Jigsaw Unintended Bias in Toxicity Classification

Pengcheng jiang[1]

[1] JiLin University, China

Introduction

Jigsaw Unintended Bias in Toxicity Classification

- given: A tagged dataset containing comments.Target 0 for malicious comments and 1 for friendly comments.
- target: detect toxic comments and minimize unintended model bias.
- evaluation: acc

text preprocess

count :Count the total number of words contained in all texts, the maximum and minimum number of words contained in a text

missing data :Check for missing data

full :Change abbreviations to full:isn't > is not(via dictionnary)

numbers :clean numbers

alphabetic :Find all non alphabetic characters and clean special chars

misspelling :Solve the problem of misspelling words

lower :lower the text

Tokenizer

- What tokenizer does is actually very simple. It divides the words it sees into spaces, and then uses numbers to correspond one by one. Then we take the first num Words is the word with the highest frequency, others are not recognized.
- First learn the dictionary of the text, and then get the corresponding relationship between words and numbers, and then convert the text into a number string through this relationship, and then use the padding method to make up the number string to the same degree, then you can proceed to the next step : embedding
- collections.counter,pytorch:torchtext.vocab

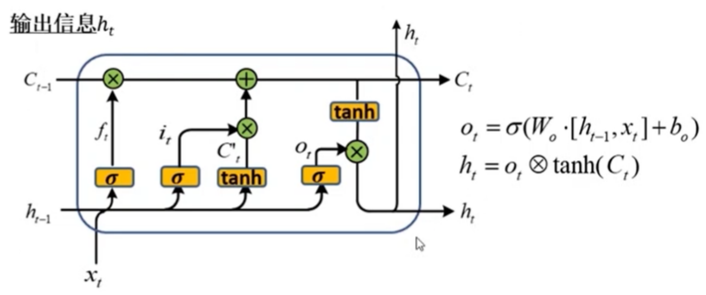
Embedding

- The embedding layer is the same as word2vec. Whether it is skip gram or cbow model, they infer each other from the context and the current, so we consider the relationship between the preceding and the following.
- glove.42B.300d.txt

decision tree

In machine learning, decision tree is a prediction model, which represents a mapping relationship between object attributes and object values. Each node in the tree represents an object, and each branch path represents a possible attribute value, while each leaf node corresponds to the value of the object represented by the path from the root node to the leaf node. The decision tree has only a single output, if you want to have complex output, you can establish an independent decision tree to deal with different outputs. Decision tree is a frequently used technology in data mining, which can be used to analyze data, and also can be used for prediction.

LSTM



BIRNN

- BiRNN:In practical problems, there are also problems that not only rely on the previous sequence, but also rely on the subsequent sequence for prediction. For those problems, we need to use bidirectional RNN (birnn)
- embed size, num hiddens, num layers = 300, 100, 2

comment_text	comment_text
This is so cool. It's like, 'would you want yo...	this is so cool it is like would you want y...
Thank you!! This would make my life a lot less...	thank you this would make my life a lot less...
This is such an urgent design problem; kudos t...	this is such an urgent design problem kudos t...
Is this something I'll be able to install on m...	is this something I will be able to install on...
haha you guys are a bunch of losers.	haha you guys are a bunch of losers

Conclusion

ACC : 0.9468