Usage recurrent neural networks for sentiment analysis of social media users' comments

A. A. Zhdanova zhdan.aleksandra@gmail.com

A. V. Kupriyanov akupr@ssau.ru

A. A. Golova nastyagolova13@gmail.com

A. S. Bulgakov charlez.wape@gmail.com

D. S. Bakanov dima.bakanov.1999@mail.ru

PROBLEMS OF TEXT SENTIMENT ANALYSIS

Some examples of sequence prediction problems include:

1. One to many: observation as input, mapped to a sequence with multiple steps as output.
2. Many-to-one: a sequence of several steps as inputs mapped to a class or quantity prediction.
3. Many-to-many: a sequence of multiple steps as input, mapped to a sequence of multiple steps as output.
4. The many-to-many problem is often referred to as sequence to sequence, or seq2seq for short.

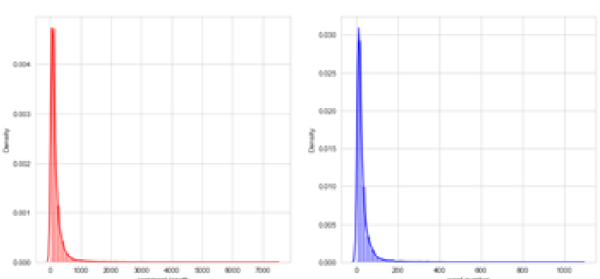
NEURAL NETWORK TRAINING

We studied 3 models and their effectiveness. The results are presented in the table.

	ROC-curve	Result
SimpleRNN	0.83	The network predicts true values quite well. It is also worth nothing that the deviations in negative responses are lower by an order of magnitude than in positive ones, indicating a fairly high efficiency.
LSTM	0.91	The neural network predicts true values well, but unlike SimpleRNN, it tends to misclassify negative sentiments as positive ones.
GRU	0.87	This neural network better identifies true sentiment values of reviews, but it tends to make more errors on negative reviews that on positive ones.

RESULTS

The training set consists of 14412 informal Russian comments from social networks with jargon and hidden subtext, making training a deep learning model more complex. The dataset was converted to lowercase and HTML tags were removed before conducting exploratory data analysis to evaluate the representativeness of the dataset and the distribution of emotional sentiment.



Comment length distribution and word number distribution in the recall

```
count     14412.000000
mean        176.525812
std         271.612376
min          21.000000
25%          57.000000
50%         101.000000
75%         197.000000
max        7404.000000
Name: comment length, dtype: float64
```

```
count    14412.000000
mean        27.946087
std         41.432195
min          1.000000
25%          9.000000
50%         16.000000
75%         32.000000
max       1078.000000
Name: word number, dtype: float64
```

Distribution of recall length and distribution of number of words in comments

As can be seen, there is no bias towards one type of text sentiment among the data, and based on the obtained numerical characteristics of the records, the text sample can be considered large enough: [21;7404] letters and [1; 1078] words. From the set of factors, it follows that the training dataset is balanced.

Of all the neural network architectures considered, the best performers were neural networks with LSTM and GRU layers, with approximately the same accuracy: 0.88 and 0.87, respectively.

Several sentences were randomly selected to display the predictions of the neural networks. Each neural network made predictions for them.

The table displays comment texts and their tonality predicted by three neural networks: SimpleRNN, GRU, and LSTM. The true values of predictions are also shown. Overall, the networks performed well in predicting tonality, with only small deviations from the true values. However, errors were made on sentences with hidden subtext, likely due to the small amount of data. GRU and LSTM networks had fewer errors compared to SimpleRNN, as they have memory blocks and apply filters.

Comment	True value	Simple RNN prediction	GRU prediction	LSTM prediction
светофоры с вызовом таймером	0	0,0001	0	0,0003
спасибо за конструктивную критику приму сведению ваши слова	0	0,0002	0,0002	0,0003
рабочий день впереди	0	0,0004	0,0001	0,0003
мать брат сват ребенок сиди пикабу вконтактике дурак	1	0,8279	0,4614	0,1727
просто похоже это горит красный это магазин иркутский фары далеке это энергетик	0	0,1031	0,0425	0,0018
яндекс мб норм проводят собеседования работать отнюдь прекрасно гугле аппле хорошие собеседования понравилось собеседование близзард например ищут джунов чаще говорим топовые компании середнячок	0	0,0049	0,0004	0,0001
южные культуры ещё существуют вроде говорили олимпиады вырубили нафиг	0	0,0002	0	0,0001
кожного дурака	1	0,955	0,9781	0,9434
шо свое лицо засвети	1	0,9981	0,9995	0,9988
моих коллег программистов сеньоры зареган соцсетях некоторые заводят резюме принципа правда отправляет резюме её запросам хороших знакомых вс равно умудряются хантить каким образом	0	0,0828	0	0,0001
фото парад уродов лол	1	0,9984	0,9997	0,9995
действительно столько всякой химозы использую путаю названия	0	0,0001	0	0,0001
спасибо большое дельный совет	0	0,0002	0	0

SAMARA UNIVERSITY

34 Moskovskoye shosse, Samara, 443086
+7 (846) 335-18-26, +7 (846) 335-18-36 (fax)
ssau@ssau.ru, www.ssau.ru

Faculty of Information Technology, Software Department
+7 (846) 267-43-61 , +7 (905) 302-39-58
zhdan.aleksandra@gmail.com