Модель отрицательно биномиального распределения в анализе категориальных последовательностей

Самарин Игорь Александрович, группа 19.Б04-мм

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель: к.ф.-м.н., доцент Алексеева Н.П.
Рецензент: биостатистик Комарова Е.С.

Санкт-Петербург
2023 г.

Введение

Предположение:

Эмоционально окрашенная лексика подчиняется отрицательному биномиальному распределению.

$$X \sim \mathrm{NB}(r, p), \quad \mathbb{P}(X = k) = \frac{\Gamma(r+k)}{k!\Gamma(r)} p^r (1-p)^k.$$

Значение параметров в лингвистике [Alexeyeva et al., 2013]:

- r — количество пропусков, неупотребления слова.
- p — вероятность неупотребления слова.

Введение. Постановка задачи

Постановка задачи:

Сравнить параметры распределений эмоционально окрашенной лексики в текстах различных тональностей.

Решение задачи:

- Тональная классификация текстов с использованием скрытой марковской модели.
- Оценка максимума правдоподобия параметров распределения и проверка гипотезы соответствия по критерию хи-квадрат.

Тональная классификация. Первичная обработка

Этапы первичной обработки:

1. *Токенизация:* разбиение сплошного текста на отдельные слова.
2. *Нормализация слов:* приведение слов к канонической форме.
3. *Стоп-слова:* удаление общих и редко встречающихся слов.
4. *Нормализация регистра:* приведение к нижнему регистру.
5. *Пунктуация:* удаление знаков пунктуации.

Тональная классификация. Латентно-семантический анализ [Landauer et al., 1998]

Рассмотрим терм-предложение матрицу $X_{m \times n}$, где m — число слов, n — число предложений.

Теорема [Eckart C., Young G., 1936]

Лучшее приближение X среди матриц ранга d — сингулярное разложение, в котором в Σ оставили d первых диагональных элементов.

Аппроксимируем X произведением трех матриц:

$$\hat{X} = U_{m \times d} \Sigma_{d \times d} V^{\mathrm{T}}_{d \times n}.$$

Выберем в качестве векторного представления слов матрицу U.

Используя метод k-средних, разделим полученные представления по k семантическим кластерам.

Скрытая марковская модель. Модель первого порядка

Определение

Модель определяется кортежем $\lambda = (\mathcal{S}, \mathcal{O}, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$.

- *Набор состояний:* $\mathcal{S} = \{\mathfrak{s}_i\}$, где $\mathfrak{s}_i \in \{1, \ldots, N\}$;
- *Набор наблюдений:* $\mathcal{O} = \{\mathfrak{o}_i\}$, где $\mathfrak{o}_i \in \{1, \ldots, M\}$;
- *Вектор начальных вероятностей:* $\boldsymbol{\pi} = \{\pi_i\}$, где $\pi_i = \mathbb{P}\{\mathfrak{s}_1 = i\}$;
- *Матрица переходов:* $\boldsymbol{A} = \{a_{ij}\}$, где $a_{ij} = \mathbb{P}\{\mathfrak{s}_{t+1} = j \mid \mathfrak{s}_t = i\}$;
- *Матрица эмиссии:* $\boldsymbol{B} = b_j(k)$, где $b_j(k) = \mathbb{P}\{\mathfrak{o}_t = k \mid \mathfrak{s}_t = j\}$.

Классические ограничения

$\sum_{i=1}^{N} \pi_i = 1, \qquad \sum_{j=1}^{N} a_{ij} = 1,\ \forall i \in \mathcal{S}, \qquad \sum_{j=1}^{N} b_j(k) = 1,\ \forall k \in \mathcal{O}.$

Набор состояний есть набор кластеров, набор наблюдений — набор слов.

Скрытая марковская модель. Оценка параметров

Параметры модели могут быть обучены двумя способами.

Обучение с учителем:

Есть обучающая выборка, полученная по методу LSA. Параметрами модели являются относительные частоты.

$$\pi_i = \frac{\mathrm{Count}(\mathfrak{s}_1{=}i)}{L},\ a_{ij} = \frac{\mathrm{Count}(\mathfrak{s}_t{=}i,\mathfrak{s}_{t+1}{=}j)}{\mathrm{Count}(\mathfrak{s}_t{=}i)},\ b_j(k) = \frac{\mathrm{Count}(\mathfrak{s}_t{=}j,\mathfrak{o}_t{=}k)}{\mathrm{Count}(\mathfrak{s}_t{=}j)},$$

где L — длина обучающей выборки.

Обучение без учителя:

Параметры задаются случайным образом. Используется алгоритм Баума-Уэлша [Baum et al., 1970] для поиска ОМП параметров.

Скрытая марковская модель. Модель высокого порядка

Предположение:

Настроение определяется цепочкой произошедших событий. Будем рассматривать модель n-го порядка ($n > 1$).

Обозначение:

$$\mathbb{P}\{\mathfrak{s}_i \mid \mathfrak{s}_{i-n}^{i-1}\} \rightleftharpoons \mathbb{P}\{\mathfrak{s}_i \mid \mathfrak{s}_{i-n}, \ldots, \mathfrak{s}_{i-1}\}.$$

Проблема:

При больших n параметры неустойчивы, будем использовать сглаживание [Chen et al., 1996].

$$\mathbb{P}_{\text{one}}(\mathfrak{s}_i|\mathfrak{s}_{i-n}^{i-1}) = \frac{\text{Count}(\mathfrak{s}_{i-n}^{i}) + \alpha\mathbb{P}_{\text{one}}(\mathfrak{s}_i|\mathfrak{s}_{i-n+1}^{i-1})}{\text{Count}(\mathfrak{s}_{i-n}^{i-1}) + \alpha},$$

где $\alpha = \gamma[n_1(\mathfrak{s}_{i-n}^{i-1}) + \beta]$, $n_1(\mathfrak{s}_{i-n}^{i-1}) = |\{\mathfrak{s}_i : \text{Count}(\mathfrak{s}_{i-n}^{i}) = 1\}|$.

Скрытая марковская модель. Декодирование

Имеем: последовательность слов $\mathfrak{O} = \{\mathfrak{o}_1, \ldots, \mathfrak{o}_h\}$.
Необходимо найти: $\hat{\mathfrak{S}} = \operatorname{argmax}_{\mathfrak{S}} \mathbb{P}\{\mathfrak{S} \,|\, \mathfrak{O}\} \propto \operatorname{argmax}_{\mathfrak{S}} \mathbb{P}\{\mathfrak{O} \,|\, \mathfrak{S}\}\mathbb{P}\{\mathfrak{S}\}$.

Используя наши предположения:

$$\hat{\mathfrak{S}} \approx \operatorname*{argmax}_{\mathfrak{S}} \prod_{t=1}^{h} \mathbb{P}\{\mathfrak{o}_t \mid \mathfrak{s}_t\} \prod_{t=1}^{h+1} \mathbb{P}\{\mathfrak{s}_t \mid \mathfrak{s}_{t-n}^{t-1}\},$$

где $\mathfrak{s}_{h+1} = \text{stop}$, а $\mathfrak{s}_{1-n}^{0} = \{*, \ldots, *\}$ — специальные символы.

Задачу поиска наиболее вероятной последовательности скрытых состояний решает алгоритм Витерби [Viterbi, 1967].

Сентиментальная скрытая марковская модель

Определение

Модель определяется множеством $G = \{g_1, \ldots, g_K\}$, где g_i — СММ высокого порядка, K — число валентности эмоции.

Параметры модели g_i оцениваются по тренировочным данным i-ой валентности.

Тональная оценка последовательности наблюдений $\mathfrak{O}$:

$$\hat{y} = \operatorname{argmax}(\mathbb{P}\{\hat{\mathfrak{S}}^{(1)} \mid \mathfrak{O}, g_1\}, \ldots, \mathbb{P}\{\hat{\mathfrak{S}}^{(K)} \mid \mathfrak{O}, g_K\}),$$

где $\hat{\mathfrak{S}}^{(i)}$ — наиболее вероятная последовательность скрытых состояний, полученная i-ой моделью.

Тональная классификация. Результаты

Классификатор был проверен на данных:

- **Movie Review Polarity Dataset** [Bo et al., 2005]. Набор из 10,662 предложений с полярными метками тональности.
- **Subjectivity Dataset** [Bo et al., 2004]. Набор из 10,000 предложений с метками модальности.

Для оценки точности был использован метод 3-fold Cross Validation.

	Clusters	Polarity Dataset, avg	Subjectivity Dataset, avg
1st order SHMM	50	0.727	0.856
2nd order SHMM	35	0.709	0.854
Ensemble SHMM	$[50, 35]$	0.731	0.865

Наибольшая точность достигается на взвешенной композиции двух моделей.

NB распределение. Вторичная обработка

- Параметры распределения оценены по методу максимального правдоподобия.

$$\hat{p} = \frac{\hat{r}}{\hat{r} + \overline{\mathbf{x}}}, \quad \hat{r} : m \ln(\hat{p}) - m\psi(\hat{r}) + \sum_{i=1}^{m} \psi(\hat{r} + X_i) = 0,$$

  где $\psi(x) = \ln' \Gamma(x)$, m — длина выборки.

- Гипотеза соответствия эмпирического закона с теоретическим по критерию хи-квадрат.

NB распределение. Экспериментальные данные

Large Movie Review Dataset [Maas et al., 2011]. Набор данных из 25,000 рецензий к фильмам. Рецензии были классифицированы на позитивные и негативные.

Рассматривались n-граммы, подчиняющиеся NB, трех видов.

- *Нейтральные:* встречающиеся как в позитивных, так и в негативных рецензиях.
- *Позитивные:* встречающиеся только в позитивных рецензиях.
- *Негативные:* встречающиеся только в негативных рецензиях.

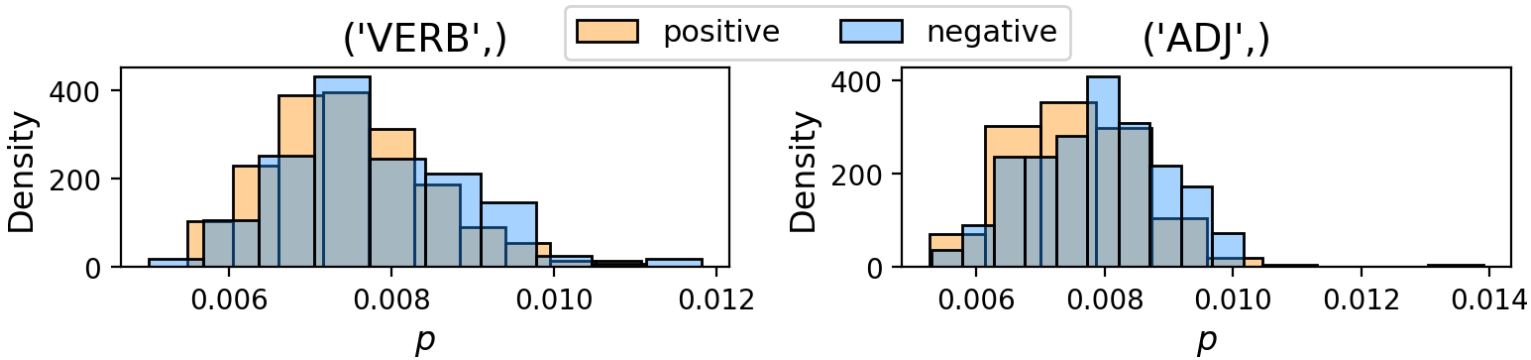
Под полярными n-граммами понимаем совокупность позитивных и негативных.

NB распределение. Частный случай

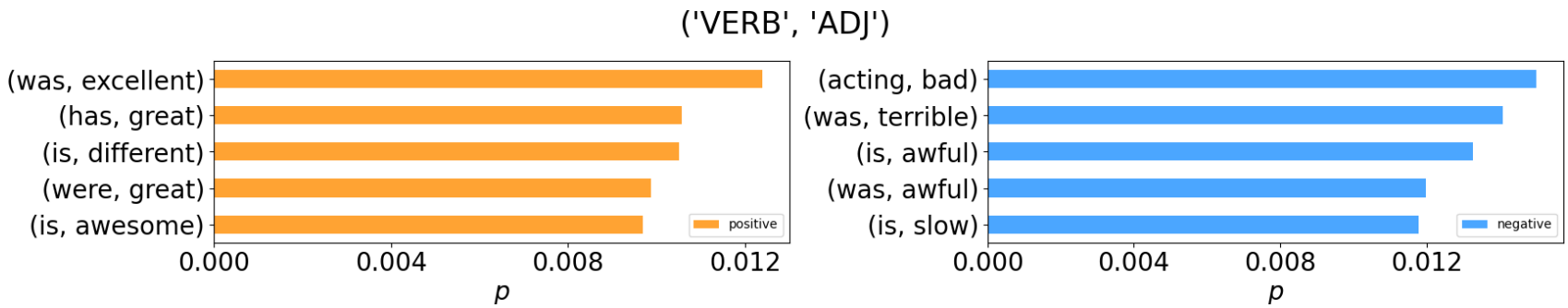
Параметр оценен по выборке $X_1, \ldots, X_m$, где X_i — позиция токена.

Различие распределений значений параметров полярных n-грамм статистически значимо, нейтральных — незначимо.

Наибольшее различие в глаголах и прилагательных (p-value < 0.008).



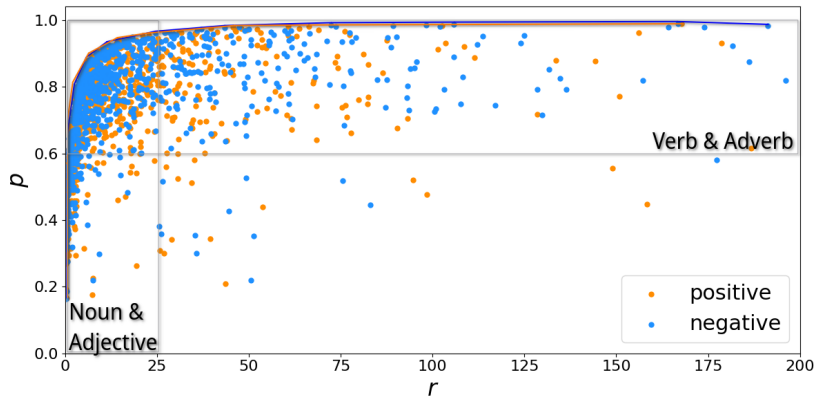
Такие n-граммы лучше всего описывают валентность материала.



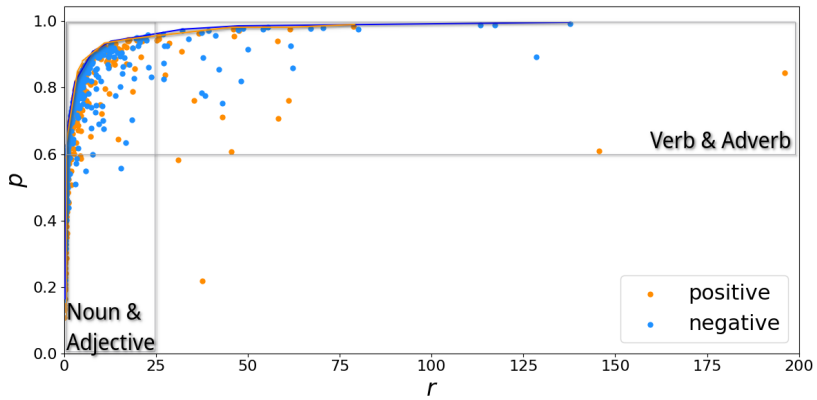
NB распределение. Общий случай

Параметры оценены по выборке $Y_1, \ldots, Y_k$, где Y_i — абсолютная частота встречаемости токена.

Огибающая кривая не зависит от тональности материала.



Нейтральные униграммы

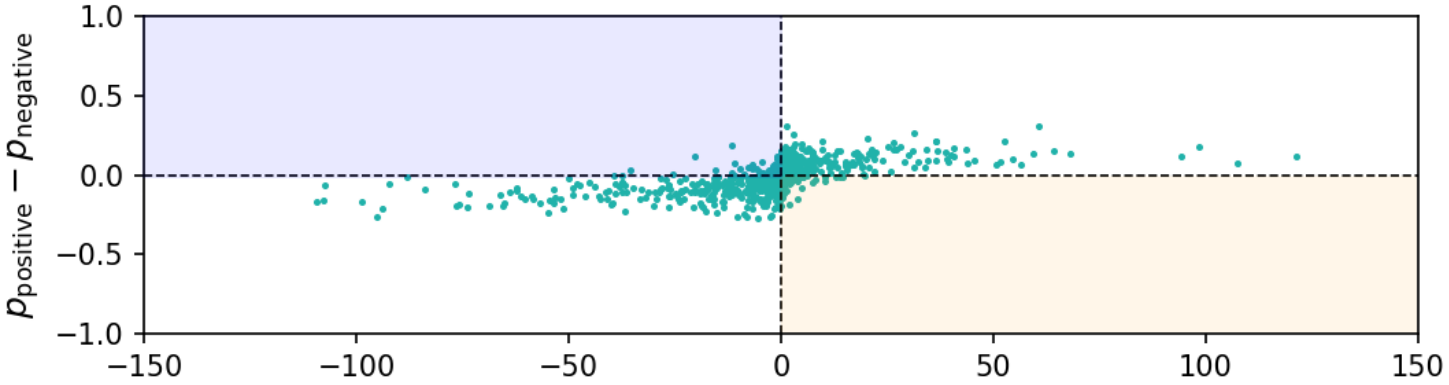


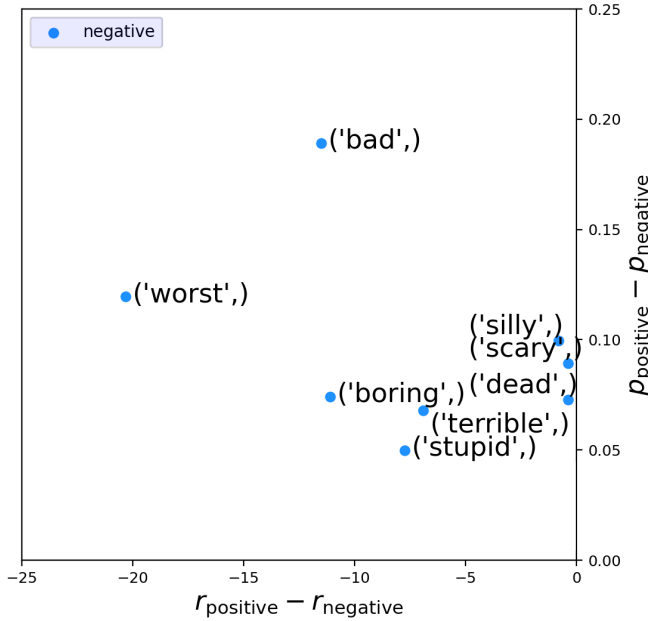
Полярные униграммы

Существительные и прилагательные имеют небольшие значения параметров, глаголы и наречия — большие.

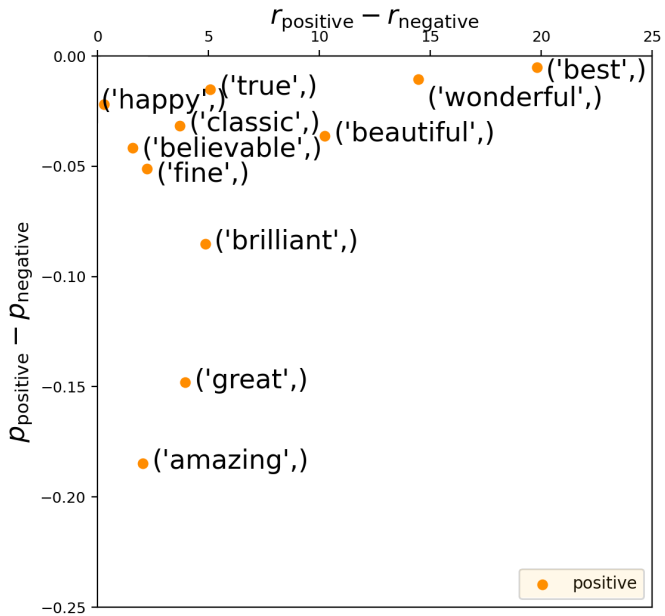
NB распределение. Разность параметров

Были рассмотрены значения разностей параметров.





Негативная лексика



Позитивная лексика

Итоги

- На основе статистической модели был построен алгоритм тональной классификации.
- Был описан метод оценки параметров отрицательного биномиального распределения.
- Была установлена принадлежность эмоционально окрашенной лексики отрицательному биномиальному распределению.
- На примере геометрического распределения, было установлено различие распределений значений параметров в текстах разных тональностей.