FOCUSING ON IMPLICIT CHINESE CYBERBULLYING : HOW TO DETECT CYBERBULLYING IN CHINESE?

Meng Zhou **Yangqian Wu**

ABSTRACT

This paper comes up with some new novel approaches to detect Cyberbullying in Chinese. With the development and spreading of social network, Cyberbullying has become a huge problem not only in Western countries but also in China. Cyberbullying's behavior could include swearing, cyberstalking, harassment, etc, which will have negative impact on Netizens. Generally speaking, there already been lots of extensive attempts and researches on Cyberbullying Detection in English, and they could do pretty well in identifying whether a comment is suspicious of bullying or not. However, there are little similar related works for Cyberbullying Detection in Chinese. Therefore, our work targets on the unique phenomenon in Chinese Bullying which is called "implicit Chinese Bullying" and innovatively comes up with effective approaches to deal with it.

***K*eywords** Natural Language Processing · CyberBullying · BERT · Implicit Chinese Expression · Homophony

1 Introduction

1.1 Background

Cyberbullying is becoming a wide-range serious phenomenon in today's cyberspace. Cyberbullying generally means that one or several bullys intentionally give offensive comments which will have negative psychological effect on the victims. Cyberbullying will not only harm receiver's psychology, but also influence other users' language style and drive the whole atmosphere into a negative tendency.

The offensive comments will appear in various ways, like swearing, cyberstalking, flaming, etc. And cyberbullying could appear in all kinds of places. It could take place in social network platforms, corresponding private communications or other network places. Moreover, sometimes a initial bullying message will result in a huge number of follow-up bullying comments and then gradually grows into a huge debate which has lots of bullying information. Therefore, sometimes it's significant to pick up the first bullying comment.

Since the wide spreading of social network, the frequency of cyberbullying is now increasing with a rapid speed. This highlights the importance of cyberbullying detection, which could detect the messages in suspicion of bullying others or conveying the very negative emotion. Then, the corresponding system could take the proper punishment to its user.

1.2 Our task: CyberBullying Detection

With the description of cyberbullying, it's of great importance to detect the bullying messages in social network.

In our project, in order to specifically train the cyberbullying Chinese model and evaluate our model, we select the Microblog as our targeted social network platform. That means we will crawl the comments from Microblog and train our cyberbullying model in this dataset and also evaluate our model in this dataset. Besides, Microblog is a big and typical social platform for Chinese users. There will be plentiful comments and messages which could represent different Chinese users' language styles. Therefore, it's reasonale to believe that the cyberbullying model we train from Microblog's data will also be suitale for the most of the Chinese platforms.

In brief, our project's task is to build a effective model which could judge a given Chinese comment is bullying or not. The difficuly is that most of Chinese sentences are quite ambiguous and perhaps have different meanings for a certain condition. Generally speaking, it's difficult to pick up implicit Chinese bullying comments because bullys are intended to express bullying information in a implicit way. Of course, that's also the main aspect we are dealing with, and till now we have come up with some useful approaches.

1.3 Our Work

In our project, we mainly achieved following works:

- Label the data crawled from Microblog which focuses on the recent hot topics. Finally, worked with other two teams, we have obtained a labeled dataset which has 10w Microblog's comments.
- Accomplish a Word Embedding+LSTM model by using Keras which has reached the accuracy of 90% in our dataset. Because there are still some room for improvement, we choose this model as our baseline and continue to try other models.
- Build a BERT model with pretrained method. By this approach, we could achieve the accuracy of 92.45%. This accuracy is quite satisfactory for our project.
- In order to seek for more innovation, we realize that Chinese comments have a lot of implicit bullying like 'homophony'. A unique feature for Chinese is its Pinyin, so we create a new RNN model with Pinyin coding inputs to be auxiliary judging model to detect the 'homophony' cyberbullying. This approach reached the accuracy of 92.56%.
- Besides, we tried to seek more unique Chinese features like character pattern, pronunciation, etc or other implicit Chinese bullying methods like 'digits', 'Network hot words', etc.

1.4 Our Contribution

Our main contribution is that we have offered some useful approaches to detect cyberbullying in Chinese.

In Chinese, there are lots of implicit bullying which is hard for existed NLP methods to detect and we propose that it's meaningful to take them into consideration. In order to do this, our project extracts some unique features for Chinese like Pinyin, character pattern, etc. And use them to bulid our new model.

Since a large proportion of explicit Chinese bullying is 'homophony' bullying. In experiment, we uses coding Pinyin as input to construct our auxiliary model and received good performance.

In the long run, we will also consider more Chinese features in detecting Chinese cyberbullying.

2 Related Work

2.1 Cyberbullying detection

With the wide spreading of social media, it has become a quite serious problem that cyberbullying appears more frequently and thus causes more negative impacts on netizens, especially on children and young adults. There have been some related works aimed to effectively detect cyberbullying.

Some papers are digging into the bullying information of a offensive sentence. They came up with some new methods to effectively present bullying information.

For instance, 'Content-Driven Detection of Cyberbullying on the Instagram Social Network' turned its view into various contents. It have studied the detection of cyberbullying in photosharing networks, with an eye on the development of earlywarning mechanisms for identifying images vulnerable to attacks. In the context of photo-sharing, they have refocused this effort on features of the images and captions themselves, finding that captions in particular can serve as a surprisingly powerful predictor of future cyberbullying for a given image.[1]

Moreover, in 'Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder', Rui Zhao and Kezhi Mao address the text-based cyberbullying detection problem, where robust and discriminative representations of messages are critical for an effective detection system. By designing semantic dropout noise and enforcing sparsity,

they have developed semantic-enhanced marginalized denoising autoencoder as a specialized representation learning model for cyberbullying detection. [2]

Besides textual information, some papers also took the relationship between users into consideration and used social net as a feature to judge cyberbullying bahavior.

In 'Cyber Bullying Detection Using Social and Textual Analysis', Qianjia Huang, Vivek K. Singh and Pradeep K. Atrey advanced the state of the art in cyber bullying detection beyond textual analysis to also consider the social relationships in which these bullying messages are exchanged. Their results indicate that social features are useful in detecting cyber bullying. In effect, it suggests that understanding the social context in which a message is exchanged is just as important as the message itself. [3]

However, most of papers are focusing on how to effectively detect cyberbullying. They are targeting on English dataset but not talking about other languages. Therefore, we tend to distinguish the difference between English and Chinese, furthermore to deal with the problem how to effectively detect Chinese cyberbullying.

2.2 Chinese features

When focusing on unique Chinese features, we discovered that there have been several papers talking about how to deal with Chinese and apply the technique in some certain different tasks. After reading these papers, we benefited a lot and came up with our own method to deal with Chinese cyberbullying.

Yun Zhang, Yongguo Liu and others proposed a novel model ssp2vec to integrate the stroke, structure and pinyin features of Chinese characters for learning Chinese word embeddings. They used the feature substring to capture the morphological and phonetic information and their relation from strokes, structures and pinyin features in ssp2vec. They validate the effectiveness of their proposed method through four evaluation tasks. [4]

More specifically, Dagao Duan, Shaohu Liang, Zhongming Han, and Weijie Yang proposed a text error correction model based on pinyin. Compared with the Chinese speech recognition error correction algorithm based on phrase translation model proposed previously, their model is improved by 20%. Moreover, they verified the effectiveness of the pinyin as coding, and confirmed the tones in the importance of the speech recognition. [5]

3 Data Processing and Early Work

3.1 Microblog dataset

Together with other two teams with the same topic, we could obtain a Microblog comment dataset with around 100000 comments. We crawled the recent commments from Microblog. Moreover, when crawling the probable bullying commment, we used some knowing bullying words like 'sb', '死绝', '臭批' and acquired some negative comments which has more potential to be bullying. In this way, we could make sure the crawled data has the similar number of bullying commments and non-bullying comments, which is very helpful to train our bullying detecting model in a balance state.

After acquiring the Microblog dataset which has around 100000 comments(about 40000 bullying comments and 60000 non-bullying comments), We labeled them by hand and finally got a labeled Microblog comment dataset with around 100000 comments.

This is the dataset for our subsequent works: Buliding model, Training model, Evaluating model, Obeserving some features, etc.

3.2 Some observations

When labeling the comment we crawled, we have some observations which might be helpful for our later work. We listed several examples as following:

- Some negative comments will appear in the format of Pinyin. That's probably because of the mistyping when trying to input a Chinese bullying phrase. Some examples are shown in Table 1
- There will also are lots of homophony bullying like '741', '司马', '草泥马', which is hard to detect in original method. Some examples are shown in Table 2

Examples
刚就有一辆外地畜生车闯红灯caonimade xianjiajun caonima xianyangbisi xianjiajunbisi daixianmajiadebisi cnm 看懂了没有? 懂脑子是干什么用的吗没有就gun吧

Table 1: Some examples

Examples
今天怼了司马孤儿室友，我反正是好了一点 造谣司马知道吗丑东西 username 睁大你的狗眼看清楚别乱咬人 排的些什么草泥马的煞笔队员，重开你妈炸了！

Table 2: Some examples

- A more challenging observation is that some bullying words perhaps don't have bullying meaning in a sentence. The real meaning should be treated in a certain condition. Some examples are shown in Table 3

Examples
动不动问候你妈，这些人都怎么了？事情没看清楚就去骂的那么难听 由衷感觉自己是个SB，活该被人造谣一万年 没有滚滚，可以省钱了

Table 3: Some examples

These are all difficult to deal with when detecting cyberbullying in Chinese. For traditional models, it might be hard to pick up homophony bullying and might also judge bullying words with non-bullying meaning to be bullying. In our work, we will give some approached to deal with it by using Pinyin as input.

4 Implicit Chinese Bullying

Our main innovation is focused on how to effectively detect the implicit Chinese Bullying. It's because when we finished the conventional bullying detection model like word embedding+LSTM and BERT which is also suitable for English bullying detection, we find that there are still some Chinese-way bullying which is hard to detect through these model.

Most of them are implicit Chinese bullying like homophony, character patterns, etc. What we need to do is to effectively pick up this form and achieve a new model which is specifically targeting on Chinese cyberbullying and performs better than the original models.

Therefore, we should know what 'implicit Chinese bullying' is and what's their representations. We will talk about three main phenomenons taken place in 'implicit Chinese bullying'.

4.1 Homophony

Homophonic substitution means that netizens don't use the existing language items, but create a homophonic form to replace the original language items. In most of time, there is no semantic relationship between the alternative homophonic form and the original language item.

There are mainly two kinds of homophonic substitution: Using Chinese characters and using digit.

4.1.1 Homophone Chinese characters

Homophone Chinese characters means that using Chinese chracters with same or similar pronunciation to replace the original words.

This phenomenon firstly appeared because of mistyping. In social network, people tend to communicte in a fast speed and thus cause the mistyping of words. However, this mistyping rapidly became popular and grew as a trend.

In today's Chinese social network, people always tend to use homophony to express bullying in a less offensive way. Some examples are shown in Table 4

Homophone Words	Original Words
尼玛	你妈
麻痹	妈逼
辣鸡	垃圾
卧槽	我操
劳资	老子
特么的	他妈的
草泥马	操你妈

Table 4: Some example for Homophony Bullying

4.1.2 Homophonic digit

Homophonic digit means that using Arabic numbers with same or similar pronunciation to replace the original words. They always have no semantic relationship.

Some examples are shown in Table 5

Homophone digit	Original Words
748	去死吧
741	气死你
0487	你是白痴
1414	要死要死
02746	你恶心死了

Table 5: Some example for Homophony Bullying

4.2 Abbreviation

In Chinese comment, the abbreviation of Pinyin from bullying words is very commonly used. Abbreviation format is a kind of omission. It derives from Pinyin of Chinese bullying phrase and combines first letter of Pinyin of different Chinese character.

Some examples are shown in Table 6

Abbreviation	Original Words
TMD	他妈的
BT	变态
NC	脑残
MD	妈的
SB	傻逼

Table 6: Some example for Chinese abbrevation Bullying

4.3 Association

A not very common implicit Chinese bullying phenomenon is through using image association. Users tend to use the same or similar things to make an analogy. Or sometimes use similar Chinese character with vivid Chinese pattern to demostrate it.

Some examples are shown in Table 7

Association Word	Original Words
白莲花	外表看上去纯洁，其实内心阴暗，一味装纯洁、装清高的人
绿茶婊	在人前装出楚楚可怜的样子，背后善于心计的女人
中央空调	同时对两个或两个以上的女性散发着温暖和爱心的男性

Table 7: Some example for Chinese association Bullying

4.4 Analysis

All these unique Chinese implicit bullying derive from netizens' tendency to express their emotion in a less offensive way. Or just want to explain their emotions in a novel way. They are all the production from the wide-spreading of social network.

However, though some of them are less offensive, they still have negative impact on other users and the whole environment. Therefore, they need to be picked up and reduced.

What we did to detect this kind of implicit Chinese bullying will be mentioned in PART V and PART VI. One of the approach is that We used Pinyin as our input and received better performance.

5 Classification Approaches

In our project, we have accomplished three models to detect cyberbullying. All of them are trained through the Microblog dataset which has around 100000 comments. They are 'Wordembedding & LSTM model', 'BERT model', 'A RNN Model Based on Pinyin'. The former two models are based on Chinese characters inputs and achieve 90% accuracy and 92% accuracy, the last one model is based on Pinyin inputs which could detect the Chinese homophony bullying and achieves 92.45% accuracy.

There is a brief introduction to three models:

- Wordembedding & LSTM model: use word index to convert a sentence into a vetcor, then through Wordembedding and LSTM to train a neural network detecting whether a sentence is bullying or not.
- BERT model: BERT is state of art model came up with in 2018. We use a pre-trained Chinese model and train it to our bullying detection task.
- A RNN Model Based on Pinyin: Our intuition for Chinese cyberbullying detection is that in order to deal with homophony bullying, we use coding Pinyin as model's input. We construct a RNN model with Pinyin input and after training, it could detect the homophony bullying words.

We will introduce three models on by one.

5.1 Classical LSTM RNN using Word Embedding

5.1.1 Word Embedding

Word embedding turns out to be one of the most important concepts in modern NLP field. It transfers the one-hot vector from vocabulary to a lower dimensional vector. Such vector representation describes 'features' of words, which will lead semantically similar words to a similar vectors. These vectors will be used in the later classification structure. Instead of just using some simple surface features like the presence or frequency of particular words(Bag of Words), words are represented as their semantic meanings, which makes the model generalize better. Therefore, it's a common choice to do text classification task in recent years.

5.1.2 Recurrenct Neural Network

When labeling our data, we found that judging whether a comment is bullying or not is a context-dependent problem. We have seen comments like "爱你，狗逼 ", which is very likely to be discriminated as bullying review if we ignore

the relationship between the bullying words and previous words. So it's reasonable to use an RNN model combined with LSTM units to build a longer dependency between words.

5.1.3 Implementation Details

Our implementation of this approach is based on Keras, we use jieba to create a vocabulary and tokenize our sentences. We shuffle our data and take the last twenty percent as our test set. We find that the model converges very fast so we only train 2 epoches. In addition, we did not utilize a pre-trained word embedding, because the pretrained word embedding is too large and we think it's unnecessary for our task.

5.1.4 Model structure

A NLP model based on Wordembedding and LSTM is a normal but effective method for a NLP task.

The model is constructed as followings: All words in a sentence are converted into a index which is extracted from all dataset. Then, the vector consisted of all indices is embedded into a 50 dimension vector through a embedding layer. It then across LSTM layer and get a 64-dim vector. It then across Dense layer, Dropout layer, Dense layer and Dense layer. Finally we got a probablity number after 'sigmoid' activation function. Whether this sentence is bullying or not is based on whether this value is bigger than 0.5 or not.

The structure is shown in Figure 3.

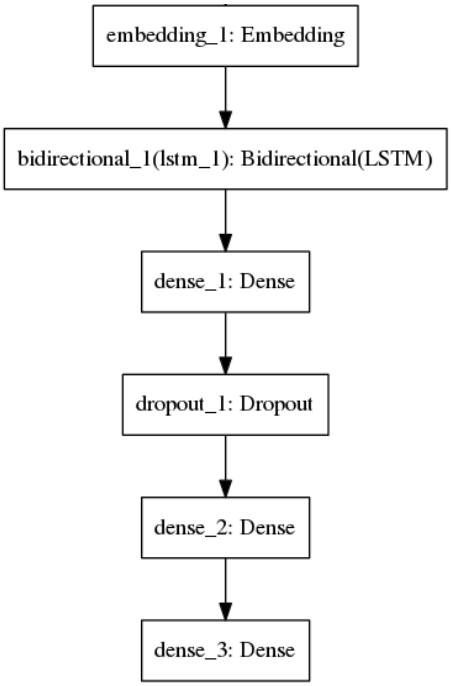

Figure 1: The Structure for WordEmbedding & LSTM model

5.1.5 Result

By accomplishing this model, we trained it in our dataset. We randomly chose 80% comments as training set and remaining 20% as testing set. Finally, we received the accuracy of 90.90%. In order to improve our model and deal with Chinese implicit bullying, we set this model as our baseline and continued to seek other approaches.

5.2 Pretrained Transformer:Bert

5.2.1 Latest NLP method

Recently, the world of NLP world has been dominated by pretrained transformers.[6] Bert[7] from Google has achieved many State-Of-The-Art results in many NLP tasks. Fortunately, Google also provides a Chinese Version so that we can

apply the cutting-edge NLP method to our own dataset. It's also an obvious trend in NLP field to pretrain a general language processing model by unsupervised learning and finetune it to our downstream tasks.

5.2.2 Self-Attention

The self-attention mechanism can be expressed as below:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

The inputs will be expressed as queries and keys of dimension d_k, and values of dimension d_v. Then We compute the dot products of the query with all keys, divide each by $\sqrt{d_k}$, and apply a softmax function to obtain the weights on the values. To express this in the matrix form, we pack a set of queries into a matrix Q. The keys and values are also packed together into matrices K and V. Then we will take a weighted sum of attentions.

Note that it's a totally different architecture from RNNs because it's suitable for parallelization and turns out to be better in capturing long-term dependency of natural languages.

5.2.3 Transformer Encoder

A transformer encoder block consists of a multi-head self attention layer and a feed forward network. There's also a residual connection between the input and feedforward network in the encoder block. The input will first be represented as vectors and then be fed into several encoder blocks and finally go through some fully-connected layers and a softman layer to do classification task.

5.2.4 Transfer Learning Implementation

The pretrained model we use is Bert-base, we also used some variants of Bert like Bert-wwm from HIT[8]. But different Bert variant models do not make much difference in our problem so we just use and tune the original Bert model. Considering the computational cost, we only use the base model, which means 12 layers, 12 attention heads, 768 hidden size and around 110M parameters totally. We download the tensorflow version from Google and use API from HuggingFace to transfer to a Pytorch model.

During our implementation, we find that the learning rate is the most important hyperparameters in transfer learning. Initially, we set a too big learning rate. After several iterations, the model always output the same value no matter what input is. It takes us a lot of time to debug bacause there is little experience resource online about doing transfer learning on Chinese dataset using Bert. We do not know the exact reason for this, but this may relate to *catastrophic forgetting*. At first, the softmax layer we add is not well connected to the previous sturcture of the model, so a big learning rate can lead to some problems.

After setting a suitable initial learning rate we warm it up which means make it ramp up in the first thousands of iterations. Besides, we schedule it by using *ReduceLROnPlateau* according to the test loss to achieve the best performance.

5.3 PinYin embedding LSTM RNN

5.3.1 Intuition

During the process of labeling our data, we find that on the Internet, many people tend to substitute the original word with PinYin to express the same hate speech. It's a special phenomenon in Chinese language because PinYin represent the pronunciation of words so they can convey exactly the same meaning to human. The reason may be just a language habit or users are trying to avoid being detected by bullying detection system based on word embedding or key words. Traditional NLP method only cares about the word. When this happen, it may be encoded as UNK as the input to the network, which can not help the model to make the right judgement.

5.3.2 Encoding Method

Firstly, we will transfer the original Chinese character into PinYin by a python library called *pypinyin*. The tone is discarded because we want the word with similar PinYin to end up with similar vectors. Then we use Byte Pair Encoding(BPE)[9] to tokenize PinYin. It's originally designed for English but we think to split the PinYin of a single word is beneficial because people do not always substitute words with the whole PinYin, maybe just part of them. In this way we get the vector representation of PinYin and we feed them into a similar Bidirectional LSTM RNN as 4.1.

5.3.3 Implementation Details

We mask part of words with a probability of 0.05 to do regularization. The model converges slower than the previous two methods. In the last layer of LSTM, we combine attention features, average features and pooling features together because we think PinYin may drop part of useful content of languages so we just add more features.

6 Experiment

In our project, we have together accomplished three cyberbullying models: 'Wordembedding & LSTM model', 'BERT model', 'A RNN Model Based on Pinyin'. In order to compare their performance obviously, we all trained them with 80% datadet and tested them with 20% dataset. All the data are shuffled to make better comparison.

The accuracy comparison is listed in Table 8

Model	Accuracy
Classical LSTM RNN using Word Embedding	90.90%
Pretrained Transformer:Bert	92.45%
PinYin embedding LSTM RNN	92.56%

Table 8: Comparison of result using different models

Comparing three models we have achieved, they all have good performance for higher than 90% accuracy. It's obvious that when using the pre-trained model(BERT), the accuracy has a rapid development. By adding a Pinyin embedding LSTM RNN model as a auxiliary model, we can detect some homophony Chinese bullying words and thus improve the accuracy.

Here is an example of how a Pinyin embedding LSTM RNN model could detect homophony Chinese bullying while BERT can't. The examples are shown in Figure **??** and Figure **??**

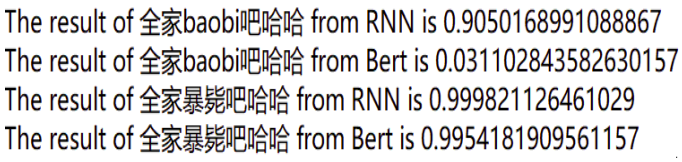

Figure 2: The example of homophony bullying detection

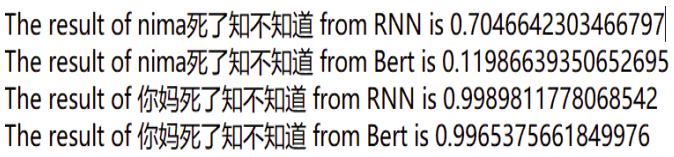

Figure 3: The example of homophony bullying detection

It seems that PinYin will not discard the most essential part of language meanings because use PinYin embedding can achieve similar results with the traditional word embedding.

7 Conclusions

7.1 Conclusion

In our project, we acquired and labeled a Microblog comment dataset with around 100000 comments(approximately 50% for bullying and non-bullying). We mainly accomplished three models to detect cyberbullying, including 'Wordembedding & LSTM', 'BERT', 'Pinyin & RNN' and finally achieve the accuracy of 92.56%. Moreover, we observed and organized some implicit Chinese bullying phenomenons which are quite hard for original model to detect.

In order to deal with it, we tried to use Pinyin as our input feature of RNN and then successfully detect homophony bullying.

From our experiments, we can see that our method is very effective in some particular hate speech detection situations. By ensembling the PinYin RNN model with the character-level Bert, we get a slightly better results. This improvement may be bigger if we have a pretrained PinYin embedding or pre-trained transformer model for PinYin.

To draw a conlusion, our project mainly consists of following tasks:

- Label the comments crawled from Microblog, construct a Microblog dataset with 100000 comments(approximately 50% for bullying and non-bullying).
- Observe that a difficulty in Chinese bullying detection is implicit Chinese bullying. Thus organize some unique Chinese features like character pattern, pronunciation, etc.
- Construct a Word Embedding & LSTM model by Keras which could reached the accuracy of 90.90% in our dataset. We set it as our baseline.
- Accomplish a BERT model which could achieve the accuracy of 92.45%. Generally, this accuracy is satisfactory for our task.
- In order to effectively detect 'homophony' cyberbullying which is the main part of implicit Chinese bullying. Create a new RNN model with Pinyin coding inputs to be an auxiliary judging model. This approach reached the accuracy of 92.56%.

Besides, cyberbullying is a really meaningful topic especially for the rapid growth of social network. Hope our project could offer some little help for real cyberbullying.

7.2 Challenges

Some challenges we meet in our project process are demostrated as below:

- Coming up with the idea that Pinyin as the input of RNN is a quite challenging process. When we firstly realized that there are lots of homophony bullying phenomenon in Microblog's comment, we actually had no idea how to deal with it. That was an intuition to think about the difference between Chinese and English, and realize that Pinyin is a quite unique feature for Chinese which is also very helpful for our Chinese bullying detection.
- There exist many bullying comments in other forms that are also difficult to detect for classical NLP method. For example, people may substitute "去死吧 " with numerals "748". We can find that bullying detection is quite a culture-related problem because some bullying words are related to the history of the culture and different language has different features. The issue described above is not so severe in English because the pronunciation is closely related to the word itself. So we may need to deal with bullying detection problem with different features for different languages.
- We argue for a benchmark in this task because when doing research about this, researchers always collect data by themselves. This will add some 'bias' to their research because language data at a specific time on a specific platform may show some tendency. It's not good to test the generalized ability of models. For us, we collect and label the dataset by ourselves, which is quite time-consuming but there may still exist some problems.

7.3 Further improvement

Although we have done loads of works in our project, our project still has a lot of improvements could achieve, some of them are listed as below:

- In our project, to the limit of time, we just tried using Pinyin as input. Actually there are more unique features for Chinese like chracter patterns, hidden meanings, etc. In the future work, we could take these factorss into consideration to construct our detecting models.
- Just as we demonstrate in PART IV, although homophony bullying is a main part of implicit Chinese bullying, there still are other forms of implicit bullying including 'Abbreviation' and 'Association'. It will be quite meaningful to deal with them in some particular method. Then, in this way, we could accomplish an outstanding model to detect cyberbullying in Chinese.

- Chinese NLP is a quite important and challenging field. Chinese cyberbullying is just a particular task in it. It's no doubt that cyberbullying is really significant for today's social network, but we still could extract the technique from our project and apply it in other Chinese NLP tasks.

7.4 Acknowledgement

In the process of our project, we have received lots of helps from Professor Gao, teaching assitant, and classmates from same topic.

Thanks for their valuable instructions and comments for our project, these suggestions actually help a lot for our improvement of models.

We are especially grateful to Professor Gao. She gave us much useful guidance, suggestions and related papers which are pretty important for our project.

References

[1] Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, Cornelia Caragea, "Content-Driven Detection of Cyberbullying on the Instagram Social Network" Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)

[2] Rui Zhao, Kezhi Mao, "Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder" DOI 10.1109/TAFFC.2016.2531682, IEEE Transactions on Affective Computing.

[3] Qianjia Huang, Vivek K. Singh, Pradeep K. Atrey, "Cyber Bullying Detection Using Social and Textual Analysis" In Proceedings of the 3rd International Workshop on Socially-Aware Multimedia (pp. 3-6). ACM.

[4] Yun Zhang, Yongguo Liu, Jiajing Zhu, Ziqiang Zheng, Xiaofeng Liu, Weiguang Wang, Zijie Chen, Shuangqing Zhai, "Learning Chinese Word Embeddings from Stroke, Structure and Pinyin of Characters " CIKM'19, November 3-7, 2019

[5] Dagao Duan, Shaohu Liang, Zhongming Han, and Weijie Yang "Pinyin as a Feature of Neural Machine Translation for Chinese Speech Recognition Error Correction" CCL 2019, LNAI 11856, pp. 651-663, 2019.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is all you need" arXiv:1706.03762

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" arXiv:1810.04805

[8] Yiming Cui et al. "Pre-Training with Whole Word Masking for Chinese BERT" arXiv:1906.08101

[9] Rico Sennrich et al. "Neural Machine Translation of Rare Words with Subword Units" arXiv:1508.07909