BRLAB

Анализ тональности: как выделить главное из тысяч отзывов и мнений

BRLAB

Виталий Горбачев
Руководитель направления R&D

vitaliy.g@brl.ru
@imbarus
vitaly.gorbachev.5

РБ Рейтинг Букмекеров

ODDS.RU

META RATINGS

BRLAB

План выступления

Что такое анализ тональности?

- Какой он бывает
- В чем сложности

Методы анализа

- Экспертные
- Обучение без учителя
- Обучение с учителем
- Инструменты

Наши результаты

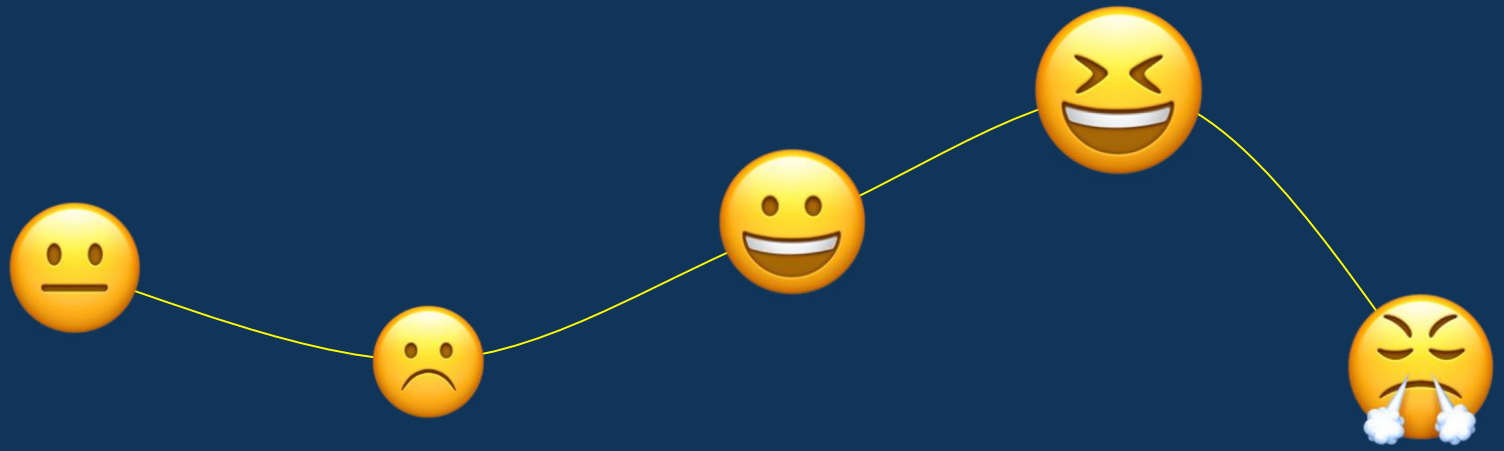
BRLAB

Как проанализировать тысячи отзывов и не свихнуться

BRLAB

BRLAB

Анализ тональности — это определение оценки, которую несёт текст или фраза.



BRLAB

Оценка текста или фраз

Бинарная

- Позитивн/негатив
- Для простой статистики

Многоклассовая

- Позитив/негатив/нейтрально/спам/etc

Числовая

- По любой шкале (0-1,0-100)
- Полезно в выявлении конфликтов

Тональность можно оценивать не по всему отзыву, а по характеристикам (аспектам)

BRLAB

Нормальный букмекер, все нормально и вводом денег и выводом. Тех поддержка ну 3,5, кэфы правда маленькие. А так все хорошо. Меньше года пользуюсь и не было проблем ещё.

Нормальный букмекер, все нормально и вводом денег и выводом. Тех поддержка ну 3,5, кэфы правда маленькие. А так все хорошо. Меньше года пользуюсь и не было проблем ещё.

Плюсы: **вывод и пополнение**

Минусы: **коэффициенты**

Нейтрально: **поддержка**

...все нормально?

… 3,5/5?

BRLAB

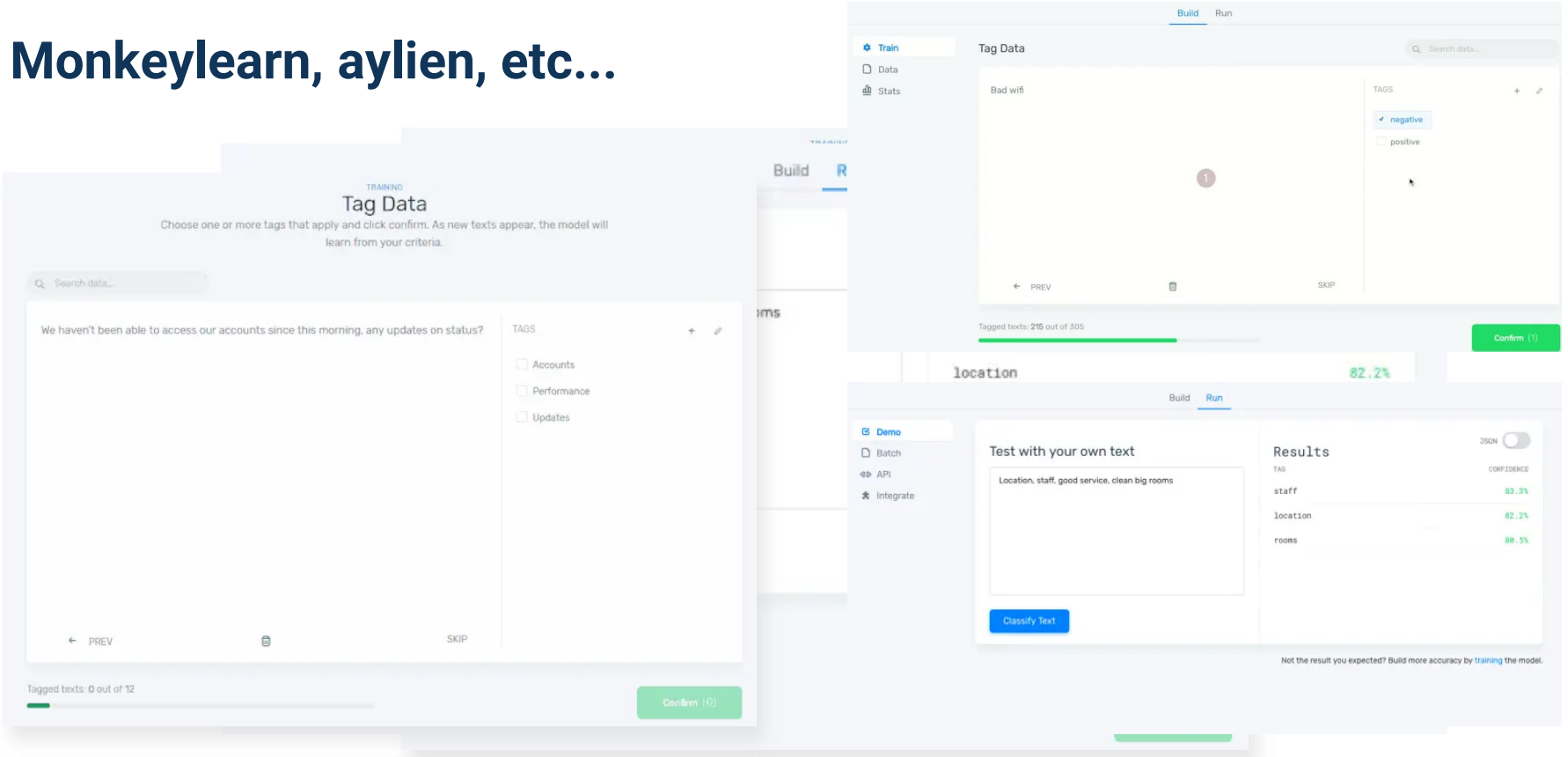
Почему?

Потому что русский язык:

- Требует предобработки из-за сложности
- Очень метафоричен
- Мало готовых инструментов

BRLAB

Monkeylearn, aylien, etc...



BRLAB

Почему?

Потому есть анафора:

“Из недостатоков отмесу то, что с недавних пор часть транясций перестала работать. Это некритично, но доставляет неудобства опередеенные. Надеюсь, это поправят скоро. А так приятная контора”

BRLAB

Почему?

Потому есть анафора:

“Первая интернет контора, где я начал играть. Раньше у них была еще классическая версия сайта, которую уже закрыли месяца 2 назад, но вот новая — просто бомба.”

BRLAB

Почему?

Потому что нет универсальности:

“...идите читать книгу!”:
Отзыв на книгу
Отзыв на фильм по книге

“... очень высокие”
Отзыв на букмекера (коэффициенты)
Отзыв на ресторан (цены)

BRLAB

Почему?

Потому что есть контекст.
“Казалось бы, должна быть идеальная букмекерская контора. Хорошие коэфициенты, широкая линия, качественная поддержка -- как в рекламе. Но на деле всё рушится.”

BRLAB

Почему?

Потому что отзывы пишут люди

“играла команда хокей : ТОРОНТО МАРЛИС -БЕЛЬВИЛЬ ПОСТАВИЛ НА П1 НА ВТОРОМ ТАЙМЕ НАЧИНАЕТСЯ ТРЕТИЙ ТАЙМ ДО КОНЦА ИГРЫ ОСТАЕТСЯ 8 МИНУТ ИГРЫ ПОКАЗЫВАЕТ ЧТО ТОРОНТО ЗАБИВАЕТ Я ЗНАЧИТ АГА МОЛОДЦЫ СДЕЛАЛИ НА ЧТО ШЕЛ РАННО РАДОВАТЬСЯ МАТЧ ВООБЩЕ ПРОПАЛ С ЛАЙВА НУ Я ЗНАЧИТ ОБНОВЛЯЮ ЕГО ОН КАК ПРОПАЛ БОЛЬШЕ НЕ ПОКАЗАЛСЯ СМОТРЮ ИСТОРИЮ ПОКАЗЫВАЕТ ПРОЙГРАЛ МАЙКОРЕ ПОСМОТРЕЛ ИГРУ ТАМ ПОКАЗАЛО ЧТО ВТОРАЯ КОМАНДА ЗАБИВАЕТ .А ОЛИМП НЕ ДАЛА ТАКУЮ ВОЗМОЖНОСТЬ МНЕ .ЕСЛИ БЫ Я ВИДЕЛ ЧТО КОГДА ЕЩЕ ШЛА ИГРА И ЗАБИЛА ВТОРАЯ КОМАНДА Я БЫ СМОГ ПРОДАТЬ СТАВКУ ЧТОБ ИГРАТЬ ДАЛЬШЕ НО ОНИ ЭТОГО НЕ ХОТЕЛИ СДЕЛАТЬ ПОТОМУ ЧТО МНОГО ЛЮДЕЙ СТАВИЛО НА ТОРОНТО ПОТОМУ ЧТО ОНИ ХОРОШО ИГРАЮТ И БЫЛО МНОГО НА НИХ ПОСТАВЛЕНО ЕСЛИ БЫ ЭТОМ МАТЧ ВИСЕЛ В ЛАЙВЕ ЛЮДИ КОТОРЫЕ ВИДЕЛИ ЧТО ИХ СТАВКА ПРОИГРЫВАЕТ ОНИ БЫ СДЕЛАЛИ КАК И Я ЖЕ ПРОДАЛИ ЕЕ ,НО ДЛЯ ОЛИМПА ЭТО НЕ ВЫГОДНО ВОТ ОНИ И УБРАЛИ МАТЧ С ГРУППЫ ИГРЫ .ПИСАЛ ИМ ОНИ МНЕ ГОВОРЯТ ВАША СТАВКА СЫГРАНО ТАК КАК ВЫ ПОСТАВИЛИ НА КОМАНДУ ПЕРВУЮ ВОЗВРАТА НЕ БУДЕТ У НАС БЫЛА ТЕХНИЧЕСКАЯ НЕ ПОЛАДКА НУ КАК ВСЕГДА ОНИ ЭТО ГОВОРЯТ Я ТОЖЕ НЕ СОВЕТУЮ ИГРАТЬ В ЭТОЙ КОНТОРЕ ЕСТЬ ХОРОШИЕ КОНТОРЫ ПО СРАВНЕНИЮ ОЛИМП И ФОНБЕТА”1

BRLAB

Почему?

Потому что отзывы пишут люди

“БК ЛЕОН БК”
“в крыму нельзя - конч енная бк апвап”
“После этого перестал играть в этой конторе. А так все очень даже неплохо у них.”

И что делать?

Свихнуться, но сделать всё красиво
и правильно.

BRLAB

Методы: экспертный

Разметка:

1. **Многоклассовая** - размечаем по классам
2. **По шкале** - размечаем по шкале либо размечаем слова внутри отзыва и высчитываем тональность для отзыва
3. **По характеристикам:**
    a. Все фразы/слова, относящиеся к характеристике + Все фразы/слова, дающие оценку этой характеристике
    b. Lite -- просто ставим тэг характеристика-тональность отзыву

BRLAB

Методы: rule-based

Смотрим на данные, придумываем правила, которые покрывают максимальное количество случаев.
В любом случае нужна экспертная разметка для:

1. **Словарей**
2. **Оценки результата**

“если сказуемое ("нравится") входит в положительный набор глаголов ("нравится", "люблю", "одобряю" ...) и в предложении не имеется отрицаний, то классифицировать тональность как "positive"

BRLAB

Методы: обучение без учителя

Автоматически выделяем из текстов слова/наборы слов по метрикам:

- **TF/IDF**
- **PMI**
- **Logreg**
- **Частоте**

В любом случае нужна экспертная разметка для:

1. **Улучшения словарей**
2. **Оценки результата**

BRLAB

Методы: обучение с учителем

1. **Берем** данные
2. **Предобрабатываем**, если необходимо
3. **Размечаем** часть как нам нужно
4. **Обучаем** классификатор.

BRLAB

Методы: обучение с учителем | предобработка

Прежде чем засовывать текст в классификаторы, нам нужно как-то преобразовать его в цифры.

Как?

1. **TF-IDF**
2. **Вектора (CountVectorizer, FastText)**
3. **К-к-комбо**

BRLAB

Методы: обучение с учителем | предобработка

Но перед этим пайплайн:

1. Убираем **стоп-слова (руками/nltk)**
2. Убираем **мусор (ссылки) (регулярками)**
3. **Лемматизируем** (переводим в начальную форму) **(pymorphy, mystem)**
4. **Ловеркейс (lower)**
5. Делаем **биграммы-триграммы,** если нужно **(руками или через nltk)**
6. **Векторизируем** (превращаем в цифры)
7. **Подаем** в классификатор**. Вы прекрасны!**

BRLAB

Методы: обучение с учителем| классификаторы

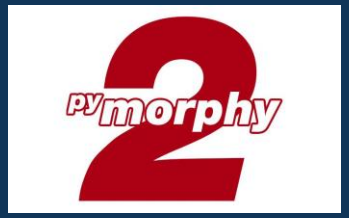
- **Алгоритмы классификации:** SVM, LinearRegression, RandomForest, GradientBoosting, LogisticRegression
- Чем **больше** классов - тем **сложнее** классифицировать
- **Маленькая** выборка **- большие** проблемы
- **Несбалансированные** классы **-** тоже **проблемы**
- Для **бинарной классификации** на общую тональность работают **хорошо,** для **аспектов - плохо**

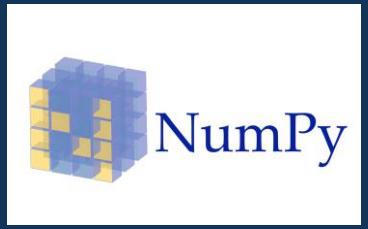
BRLAB

Методы: обучение с учителем| результаты

- Разбиваем на **train и test,** смотрим**:**
  a. **Accuracy**
  b. **Precision**
  c. **Recall**
  d. **F1**
- Сравниваем с **бенчмарками.**

Инструменты, которые вам понадобятся:





Инструменты, которые вам понадобятся:

Словари тональности:

- http://linis-crowd.org/about/
- **Четверкина-Лукашевич (по запросу)**
- http://web-corpora.net/wsgi/senti_game.wsgi/about

Инструменты, которые вам понадобятся:

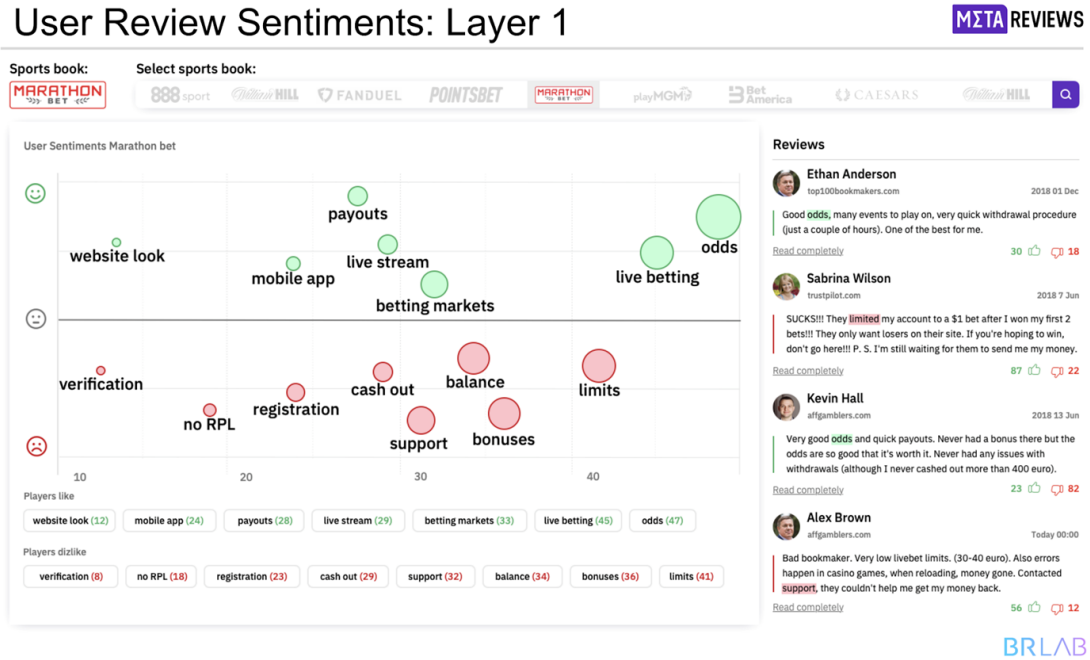
Открытые корпусы для тренировок:

- https://github.com/oldaandozerskaya/auto_reviews
- https://study.mokoron.com/
- http://alt.qcri.org/semeval2016/task5/data/uploads/trial-data/restaurants_trial_russian.xml

Или достаем сами с помощью парсеров

BRLAB

Для чего это нам



Что получилось у нас:

Бинарная классификация для общей тональности:

```
tf-idf + LogisticRegression = precision/recall/f-measure ~ 0.92-0.93
```

Что получилось у нас:

Классификация по аспектам
(15 аспектов, 2 тональности: позитив, негатив):

```
deep memory network with attention model: test-acc=0.70-0.71
```

СПАСИБО!

… и не забудьте оставить отзыв!

BRLAB

BRLAB

Виталий Горбачев
Руководитель направления R&D

vitaliy.g@brl.ru
@imbarus
vitaly.gorbachev.5

РБ Рейтинг Букмекеров

ODDS.RU

META RATINGS