Санкт-Петербургский государственный университет

Самарин Игорь Александрович

Выпускная квалификационная работа

Модель отрицательно биномиального распределения в анализе категориальных последовательностей

Уровень образования: бакалавриат

Направление 01.03.02 «Прикладная математика и информатика»

Основная образовательная программа СВ.5004.2019 «Прикладная математика и информатика»

Научный руководитель:

Доцент, кафедра статистического моделирования

к. ф.-м. н., доцент Н. П. Алексеева

Рецензент:

Биостатистик, ООО «Смуз Драг Девелопмент» Е. С. Комарова

Санкт-Петербург

2023

Saint Petersburg State University
Applied Mathematics and Computer Science

Samarin Igor Aleksandrovich

Graduation Project

Negative Binomial Distribution Model in Categorical Sequence Analysis

Scientific Supervisor:
Associate Professor, Department of
Statistical Modelling N. P. Alekseeva

Reviewer:
Biostatistician, LLC "Smooth Drug
Development" E. S. Komarova

Saint Petersburg
2023

Оглавление

Введение

Анализ и установление взаимосвязей между категориальными переменными, описывающими объект или явление, является одной из важнейших задач современной математической статистики. Основная цель анализа заключается в группировке значений по взаимоисключающим категориям. Наиболее интересен случай, когда значения внутри категорий подчиняются известному закону распределения. Тогда решение множества прикладных задач сводится к решению задачи проверки гипотезы принадлежности случайной величины к соответствующему распределению.

В данной работе рассматриваются распределения слов в текстах различных тональностей. Предполагается, что эмоционально окрашенная лексика будет подчиняться известному закону распределения, в частности отрицательному биномиальному. Решение поставленной задачи может быть разделено на две части. Во-первых, необходимо уметь классифицировать тексты. Во-вторых, необходимо уметь оценивать параметры распределения и проверять гипотезу согласия эмпирического закона распределения с теоретическим.

Сделаем краткий обзор содержания данной работы. В первой главе рассмотрен алгоритм тональной классификации. В Разделах 1.2 и 1.3 содержится описание компонент классификатора. В Разделе 1.4 описана идея и модель классификатора, а в Разделе 1.6 приведена его точность в задачах тональной и модальной классификации. Вторая глава посвящена поиску эмоционально окрашенной лексики в позитивных и негативных текстах. Так, в Разделе 2.1 описан метод максимального правдоподобия, применимый к задаче оценивания параметров отрицательного биномиального распределения, и критерий проверки согласия распределений хи-квадрат. Раздел 2.2 содержит результаты и выводы по распределениям рассматриваемой лексики.

Глава 1

Тональная классификация

Анализ тональности (Sentiment Analysis, SA) — класс методов обработки естественного языка, предназначенный для автоматизированного выявления эмоционально окрашенной лексики и эмоциональной оценки мнений авторов по отношению к объектам или событиям, описываемым в тексте. Задача определения тональности может рассматриваться как задача классификации. В частности, текстовые данные могут быть разделены по двум полярным семантическим классам: позитивные и негативные элементы.

В зависимости от контекста использования и ожидаемого результата, существуют различные подходы решения задачи тональной классификации. Наиболее распространены методы, использующие алгоритмы машинного обучения (Machine Learning, ML), статистики, подходы обработки естественного языка (Natural Language Processing, NLP), а также методы, основанные на правилах и словарях (Bag-of-Words, BoW).

Семейство алгоритмов машинного обучения, называемое алгоритмами Глубокого обучения (Deep Learning, DL), представляет наибольший интерес у исследователей и компаний за концептуальную идею имитации работы человеческого мозга. На больших объемах данных такие методы показывают преимущественно хорошие результаты в сравнении с традиционными подходами. Однако использование многослойных (глубоких) нейронных сетей делает систему плохо интерпретируемой, представляющей собой черный ящик, внутреннее устройство которого сложно или вовсе неизвестно.

Статистические же методы напротив имеют понятную, последовательную структуру, сохраняя при этом приемлимую точность результатов. Так, например, скрытая марковская модель (Hidden Markov Model, HMM) использует последовательный анализ категориальных переменных, что позволяет видеть как изменяется настроение по ходу исследуемого материала. Учитывая особенности задачи тональной классификации, использование HMM является логичным и эффективным способом повышения производительности и интерпретируемости результатов. Однако стандартная модель не использует семантику слова, поэтому предлагается строить модель на латентно-семантических кластерах (см. раздел 1.2).

1.1. Предобработка данных

Первичная обработка является важным этапом интеллектуального анализа данных. Полезная информация, полученная на этом этапе, напрямую влияет на способность модели к обучению и, в конечном итоге, на её точность. Далее представлены методы предобработки в контексте рассматриваемой задачи анализа тональности.

1. *Нормализация регистра.* Приведение материала к нижнему регистру.
2. *Токенизация.* Разбиение материала на отдельные слова и предложения.
3. *Удаление стоп-слов.* Удаление общих и редко употребляемых слов.
4. *Нормализация слов.* Приведение слов к начальной форме.

Нормализация регистра. Как правило, тексты написаны в смешанном регистре, т.е. нам могут встречаться как прописные, так и строчные буквы. Поскольку регистр не изменяет валентность эмоции, а лишь указывает на её интенсивность, приведение текста к одному регистру позволит избежать ситуаций с его чередованием.

Токенизация. Задача токенизации, иначе сегментации текста, заключается в разделении линейной последовательности символов на отдельные лингвистические токены. Под токеном будем понимать произвольную n-грамму ($n \geq 1$), т.е. последовательность из n последовательных слов.

Удаление стоп-слов. Стоп-слова, иначе шумовые слова, являются тонально нейтральными словами, не несущими какого-либо дополнительного смысла. Будем рассматривать три группы стоп-слов: общие (предлоги, суффиксы, частицы и т.п.), редко встречающиеся и часто встречающиеся слова. Под редко (часто) встречающимися словами будем понимать слова, встречаемость которых строго ниже (выше) заданного уровня. Удаление стоп-слов не только экономит место, но и повышает качество, производительность будущих алгоритмов.

Нормализация слов. Необходимо привести каждое слово к его канонической форме. Это позволит не только исправить грамматические ошибки, но и сократить количество синонимов. Рассматриваются два подхода канонизации: лемминг и стемминг. Первый использует словарь и морфологический анализ для определения начальной формы.

Второй удаляет префиксы, суффиксы и окончания, оставляя лишь корни слов. Лемминг является более мощной операцией, результаты которой оказываются предпочтительней стемминга.

1.2. Латентно-семантический анализ

Латентно-семантический анализ (Latent semantic analysis, LSA) — это статистический метод обработки естественного языка, позволяющий анализировать отношения между набором предложений и содержащихся в них терминами. Метод использует принцип факторного анализа, полагая, что слова, близкие по значению, встречаются в схожих фрагментах текста. Отличительной особенностью подхода является его интерпретируемость и простота реализации. LSA способен обеспечить достойные результаты, в сравнении с другими, более сложными, моделями [1].

В качестве исходных данных, метод использует терм-предложение матрицу X размерности $m \times n$, где m — число слов, а n — количество предложений. В данной работе, элементами матрицы являются значения статистики TF-IDF [2], учитывающие частоты употребления слова внутри отдельного предложения и его участия в других.

На следующем шаге используем сингулярное разложение исходной матрицы X.

Определение 1.2.1. ***Сингулярным разложением*** *(Singular Value Decomposition, SVD) вещественной матрицы $X_{m\times n}$ называется разложение вида:*

$$X = U\Sigma V^{\mathrm{T}},$$

где $U_{m\times m}$ — матрица левых сингулярных векторов, а $V^{\mathrm{T}}_{n\times n}$ — матрица правых сингулярных векторов. Матрица $\Sigma_{m\times n}$ — диагональная. Значения главной диагонали матрицы Σ называются сингулярными числами.

Такое разложение имеет отличительную особенность. Так, если в матрице Σ оставить первые d наибольшие сингулярные значения, а в матрицах U, V соответствующие им сингулярные векторы, то по теореме Эккарта-Янга (теореме 1.2.1) произведение усеченных матриц будет наилучшим образом, с точки зрения нормы Фробениуса, аппроксимировать исходную терм-предложение матрицу X.

Теорема 1.2.1 ([3]) *Наилучшим приближением матрицы X среди матриц ранга d является сингулярное разложение, в котором в Σ были оставлены первые d диагональных элемента (если они упорядочены по невозрастанию):*

$$X \cong \widetilde{X} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^{\mathrm{T}}.$$

Таким образом, выбрав за векторное представление слов строки матрицы $\widetilde{U}$, используя метод k-средних [4] можно разделить слова предложений по k латентно-семантическим кластерам.

1.3. Скрытая марковская модель

Определение 1.3.1. *Последовательность случайных величин $\{x_n\}_{n\geq 1}$ со значениями в I называется* ***марковской цепью****, если $\forall n \geq 1$ и любых $i_1, \ldots, i_{n+1} \in I$, выполняется*

$$\mathbb{P}(x_{n+1} = i_{n+1} \mid x_n = i_n, \ldots, x_1 = i_1) = \mathbb{P}(x_{n+1} = i_{n+1} \mid x_n = i_n).$$

Определение 1.3.2. *Последовательность случайных величин $\{x_n\}_{n\geq 1}$ со значениями в I называется* ***однородной марковской цепью****, если она определяется следующими компонентами*

- *Вектор начальных вероятностей $\boldsymbol{\pi} = \{\pi_i\}_{1\leq i\leq N}$, где $\pi_i = \mathbb{P}\{x_1 = i\}$;*
- *Матрица переходов $\boldsymbol{A} = \{a_{ij}\}_{1\leq i,j\leq N}$, где $a_{ij} = \mathbb{P}\{x_{n+1} = j \mid x_n = i\}$.*

1.3.1. Модель первого порядка

Марковская цепь полезна, когда необходимо вычислить вероятность заданной последовательности состояний. Однако во многих задачах последовательность состояний скрыта. В частности, в задаче тональной классификации, мы наблюдаем лишь последовательность слов в тексте и не имеем представления об их эмоциональной окраске. Скрытая марковская модель позволяет говорить как о наблюдаемых, так и о скрытых состояниях одновременно. Модель определяется кортежем $\lambda = (\mathcal{O}, \mathcal{S}, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ [5].

(a) Набор состояний: $\mathcal{S} = \{\mathfrak{s}_i\}$, где $\mathfrak{s}_i \in \{1, \ldots, N\}$;

(b) Набор наблюдений: $\mathcal{O} = \{\mathfrak{o}_i\}$, где $\mathfrak{o}_i \in \{1, \ldots, M\}$;

(c) Вектор начальных вероятностей: $\boldsymbol{\pi} = \{\pi_i\}_{1\leq i\leq N}$, где $\pi_i = \mathbb{P}\{\mathfrak{s}_1 = i\}$;

(d) Матрица переходов: $\boldsymbol{A} = \{a_{ij}\}_{1 \le i,j \le N}$, где $a_{ij} = \mathbb{P}\{\mathfrak{s}_{t+1} = j \mid \mathfrak{s}_t = i\}$;

(e) Матрица вероятностей эмиссии: $\boldsymbol{B} = \{b_j(k)\}_{1 \le k \le M}^{1 \le j \le N}$, где $b_j(k) = \mathbb{P}\{\mathfrak{o}_t = k \mid \mathfrak{s}_t = j\}$.

Классические ограничения:

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

Рассмотренная модель является языковой моделью биграмм. Это означает, что состояние наблюдаемого элемента последовательности зависит непосредственно от его предшествующего и не зависит от всех предыдущих. Под наблюдениями будем понимать отдельные токены (слова), под состояниями — полученные ранее латентно-семантические кластеры (см. Раздел 1.2).

1.3.2. Оценка параметров модели

Как и любой другой метод машинного обучения, скрытая марковская модель может быть обучена двумя традиционными способами: с учителем и без учителя.

Обучение с учителем

В контексте задачи тональной классификации, данный подход подразумевает использование латентно-семантических кластеров. Предположим, что у нас имеется L предложений, иначе говоря, последовательностей наблюдений. Каждому предложению поставлена в соответствие скрытая последовательность состояний (кластеров). Тогда параметрами модели будут соответствующие относительные частоты.

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

где $\mathfrak{o}_t$ — наблюдение на t-ом месте последовательности, $\mathfrak{s}_t$ — скрытое состояние, соответствующее наблюдению на t-ом месте последовательности.

Обучение без учителя

В данном случае рассматриваются только последовательности наблюдений. Наиболее распространенным алгоритмом оценки неизвестных параметров является алгоритм прямого-обратного хода, известный также как алгоритм Баума-Уэлша [6]. Данный алгоритм находит оценку максимального правдоподобия параметров модели. Иначе говоря,

алгоритм решает задачу (1.1).

$$\mathbb{P}\{\mathfrak{O} \mid \lambda\} = \sum_{\mathfrak{S}} \mathbb{P}\{\mathfrak{S}, \mathfrak{O} \mid \lambda\} = \sum_{\mathfrak{S}} \mathbb{P}\{\mathfrak{O} \mid \mathfrak{S}, \lambda\}\mathbb{P}\{\mathfrak{S} \mid \lambda\} \xrightarrow[\lambda]{} \max. \tag{1.1}$$

Как и любой алгоритм восхождения к вершине, алгоритм Баума-Уэлша способен зайти в тупик и "застрять" в локальном максимуме, поэтому нельзя полагать, что найденное решение будет соответствовать глобальному максимуму оценки правдоподобия.

Для начала работы алгоритма необходимо задать начальные распределения параметров. Как правило, вектор начальных вероятностей и матрица переходов задается случайным образом, либо заполняются исходя из субъективного опыта эксперта. Далее, обозначим $\alpha_t(i) = \mathbb{P}(\mathfrak{o}_1, \ldots, \mathfrak{o}_t, \mathfrak{s}_t = i \mid \lambda)$ как вероятность нахождения в состоянии i в момент времени t с наблюдаемыми $\mathfrak{o}_1, \ldots, \mathfrak{o}_t$ наблюдениями. Тогда значение прямой процедуры $\alpha_t(i)_{1\le i\le N}^{1\le t\le T}$ вычисляется итеративно по формулам (1.2) и (1.3).

$$\alpha_1(i) = \pi_i \cdot b_i(\mathfrak{o}_1), \tag{1.2}$$

$$\alpha_{t+1}(j) = b_j(\mathfrak{o}_{t+1}) \sum_{i=1}^{N} \alpha_t(i) \cdot a_{ij}. \tag{1.3}$$

Обозначим $\beta_t(i) = \mathbb{P}\{\mathfrak{o}_{t+1}, \ldots, \mathfrak{o}_T \mid \mathfrak{s}_t = i, \lambda\}$ как вероятность нахождения в состоянии i в момент времени t с будущими $\mathfrak{o}_{t+1}, \ldots, \mathfrak{o}_T$ наблюдениями. Тогда значение обратной процедуры $\beta_t(i)_{1\le i\le N}^{1\le t\le T}$ вычисляется итеративно по формулам (1.4) и (1.5).

$$\beta_T(i) = 1, \tag{1.4}$$

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) \cdot a_{ij} b_j(\mathfrak{o}_{t+1}). \tag{1.5}$$

Используя полученные значения, находим по формуле (1.6) вероятность нахождения в состоянии i в момент времени t при заданной последовательности наблюдений.

$$\gamma_t(i) = \mathbb{P}\{\mathfrak{s}_t = i \mid \mathfrak{O}, \lambda\} = \frac{\mathbb{P}\{\mathfrak{s}_t = i, \mathfrak{O} \mid \lambda\}}{\mathbb{P}\{\mathfrak{O} \mid \lambda\}} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}. \tag{1.6}$$

По формуле (1.7) вычисляем вероятность перехода в j-ое состояние в момент времени $t+1$ из состояния i в момент времени t при заданной последовательности наблюдений.

$$\xi_t(i,j) = \mathbb{P}\{\mathfrak{s}_t = i, \mathfrak{s}_{t+1} = j \mid \mathfrak{O}, \lambda\} = \frac{\alpha_t(i) a_{ij} b_j(\mathfrak{o}_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(\mathfrak{o}_{t+1}) \beta_{t+1}(j)}. \tag{1.7}$$

Тогда формулы пересчета значений параметров модели будут иметь вид (1.8).

$$\hat{\pi}_i = \gamma_1(i), \quad \hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad \hat{b}_j(k) = \frac{\sum_{t=1}^{T} \mathbb{1}_{[\mathfrak{o}_t = \mathcal{O}_k]} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}. \tag{1.8}$$

1.3.3. Модель высокого порядка

Скрытая марковская модель первого порядка довольно сильно ограничена в своих возможностях. Для некоторых последовательностей мы ожидаем, что значение текущего состояния зависит от нескольких предыдущих. Так, в задаче тональной классификации мы предполагаем, что настроение не изменяется внезапно, а является результатом прошлых событий. Таким образом, имеет смысл рассматривать модели более высокого n-го порядка ($n > 1$). Далее будем использовать сокращение: $\mathbb{P}\{\mathfrak{s}_i \mid \mathfrak{s}_{i-n}, \ldots, \mathfrak{s}_{i-1}\} \rightleftharpoons \mathbb{P}\{\mathfrak{s}_i \mid \mathfrak{s}_{i-n}^{i-1}\}$, где $\mathfrak{s}_i$ — скрытое состояние на i-ом месте последовательности.

Параметры моделей высокого порядка могут быть неустойчивыми, а в некоторых случаях и вовсе обнуляться. Решением проблемы является процесс сглаживания.

Определение 1.3.3. ***Сглаживание*** *— это процесс выравнивания распределения вероятностей.*

Наиболее распространенным и легким в реализации методом сглаживания является λ-сглаживание, где λ — некоторый заданный наперед параметр.

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

где N — число состояний модели. Наиболее же эффективным методом сглаживания является one-count сглаживание. В большей степени, метод учитывает параметры модели меньших порядков, что позволяет не тратить времени на поиск оптимального множителя в формуле.

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

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

1.3.4. Декодирование

Задача декодирования заключается в поиске наиболее вероятной последовательности скрытых состояний. Пусть дана последовательность наблюдений $\mathfrak{O} = \{\mathfrak{o}_1, \ldots, \mathfrak{o}_h\}$, обозначим соответствующую ей наиболее вероятную последовательность скрытый состояний $\hat{\mathfrak{S}} = \{\hat{\mathfrak{s}}_1, \ldots, \hat{\mathfrak{s}}_h\}$.. Тогда формально задача декодирования примет вид:

$$\hat{\mathfrak{S}} = \underset{\mathfrak{S}}{\text{argmax}}\, \mathbb{P}\{\mathfrak{S} \mid \mathfrak{O}\} = \underset{\mathfrak{S}}{\text{argmax}}\, \frac{\mathbb{P}\{\mathfrak{O} \mid \mathfrak{S}\}\mathbb{P}\{\mathfrak{S}\}}{\mathbb{P}\{\mathfrak{O}\}} \propto \underset{\mathfrak{S}}{\text{argmax}}\, \mathbb{P}\{\mathfrak{O} \mid \mathfrak{S}\}\mathbb{P}\{\mathfrak{S}\},$$

где $\mathbb{P}\{\mathfrak{O} \mid \mathfrak{S}\}$ — условная вероятность наблюдаемой последовательности $\mathfrak{O}$ при заданной последовательности состояний $\mathfrak{S}$, а $\mathbb{P}\{\mathfrak{S}\}$ — вероятность последовательности состояний. Используя наши предположения, получаем:

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

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

Для поиска наиболее вероятной последовательности скрытых состояний будем использовать алгоритм Витерби [7]. Идея алгоритма заключается в том, чтобы из всех последовательностей состояний рассматривать только наиболее вероятные. В отличие от метода полного перебора, сложность алгоритма Витерби равна $O(h \times N^{n+1})$, где h — длина последовательности наблюдений, N — количество возможных состояний.

1.4. Сентиментальная скрытая марковская модель

Будем строить сентиментальную скрытую марковскую модель (Sentimental Hidden Markov Model, SHMM) для поиска наиболее вероятной тональной последовательности скрытых состояний. Модель определяется множеством $G = \{g_1, \ldots, g_K\}$, где g_i — скрытая марковская модель n-го порядка, K — число валентности эмоции. Алгоритмы обучения (см. Алгоритм 1) и поиска наиболее вероятной тональной последовательности скрытых состояний (см. Алгоритм 2) представлены ниже.

Algorithm 1: SHMM: Supervised Learning

Input: $\{(x_i, y_i)\}_{i=1}^{L}$, где $x_i \in X$, $y_i \in Y = \{1, \ldots, K\}$.

Definition: X — набор предложений, Y — набор меток тональных классов.

for $k = 1, \ldots, K$ **do**

  $\bar{\mathfrak{O}}^{(k)} \leftarrow \{x_{k_1}, \ldots, x_{k_{n_k}}\}$, где $y_{k_i} = k,\ i = \overline{1, n_k}$.

  if LSA **then**

    $\bar{\mathfrak{S}}^{(k)} \leftarrow \text{LSA}(\bar{\mathfrak{O}}^{(k)})$ — находим латентно-семантические кластеры.

    Обучаем модель g_k на данных $(\bar{\mathfrak{O}}^{(k)}, \bar{\mathfrak{S}}^{(k)})$.

  else

    Обучаем модель g_k на данных $\bar{\mathfrak{O}}^{(k)}$ по алгоритму Баума-Уэлша.

  end

end

return SHMM$[g_1, \ldots, g_k]$.

Algorithm 2: SHMM: Decoding

Input: $\{x_i\}_{i=1}^{L}$, где $x_i \in X$.

Definition: X — набор предложений.

for $i = 1, \ldots, L$ **do**

  for $k = 1, \ldots, K$ **do**

    $\hat{\mathfrak{S}}_i^{(k)} \leftarrow \text{ViterbiDecode}(x_i \mid g_k)$.

  end

  $\hat{y}_i \leftarrow \text{argmax}(\mathbb{P}\{\hat{\mathfrak{S}}_i^{(1)} \mid x_i, g_1\}, \ldots, \mathbb{P}\{\hat{\mathfrak{S}}_i^{(K)} \mid x_i, g_K\})$

end

Будем использовать модель SHMM для решения задачи тональной классификации. Тональным классом предложения будет номер скрытой марковской модели с наиболее вероятной тональной последовательностью скрытых состояний.

1.5. Ансамбль моделей

Вариация числа кластеров по разному описывает предложение. Так, для получения более универсального классификатора, имеет смысл рассматривать не одну модель, а сразу несколько. Для построения композиции, иначе говоря, ансамбля моделей будем использовать процедуру бустинга.

Определение 1.5.1. *Бустинг (англ. Boosting) — процедура построения композиции алгоритмов, каждый следующий алгоритм которой стремится компенсировать недостатки предыдущих.*

Далее будем рассматривать алгоритм адаптивного бустинга [8] для задачи построения бинарного классификатора.

1.5.1. Адаптивный бустинг

Будем рассматривать взвешенную тренировочную выборку из Таблицы 1.1, где $y_i \in \{-1, +1\}$, а $\omega_i \geq 0$ и $\sum_{i=1}^{L} \omega_i = 1$. Хотим построить классификатор вида:

$$\hat{y}_i = \text{argmax}\left(\sum_{k=1}^{K} \alpha_k \cdot \mathbb{P}\{\hat{\mathfrak{S}}_i^{(-,k)} \mid x_i, g_-^{(k)}\}, \sum_{k=1}^{K} \alpha_k \cdot \mathbb{P}\{\hat{\mathfrak{S}}_i^{(+,k)} \mid x_i, g_+^{(k)}\}\right),$$

где α_k — веса классификаторов композиции, а g_-, g_+ — скрытые марковские модели n-го порядка.

Таблица 1.1. Взвешенная выборка

Предложение	Вес	Валентность
x_1	ω_1	y_1
x_2	ω_2	y_2
...	...	...
x_L	ω_L	y_L

Изначально веса выборки инициализируются равномерно. На каждой итерации выбирается случайное, либо заданное наперед, количество латентно-семантических кластеров, на которых обучается k-ая модель SHMM. По тренировочным данным находится взвешенная ошибка k-го классификатора. Чем точнее классификатор, тем больше его вес в композиции. На последнем этапе пересчитываются веса предложений тренировочной выборки. Так, веса неправильно классифицированных предложений увеличиваются, а веса правильно классифицированных уменьшаются. Это позволяет следующему алгоритму композиции сосредоточиться на плохо классифицируемых объектах.

Algorithm 3: Ensemble SHMM

Input: $\{(x_i, \omega_i, y_i)\}_{i=1}^{L}$, где $x_i \in X$, $y_i \in Y = \{-1, +1\}$.

Initialization: $\omega_i = \frac{1}{L}$.

for $k = 1, \ldots, K$ **do**

Обучаем SHMM со случайным числом кластеров.

Находим $\hat{y}_i^{(k)} = \text{argmax}(\mathbb{P}\{\hat{\mathfrak{S}}_i^{(-,k)} \mid x_i, g_-^{(k)}\}, \mathbb{P}\{\hat{\mathfrak{S}}_i^{(+,k)} \mid x_i, g_+^{(k)}\})$.

Вычисляем ошибку классификатора: $\epsilon_k = \sum_{i=1}^{L} \omega_i^{(k)} \mathbb{1}_{\{\hat{y}_i^{(k)} \neq y_i\}}$.

Вычисляем вес классификатора: $\alpha_k = \frac{1}{2} \ln \frac{1-\epsilon_k}{\epsilon_k}$.

Обновляем веса обучающей выборки: $\omega_i^{(k+1)} = \frac{\omega_i^{(k)} \exp(-\alpha_k y_i \hat{y}_i^{(k)})}{\sum_{j=1}^{L} \omega_j^{(k)} \exp(-\alpha_k y_j \hat{y}_j^{(k)})}$.

end

return Ensemble-SHMM$[\alpha_1, \ldots, \alpha_K, g_-^{(1)}, g_+^{(1)}, \ldots, g_-^{(K)}, g_+^{(K)}]$.

1.6. Результаты

Алгоритм классификации был проверен в задачах двух типов: тональная и модальная классификация. Под модальной классификацией будем понимать субъективную и объективную идентификацию. Субъективные предложения описывают мнения и чувства по отношению к объекту, объективные — конкретную информацию. Далее, в контексте каждой задачи, описаны экспериментальные данные.

- **Movie Review Polarity Dataset v1.0 [9].** Набор данных из 10,662 предложений с полярными метками тональности: 5,331 позитивных, 5,331 негативных. Предложения взяты из рецензий на фильмы с веб-ресурса Rotten Tomatoes.

- **Subjectivity Dataset [10].** Набор данных из 10,000 предложений с метками модальности: 5,000 субъективных, 5,000 объективных. Субъективные предложения взяты из рецензий на фильмы с веб-ресурса Rotten Tomatoes, объективные — из сюжетных аннотаций к фильмам с веб-ресурса IMDb.

Для оценки точности модели был использован метод перекрестной проверки (англ. k-fold Cross-Validation). Метод позволяет получить оценку качества модели по всем имеющимся данным. Так, эксперементальные данные были разделены на три части. Две части использовались для обучения модели, оставшиеся для тестирования. Процедура повторялась три раза. Таким образом, каждая из трех частей один раз использовалась для тестирования. Точность результатов классификации представлена в Таблице. 1.2.

Таблица 1.2. Точность классификаторов

	Clusters	Smoothing	Movie Review Polarity Dataset, avg	Movie Review Polarity Dataset, max	Subjectivity Dataset, avg	Subjectivity Dataset, max
1st order SHMM	50	–	0.7157	0.7211	0.8503	0.8578
1st order SHMM	50	one-count	0.7265	0.7307	0.8561	0.8637
2nd order SHMM	35	–	0.6978	0.7036	0.8475	0.8501
2nd order SHMM	35	one-count	0.7091	0.7160	0.8540	0.8575
Ensemble SHMM	$[50, 35]$	–	0.7153	0.7214	0.8593	0.8632
Ensemble SHMM	$[50, 35]$	one-count	0.7314	0.7365	0.8652	0.8758

Глава 2

Распределение эмоционально окрашенной лексики

Рассмотрев алгоритм тональной классификации, можем приступать к выделению эмоционально окрашенной лексики в текстах различных тональностей. В частности, нас будут интересовать слова, либо произвольные n-граммы, подчиняющиеся отрицательному биномиальному распределению.

2.1. Отрицательное биномиальное распределение

Определение 2.1.1 ([11]) ***Отрицательное биномиальное распределение*** — *это распределение дискретной случайной величины, равной числу произошедших неудач с вероятностью успеха p, проводимых до r-го успеха.*

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

Для приложений данного распределения в анализе текстов или других категориальных последовательностей удобнее использовать гамма-пуассоновскую модель [12], которая ранее применялась в паразитологии для регуляции распределения личинок подкожного овода на теле крупного рогатого скота [13]. Нетрудно убедиться, что

$$P\{X = k\} = \int_0^\infty \mathcal{P}(k|\lambda q)\gamma(\lambda p|r)d(\lambda p), \tag{2.1}$$

где p — вероятность гибели, то есть неупотребления слова, r — сколько раз слово было потеряно, $\mathcal{P}(k|\lambda q) = \frac{(\lambda q)^k}{k!}e^{-\lambda p},\ q = 1-p,\ k = 0, 1, \ldots$, пуассоновский закон случайного числа "выживших" слов в тексте, а $\gamma(\lambda p|r) = \frac{(\lambda p)^{r-1}}{\Gamma(r)}e^{-\lambda p}$ плотность распределения интенсивности r-кратной гибели слова. В результате в (2.1) имеем вероятность того, что при наблюдаемых k употреблениях некоторого слова r раз оно было уничтожено и в текст в результате не вошло. Например, имена собственные, которые трудно заменяемы, имеют невысокие параметры p и r. Нас интересует вопрос, влияет ли тональность текста на параметры $\mathrm{NB}(r, p)$.

2.1.1. Оценивание параметров распределения

Будем оценивать параметры распределений токенов по методу максимума правдоподобия. Предварительно текст должен быть приведен к удобной для анализа форме. Так, основные методы первичной обработки были описаны ранее, в Разделе 1.1. Удаление общих, мало и редко встречающихся слов будем считать нецелесообразным. Как правило, такие слова подчиняются рассматриваемому распределению и влияют на параметры распределений своих "соседей".

Метод максимального правдоподобия

Определение 2.1.2 ([14]) ***Метод максимального правдоподобия*** — *метод оценивания неизвестного параметра путем максимизации функции правдоподобия.*

Рассматриваем выборку $X_1, \ldots, X_m$. Если предполагать, что выборка взята из отрицательного биномиального распределения с вектором параметров $\theta = (r, p)$, тогда оценка по методу максимального правдоподобия (ОМП) будет иметь вид:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(X_1, \ldots, X_m \mid \theta) = \underset{\theta}{\operatorname{argmax}}\, l(X_1, \ldots, X_m \mid \theta),$$

где $L(\mathbf{x} \mid \theta)$ — функция правдоподобия, а $l(\mathbf{x} \mid \theta) = \ln L(\mathbf{x} \mid \theta)$.

Запишем логарифм функции правдоподобия для нашей выборки:

$$l(\mathbf{x} \mid \theta) = mr \ln(p) - m \ln(\Gamma(r)) + \sum_{i=1}^{m} \ln(\Gamma(r + X_i)) + X_i \ln(1 - p) - \ln(X_i!).$$

Продифференцируем $l(\mathbf{x} \mid \theta)$ по параметру p и приравняем к нулю:

$$\frac{\partial l(\mathbf{x} \mid \theta)}{\partial p} = \frac{mr}{p} - \sum_{i=1}^{m} \frac{X_i}{1 - p} = 0 \Rightarrow \hat{p} = \frac{\hat{r}}{\hat{r} + \overline{\mathbf{x}}}, \text{ где } \overline{\mathbf{x}} = \frac{1}{m} \sum_{i=1}^{m} X_i$$

Продифференцируем $l(\mathbf{x} \mid \theta)$ по параметру r и приравняем к нулю:

$$\frac{\partial l(\mathbf{x} \mid \theta)}{\partial r} = m \ln(p) - m\psi(r) + \sum_{i=1}^{m} \psi(r + X_i) = 0,$$

где $\psi(x) = \ln' \Gamma(x)$ — производная логарифма гамма-функции. Подставляя полученную оценку $\hat{p}$, получаем уравнение с одной переменной. Его корень может быть найден численными методами, например, методами Ньютона [15].

2.1.2. Проверка гипотезы согласия

Для проверки гипотезы соответствия эмпирического закона распределения с гипотетическим будем использовать критерий согласия хи-квадрат, обобщенный на случай оценки параметров распределения по выборке.

Критерий Хи-квадрат

Критерий хи-квадрат предусматривает группирование выборки. Область определения с.в. разбивается на k непересекающихся интервалов $(X_{(0)}, X_{(1)}], \ldots, (X_{(k-1)}, X_{(k)}]$. Для каждого из интервалов считаются эмпирические частоты m_i и теоретические вероятности p_i. Необходимым условием применения является неравенство $mp_i \geq 5$. Если неравенство нарушается, требуется выбрать другое, более плотное, интервальное разбиение. Статистика критерия имеет вид:

$$\chi^2 = m \sum_{i=1}^{k} \frac{(m_i/m - p_i)}{p_i},$$

где m — количество наблюдений в выборке.

В случае проверки сложной гипотезы, когда значения параметров теоретического распределения оцениваются по выборке, статистика χ^2 подчиняется χ^2_t-распределению с $t = k - h - 1$ степенями свободы, где h — число оцененных по выборке параметров.

2.2. Результаты

Были рассмотрены параметры распределений эмоционально окрашенной лексики в следующих экспериментальных данных.

- **Large Movie Review Dataset [16]**. Набор данных из 25,000 рецензий к фильмам. На каждый фильм допускается не более 30-ти отзывов, что позволяет данным не коррелировать между собой.

Рассматривались n-граммы, подчиняющиеся отрицательному биномиальному распределению, трёх видов.

1. *Нейтральные.* Встречающиеся как в позитивных, так и в негативных рецензиях.
2. *Позитивные.* Встречающиеся только в позитивных рецензиях.

3. *Негативные.* Встречающиеся только в негативных рецензиях.

Под полярными n-граммами понимаем совокупность позитивных и негативных n-грамм. На точечных диаграммах по оси абсцисс отложен параметр отрицательного биномиального распределения $r \in [0, +\infty)$, по оси ординат параметр $p \in [0, 1]$.

2.2.1. Геометрическое распределение

При $r = 1$ отрицательное биномиальное распределение вырождается в геометрическое. Параметр распределения был оценен по выборке $X_1, \ldots, X_m$, где X_i — позиция токена (n-граммы) в тексте, m — число вхождений токена во все тексты. В данном случае параметр p описывает вероятность встретить токен.

Были рассмотрены параметры униграмм. Различие распределений значений параметров нейтральных униграмм статистически не значимо, полярных униграмм — значимо. Наибольшее различие наблюдается в глаголах и прилагательных (p-value < 0.008).

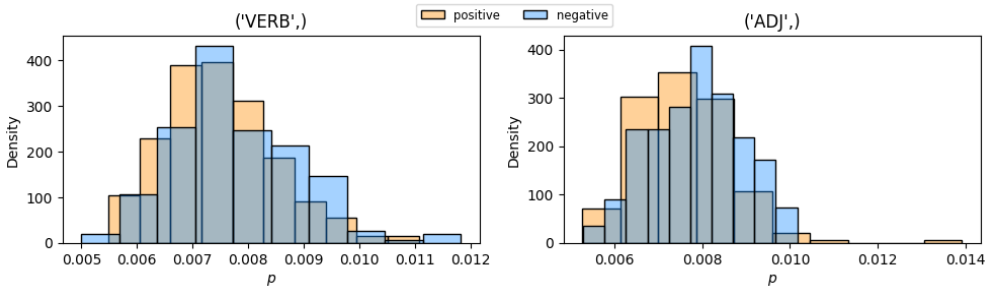

Рис. 2.1. Распределения параметров полярных униграмм

Также были рассмотрены параметры биграмм. Аналогично предыдущему, различие распределений значений параметров нейтральных биграмм не значимо, полярных биграмм — значимо (p-value $< 10^{-4}$). На Рис. 2.2 представлено по пять биграмм с наибольшими значениями параметра p в текстах различных тональностей. Нетрудно заметить, что полярные биграммы наилучшим образом описывают валентность исследуемого материала.

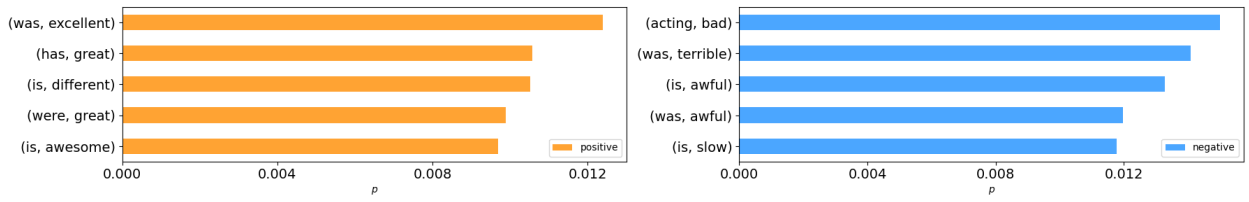

Рис. 2.2. Биграммы (Verb, Adjective)

2.2.2. Отрицательное биномиальное распределение

В общем случае, экспериментальные данные были объединены в k текстов с одинаковым количеством токенов в каждом тексте. Параметры распределения были оценены по выборке $Y_1, \ldots, Y_k$, где Y_i — абсолютная частота встречаемости токена в тексте, k — число текстов. Параметры имеют лингвистическую интерпретацию, описанную в Разделе 2.1.

В первую очередь, были рассмотрены параметры униграмм. Как видно из Рис. 2.3 и Рис. 2.4, огибающая кривая, построенная по точкам $\{(r, p) : f(r) = \max(p \mid r)\}$, не зависит от тональности исследуемого материала.

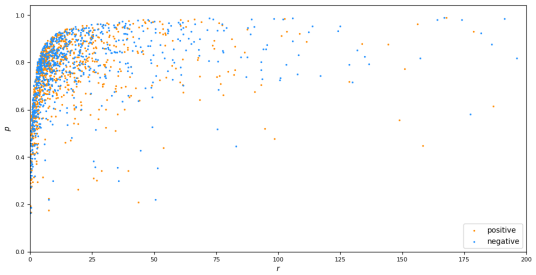

Рис. 2.3. Нейтральные униграммы

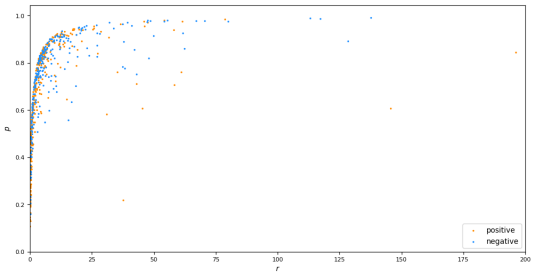

Рис. 2.4. Полярные униграммы

Более подробно были рассмотрены параметры нейтральных униграмм отдельных частей речи. Как видно из Рис. 2.5 - 2.8, существительные и прилагательные преимущественно имеют небольшие значения параметров, их точки группируются в левой части диаграммы. Глаголы и наречия, наоборот, имеют большие значения параметров, их точки сосредоточены в верхней части диаграммы. Статистически значимых различий, между значениями параметров токенов различных частей речи в текстах разных тональностей, за исключением глаголов (p-value < 0.007), не обнаружено.

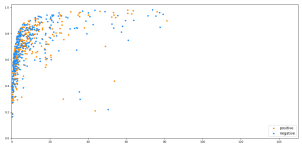

Рис. 2.5. Noun

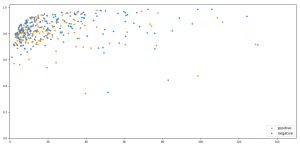

Рис. 2.6. Verb

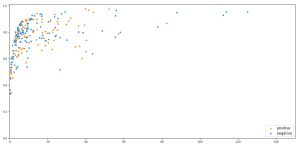

Рис. 2.7. Adjective

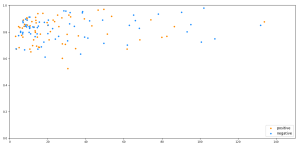

Рис. 2.8. Adverb

Интересная зависимость наблюдается между значениями разностей параметров. Так, из Рис. 2.9 видно, что из величин одного параметра следует отношение величин другого параметра. Иначе говоря, из того, что $r_{\text{pos}} \gg r_{\text{neg}}$ следует $p_{\text{pos}} > p_{\text{neg}}$ и наоборот. Эмоционально окрашенная лексика же группируется в зонах, так называемой,

неопределенности, а именно, в правом нижнем и левом верхнем секторах диаграммы. В левый верхний сектор попадают негативно окрашенные слова (см. Рис. 2.10), в правый нижний — позитивные (см. Рис. 2.11).

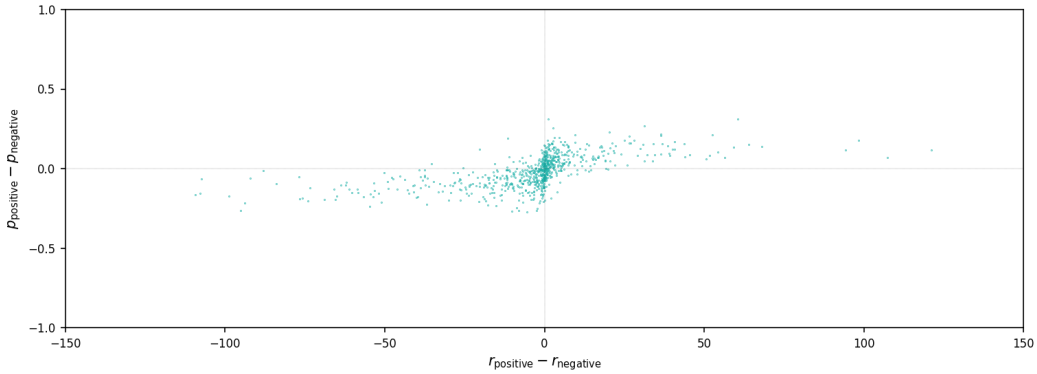

Рис. 2.9. Разность параметров распределения

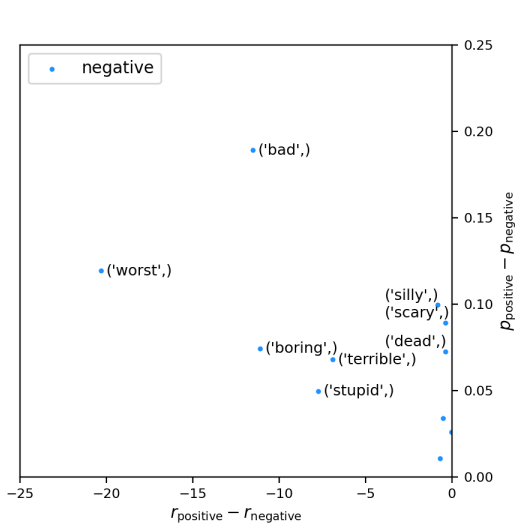

Рис. 2.10. Левый верхний сектор, Adjective

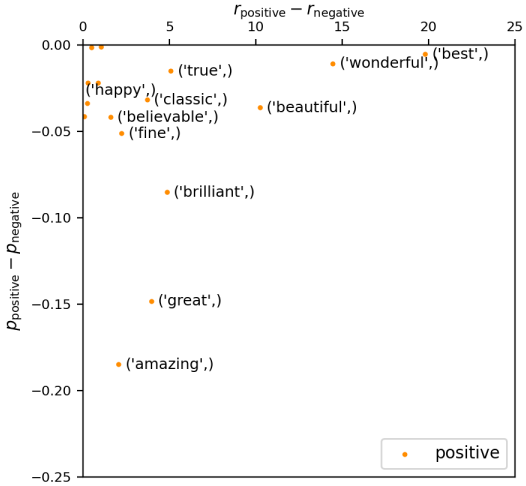

Рис. 2.11. Правый нижний сектор, Adjective

В приложении модели к описанию процесса регуляции в паразитологии, низкие значения числа погибших личинок в сочетании с высокими значениями вероятности гибели означали невысокий уровень инвазии. В нашем случае, негативно окрашенная лексика в позитивных текстах имеет высокую вероятность неупотребления при малом числе неупотреблений, что говорит о невысоком уровне встречаемости. Позитивно окрашенная лексика же, наоборот, имеет малое значение вероятности p, но большее значение параметра r, что говорит о высоком уровне встречаемости.

Заключение

Таким образом, в работе были получены следующие результаты.

- На основе статистической модели был построен алгоритм тональной классификации. Классифицированные тексты могут быть использованы в задаче проверки распределений слов в текстах различных тональностей.
- Был описан метод оценивания параметров отрицательного биномиального распределения, а также критерий проверки гипотезы согласия распределений.
- На примере геометрического распределения, было установлено различие значений параметров распределений полярных слов в позитивных и негативных текстах.
- Установлена принадлежность эмоционально окрашенной лексики отрицательному биномиальному распределению.

Полученные результаты дают представление об использовании ключевых слов в текстах различных тональностей, а значит могут быть использованы в решении задачи тональной классификации. Исходный код может быть найден на Zenodo [17, 18].

Список литературы

1. Landauer T. K., Foltz P. W., Laham D. An introduction to latent semantic analysis // Discourse Processes. — 1998. — Vol. 25, no. 2–3. — P. 259–284. DOI: 10.1080/01638539809545028.
2. Rajaraman A., Ullman J. D. Data Mining // Mining of Massive Datasets. — Cambridge University Press, 2011. — P. 1—-17. DOI: 10.1017/CBO9781139058452.002.
3. Eckart C., Young G. The approximation of one matrix by another of lower rank // Psychometrika. — 1936. — P. 211–218. DOI: 10.1007/BF02288367.
4. MacQueen J. Some methods for classification and analysis of multivariate observations // Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. — 1967. — Vol. 1. — P. 281–297.
5. Lawrence R. R. A Tutorial on Hidden Markov Models and Selected Applications // Proceedings of the IEEE. — 1989.
6. Dugad R., Desai U. A Tutorial On Hidden Markov Models. — 1996.
7. Forney G.D. The viterbi algorithm // Proceedings of the IEEE. — 1973. — Vol. 61, no. 3. — P. 268–278. DOI: 10.1109/PROC.1973.9030.
8. Freund Y., Schapire R. E. A desicion-theoretic generalization of on-line learning and an application to boosting // Computational Learning Theory. — Springer Berlin Heidelberg, 1995. — P. 23–37.
9. Bo P., Lillian L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales // Proceedings of the ACL. — 2005.
10. Bo P., Lillian L. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts // Proceedings of the ACL. — 2004.
11. DeGroot M.H., Schervish M.J. Probability and Statistics. — Pearson Education, 2013. — P. 297–298.
12. Alexeyeva N., Sotov A. The Negative Binomial Model of Word Usage // Electronic Journal of Applied Statistical Analysis. — 2013. — Vol. 6, no. 1.
13. Барт А. Г. Анализ медико-биологических систем. Метод частично-обратных функций. — Издательство Санкт-Петербургского государственного университета, 2003.
14. Le Cam L. Maximum likelihood: An Introduction. — International Statistical Institute, 1990.

15. Süli E., Mayers D. F. An Introduction to Numerical Analysis. — Cambridge University Press, 2003.
16. Maas A. L., Daly R. E., Pham P. T. Learning Word Vectors for Sentiment Analysis // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. — Association for Computational Linguistics, 2011. — June. — P. 142–150. — URL: http://www.aclweb.org/anthology/P11-1015.
17. Samarin I. Sentimental Hidden Markov Model. — 2023. DOI: 10.5281/zenodo.7957936.
18. Samarin I. Sentiment Distribution. — 2023. DOI: 10.5281/zenodo.7958076.