods for automatically creating prompts for various tasks; Continuous-prompt [13]: Solve the problem that discrete prompts will modify model parameters, etc.

3. Dataset

The data sets of subtasks 1 and 2 are trained and validated by the same corpus. The number of training samples is only 3360. Subtask 1 is a binary classification task, and each text is marked as related to or not related to violence. The distribution of the two labels tends to be 1:1. However, as an extension of subtask 1, labels in subtask 2 include Accident, Homicide, Non-Violent-incident, Robbery and Kidnapping. Except for "non-violent incidents", other labels are subdivisions of the labels belonging to violent incidents in subtask 1, which will bring challenges of data imbalance.

Table 1
The number of labels in the training set of DAVINCIS subtask 2.

Label ID	Category	Training Set	Co-present Labels
0	Accident	1124	1
1	Homicide	206	0,3,4
2	Non-Violent-incident	1798	-
3	Robbery	179	1
4	Kidnapping	45	1

Table 2
The number of labels in the validation set of DAVINCIS subtask 2.

Label ID	Category	Training Set	Co-present Labels
0	Accident	11	-
1	Homicide	5	4
2	Non-Violent-incident	27	-
3	Robbery	5	-
4	Kidnapping	2	1

We analyzed the label distribution of subtask 2 in the DA-VINCIS dataset, as is shown in table 1 and table 2, we found that the dataset had the following data imbalances [14]: (1) the distribution of labels is unbalanced; (2) the combination of labels is unbalanced; (3) the interior of labels is unbalanced. At the same time, we found that the labels are not only related to the described event, but also related to the context, and dev set has multi-label combination sample that is not found in the train set. Therefore, in order to obtain better performance, we need the model not only to be able to learn text features well, but also to have the ability to learn with few-shot learning and zero-shot learning.

4. Methodology

The framework we propose mainly consists of three parts, namely prompt learning module, feature extraction module and multi-label classification module. In the first part, we transform the multi-label classification problem into the span prediction problem with prompt learning that jointly captures the relevant semantic representation of input text and label tokens. In the second part, we extract and fuse the context feature and local feature of the representation input text. In the last part, we combine all the features learned in the previous part as input of the classifier, then perform the multi-label classification.

4.1. Prompt learning module

Prompt learning is more similar to the task of pretrain language models (PLM), which can decrease the gap of training objectives between PLMs and downstream fine-turnings, and can also utilize more knowledge PLMs learned. Many works of prompt learning would transform problems as MLM tasks, e.g., for subtask 1, we can reconstruct the input text as: "*x*, violence is _ to this sentence", and ask PLMs to fill the blank as 'relevant' or 'irrelevant'. In this way to build prompt for subtask 2, we must build as many prompts as labels and can implement them in two ways: (i) combine the prompts into one; (ii) perform *N* times prediction with *N* prompts for each sample. However, this kind of prompts are inefficiency for multi-label classification in subtask 2. The first method brings too many duplicate prompt tokens and occupies a lot of input space in PLMs that makes the problem complicated. The second method takes longer to compute.

Binary classification prompt:	***x***, *violence* is __ to this sentence
Multi-lable classification prompt 1	***x***, *accident* is __ to this sentence, *murder* is __ to this sentence, ·····
Multi-lable classification prompt 2	Input 1: ***x***, *accident* is __ to this sentence Input 2: ***x***, *murder* is __ to this sentence .....
Our prompt:	**Accident murder peace rob or kidnap? *x***

Figure 1: *x* is the origin input text sequence; orange or blue fonts are prompts.

As is shown in Figure 1, different from the transformation of binary classification into prompt learning, we transform subtask 2 into a span prediction problem. For instance, we use specific tokens to represent labels: 'Accident', 'Asesinato', 'Paz', 'Robo' and 'Secuestro' correspond to Accident, Homicide, Non-Violent-incident, Robbery and Kidnapping labels. To adapt the dictionary of PLM and to make the non-violent labels more distinct from other labels, especially, we take 'Paz'(peace) as the non-violent-incident label's token which has a large semantic space distance from the other four label tokens. Following SpanEmo [15], we adopt a weak prompt that mainly consists of label tokens. To be specific, we transform the input text x into $\hat{x}$ =

accidente asesinato paz robo o secuestro? x, which is prompt "Accident murder peace robbery or kidnapping? x " in Spanish. We get text and label token representation via BETO [16], a BERT based Spanish PLM. For an origin text input $X = \{x_1, x_2, ..., x_n\}$, the prompt would be $\hat{X} = \{y_1, y_2, \dots, y_4, p_1, y_5, p_2, x_1, x_2, \dots, x_n\}$, where $p_i \in P$ is prompt token, and $y_j \in Y$ is label token.

The input of BERT needs its' special tokens: [CLS] and [SEP]. Formally, we let the token sequence $\{[CLS], y_1, y_2, ..., y_4, p_1, y_5, p_2, [SEP], x_1, x_2, ..., x_n, [SEP]\}$ as input of PLM, and then we get the representation of the token sequence $\left\{h_{[CLS]}, h_{y_1}, \dots, h_{x_n}, h_{[SEP]}\right\}$, each token representation $h_i \in \mathbb{R}^k$, where k indicates the feature size of PLM output. Next, we extract the required labels' token representation $H_Y = \left\{ h_{y_1}, h_{y_2}, \dots, h_{y_5} \right\}$ and text token representation $H_X = \left\{ h_{x_1}, h_{x_2}, \dots, h_{x_n} \right\}$.

It's worth noting that the prompt mentioned above is naturally suitable for multi-label classification, for we can get label representation from that way, besides, the model can learn label-wise attention, label-text attention, text-label attention. Co-occurrence information on labels would be learnt by PLM as label-wise attention.

4.2. Feature extraction module

Context feature learning: To get more context information in input text for the model, we employ a bidirectional LSTM [17] layer to learn the context feature in the text representation H_X. At the time-step t in LSTM the hidden state $g_t \in \mathbb{R}^k$ updated with the current input and the $(t-1)$th step hidden state:

$$\overrightarrow{g_t} = LSTM\left(\overrightarrow{g_{t-1}}, h_{x_{t-1}}\right) \tag{1}$$

$$\overleftarrow{g_t} = LSTM\left(\overleftarrow{g_{t-1}}, h_{x_{t-1}}\right) \tag{2}$$

where g_t, $g_t \in \mathbb{R}^k$, indicate two different direction of LSTM hidden state at the time-step t respectively. Then we can get $G \in \mathbb{R}^{n \times k}$, the text representation with context information, where n is the length of text:

$$\overleftarrow{G} = \left(\overleftarrow{g_1}, \overleftarrow{g_2}, \dots, \overleftarrow{g_n}\right) \tag{3}$$

$$\overrightarrow{G} = \left(\overrightarrow{g_1}, \overrightarrow{g_2}, \dots, \overrightarrow{g_n}\right) \tag{4}$$

$$G = \left(\overrightarrow{G}, \overleftarrow{G}\right) \tag{5}$$

Label-Text Fusion with Context Information Injecting: To adapt the task of the prompt learning: predicting the probability of right label tokens, we first incorporate the context feature learn from Bi-LSTM into label representation via dot product:

$$D = GH_Y^T \tag{6}$$

where $D \in \mathbb{R}^{n \times 5}$. Then we utilize CNN to extract label-word co-relation feature with ReLU activation and max-pooling in the function Φ, and we get a vector $a \in \mathbb{R}^5$ which contents a measure of each label associated with the input text:

$$a = tanh\left(\Phi(D)\right) \tag{7}$$

4.3. Multi-label Classification module

We use a Fully Connected Layer (FC) as our framework's classifier after fusing the features we learnt in section 4.1 and section 4.2. To be specific, we take a as an attention score vector and we built the input of Fully Connected Layer K after finding the cross product of the labels' token representation H_Y and a:

$$H_Y^{'} = H_Y \times a \tag{8}$$

$$K = H_Y + H_Y^{'} \tag{9}$$

Finally, we obtain the label prediction from the formula as follows:

$$\hat{y} = sigmoid\left(FC(K)\right) \tag{10}$$

where $\hat{y} \in \mathbb{R}^5$ contains prediction of 5 labels. Note that we employ a sigmoid function to determine whether the $label_i$ is predicted by the rule: $\hat{y}_i \geq threshold = 0.5$.

4.4. Asymmetric Loss

We use Asymmetric Loss [18] as the loss function, the formula is as below:

$$ASL = \begin{cases} L_+ = (1-p)^{\gamma_+} + \log(p) \\ L_- = (p_m)^{\gamma_-} + \log\left(1 - p_m\right) \end{cases} \tag{11}$$

where p is the prediction probability, γ_- and γ_+ is the positive and negative focusing parameters respectively, and set $\gamma_- > \gamma_+$ to emphasize the contribution of positive sample. In addition, p_m is the shifted probability which performs hard thresholding of very easy negative samples that can discard negative samples with very low probability in training.

5. Results and discussion

Due to the limitation of submission and time, we just submitted three results in Final Subtask 2. The task result was measured by the macro-F1 score, two of our submissions had the best performance among the participants. In order to tackle the noisy dataset problem, we tried to introduce FGM [19] to improve model robustness and generalization, but it brings an unsatisfactory result. Comparison of the baseline and our methods is shown in table 3.

Note that Label-Correlation Aware Loss (LCA) is proposed by Yeh et al. [20], which is implemented with binary cross-entropy (BCE). Because of similar performance, we analyze the result rationality of the two best methods in Table 4. It's intuitively obvious that label 2 (non-violent-incident) and other labels are mutual exclusion, so we counted the number of predictions of label 2 and its co-occurrence with other labels. We found that the method

Table 3
Results of methods on DAVINCIS subtask 2.

Method	Macro-F1	Recall	Precision
Baseline	0.4981	0.46	0.57
Prompt+LCA Loss+FGM	0.509469	0.541333	0.489858
Prompt+BiLSTM+LCA Loss	0.551313	**0.581856**	0.542372
Prompt+BiLSTM+CNN+AsyLoss	**0.554281**	0.562260	**0.550030**

Table 4
Statistics for label 2 in the prediction results of the two optimal models.

Method	Total	Non-Exclusive	Reasonable Rate
Prompt+BiLSTM+LCA Loss	747	65	0.9130
Prompt+BiLSTM+CNN+AsyLoss	774	87	0.8876

with LCA brings a result more in accordance with the actual situation, thus, it is necessary to introduce LCA or a module to learn label relation based on the importance of label relation in the dataset.

6. Conclusion

In this paper, we introduce a multi-label classification framework based on prompt learning. We built a module for context feature and local feature extraction and fusion based on Bi-LSTM and CNN. We utilize AsyLoss as our loss function to alleviate data imbalance. Our method achieves the best result in subtask 2 of DAVINCIS@IberLEF 2022. In addition, we analyze the rationality of results by different methods and find that it is necessary to introduce a module for learning labels relation in multi-label classification. In future work, we will focus on the ensemble of multi-label classification and labels co-occurrence prediction.

Acknowledgments

This work is partially supported by the Ministry of education of Humanities and Social Science project [grant numbers 19YJAZH128 and 20YJAZH118], Science and Technology Plan Project of Guangzhou [grant number 202102080305].

References

[1] L. V.-P. M. M. y. G. F. S.-V. Luis Joaquín Arellano, Hugo Jair Escalante, Overview of da-vincis at iberlef 2022: Detection of aggressive and violent incidents from social media in spanish, Procesamiento del Lenguaje Natural 69 (2022).

[2] N. Yuvaraj, K. Srihari, G. Dhiman, K. Somasundaram, A. Sharma, S. Rajeskannan, M. Soni, G. S. Gaba, M. A. AlZain, M. Masud, Nature-inspired-based approach for automated

cyberbullying classification on multimedia social networking, Mathematical Problems in Engineering 2021 (2021).
[3] D. Won, Z. C. Steinert-Threlkeld, J. Joo, Protest activity detection and perceived violence estimation from social media images, in: Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 786–794.
[4] V. Lam, S. Phan, D.-D. Le, D. A. Duong, S. Satoh, Evaluation of multiple features for violent scenes detection, Multimedia Tools and Applications 76 (2017) 7041–7065.
[5] F.-M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, Detecting misogyny and xenophobia in spanish tweets using language technologies, ACM Transactions on Internet Technology (TOIT) 20 (2020) 1–19.
[6] T. Schick, H. Schütze, Exploiting cloze questions for few shot text classification and natural language inference, arXiv preprint arXiv:2001.07676 (2020).
[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[8] L. Xiao, X. Huang, B. Chen, L. Jing, Label-specific document representation for multi-label text classification, 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), 2019, pp. 466–475.
[9] A. Pal, M. Selvakumar, M. Sankarasubbu, Multi-label text classification using attention-based graph neural network, arXiv preprint arXiv:2003.11644 (2020).
[10] X. Zhang, Q.-W. Zhang, Z. Yan, R. Liu, Y. Cao, Enhancing label correlation feedback in multi-label text classification via multi-task learning, arXiv preprint arXiv:2106.03103 (2021).
[11] D. A. Huerta-Velasco, H. Calvo, Using lexical resources for detecting offensiveness in mexican spanish tweets., in: IberLEF@ SEPLN, 2021, pp. 240–250.
[12] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, S. Singh, Autoprompt: Eliciting knowledge from language models with automatically generated prompts, arXiv preprint arXiv:2010.15980 (2020).
[13] X. L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, arXiv preprint arXiv:2101.00190 (2021).
[14] A. N. Tarekegn, M. Giacobini, K. Michalak, A review of methods for imbalanced multi-label classification, Pattern Recognition 118 (2021) 107965.
[15] H. Alhuzali, S. Ananiadou, Spanemo: Casting multi-label emotion classification as span-prediction, arXiv preprint arXiv:2101.10038 (2021).
[16] J. Canete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, Pml4dc at iclr 2020 (2020) 1–10.
[17] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.
[18] E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, arXiv preprint arXiv:2009.14119 (2020).
[19] T. Miyato, A. M. Dai, I. Goodfellow, Adversarial training methods for semi-supervised text classification, arXiv preprint arXiv:1605.07725 (2016).
[20] C.-K. Yeh, W.-C. Wu, W.-J. Ko, Y.-C. F. Wang, Learning deep latent space for multi-label classification, in: Thirty-first AAAI conference on artificial intelligence, 2017.