Classification, Extraction, and Normalization : CASIA_Unisound Team at the Social Media Mining for Health 2021 Shared Tasks

Tong Zhou[1,3,*,†], Zhucong Li[1,2,*], Zhen Gan[1,4,*,†], Baoli Zhang[1], Yubo Chen[1,2], Kun Niu[3]
Jing Wan[4], Kang Liu[1,2], Jun Zhao[1,2], Yafei Shi[5], Weifeng Chong[5], Shengping Liu[5]
[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences
[2] School of Artificial Intelligence University of Chinese Academy of Sciences
[3] Beijing University of Posts and Telecommunications
[4] Beijing University of Chemical Technology
[5] Beijing Unisound Information Technology Co., Ltd
{tongzhou21, niukun}@bupt.edu.cn
{zhucong.li,baoli.zhang,yubo.chen,kliu,jzhao}@nlpr.ia.ac.cn
{ganzhen, wanj}@mail.buct.edu.cn
{shiyafei,chongweifeng,liushengping}@unisound.com

Abstract

This is the system description of the CASIA_Unisound team for Task 1, Task 7b, and Task 8 of the sixth Social Media Mining for Health Applications (SMM4H) shared task in 2021. To address two shared challenges among those tasks, the colloquial text and the imbalance annotation, we apply customized pre-trained language models and propose various training strategies. Experimental results show the effectiveness of our system. Moreover, we got an F1-score of 0.87 in task 8, which is the highest among all participates.

1 Introduction

Enormous data in social media has drawn much attention in medical applications. With the rapid development of health language processing, effective systems in mining health information from social media were built to assist pharmacy, diagnosis, nursing, and so on (Paul et al., 2016) (Yang et al., 2012) (Zhou et al., 2018).

The health language processing lab at the University of Pennsylvania organized the Social Media Mining for Health Applications (SMM4H) shared task 2021 (mag), which provided an opportunity for fair competition among state-of-the-art health information mining systems customized in the social media domain. We participated in task 1, subtask b of task 7, and task 8.

Task 1 consists of three subtasks in a cascade manner: (1) identifying whether a tweet mentions adverse drug effect; (2) mark the exact position that mentions ADE in the tweet; (3) normalization ADE mentions to standard terms. Subtask b of task 7 (Miranda-Escalada et al., 2021) is designed to identify professions and occupations (ProfNER) in Spanish tweets during the COVID-19 outbreak. Task 8 is targeting the classification of self-reported breast cancer posts on Twitter.

The ubiquitous two challenges of all the SMM4H shared tasks are (1) how to properly model the colloquial text in tweets; (2) avoid prediction bias caused by learning from unbalanced annotated data. The tweet's text, mixing with informal spelling, various emojis, usernames mentioned, and hyperlinks, will hinder the real semantic comprehension by a common pre-trained language model. Meanwhile, medical concepts are imbalanced in the real world due to the imbalanced morbidity of various diseases, and this phenomenon is also reflected in social media data. Training with imbalanced data will induce the model to pay much attention to the major classes and neglect the tail classes, which hinders the model's robustness and generalization.

To address the challenges above, we utilize a language model pre-trained on tweet data as the backbone and introduce multiple data construction methods in the training process. In the following, we will describe our methods and corresponding experiments for each task separately. At last, we summary this competition and discuss future directions.

2 Task 1: English ADE Tweets Mining

Adverse drug effect (ADE) is among the leading cause of morbidity and mortality. The collection of those adverse effects is crucial in prescribing and new drug research.

* Equal contribution.
† Works are done during internship at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences.

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 77–82

Tweet	MedDRA Term
vyvanse completely gets rid of my appetite. not quite sure how to feel about this.	10003028 appetite lost

Table 1: An example of tweets labeled ADE in Task 1. The ADE span is colored red, and the corresponding MedDRA term id is 10003028.

Model	Precision	Recal	F1
Ours	0.592	0.417	0.49
Ours w/o pseudo data	0.552	0.325	0.41
Average scores	0.505	0.409	0.44

Table 2: Results on the SMM4H Task 1a test set.

This task's objective is to find the tweet containing ADE, locate the span, and finally map the span to concepts in standard terms.

2.1 Classification

The goal of this subtask is to distinguish whether a tweet mentions adverse drug effects. As shown in Table 1, "rid of my appetite" is an ADE mention, so this tweet is labeled on "ADE". In this dataset, the training set consists of 17385 tweets (16150 NoADE and 1235 ADE tweets), the validation set consists of 914 labeled tweets (849 NoADE and 65 ADE tweets), and the test set consists of 10984 tweets. Since only about 7% of the tweets contain ADEs, we target this class imbalance issue with a customized pseudo data construction strategy.

2.1.1 Method

Pseudo Data: A human may differentiate ADE tweets by some complaints trigger words like verb "feel" "think" or some negative sentiment words like "gets rid of", but a more precise way is discerning ADE mention. The mention in the tweet indicating ADE is a colloquial MedDRA term, and they express the same semantic. We construct ADE tweet for training in two ways: (1) randomly inserting the text description of a standard term in a tweet; (2) regarding the text description of a standard term as an ADE tweet. With those pseudo training data, a model should pay more attention to ADE mention in a tweet and more robust to diversified and unseen context.

Model: We apply the BERTweet (Nguyen et al., 2020), a RoBERTa (Liu et al., 2019) language model pre-trained on Twitter data, to encode tweet text and make a binary prediction according to the corresponding pooling vector.

2.1.2 Experiments

We set the batch size to 32 and using AdamW (Loshchilov and Hutter, 2018) optimizer for optimizing. For BERTweet parameters, we set a learning rate of 3e-5, the weight of L2 normalization is 0.01; for other parameters, we set the learning rate to 3e-4, the weight of L2 normalization is 0. We finetune all models using 5-fold cross-validation on the training set for 50 epochs. The amount of pseudo data is equal to 85.80% of the origin training data to balance the two classes. The experimental results are shown in Table 2, and indicate the advantage of our data construction strategies.

2.2 Extraction

This subtask aims to extract ADE entities from English Twitter texts containing ADE. The dataset includes training set, validation set, and test set containing 17385, 915, and 10984 tweets respectively. The proportion of tweets involving ADE mentions in the training set and the validation set is about 7.1%.

2.2.1 Method

Preprocessing: To reflect real semantic properly, we preprocess tweets in customized manners. (1) Since most user names are outside the vocabulary, We change all user names behind @ to "user". (2) There are some escape characters in the Twitter text, such as """, "&", "<", ">", and we replace them with the original characters: """, "&", "<", ">" respectively.

Training: During the training stage, We use a five-fold cross-training fusion system, which include 7 different pre-training models. We ensemble them through average weighted voting to weaken the fluctuations of performance of single model.

Model: We use seven pre-training models: bertweet-base, bertweet-covid19-base-cased, bertweet-covid19-base-uncased, bert-base-cased, bert-base-uncased, bert-large-cased, and bert-large-uncased.

2.2.2 Experiments

The models we choose and their learning rates are shown in Table 3. Each model has two learning rates, the former is the learning rate of BERT, and the latter is the learning rate of BiLSTM(Ma and Hovy, 2016)+CRF(Lafferty et al., 2001). Each BERT model is finetuned for 50 epochs with the dropout (Srivastava et al., 2014) of 0.3 using AdamW (Loshchilov and Hutter, 2018) optimizer.

Model	Learning Rate
bertweet-base+BiLSTM+CRF	[5e-5, 5e-3]
bertweet-covid19-base-cased+ BiLSTM+CRF	[5e-5, 5e-3]
bertweet-covid19-base-uncased+ BiLSTM+CRF	[5e-5, 5e-3]
bert-base-cased+BiLSTM+CRF	[5e-5, 5e-3]
bert-base-uncased+BiLSTM+CRF	[4e-5, 4e-3]
bert-large-cased+CRF	[1e-5, 1e-3]
bert-large-uncased+CRF	[7e-6, 7e-4]

Table 3: Implementation details of our models of the SMM4H Task 1b.

Model	Precision	Recal	F1
Ours	0.381	0.475	0.42
Average scores	0.493	0.458	0.42

Table 4: Results on the SMM4H Task 1b test set.

We set the batch size of bert-large-cased and bert-large-uncased to 8, and the others are 64. The experimental results are shown in Table 4. The Recall of our result is close to two percentage points higher than the average, but our Precision is about 11 percentage points lower than the average. Therefore, our model recalls more correct entities, but it also recalls a lot of wrong entities. So this may be a direction in which our method can be optimized.

2.3 Normalization

MedDRA (Brown et al., 1999) is a rich and highly specific standardized medical terminology to facilitate sharing regulatory information internationally for medical products used by humans. This subtask aims to normalize ADE mention to standard MedDRA term based on the result of span detection.

2.3.1 Method

Our model's inference process consists of a classification phase and a compare phase, responsible for recall and rank, respectively. We train the above two phrases with shared parameters and optimizing with the combined supervising signal.

Recall: In view of the representation process of ADE's mention could be benefited from its context, we utilize BERTweet for complete tweet representation. Since we have a specific position of mention in a tweet from subtask b, we first truncate mention's representations and calculate out the mean vector as the mention representation. Next, we calculate the dot product between mention representation and term embedding. Each vector in the term embedding is initialized according to its corresponding mean BERTweet representation of standard term text description. Finally, a softmax operation is added to convert the dot product value to conditional probabilities. A cross-entropy loss function responsible for supervising this process.

Model	Precision	Recal	F1
Ours* (recall)	0.244	0.305	0.271
Ours* (recall + rank)	0.248	0.311	0.276
Ours (recall + rank)	0.129	0.403	0.195
Average scores	0.231	0.218	0.22

Table 5: Results on the SMM4H Task 1c test set, * denotes the results of our method based on our best prediction in subtask b.

Rank: Since the MedDRA term's description is a normalized expression of its corresponding ADE mention, the global semantic of a tweet should remain unchanged after exchanging the colloquial ADE mention and correct term description. On the contrary, the global semantic should have an offset after exchanging with a wrong term. Based on the above assumption, we add an additional supervising signal. A tweet's global representation is obtained from BERTweet's mean pooling vector. The model calculates triplet loss among the following global representations: (a) origin tweet (b) replace the mention with target term's description (c) replace the mention with a wrong term's description. The wrong term is firstly obtained by random selection from the whole term set, and with the procedures of the training process, it is randomly selected from the classification model's top K prediction. The triplet loss intends to maximize the similarity of the global representation of (a) and (b); meanwhile, it minimizes the similarity of (a) and (c).

Inference: In the inference stage, first, we obtain the top K terms based on the prediction of the recall procedure. Then we exchange the candidate K terms with the mention in the origin tweet and calculate the similarity of global representation with the origin tweet. The similarity score is the base of term ranking. Finally, we retain the top 1 as the final prediction.

2.3.2 Experiments

Our hyperparameter setting is identical to subtask a. Besides, we set K to 10, and for the combination of cross-entropy loss and triplet loss, we set equal weights. The experimental results are shown in Table 5, and indicate the advantage of the compare-based rank procedure.

3 Task 7: ProfNER for Spanish Tweets

3.1 Extraction

This subtask aims to detect the spans of professions and occupations entities in each Spanish tweet. The corpus contains four categories, but participants will only be evaluated to predict two of them: PROFESSION [profession] and SITUACION_LABORAL [working status]. The dataset includes a training set, validation set, and test set containing 6000, 2000, and 27000 tweets, respectively.

3.1.1 Method

Preprocessing: According to the characteristics of the competition's Spanish Twitter data and the competition requirements, we preprocess data to improve the model's ability to capture text information. (1) Since most user names are outside the vocabulary, We change all user names behind @ to "usuario". (2) The corpus contains four kinds of labels, but we will only be evaluated in the prediction of 2 of them: PROFESSION and SITUACION_LABORAL, so we removed the other two labels ACTIVIDAD and FIGURATIVA.

Training: Similar to subtask b of task 1, we make predictions on the multiple trained models and perform a simple voting scheme to get the final result.

Model: We use three BERT-based (Devlin et al., 2018) pre-training models: bert-base-spanish-wwm-cased, bert-spanish-cased-finetuned-ner, and bert-spanish-cased-finetuned-pos.

3.1.2 Experiments

For this subtask, each BERT model is finetuned for 50 epochs with the learning rate of 5e-5 using AdamW optimizer, and for the BiLSTM+CRF module, our learning rate is 5e-3, and the batch size is 64. The experimental results are shown in Table 6. The Model_ensemble0(noLSTM) is the result of the fusion of fifteen models without the BiLSTM modules, and The Model_ensemble1(LSTM) is the result of the fusion of fifteen models with the BiLSTM modules. The Ours is the final result, which is the voting fusion result of 30 models. From the experimental results, we can see that the F1 score of the fusion record on the validation set is superior, but the test set score has dropped. According to our analysis, this is probably related to a large amount of test data.

https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

https://huggingface.co/mrm8488

Model	Validation F1	Test F1
bert_spanish_cased	0.732	-
bert_spanish_ner	0.736	-
bert_spanish_pos	0.723	-
Model_ensemble0(noLSTM)	0.742	0.725
Model_ensemble1(LSTM)	0.744	0.731
Ours	-	0.733

Table 6: Results on the SMM4H Task 7b Validation and test set.

Tweet	**Label**
Excellent cause! I hope you are doing well. I had breast cancer too. I'm into my 3rd year of Tamoxifen.	S
OH MY GOD i just remembered my dream from my nap earlier i understand now why i felt so bad when i woke up i literally dreamt that i had breast cancer	NR

Table 7: Two examples of tweets and corresponding labels in Task 8.

4 Task 8: Self-reported Patient Detection

The adverse patient-centered outcomes (PCOs) caused by hormone therapy would lead to breast cancer patients discontinuing their long-term treatments (Fayanju et al., 2016). The research on PCOs is beneficial to reducing the risk of cancer recurrence. However, PCOs are not detectable through laboratory tests and are sparsely documented in electronic health records. Social media is a promising resource, and we can extract PCOs from the tweet with breast cancer self-reporting (Freedman et al., 2016). First and foremost, the PCO extraction system requires the accurate detection of self-reported breast cancer patients. This task's objective is to identify tweets in the self-reports category. In this dataset, the training set consists of 3513 tweets (898 self-report and 2615 non-relevant tweets), the validation set consists of 302 tweets (77 self-report and 225 non-relevant tweets), and the test set consists of 1204 tweets.

4.1 Method

Preprocessing: We preprocess the data to fit the tokenizer of the pre-trained RoBERTa model BERTweet, which is customized in tweet data. (1) The BERTweet's tokenizer transform the URL string in tweet to a unified special token by matching "http" or "www". For the tokenizer to effectively identify the URL, we insert "http://" before

Model	Precision	Recal	F1
Ours w/o preprocessing, w/o robust training	0.8571	0.8571	0.8571
Ours w/o robust training	0.8844	0.8442	0.8637
Ours	0.8701	0.8701	0.8701
Average scores	0.8701	0.8377	0.85

Table 8: Results on the SMM4H Task 8 test set.

"pic.twitter.com" in tweets. (2) The emoji in tweets is expressed as UTF-8 bytes code in string form. We match the "\x" and transform the code into its corresponding emoji.

Training: Although the generalization ability of the pre-trained language model finetuned in text classification tasks has been proved, it could still seize the wrong correction between specific tokens and the target label, turn out to neglect the crucial semantic. As shown at the top of Table 7, "I had breast cancer" is convincing evidence to a positive prediction. A model can make the right decision on the example at the bottom of Table 7 only if it takes the context into consideration. To avoid this wrong correction and improve our model's robustness, we apply two strategies on the training stage exert in data level and model level, respectively.

(1) Noise: Each word in a tweet has a probability p to be replaced by a random word, and the target label has a probability p to reverse.

(2) FGM: Following the fast gradient method (Miyato et al., 2016), we move the input one step further in the direction of rising loss, which will make the model loss rise in the fastest direction, thus forming an attack. In response, the model needs to find more robust parameters in the optimization process to deal with attacks against samples.

Model: Similar to subtask a in Task 1, we apply the BERTweet to encode tweet text and make a binary prediction according to the corresponding pooling vector.

4.2 Experiments

We set the batch size to 32 and using AdamW optimizer for optimizing. For BERTweet parameters, we set a learning rate of 3e-5, the weight of L2 normalization is 0.01; for other parameters, we set the learning rate to 3e-4, the weight of L2 normalization is 0. We set the noise rate to 0.025 and the epsilon of FGM to 0.5. We finetune all models using 5-fold cross-validation on the training set for 50 epochs. The experimental results are shown in Table 8. Our method has obtained the highest F1 score in this task. Furthermore, the ablation results indicate the advantage of the customized data preprocessing procedure and the robust training strategies.

5 Conclusion and Future Work

This work explores various customized methods in tasks of classification, extraction, and normalization of health information from social media. We have empirically evaluated different variants of our system and demonstrated the effectiveness of the proposed methods. As future work, we intend to introduce the medical domain's knowledge graph to improve our system further.

Acknowledgements

This work is supported by the National Key RD Program of China (2020AAA0106400), the National Natural Science Foundation of China (No.61806201) and the Key Research Program of the Chinese Academy of Sciences (Grant NO. ZDBS-SSW-JSC006).

References

Elliot G Brown, Louise Wood, and Sue Wood. 1999. The medical dictionary for regulatory activities (meddra). *Drug safety*, 20(2):109–117.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Oluwadamilola M Fayanju, Tinisha L Mayo, Tracy E Spinks, Seohyun Lee, Carlos H Barcenas, Benjamin D Smith, Sharon H Giordano, Rosa F Hwang, Richard A Ehlers, Jesse C Selber, et al. 2016. Value-based breast cancer care: a multidisciplinary approach for defining patient-centered outcomes. *Annals of surgical oncology*, 23(8):2385–2390.

Rachel A Freedman, Kasisomayajula Viswanath, Ines Vaz-Luis, and Nancy L Keating. 2016. Learning from social media: utilizing advanced data extraction techniques to understand barriers to breast cancer treatment. *Breast cancer research and treatment*, 158(2):395–405.

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

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 preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.

Antonio Miranda-Escalada, Eulàlia Farré-Maduell, Salvador Lima López, Luis Gascó-Sánchez, Vicent Briva-Iglesias, Marvin Agüero-Torales, and Martin Krallinger. 2021. The profner shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Michael J Paul, Abeed Sarker, John S Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L Smith, and Graciela Gonzalez. 2016. Social media mining for public health monitoring and surveillance. In *Biocomputing 2016: Proceedings of the Pacific symposium*, pages 468–479. World Scientific.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Christopher C Yang, Haodong Yang, Ling Jiang, and Mi Zhang. 2012. Social media mining for drug safety signal detection. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 33–40.

Lina Zhou, Dongsong Zhang, Christopher C Yang, and Yu Wang. 2018. Harnessing social media for health information management. *Electronic commerce research and applications*, 27:139–151.