Polarization detection of texts in the news flow

A Preprint

Роман Авдеев
Студент 4 курса 417 гр.
Факультет ВМК
Кафедра ММП
МГУ имени М. В. Ломоносова
roma.avdeyev@gmail.com

Константин Вячеславович Воронцов
Профессор РАН, д.ф.-м.н., проф., зав.каф.ММП
Факультет ВМК
Кафедра ММП
МГУ имени М. В. Ломоносова

Abstract

В данной работе предлагается способ определения поляризации текстов в новостном потоке. Решение основано на методах машинного обучения без учителя, что позволяет работать как с малыми, так и с большими наборами текстов. Решается задача разделения множества новостных сообщений на кластеры-мнения, выделения отдельных кластеров нейтральных и нерелевантных документов. Предложены метрики оценивания качества отсева нерелевантных и нейтральных сообщений. Реализована модель, работающая в среднем не хуже, чем разметчики. Эксперименты проводились на датасете, состоящем из 30 корпусов новостных сообщений по темам «политика» и «происшествия».

Keywords Polarization · opinion mining · sentiment analysis

1 Введение

Задача кластеризации текстов (то есть разбиения множества документов на подмножества тематически близких) остается актуальной на протяжении ряда последних лет. Знания в этой области полезны не только для того, чтобы корректно идентифицировать мнения клиентов в отзывах на какой-либо продукт, но и для более сложных задач. Например, для сбора общественного мнения в рамках анализа ценообразования, конкурентной разведки, прогнозирования рынка, прогнозирования выборов и выявления рисков в банковских системах.

Решаемая мной проблема относится к областям Opinion Mining и Sentiment Analysis. В задачах первой упомянутой области происходит поиск (в блогах, форумах, интернет-магазинах и пр.) мнений пользователей о товарах и других объектах, а также производится анализ этих мнений. Вторая область близка к классической задаче контент-анализа текстов массовой коммуникации, в ней оценивается общая тональность высказываний и текста в целом.// Существуют разные подходы к выявлению поляризации мнений в тексте. Многие из них основываются на работе нейронных сетей [10]. Также существует подход, при котором используется обучение с учителем, что не всегда является оптимальным из-за разного размера текстов и объема обучающей выборки в целом. Работы [1] и [3] базировались на тематическом моделировании. Их авторы используют такие модальности как SPO триплеты (subject-predicate-object), семанические роли (по Филлмору), тональность слов, социально-демографические показатели. Также было показано, что включение семантической близости текстов в модель улучшает качество, а социально-демографические показатели не вносят особого вклада.

В данной работе учтен опыт предыдущих работ, проведены эксперименты со структурой модели, с разными алгоритмами. Предложены метрики оценивания различных аспектов качества модели.

2 Данные

Используемый датасет представляет собой набор из 30-ти корпусов новостных сообщений из рубрик «Политика» и «Происшествия». В каждом корпусе от 8-ми до 33-х документов. Всего 452 документа.
Разметкой будем называть набор меток $X = \{x_1, ..., x_n\}$, каждый элемент которого x_i является

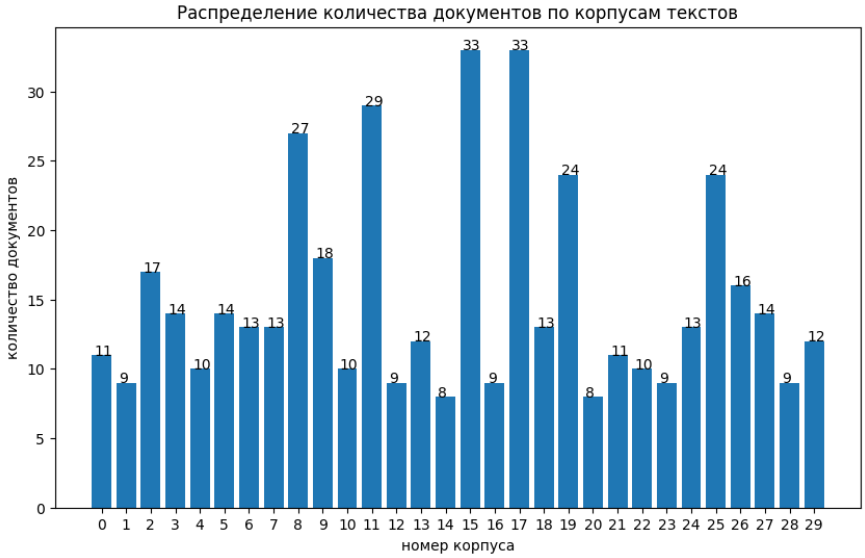

Рис. 1: Количество документов в каждом корпусе

меткой i-го сообщения в корпусе. Предполагается, что каждый корпус содержит сообщения об одном событии или теме, однако может содержать и посторонние сообщения. Метки $\{1, 2, 3, ...\}$ соответствуют кластерам-полюсам общественного мнения по данной теме. Метка <0> означает, что сообщение является нейтральным, то есть не содержит субъективного мнения. Метка <-1> означает, что документ нерелевантен, то есть не относится к общей для всех сообщений теме.

Для разметки данного датасета использовался сервис Яндекс.Толока.

3 Метрики

Наша задача заключается в корректной кластеризации документов внутри каждой темы. Нужно не только верно выделить мнения, но и проверить, существует ли отдельная группа нейтральных/ нерелевантных документов.
Для сравнения разметки $X = \{x_1, ..., x_n\}$ с «золотым стандартом» — экспертной разметкой $Y = \{y_1, ..., y_n\}$, используются BCubed-версии точности и полноты поиска.
Для контроля качества алгоритма введем несколько критериев:

Критерий M1: точность и полнота кластеризации мнений
Точность и полнота сначала определяются относительно каждого объекта x_i, затем усредняются по всем объектам с меткой мнения:

$$P = x_i > 0 avr P_i; \qquad P_i = \frac{\sum_k [x_k = x_i \ and \ y_k = y_i]}{\sum_k [x_k = x_i]}$$

$$R = y_i > 0 avr R_i; \qquad R_i = \frac{\sum_k [x_k = x_i \ and \ y_k = y_i]}{\sum_k [y_k = y_i]}$$

(если знаменатель дроби оказывается равным нулю, то считается, что дробь равна нулю).
Агрегированный критерий (F1-мера) определяется как среднее гармоническое:

$$M_1(X,Y) = \frac{2PR}{P+R}$$

Критерий M2: точность и полнота отсева нейтральных документов
Определим точность и полноту отсносительно метки c (c=0):

$$P_c = \frac{\sum_k [x_k \neq c \ and \ y_k \neq c]}{\sum_k [x_k \neq c]}$$

$$R_c = \frac{\sum_k [x_k \neq c \ and \ y_k \neq c]}{\sum_k [y_k \neq c]}$$

Глядя на формулу, можно сказать, что по сути мы оцениваем, как хорошо модель отделяет субъективные позиции от нейтральной констатации фактов. Агрегированный критерий (F1-мера) определения нейтральных сообщений:

$$M_2(X,Y) = \frac{2P_0R_0}{P_0+R_0}$$

blueКритерий M3: точность и полнота отсева нерелевантных документов
Здесь по аналогии с нейтральными документами мы оцениваем корректность выявления релевантных, отделяем нужные нам мнения от шума. Агрегированный критерий (F1-мера) определения релевантных сообщений (c=-1):

$$M_3(X,Y) = \frac{2P_{-1}R_{-1}}{P_{-1}+R_{-1}}$$

Критерий M4: точность определения числа мнений
Обозначим через K_x и K_y число различных мнений в разметках X и Y соответственно.

$$M_4(X,Y) = \frac{min\{K_x,K_y\}}{max\{K_x,K_y\}}$$

4 Headings: first level

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula. See Section 4.

4.1 Headings: second level

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j | y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})} \tag{1}$$

4.1.1 Headings: third level

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor,

Таблица 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Paragraph Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

5 Examples of citations, figures, tables, references

5.1 Citations

Citations use `natbib`. The documentation may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Here is an example usage of the two main commands (`citet` and `citep`): Some people thought a thing [??] but other people thought something else [?]. Many people have speculated that if we knew exactly why ? thought this...

5.2 Figures

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi. See Figure 2. Here is how you add footnotes. [1] Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

5.3 Tables

See awesome Table 1.

The documentation for `booktabs` ('Publication quality tables in LaTeX') is available from:

`https://www.ctan.org/pkg/booktabs`

5.4 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.

[1] Sample of the first footnote.

Рис. 2: Sample figure caption.

- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

Список литературы