Московский авиационный институт (национальный исследовательский университет)
Институт № 8 «Компьютерные науки и прикладная математика»
Кафедра № 804 «Теория вероятностей и компьютерное моделирование»

Распознавание комментариев ботов и троллей в социальной сети Reddit.

Выпускная квалификационная работа бакалавра

Студент группы М8О-401Б-19: Ермакова Анна Николаевна
Научный руководитель: к.ф.-м.н. доцент каф. 804 М. В. Лебедев

Москва — 2023

Общая постановка задачи

X - множество описаний объектов. $Y = \{\text{бот}, \text{тролль}, \text{пользователь}\}$ - множество классов.
Существует неизвестная целевая зависимость $y^* : X \to Y$, значения которой известны только на объектах обучающей выборки $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$

Требуется построить алгоритм $a : X \to Y$, способный классифицировать $x \in X$.
Метрика качества

$$L(a, X) \to \max_a$$

.

Обработка естественного языка

Предобработка:

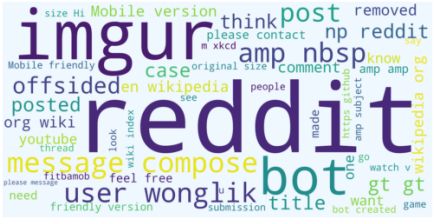
- Перевод текста в нижний регистр
- Удаление специальных символов и стоп-слов
- Лемматизация

Признаки:

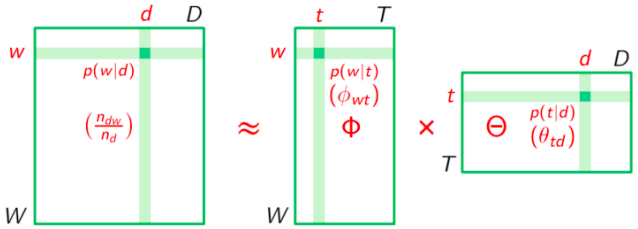
- Оценка тональности:
  Полярность / Субъективность
- Векторизация: для элемента i в документе j:

$$w_{i,j} = tf_{i,j} \times log\left(\frac{N}{1 + df_i}\right),$$

где $tf_{i,j}$ - количество появлений элемента i в j, df_i - количество документов содержащих i, N - общее количество документов.



Латентное Размещение Дирихле



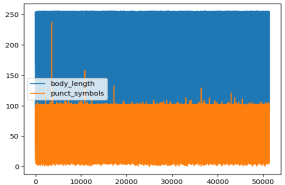
$w \in W$ - терм, $t \in T$ - тема, $d \in D$ - документ. n_{dw} - количество терма w в документе d, n_d - количество термов в документе d. P - матрица частот слов w в документах d. Φ - матрица вероятностей термов w в каждой теме t, Θ - матрица вероятностей тем t в документах d.

Максимизация логарифма правдоподобия:

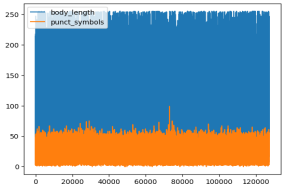
$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w|d)p(d) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \to \max_{\phi,\theta}.$$

Дополнительные числовые признаки

- Длина комментария
- Количество дубликатов комментариаев в тренировочной выборке
- Количество символов "смайл"в комментариях
- Доля знаков пунктуаций в документе. Для признака была проверена гипотеза о равенстве средних двух выборок: выборки с комментариями ботов (математическое ожидание μ_1) и выборки с остальными комментариями (математическое ожидание μ_2).
  $H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$

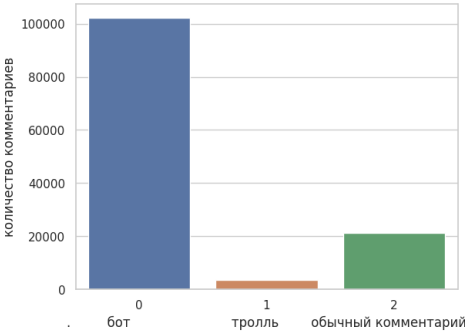


боты

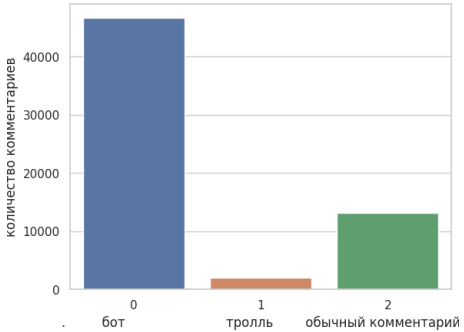


не боты

Данные



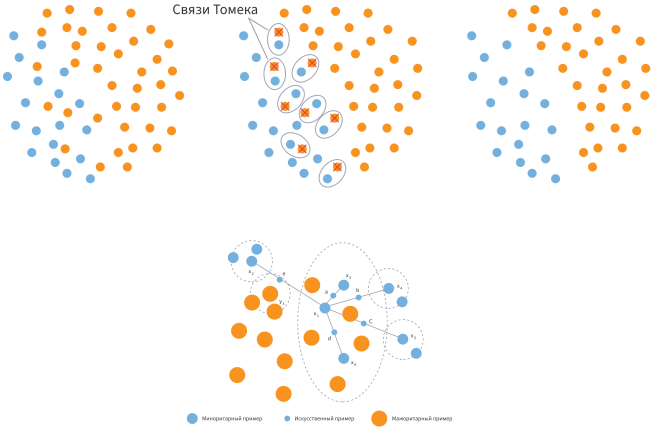
train



test

Балансировка

- Недосемплирование: стратегия метода Tomek links – удалить объекты большого класса, образующие связи Томека.
- Пересемплирование: метод SMOTE для точки малого класса выбирает один из k ближайших соседей и на отрезке между ними случайно генерирует новый объект.



Случайный лес

Случайный лес - ансамбль алгоритмов, в котором базовый алгоритм - это решающее дерево. Решающее дерево представляет собой бинарное дерево.
Критерий ветвления в вершине - Критерий Джини:

$$H(X_m) = \sum_{k=1}^{K} p_k(1 - p_k)$$

Доля объектов класса k в текущей вершине X_m:

$$p_k = \frac{1}{|X_m|} \sum_i \mathbb{I}[y_i = k]$$

- Для построения дерева выбирается случайная подвыборка с повторением.
- Во время обучения для каждой вершины каждого дерева случайно выбирается n признаков и для них ищется оптимальное разделение.
- Предсказание всего ансамбля - самый частый класс.

Многоклассовая логистическая регрессия

$X_{N\times M} \to Y_N$, $X_{N\times M}$ - матрица объект-признак, Y_N - вектор с метками классов.
Прямой проход от строки x матрицы X к элементу вектора Y будет состоять из трех этапов:

$$z = -x \times W,\ p = Softmax(z),\ y = Argmax(p)$$

$W_{M\times C}$ - матрица весов, z - вектор-строка оценок для одного объекта,p - вектор-строка вероятностей принадлежности к классу, y - номер класса.
Проход в обратную сторону от известного класса и объекта:

$$p(y|x, W) = p_{k=y} = Softmax(z_{k=y}) = \frac{\exp(z_{k=y})}{\sum_{k=0}^{C}\exp(z_k)} = \frac{\exp(-xW_{k=y})}{\sum_{k=0}^{C}\exp(-xW_k)}.$$

Далее применяется метод макимального правдоподобия:

$$L(W) = -\frac{1}{N}\log p(Y|X, W) = \frac{1}{N}\sum_{i=1}^{N}\left(X_i W_{k=Y_i} + \log\sum_{k=0}^{C}\exp(-X_i W_k)\right) \to \min_W.$$

Оценка предсказаний

F1-мера:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Микро усреднение:

$$Precision = \overline{TP}/(\overline{TP} + \overline{FP}), \quad Recall = \overline{TP}/(\overline{TP} + \overline{FN})$$

Макро усреднение:

$$F1_{macro} = \overline{F1}, \quad F1_{weighted} = \sum_{i=1}^{3} F1_i \times \frac{N_i}{N}$$

$$Accuracy = \overline{Recall}$$

Площадь под roc-кривой:

$$AUC = \int TPR\, d(FPR), \quad TPR = TP/(TP + FN), \quad FPR = FP/(FP + TN)$$

Результаты

Оценка предсказаний моделей с разным набором признаков:

	f1_score_macro	f1_score_micro	f1_score_weighted	balanced_accuracy_score	roc_auc_score ovr	roc_auc_score ovo
ups,score,body_length	0.4324	0.6345	0.7	0.5846	0.741	0.754
ups,score,body_length,count	0.4429	0.6345	0.6987	0.6068	0.7838	0.7873
+TF-IDF+pol_and_sub	0.509	0.7135	0.7629	0.6574	0.8654	0.8577
+LDA	0.547	0.7514	0.7905	0.6674	0.8852	0.876
GS_RF+new_features	0.5504	0.8671	0.8551	0.5653	0.9523	0.9201
GS_LogReg+new_features	0.5733	0.7753	0.8073	0.6842	0.9001	0.8888
LogReg+RandomOverSampler	0.5781	0.7837	0.8063	0.6399	0.8988	0.8849
LogReg+SMOTE	0.5836	0.7872	0.8123	0.6592	0.9051	0.8878
LogReg+ADASYN	0.5858	0.79	0.8122	0.6497	0.9072	0.8894
LogReg+RandomUnderSampler	0.5555	0.7681	0.798	0.652	0.8913	0.8763
LogReg+NearMiss	0.2705	0.2607	0.2812	0.338	0.5017	0.4808
LogReg+NearMiss	0.4783	0.636	0.6886	0.5913	0.8572	0.8376
LogReg+TomekLinks	0.5673	0.7705	0.8043	0.6847	0.8976	0.8869
LogReg+EditedNearestNeighbours	0.5109	0.7326	0.7735	0.682	0.8897	0.8817

Оценка предсказаний тривиального классификатора:

	f1_score_macro	f1_score_micro	f1_score_weighted	balanced_accuracy_score	roc_auc_score ovr	roc_auc_score ovo
0	0.2869	0.7553	0.65	0.3333	0.5	0.5

Результаты

	f1_score_macro	f1_score_micro	f1_score_weighted	balanced_accuracy_score	roc_auc_score ovr	roc_auc_score ovo
ups,score,body_length	0.4324	0.6345	0.7	0.5846	0.741	0.754
ups,score,body_length,count	0.4429	0.6345	0.6987	0.6068	0.7838	0.7873
+TF-IDF+pol_and_sub	0.509	0.7135	0.7629	0.6574	0.8654	0.8577
+LDA	0.547	0.7514	0.7905	0.6674	0.8852	0.876
GS_RF+new_features	0.5504	0.8671	0.8551	0.5653	0.9523	0.9201
GS_LogReg+new_features	0.5733	0.7753	0.8073	0.6842	0.9001	0.8888
Log_reg+SMOTE	0.5836	0.7872	0.8123	0.6592	0.9051	0.8878
Log_reg+TomekLinks	0.5673	0.7705	0.8043	0.6847	0.8976	0.8869

Вывод

- Balanced accuracy и f1 macro в данной задаче оказались самыми релевантными метриками качества в условиях решения задачи с несбалансированными данными.
- Все добавленные признаки улучшили оценки предсказаний.
- Логистическая регрессия показала лучшие результаты в сравнении со Случайным лесом.
- Алгоритмы SMOTE и Tomek links были выбраны как лучшие алгоритмы перебалансировки.
- Лучший показатель метрики f1 macro - 0.5863, Balanced accuracy - 0.6847.