Многозначная классификация текстов

Нейросети для анализа текстов

Задачи классификации

Бинарная классификация (binary classification)

- Два класса объектов
- Объект может принадлежать только одному классу
- Положительная или отрицательная тональность в IMDB или YELP

Многоклассовая классификация (multiclass classification):

- Несколько классов объектов
- Объект может принадлежать только одному классу
- Темы новостей AG News

Многозначная классификация (multilabel classification)

- Несколько классов объектов
- Каждый объект может принадлежать нескольким классам
- Токсичные комментарии (https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge)

Соревнования по определению токсичности комментариев

kaggle

Search

Home
Compete
Data
Notebooks
Discuss
Courses
More

Recently Viewed
Keras - averaging runs...
GINI + Keras callback E...
AUC-ROC metric for k...
About my 0.9872 singl...
Toxic Comment Classif...

Featured Prediction Competition

Toxic Comment Classification Challenge

Identify and classify toxic online comments

$35,000
Prize Money

Jigsaw/Conversation AI · 4,550 teams · 2 years ago

Overview | Data | Notebooks | Discussion | Leaderboard | Rules | Team | My Submissions | Late Submission

Overview

Description
Evaluation
Timeline
Prizes

Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

The Conversation AI team, a research initiative founded by Jigsaw and Google (both a part of Alphabet) are working on tools to help improve online conversation. One area of focus is the study of negative online behaviors, like toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). So far they've built a range of publicly available models served through the Perspective API, including toxicity. But the current models still make errors, and they don't allow users to select which types of toxicity they're interested in finding (e.g. some platforms may be fine with profanity, but not with other types of toxic content).

In this competition, you're challenged to build a multi-headed model that's capable of detecting different types of of toxicity like threats, obscenity, insults, and identity-based hate better than Perspective's current models. You'll be using a dataset of comments from Wikipedia's talk page edits. Improvements to the current model will hopefully help online discussion become more productive and respectful.

Классы токсичности комментариев

Название класса на английском	Название класса на русском
toxic	токсичный
severe_toxic	существенно токсичный
obscene	обсценная лексика
threat	угроза
insult	оскорбление
identity_hate	личная ненависть

Формат данных

Комментарий	toxic	severe_toxic	obscene	threat	insult	identity_hate
Hi! I am back again! Last warning! Stop undoing my edits or die!	1	0	0	1	0	0
Would you both shut up, you don't run wikipedia, especially a stupid kid.	1	0	0	0	1	0
COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK	1	1	1	0	1	0
Your vandalism to the Matt Shirvington article has been reverted. Please don't do it again, or you will be banned.	0	0	0	0	0	0

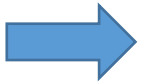
Архитектура нейронной сети

```
model = Sequential()
model.add(Embedding(10000, 128, input_length=50))
model.add(SpatialDropout1D(0.5))
model.add(LSTM(40, return_sequences=True))
model.add(LSTM(40))
model.add(Dense(6, activation='sigmoid'))
```

Пространственный Dropout

I	0.1	0.8	-0.5	0.3
was	-0.2	0.15	0.99	0.01
very	0.75	-0.1	-0.17	0.64
touched	-0.1	-0.2	0.23	0.13

Исходные данные



I	0.1	0.8	-0.5	0.3
was	0.0	0.0	0.0	0.0
very	0.75	-0.1	-0.17	0.64
touched	-0.1	-0.2	0.23	0.13

SpatialDropout

Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, Christoph Bregler. Efficient Object Localization Using Convolutional Networks. 2015. https://arxiv.org/pdf/1411.4280.pdf

Функция ошибки

```
model.compile(optimizer='adam',
              loss='binary_crossentropy',
              metrics=['accuracy', 'AUC'])
```

Результаты работы сети

```
comment = "X-BOX 360 SUKCS BIG BUMM AND LIKES IT UP THE ASS"
sequence = tokenizer.texts_to_sequences([comment])
data = pad_sequences(sequence, maxlen=50)
result = model.predict(data)
array([[0.93577635, 0.05646498, 0.63453263, 0.02494773,
0.59398603, 0.07905076]], dtype=float32)
```

Правильный ответ:

```
[1,0,1,0,1,0]
```

Функции активации и ошибки для задач классификации

Задача классификации	Функция активации выходного слоя	Функция ошибки
Бинарная классификация (binary classification)	sigmoid	binary_crossentropy
Многоклассовая классификация (multiclass classification)	softmax	categorical_crossentropy
Многозначная классификация (multilabel classification)	sigmoid	binary_crossentropy