

Humor detection

(на материале русского языка)



ПОДГОТОВИЛИ
Арсений Анисимов и
Александра Нужненко,
БКЛ-202

Задача

Построить языковую модель, которая способна **делить короткие отрывки текста на смешные и несмешные** (бинарная классификация).

Почему это актуально?

В шутках кодируются сложные эмоции и отношение говорящего к ситуации, поэтому **распознавание юмора является частью задачи sentiment analysis** и помогает автоматическим системам лучше “понимать” человека.

Датасет

Процесс сбора шуток:

- сайт [Некдо](#)
- проход по страницам и парсинг html-кода с помощью библиотеки **BeautifulSoup**

пример шутки из датасета:

Черная кошка перешла дорогу женщине с пустым ведром. Теперь у обеих проблемы.

Итог: **20615 коротких анекдотов**
(длина от 4 до 14 слов).

Процесс сбора не-шуток:

- Википедия
- обход страниц ресурса и сбор предложений с помощью библиотеки **wikipedia**

пример не-шутки из датасета:

Основной одеждой женщины были льняная сорочка и шерстяной сюртук.

Итог: **20615 предложений**
(длина от 4 до 14 слов).

Датасет

Почему Википедия?

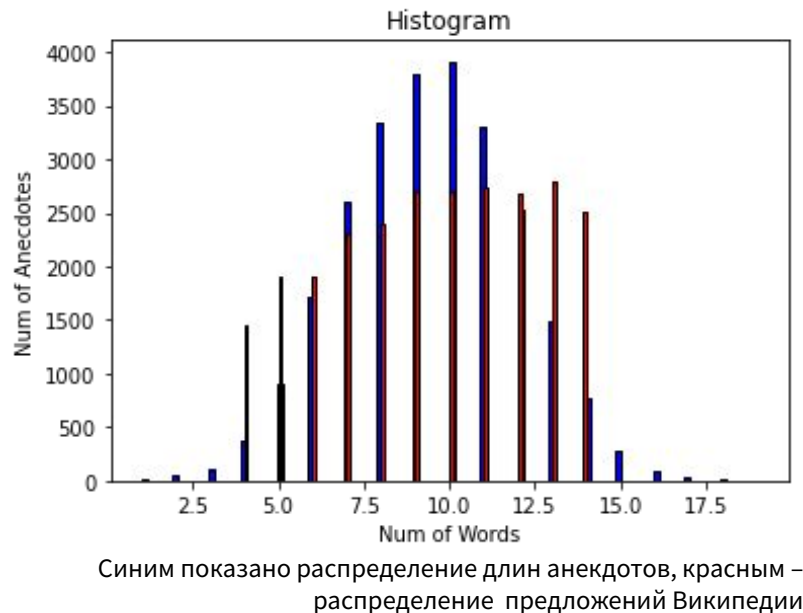
- В статьях точно нет шуток, в то время как в книгах они могут быть;
- Ресурс составлен разными людьми, поэтому нет влияния стиля одного/нескольких конкретных авторов;
- (По идее) нейтральный стиль.

Как собирали предложения:

- задали случайный начальный список статей Википедии,
- извлекали тексты статей и ссылки на другие статьи Википедии, содержащиеся в этих текстах,
- искали предложения, переходя по найденным ссылкам, пока не набрали нужный объем
- делили тексты на предложения с помощью nltk,
- фильтровали данные (напр., не собирали предложения, если дошли до разделов “Литература”, “Примечания” или “Ссылки”; ограничивали длину (от 4 до 14 слов))

Датасет (проблемы)

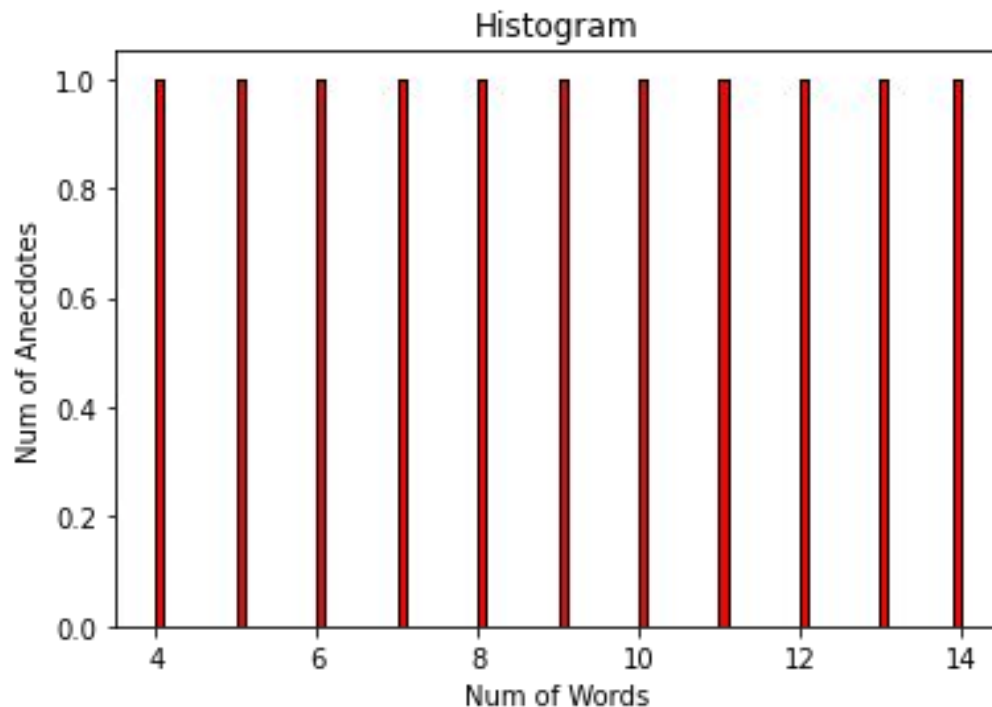
- Разное распределение длин предложений в шутках и не-шутках
=> Произвели выравнивание длин
- Стиль предложений заметно отличается от шуток: тексты более “академичные” (больше пассивных залогов, точных данных)
- sent_tokenizer из nltk плохо делит на предложения, в результате данные состоят из обрывков фраз, элементов списков и т.п.



пример плохих данных из Википедии:

- España, МФА: [es'paɾa]), официально Королёвство Испания (исп.*
- По состоянию на 2016 год средний размер оплаты труда в Испании составляет €2189.*

Датасет



Взято равное количество данных (и шуток, и не-шутки) с всеми возможными длинами.

Методы

Для решения задачи классификации была выбрана языковая модель MultiBERT

Почему BERT?

- BERT учитывает более широкий контекст, чем, например, word2vec,
- в отличие от word2vec, BERT умеет учитывать позицию слова в предложении и запоминать эту информацию в эмбедингах,
- в отличие от word2vec, BERT умеет работать со словами, которых не было в обучающей выборке,
- в целом это более продвинутая модель, показывающая лучшие результаты на многих задачах

Результаты

Мы смотрим преимущественно на **precision**, поскольку на практике (например, при общении бота с человеком) кажется более важным, чтобы бот не начал считать шутками серьезные предложения, чем не-улавливание ста процентов юмора.

– Validation accuracy (epoch): 0.9406

	precision	recall	f1-score	support
0.0	0.93	0.96	0.94	4077
1.0	0.96	0.92	0.94	4169
accuracy			0.94	8246
macro avg	0.94	0.94	0.94	8246
weighted avg	0.94	0.94	0.94	8246

Показаны значения после первой эпохи

Точность не такая высокая, как хотелось бы, быстро началось переобучение (показан результат после первой эпохи).

Датасет2

В нем мы постарались улучшить собранные из Википедии данные



боремся с мусором:

- исключаем предложения с несовпадением скобок и кавычек (такие предложения точно оборванные),
- исключаем предложения, которые начинаются и заканчиваются так, как не могут начинаться/заканчиваться русские предложения (напр., которые заканчиваются на “тыс.”, “млн.”).

боремся со стилем:

- исключаем предложения с числительными (чтобы было меньше точных фактов),
- оставляем только предложения с символами русского алфавита (чтобы избежать попадания в датасет переводов, транскрипций и т.п.),
- задаем начальный список статей для обхода, отталкиваясь от тем, которые часто обыгрываются в анекдотах.


В новом датасете также выровняли длины шуток и предложений по количеству слов.

Результаты на датасете2

Взяли RuBERT вместо
MultiBERT

Результаты на второй эпохе
обучения (далее модель
начинает переобучаться, т.к.
увеличивается train loss)

Validation accuracy (epoch): 0.9766



	precision	recall	f1-score	support
0.0	0.97	0.99	0.98	4159
1.0	0.99	0.97	0.98	4087
curacy			0.98	8246
ro avg	0.98	0.98	0.98	8246
ed avg	0.98	0.98	0.98	8246

В результате точность выросла!

Наша команда



Арсений Анисимов

Делал:

- сбор шуток,
- обучение BERT'a



Александра Нужненко

Делала:

- сбор и фильтрация не-шутков,
- презентация

Спасибо за внимание!