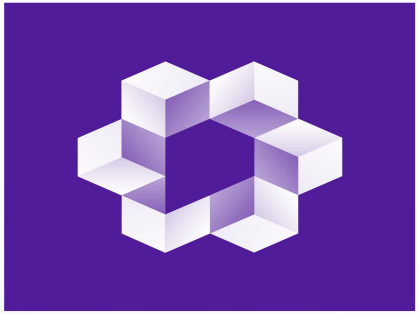
kaggle



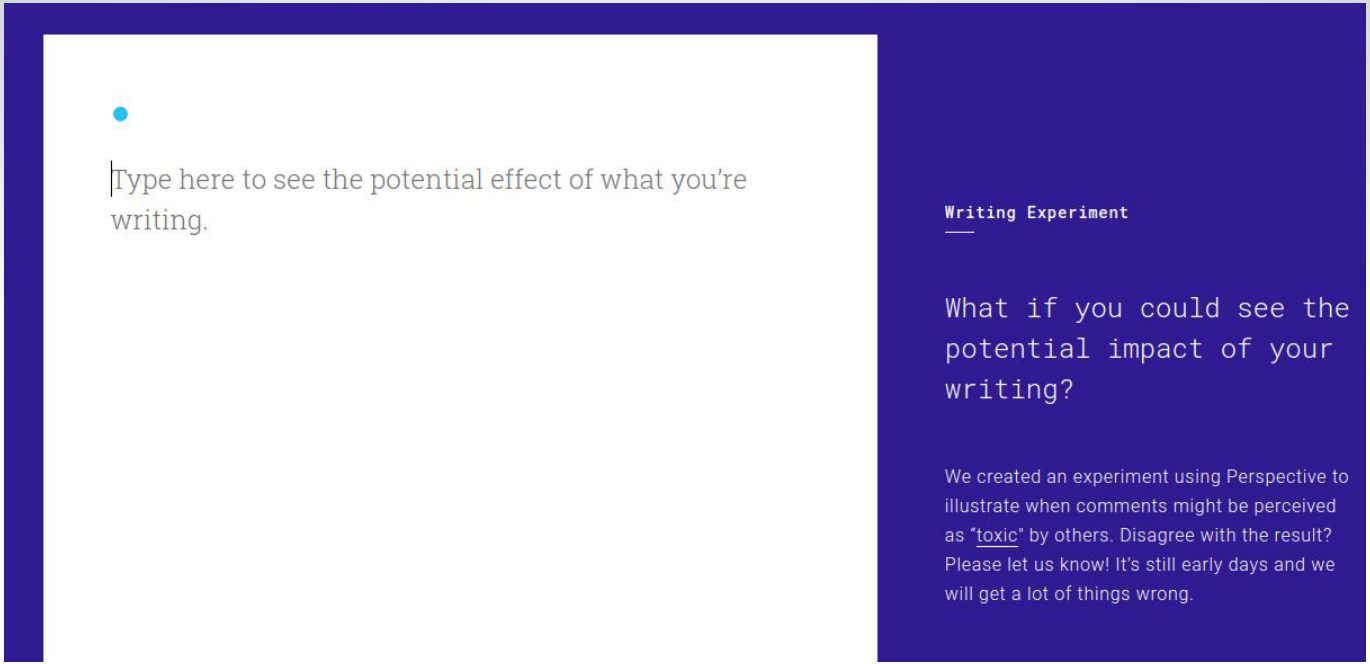
What if technology could help improve conversations online?

Toxic Comment Classification Challenge

(Выявление и классификация токсичных комментариев)

DecisionGuys: 10 out of 4551

Perspective API



www.perspectiveapi.com

Постановка задачи

Имеем следующую задачу **NLP** – научиться детектировать различные классы токсичных комментариев.

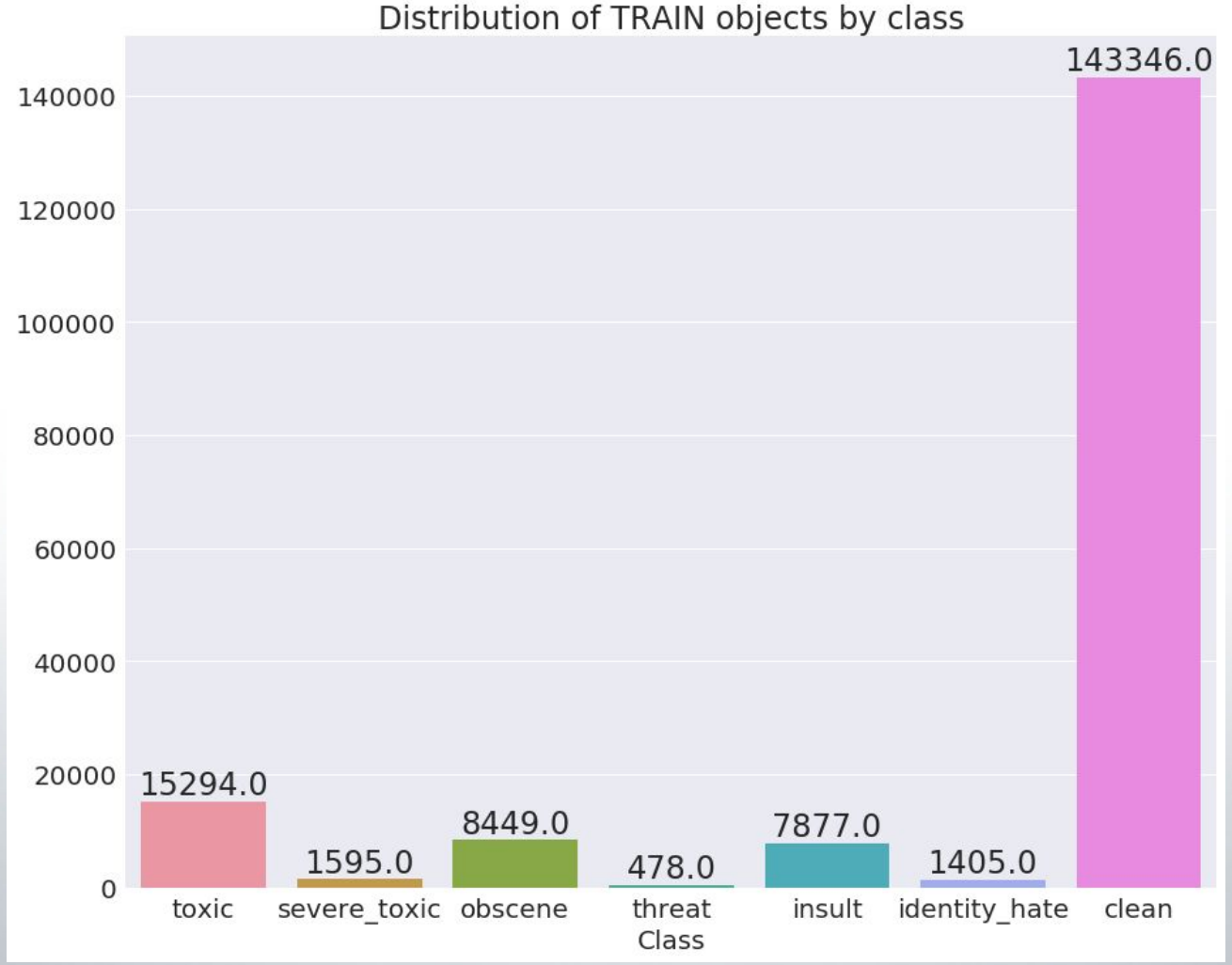
Train: ~160к комментариев Test: ~153к комментариев

Постановка задачи

Классы токсичности:

- угроза (класс **toxic**)
- едкая угроза (класс **severe_toxic**)
- непристойность (класс **obscene**)
- ещё какой-то тип угрозы (класс **threat**)
- оскорбление (класс **insult**)
- ненависть к личности (класс **identity_hate**)

Метрика качества: **mean column-wise ROC AUC**


Distribution of TRAIN objects by class
140000
120000
100000
80000
60000
40000
20000
0
15294.0
1595.0
8449.0
478.0
7877.0
1405.0
143346.0
toxic
severe_toxic
obscene
threat
insult
identity_hate
clean
Class

Предобработка данных

Базовая предобработка данных:

- преобразование вида:
- "W H A T A F ■ C K M A N" → "WHAT A F■CK MAN"
- приведение текста к нижнему регистру
- удаление ссылок, ip
- удаление цифр
- удаление пунктуаций (кроме апострофов)

Предобработка данных

Доп предобработка данных:

- замена смайликов на соответствующие слова
- расшифровка сокращений
- исправление опечаток в ненормативной лексике
- приведение различного сорта мата к одному и тому же ввиду (например, "fc*k": "f■ck", "fu**": "f■ck")
- удаление изображений

$$\begin{pmatrix}\text{разброс}\\ \text{композиции}\end{pmatrix} = \frac{1}{N}\begin{pmatrix}\text{разброс одного}\\ \text{базового алгоритма}\end{pmatrix} + \begin{pmatrix}\text{корелляция между}\\ \text{базовыми алгоритмами}\end{pmatrix}$$

Bag of Words

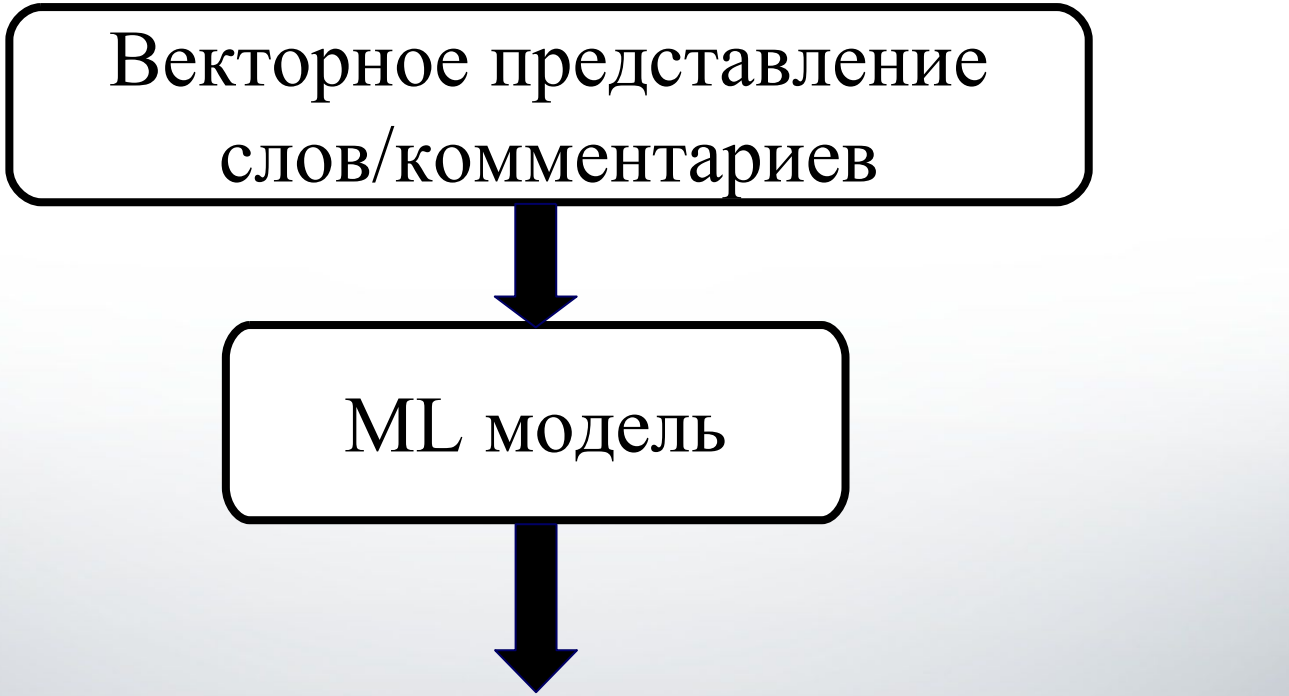
«Порядок слов имеет значение?»

«Значение слов имеет порядок?»

«Порядок имеет значение слов?»

Хм…

Общий подход



Стратифицированная кросс-валидация по 10 фолдам

CV: 0.9785

LB: 0.9714 **(private)**

SO LONG.

arxiv.org/abs/1405.4053

кросс-валидация по 10 фолдам

LB: 0.9854
(public)

LB: 0.9851
(private)

One of the best single model

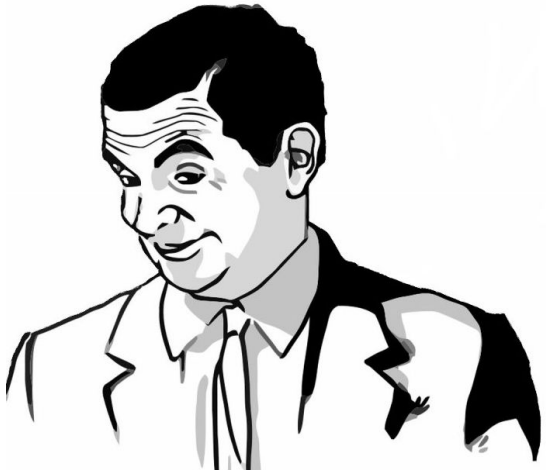
arxiv.org/abs/1607.04606

Рецепт успеха

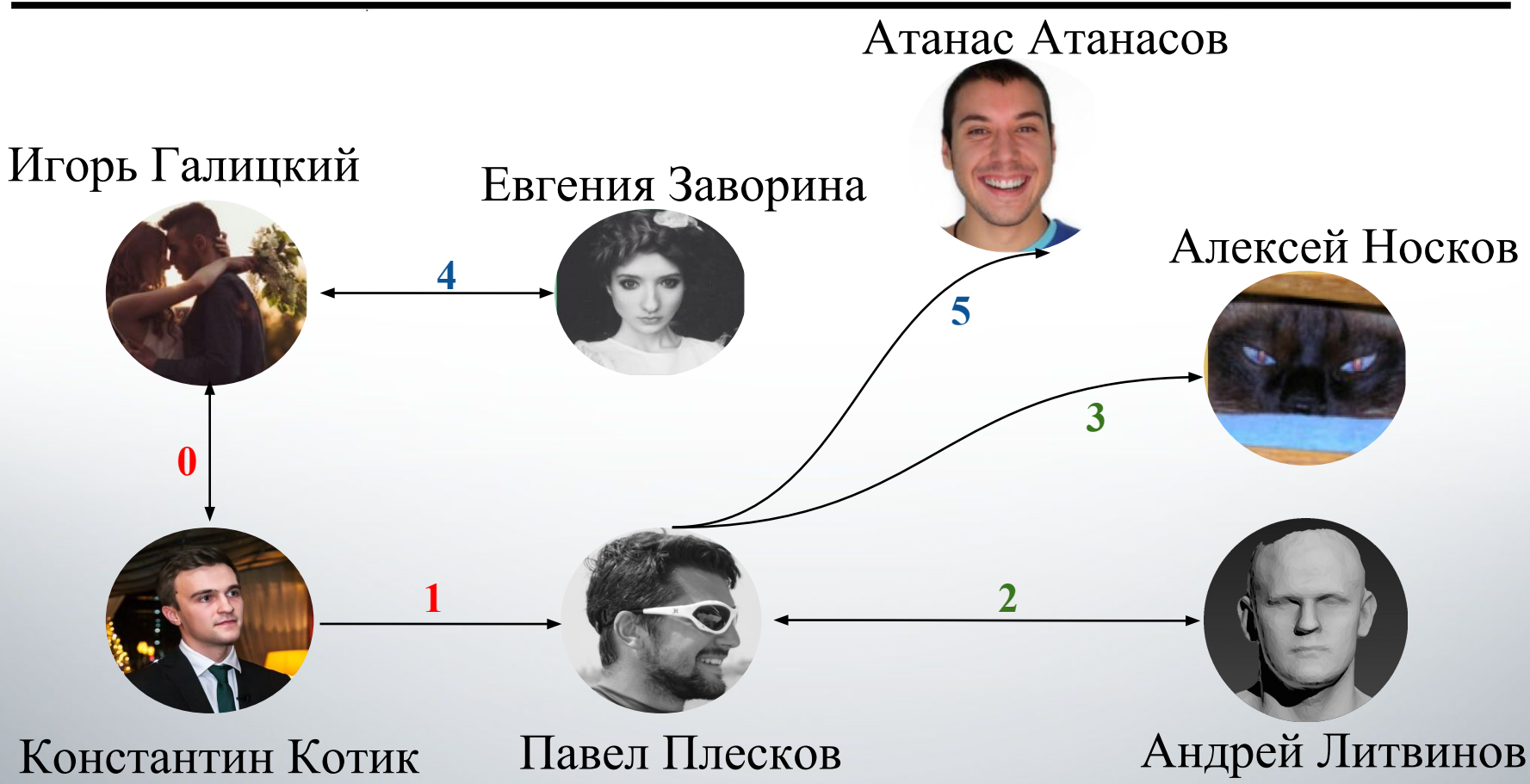
BLENDING

STACKING

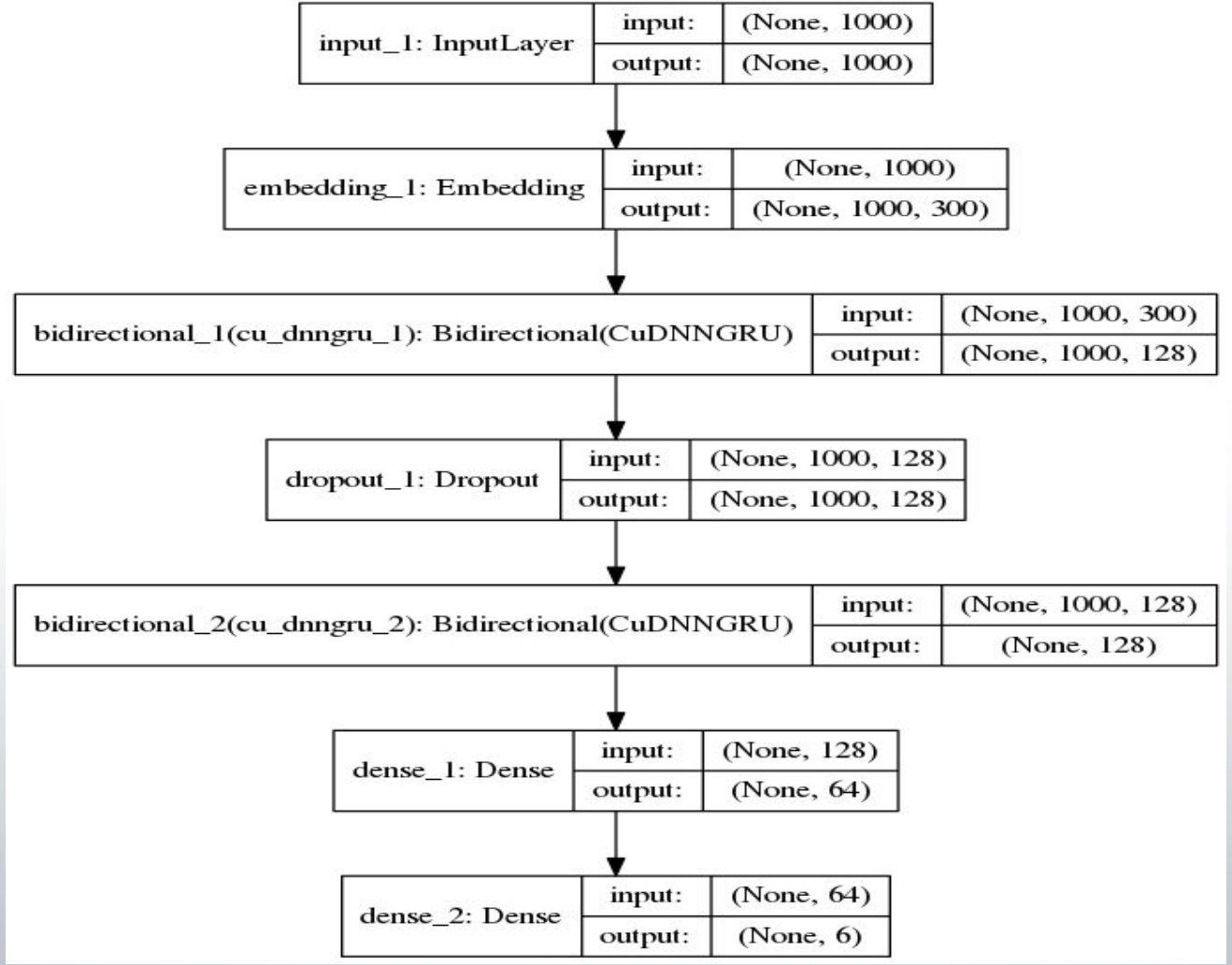
NETWORKING



NETWORKING


Атанас Атанасов
Игорь Галицкий
Евгения Заворина
Алексей Носков
4
5
3
0
1
2
Константин Котик
Павел Плесков
Андрей Литвинов


Neural Network: Epoch 1


input_1: InputLayer
input: (None, 1000)
output: (None, 1000)
embedding_1: Embedding
input: (None, 1000)
output: (None, 1000, 300)
bidirectional_1(cu_dnngru_1): Bidirectional(CuDNNGRU)
input: (None, 1000, 300)
output: (None, 1000, 128)
dropout_1: Dropout
input: (None, 1000, 128)
output: (None, 1000, 128)
bidirectional_2(cu_dnngru_2): Bidirectional(CuDNNGRU)
input: (None, 1000, 128)
output: (None, 128)
dense_1: Dense
input: (None, 128)
output: (None, 64)
dense_2: Dense
input: (None, 64)
output: (None, 6)

Павел Плесков

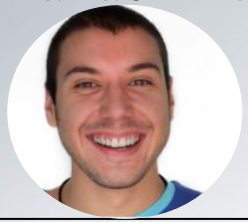
"У меня bi-gru на fastext и glove на 10 фолдах с кастомным препроцессингом, и чуть допиленный svm Джереми
pseudo labeling ещё юзаю, остальное публичное"

LB: 0.9861
(private)

LB: 0.9867
(private)

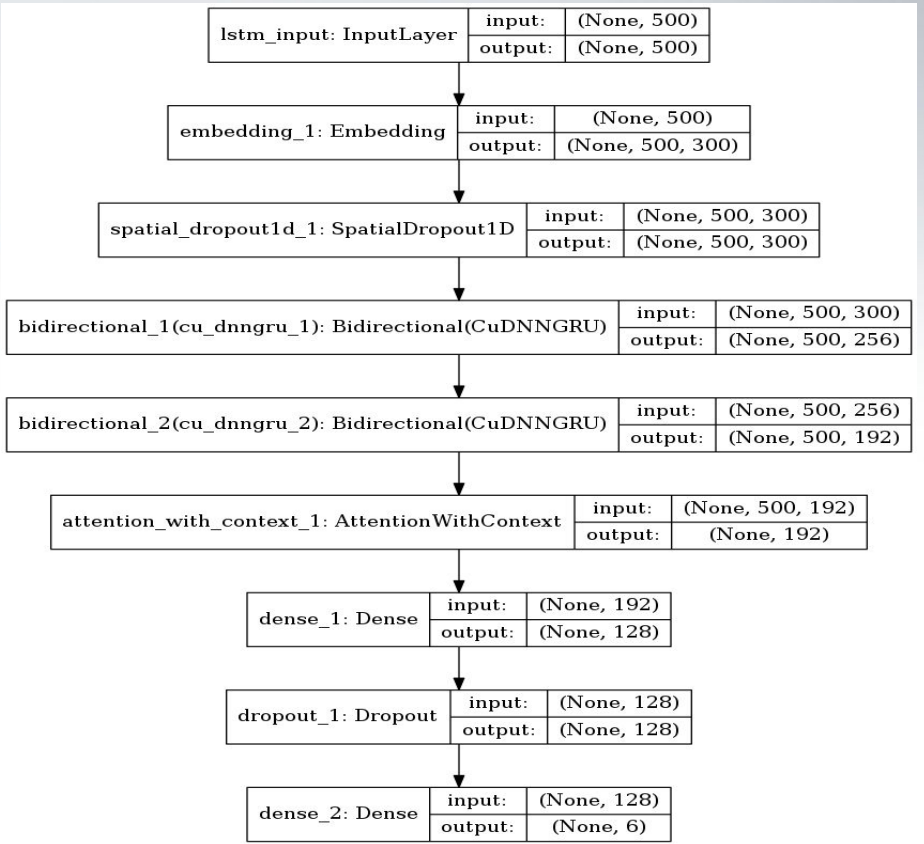
Vanilla Bi-Gru combined with AttentionWithContext

Атанас Атанасов



Standard 2-layer Bi-Gru Model + AttentionWithContext.

Implementation Framework: Keras
Preprocessing: Yes
Sentence Length: 500
Number of Folds: 10
Word Embeddings: FastText Common Crawl (600B tokens)
Tunning Hyperparameters: Hyperopt
CV Score: 0.9886
Private LB: 0.9848

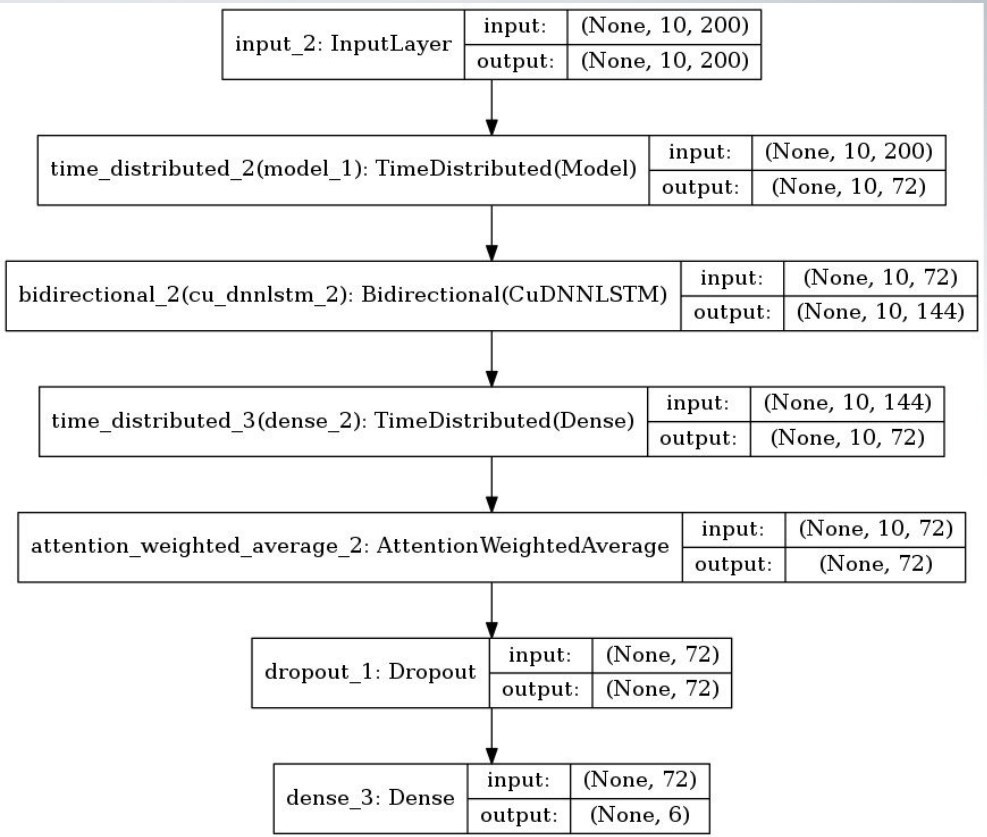


HANN

Classical implementation of Hierarchical Attention Neural Networks using Bi-LSTMs

Implementation Framework: Keras
Preprocessing: Yes
Max Text Length: 200
Max Sentences: 10
Max Features: 150000
Number of Folds: 10
Word Embeddings: FastText Common Crawl (600B tokens)
Tunning Hyperparameters: Hyperopt

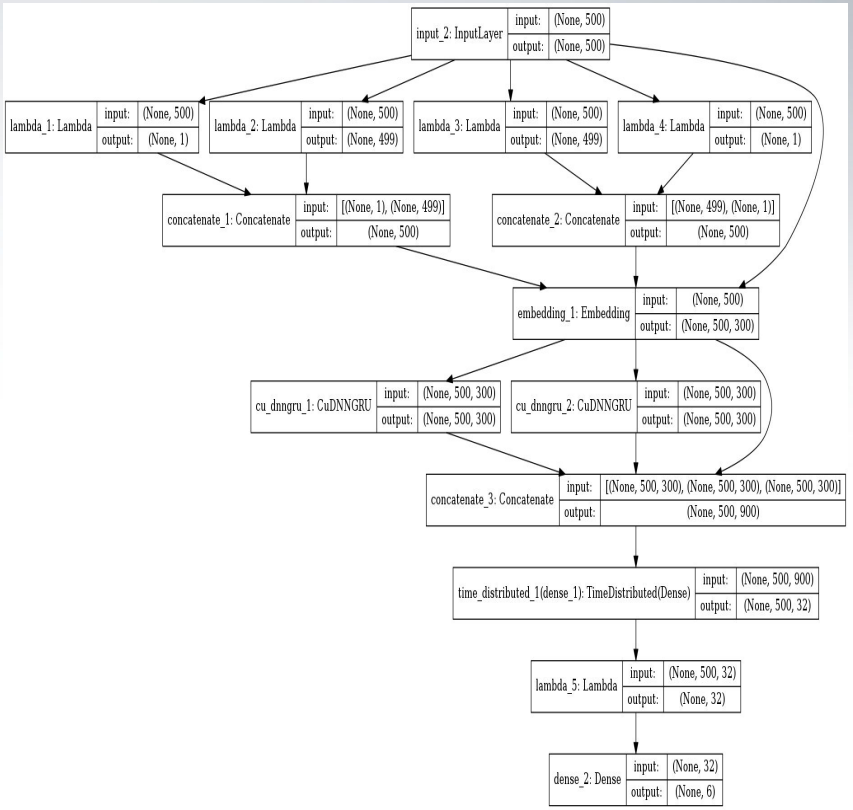
CV Score: 0.9869
Private LB: 0.9833



Recurrent CNN for Text Classification

Bi-directional (GRU) recurrent structure that reduces noise and captures semantic information to the greatest extent possible.

Implementation Framework: Keras
Preprocessing: Yes
Sentence Length: 500
Number of Folds: 10
Word Embeddings: FastText Common Crawl (600B tokens)
Tunning Hyperparameters: Hyperopt
CV Score: 0.9874
Private LB: 0.9832



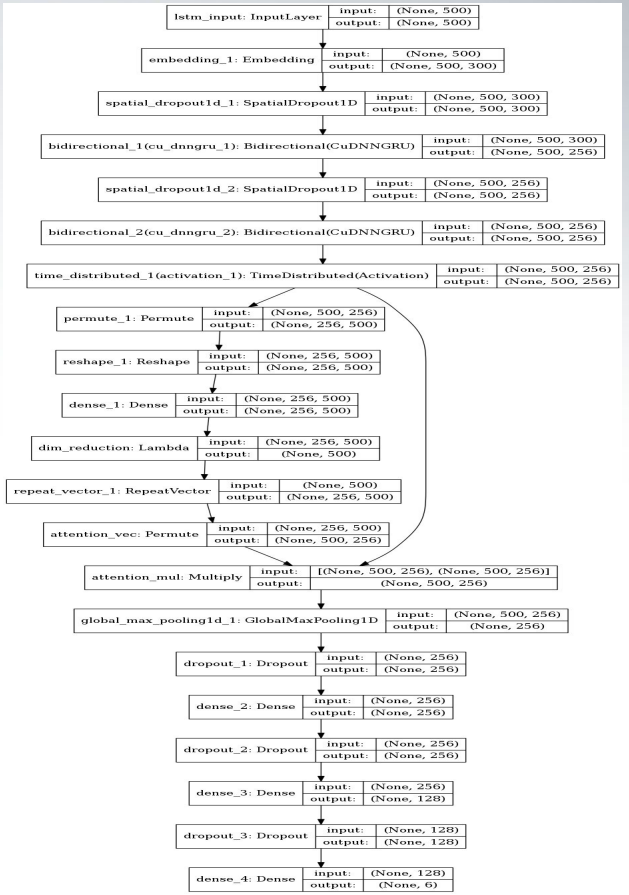
Paper Link:https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/viewFile/9745/9552

Vanilla Bi-GRU with attention block

2-layer Bi-Gru Model + attention block.

Implementation Framework: Keras
Preprocessing: Yes
Sentence Length: 500
Number of Folds: 10
Word Embeddings: FastText Common Crawl (600B tokens)
Tunning Hyperparameters: Hyperopt

CV Score: 0.9890
Private LB: 0.9846

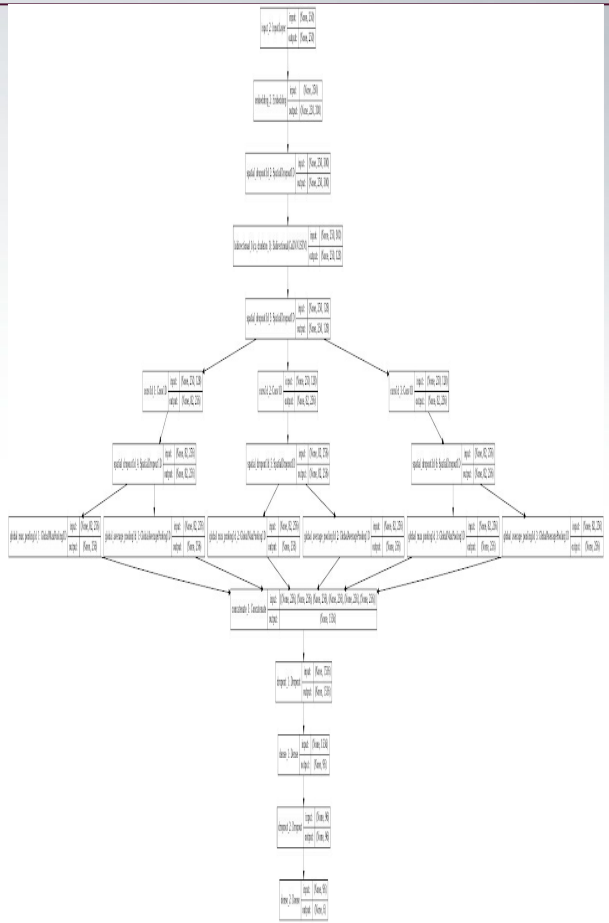


LSTM combined CNN (max, average) poolings

LSTM concatenated with 3 layers of CNN[Convolutional1D, GlabalMaxPooling1D, GlobalAveragePooling1D]

Implementation Framework: Keras
Preprocessing: Yes
Sentence Length: 250
Number of Folds: 10
Word Embeddings: FastText Common Crawl (600B tokens)
Tunning Hyperparameters: Hyperopt

CV Score: 0.9866
Private LB: 0.9842

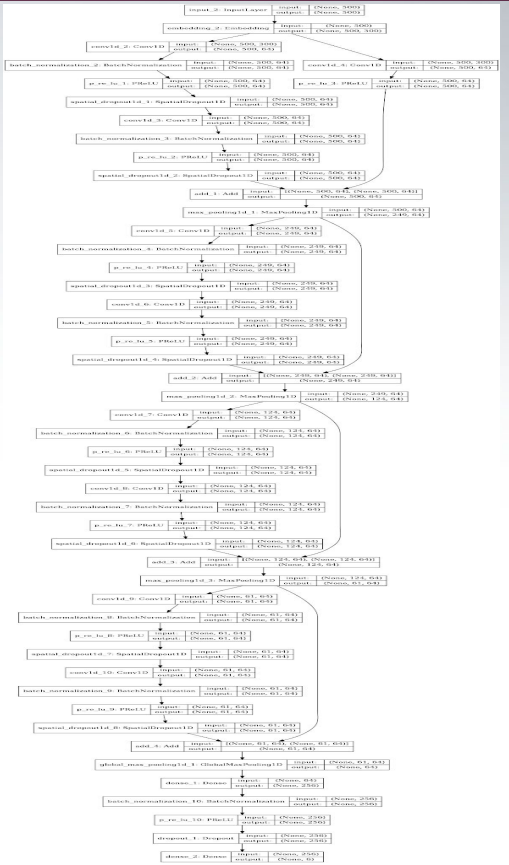


DPCNN

Word-level deep convolutional neural network (CNN) architecture for text categorization that can efficiently represent longrange associations in text.

Implementation Framework: Keras
Preprocessing: Yes
Sentence Length: 500
Number of Folds: 10
Word Embeddings: FastText Common Crawl (600B tokens)
Tunning Hyperparameters: Hyperopt

CV Score: 0.9850
Private LB: 0.9838



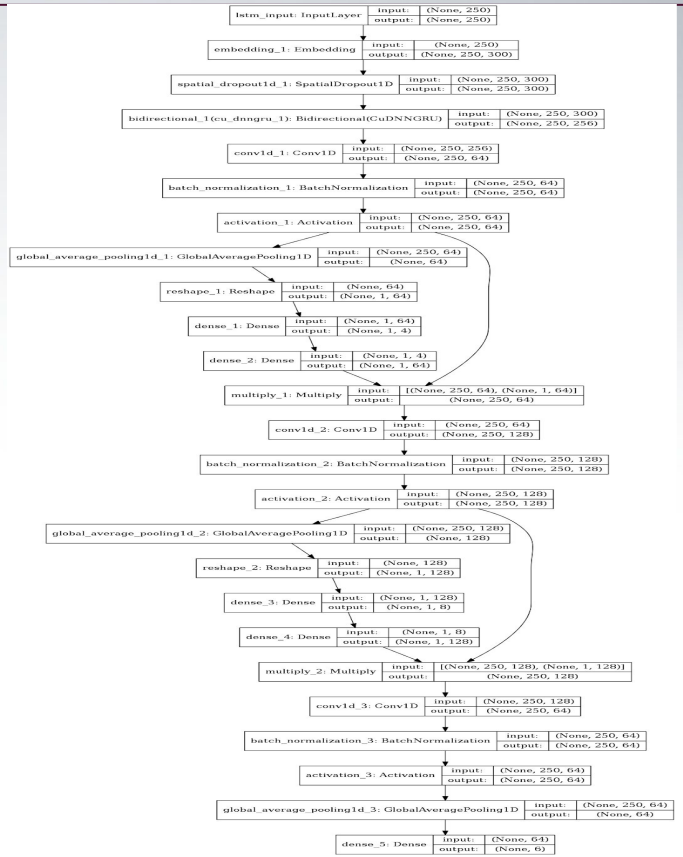
Paper Link: http://ai.tencent.com/ailab/media/publications/ACL3-Brady.pdf

Squeeze and Excitation Networks (adapted for Text)

SE block adapted for text that recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels.

Implementation Framework: Keras
Preprocessing: Yes
Sentence Length: 250
Number of Folds: 10
Word Embeddings: FastText Common Crawl (600B tokens)
Tunning Hyperparameters: Hyperopt

CV Score: 0.9902
Private LB: 0.9845



Paper Link: Paper Link: https://arxiv.org/abs/1709.01507

AC-BLSTM

Asymmetric Convolutional Bidirectional LSTM Networks for Text Classification

Implementation Framework: Keras
Preprocessing: Yes
Sentence Length: 600
Number of Folds: 10
Word Embeddings: FastText Common Crawl (600B tokens)
Tunning Hyperparameters: Hyperopt

CV Score: 0.9904
Private LB: 0.9845

Paper Link: https://arxiv.org/abs/1611.01884

Stacker 1:

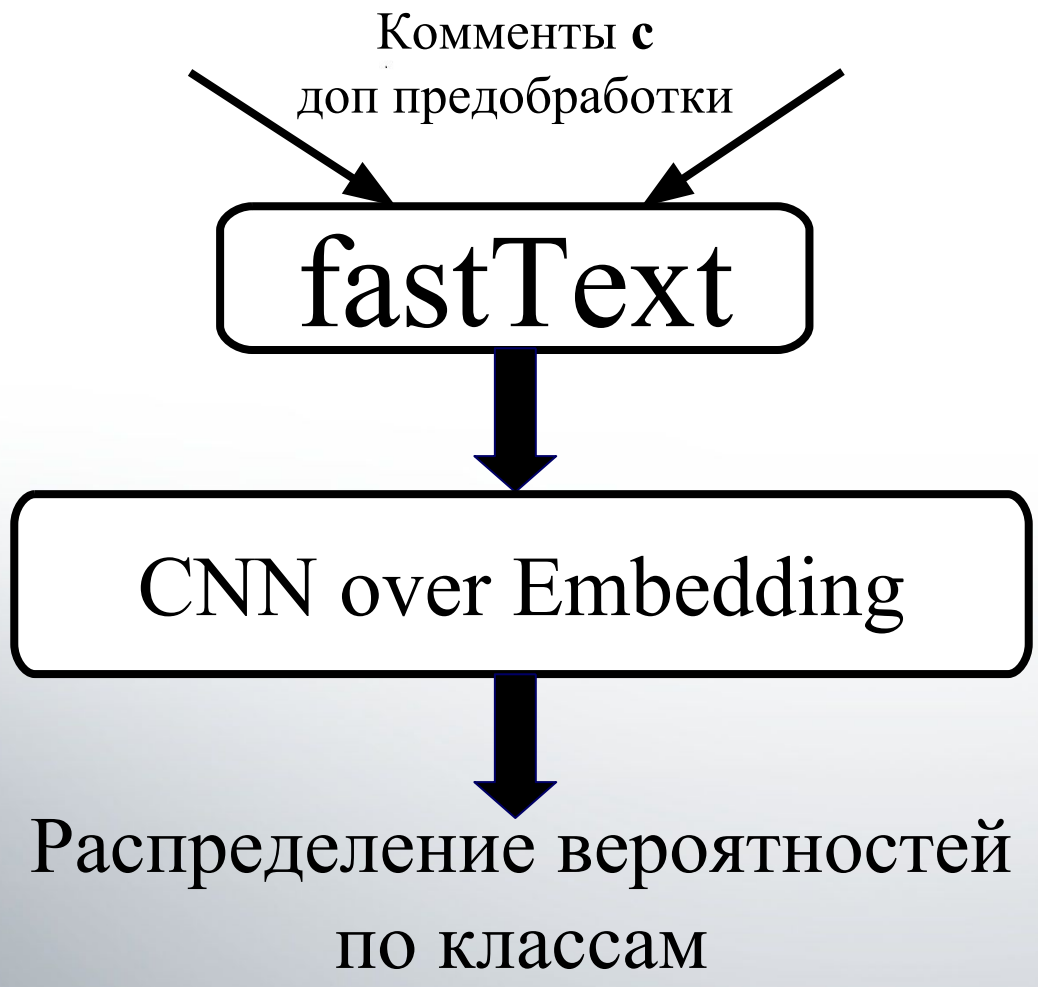
- All of these models stacked with LightGBM with the following features:

len, uppercase_freq, number_of_f_words,number_of_unique_words, number_of_you_words, number_of_punct, number_of_mother, number_of_nig[redacted]er, total_length, num_symbols, num_smilies etc...

- CV = 5
- CV Score: 0.99211384493568
- PLB : 0.9867

Stacker 2:

- All of these models (pretrained with Glove) stacked with (MLP + XGBoost)
- Used additional features as auxiliary input for the models (punctuation, word_len, mean_word_len)
- CV Score: 0.9103829829292
- CV = 10
- PLB: 0.9863

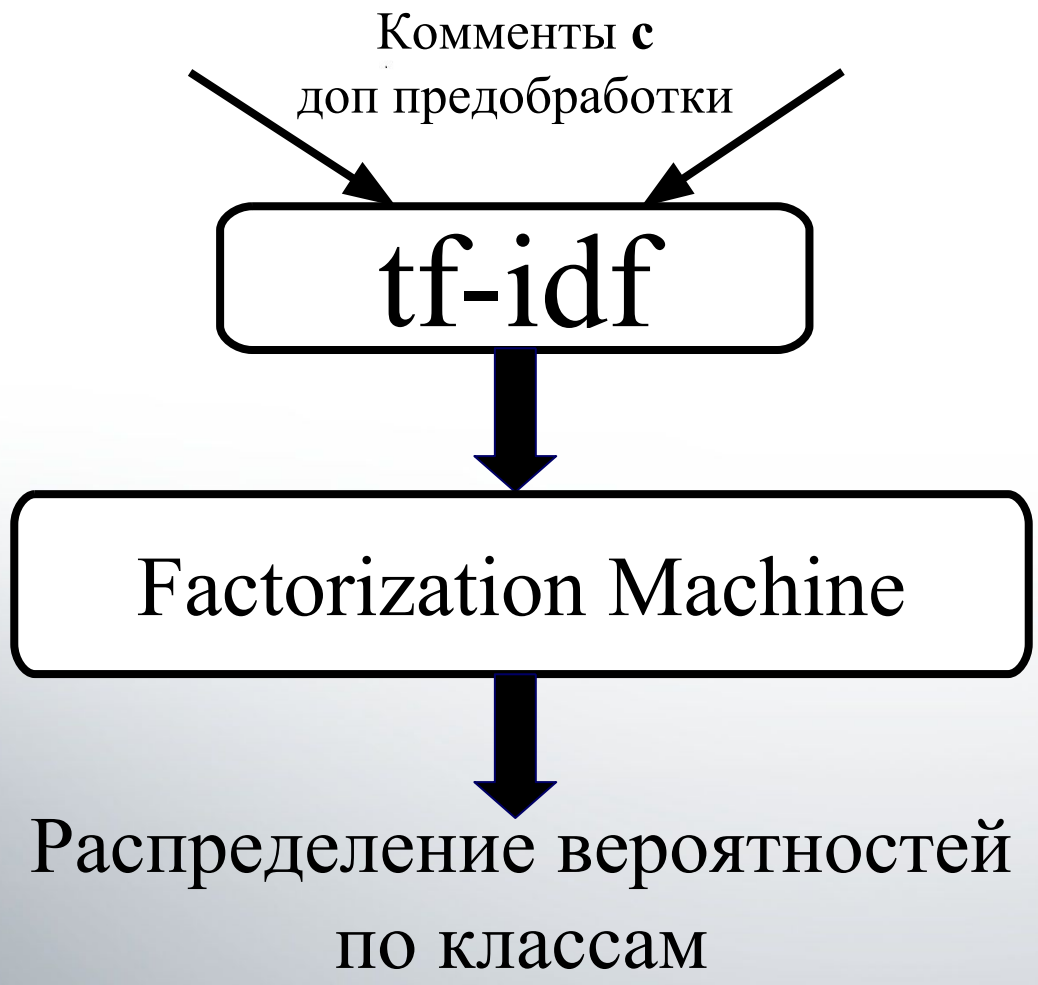


Стратифицированная кросс-валидация по 10 фолдам

CV: 0.9819

LB: 0.9816
(private)

arxiv.org/abs/1405.4053



Стратифицированная кросс-валидация по 10 фолдам

CV: 0.9824

LB: 0.9820
(private)

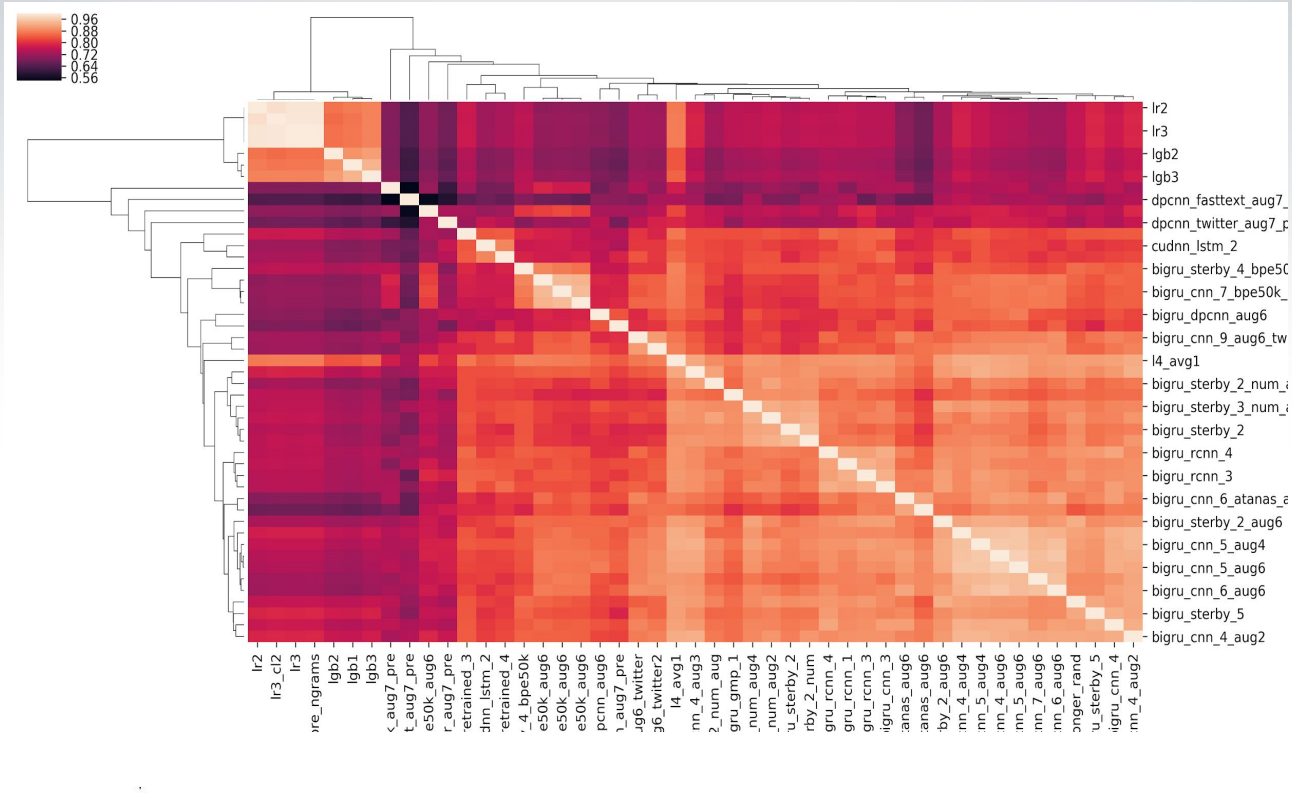
arxiv.org/abs/1405.4053


STACKING

Общий сетап

- Кросс-валидация на 10-KFold
- Модели с различными векторами и препроцессингом
- Аугментация данных
- Два подхода:
  a. Обучил N моделей - посмотрел ошибки на CV - исправил препроцессинг
  b. Обучил N моделей - посмотрел на корреляции, усилил слабо скоррелированные

Корреляции Спирмена между моделями



Аугментации

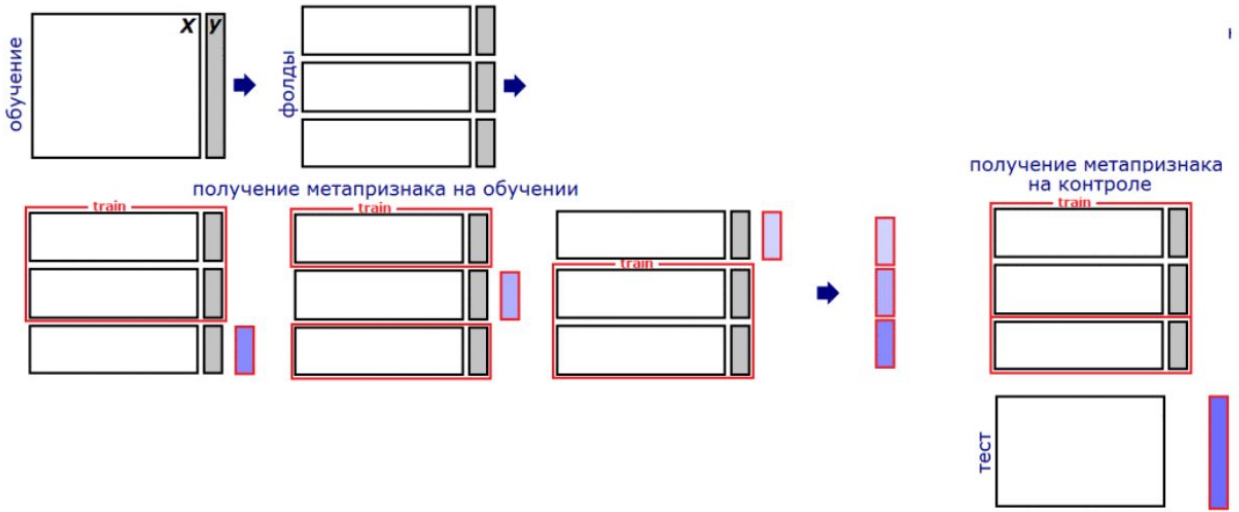
- **Переводы**
  - Идея Павла Остякова
  - https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/48038
  - Перевод комментария на какой-то язык и обратно
- **Конкатенация**
  - Склеиваем два комментария, возьмем в качестве лейблов их объединение
- Каждую эпоху применяем к новому подмножеству комментариев

Модели на BPE

- SentencePiece - токенайзер Google для нейросетей
  - Строит фиксированный словарь заданного размера без UNK
  - Hello_World. => [Hello] [_Wor] [ld] [.]
  - https://github.com/google/sentencepiece
  - Два алгоритма - BPE и Unigram
- Pretrained BPE embeddings
  - https://github.com/bheinzerling/bpemb
- Неплохо заработал словарь 50к

Stacking

Классическая схема



Ансамбли

- **Усреднение N x 10 моделей**
  - Практически до самого конца ничего стабильно лучше не получалось сделать
- **LightGBM**
  - Усреднение групп схожих моделей первого уровня
  - Обучение на 10 фолдах CV, усреднение перед генерацией сабмишна
  - Баггинг 20 запусков
  - Метафичи

API

- Было разрешено использовать API существующей системы
  - https://www.perspectiveapi.com/
- Сходный (но отличный) набор лейблов
- Хорошо заходит в качестве метафичей на втором уровне
  - Особенно вытягивает класс TOXIC

0.99275 => 0.99314 CV

0.9879 => 0.9882 LB

Финальные сабмишны

- Бленд моделей без API
  - Public 0.9882
  - Private 0.9876

- Бленд моделей без API + LightGBM с API
  - Public 0.9880
  - Private 0.9874

DecisionGuys — 10 out of 4551

Алексей Носков

alexeynoskov (Leader)

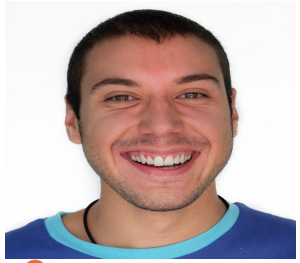
Павел Плесков

ppleskov

Константин Котик

kotikkonstantin

Атанас Атанасов



atanasova

Игорь Галицкий

igeti

Андрей Литвинов

lao777

Евгения Заворина

panamka