

пусть англ, а дайте текст на русском?

Polarization and neutrality 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

не фидено ниже на текст

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. Эксперименты проводились на датасете, состоящем из 30 корпусов новостных сообщений по темам «политика» и «происшествия».

Keywords Поляризация · sentiment analysis · opinion mining · нейтральность

1 Введение

Все пояснения со скобок убрал.

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.

} ссылок

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), семантические 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.

like

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

подробнее про этот метод.

2 Данные ← *оо уже эксперимент. где постанова задана?*

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 $X = \{x_1, \dots, x_n\}$, каждый элемент которого x_i является

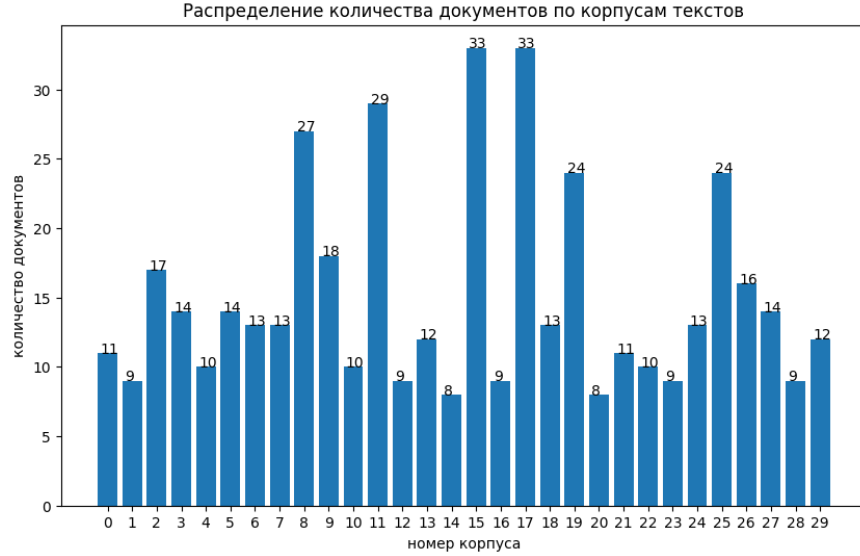


Рис. 1: Количество документов в каждом корпусе

меткой i -го сообщения в корпусе. Предполагается, что каждый корпус содержит сообщения об одном событии или теме, однако может содержать и посторонние сообщения. Метки $\{1, 2, 3, \dots\}$ 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 $X = \{x_1, \dots, x_n\}$ с «золотым стандартом» — экспертной разметкой $Y = \{y_1, \dots, y_n\}$, используются VCubed-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:

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

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

(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 относительно метки c ($c=0$):

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

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

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:

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

Критерий 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 ($c=-1$):

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

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

Обозначим через K_x и K_y число различных мнений в разметках X и Y соответственно.

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

3.1 Эксперименты с метриками

Описанные выше метрики являются результатом нескольких экспериментов. Критерии M1 и M4 стабильно давали понятные и хорошо интерпретируемые результаты как для модели, так и для разметчиков. Проблемы возникали с M2 и M3. Количество найденных нерелевантных документов сильно отличается от корпуса к корпусу, но, как мы видим на рисунке №2, доля согласия у разметчиков довольно высокая. Сложнее ситуация обстоит с нейтральными документами. Во-первых, из рисунка №3 следует, что весь корпус документов может быть нейтральным (в отличие от нерелевантных). Во-вторых, доля согласия между ассессорами существенно ниже. Изначально точность и полнота для M2 и M3 задавались формулами:

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

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

При таком определении метрик возникало сразу несколько сложностей:

1. неясно, как определять P в случае нулевого знаменателя, что случалось часто. Для precision нулевой знаменатель означает, что модель не нашла нерелевантные/нейтральные документы, что автоматически обнуляет и числитель дроби. Получили неопределенность вида $\frac{0}{0}$.
2. аналогичная проблема возникает и с R . Нуль в знаменателе recall говорит о том, что в разметке нет нейтральных/нерелевантных документов. Тогда возникает вопрос, можем ли мы оценивать корпус для таких X и Y . При этом, оставив NaN и отказавшись от оценивания, мы теряем информацию о качестве модели.
3. для M2 получаем очень низкие значения из-за довольно слабого согласия между экспертами.

Результаты метрик при таком задании precision и recall представлены ниже:

	M2	M3
Разметка	0.08	0.49

Оценим качество модели и уровень согласия разметчиков с помощью каппа-статистики Коэна:
 $\kappa = \frac{p_0 - p_e}{1 - p_e}$; $\kappa \in [0, 1]$, где p_0 - относительное наблюдаемое согласие среди оценщиков (ассигасу), p_e - гипотетическая вероятность случайного совпадения.

Значения κ носят следующий смысл ([11]):

каппа	интерпретация
<0	согласие меньше, чем случайная вероятность
0	нет согласия
0.1-0.2	незначительное согласие
0.21-0.4	"удовлетворительное" согласие
0.41-0.6	умеренное согласие
0.61-0.8	существенное согласие
0.81-0.99	почти идеальное согласие
1	полное согласие

Применив данную метрику к нерелевантным документам, получил, что средний уровень согласия разметчиков - 0.35, а качество модели - 0.2. Вызвано это тем, что каппа-статистика сильно штрафует неверные ответы для класса с меньшим количеством объектов. Но от одного корпуса к другому «меньшим классом» может являться как множество нерелевантных/нейтральных, так и множество кластеров-мнений (см. рис.2 и рис.3). Таким образом, штрафы за ошибки разных типов «смешиваются», и мы не получаем реальной оценки качества работы конкретного блока алгоритма.

Поэтому в качестве итоговых метрик были выбраны критерии, описанные в разделе ??.

4 Модель

4.1 Данные

Опишем данные с технической точки зрения. На вход поступил JSON-файл, содержащий корпуса документов. Каждый документ имеет уникальный идентификационный номер. Также в имеющемся датасете присутствует текст каждой новости и выделенные именованные сущности с промаркированной тональностью.

Пример: {'neg': {'ТАСС': 1, 'Приморском крае': 1}, 'pos': {'Орловской области': 1}}

В ходе работы над моделью понадобились лемматизированные предложения, поэтому текст каждого документа был соответствующим образом обработан и добавлен в датасет.

4.2 Эксперименты

Сначала была предпринята попытка решить поставленную задачу с помощью методов обучения с учителем. Использовался RandomForestClassifier, работающий на TF-IDF (term frequency - inverse document frequency) представлении лемматизированного текста. Лучшим результатом являлась F1-мера, равная 0.95 для кластеризации мнений и 0.67 для выделения нейтральных. Однако при уменьшении обучающей выборки и увеличении тестовой данные значения стремительно падали. От такого варианта реализации было решено отказаться, так как в общем случае нам неизвестен размер приходящего датасета; кроме того, в реальной задаче отсутствует разметка, поэтому supervised методы не подойдут.

Далее был сформирован общий вид модели - алгоритм, имеющий трехступенчатую структуру: на 1-ом шаге проводится отсев нерелевантных документов, на 2-ом - нейтральных, а на 3-ем шаге выполняется

деление на кластеры-мнения.

4.2.1 Выделение нерелевантных документов

Выделение нерелевантных документов происходит при помощи подсчета семантической близости лемматизированных предложений в каждом документе. Для этого используется библиотека SpaCy с готовым пайплайном `ru_core_news_sm` для русского языка; данный пайплайн обучен на текстах новостей. Так как семантическая близость считается для пар документов, получаем $C_{docs_amount}^2$ пар. Как будет описано ниже, мы не используем семантическую близость всех пар сообщений. Проведя несколько экспериментов, было решено брать только 8 самых близких к конкретному документу сообщений. А характеристикой документа является среднее значение семантической близости его 8-ми соседей. Подробное описание данных экспериментов представлено ниже.

OPTICS

Сначала было решено применить unsupervised метод OPTICS (Ordering Points To Identify the Clustering Structure). Можно сказать, что OPTICS является обобщением более известного DBSCAN. Есть несколько существенных отличий и особенностей:

1. OPTICS строит дендрограмму по ближайшим соседям точек
2. также внутри OPTICS строится график достижимого расстояния (расстояния до ближайшего соседа) от номера точки, а кластеры определяются как «долины» на графике

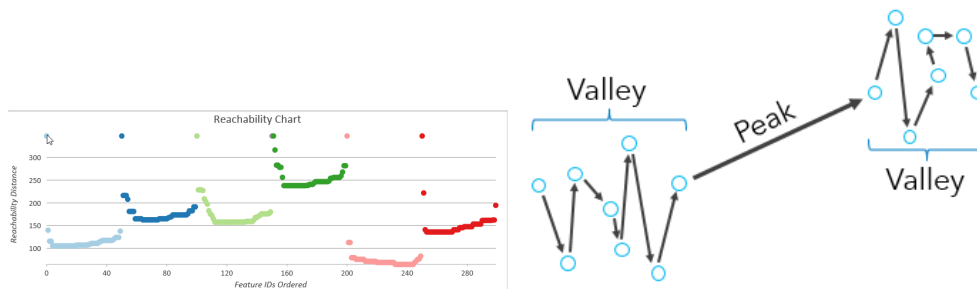


Рис. 2: Принцип работы OPTICS

3. как и DBSCAN, OPTICS выделяет шумовые точки
4. технология OPTICS требует больше памяти, так как хранит большое количество расстояний до ближайших точек для определения расстояния достижимости
5. главным достоинством метода OPTICS является то, что он не требует указания параметра ϵ , что приводит к существенному сокращению процесса настройки параметров

Данный подход дал следующие результаты:

	M3*
Алгоритм	0.39
Разметка	0.49

* - метрика M3 вычислялась как агрегированный критерий поиска Нерелевантных (раздел 3.1)

Далее было решено протестировать методы, ориентированные именно на выявление аномалий.

IsolationForest

Попробовал применить алгоритм Isolation Forest, используя разное число деревьев (`n_estimators`).

Результаты при использовании финальной реализации метрики M3 (раздел ??):

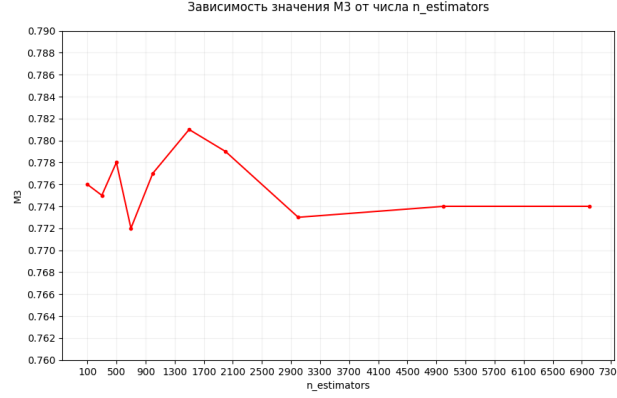


Рис. 3: Зависимость качества выделения нерелевантных от числа деревьев

Максимальное значение $M3=0.781$ было достигнуто при $n_estimators=1500$ (при $M3_{expert} = 0.93$).

Видим, что в сравнении с OPTICS (по «старой» метрике) качество повысилось. Однако существенным недостатком Isolation Forest является то, что он предполагает наличие лишь нескольких аномалий, а, глядя на разметку, мы видим, что в корпусе может присутствовать до 79% нерелевантных документов. Поэтому было принято решение протестировать еще одну модель.

OneClassSVM

Данный алгоритм позволяет осуществить более точную настройку выделения аномалий, также здесь по умолчанию стоит ограничение в 50% на долю выбросов среди всей выборки, но этот параметр можно указать любым.

Сначала проверим три бейзлайна с разными ядрами:

1. 'linear' - линейное ядро
2. 'rbf' - радиальная базисная функция
3. 'poly' - полиномиальное ядро

	linear	rbf	poly
M3 (old)	0.562	0.45	0.58
M3 (new)	0.68	0.632	0.682

Видим, что два из трех бейзлайнов превышают «старую» экспертную метрику $M3$ (0.49), а при повышении точности (tol) разница становится еще больше. Рассмотрев подробнее метки объектов, был замечен очень низкий precision при высоком recall. Такие показатели доказывают, что данная метрика действительно некорректно отражает качество выделения аномалий. Таким образом, подтвердились проблемы, указанные в разделе 3.1, поэтому с текущего момента будем оценивать все модели только по новой реализации критерия $M3$ (раздел ??).

Что касается новой версии критерия $M3$, то лучшее значение достигается при полиномиальном ядре. Постараемся подобрать некоторые другие гиперпараметры для улучшения качества.

Попробуем найти оптимальную степень полинома:

	1	2	3	4	5	6	7	8
M3	0.666	0.666	0.682	0.672	0.671	0.684	0.68	0.669

Заметим, что значения практически не отличаются (различие в пределах шума).

Теперь обратим внимание на другой важный гиперпараметр OneClassSVM. Параметр ν отвечает за верхний порог ожидаемой доли аномалий в датасете. По умолчанию он равен 0.5. Посмотрим на значение метрики M3 при разных ν :

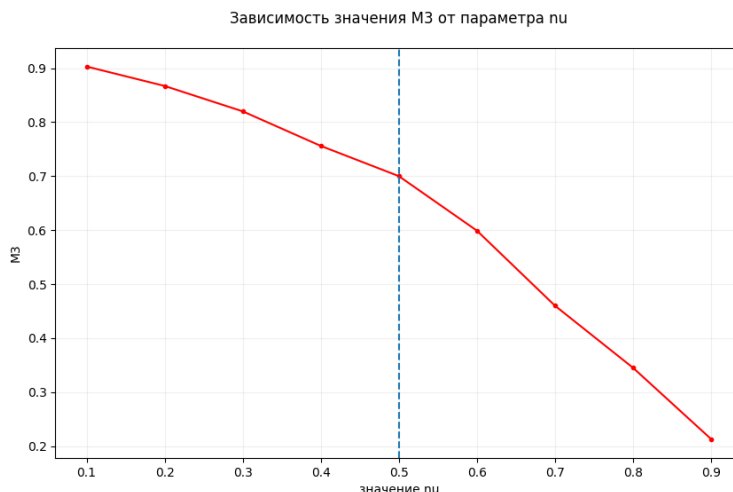


Рис. 4: Зависимость значения M3 от верхнего порога доли аномалий

Качество ожидаемо снижается при увеличении верхнего порога. В разделе ?? было показано, что в данном датасете ассесоры нашли около 10% нерелевантных документов. Поэтому наиболее высокое качество достигается именно при таком пороге. Но, используя данную информацию, мы фактически пользуемся разметкой, поэтому введем модифицированный алгоритм выделения нерелевантных сообщений.

Что касается остальных гиперпараметров OneClassSVM, их подбор сильно зависит от конкретного датасета, и, так как мы решаем поставленную задачу в общем виде, их перебирать не будем.

Модифицированный алгоритм

Сначала коротко опишем подготовку данных в предыдущих экспериментах. Как было сказано в начале текущего раздела, мы не используем значения семантической близости всех пар документов, хотя на начальных этапах был вариант алгоритма, где для каждого документа считалась сумма расстояний до всех остальных сообщений. Такая модель достигла значения $M3=0.685$ при полиномиальном ядре с $\text{degree}=6$ и $\nu=\text{default}$. Такой подход можно существенно улучшить. Теперь будем брать некоторое число «соседей» (наиболее близких сообщений) и усреднять эти значения.

Схематично алгоритм выглядит так:

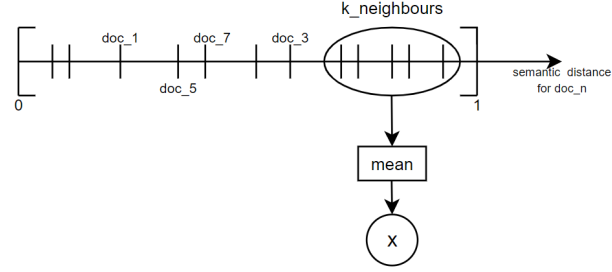


Рис. 5: Обработка семантической близости пар документов

Теперь попробуем разбить данные на две части и анализировать их отдельно. Мы можем сразу выделить 50% документов с наибольшими усредненными значениями семантической близости. Эти сообщения будем считать заведомо релевантными, так как если в корпусе находится более половины шумовых объектов, то, возможно, они и являются «чистыми» данными, а меньшая часть - это аномалии. Далее к 50% документов с меньшими показателями применяем OneClassSVM/Isolation Forest. Здесь использование Isolation Forest уже более корректно, так как поиск нерелевантных происходит в подвыборке, и проблема, описанная в 4.2.1 (IsolationForest), становится не такой серьезной.

Посмотрим на значения МЗ при разном числе "соседей":

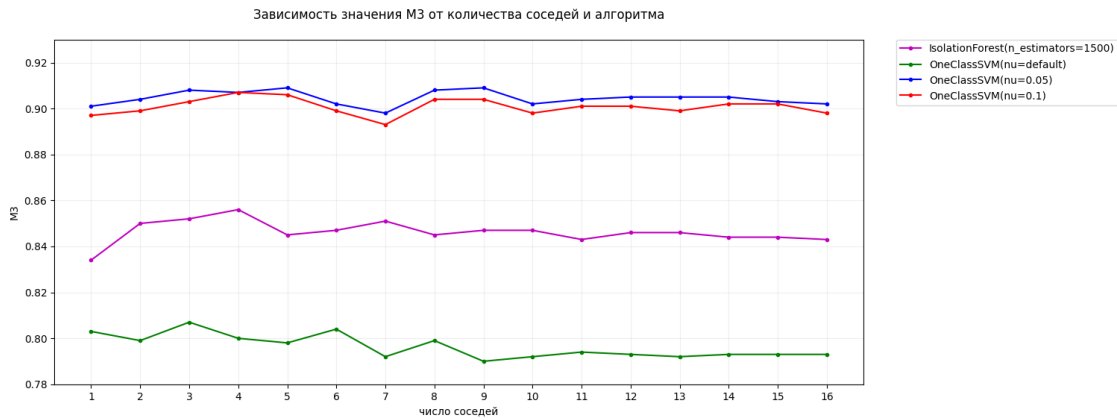


Рис. 6: Зависимость значения МЗ от числа соседей

Из четырех рассмотренных моделей IsolationForest и OneClassSVM(nu=default) являются общими решениями, в то время как другие две вариации OneClassSVM используют информацию из разметки, но при этом демонстрируют более высокое качество.

Теперь рассмотрим случай, когда число соседей зависит от размера корпуса. На приведенном ниже графике n_docs обозначает количество документов в корпусе.

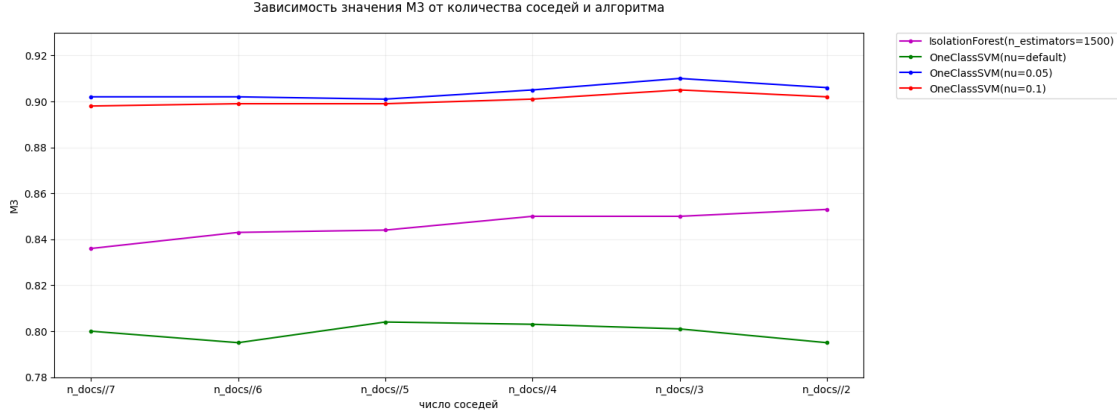


Рис. 7: Зависимость значения МЗ от числа соседей

Заметим, что данный график фактически представляет собой усредненную и сглаженную версию предыдущего. Оптимальным по суммарным результатам четырех графиков является $n_docs//3$. Наверное, для небольших корпусов сообщений такой подход будет работать хорошо, но для многочисленных блоков документов лучше брать фиксированное число соседей.

4.2.2 Выделение нейтральных документов

Нейтральный документ представляет собой текст без какой-либо субъективной позиции и/или эмоциональной окраски. Поэтому в качестве основного инструмента для определения сообщений такого вида было решено взять эмоционально окрашенные именованные сущности, имеющиеся в датасете (раздел 4).

Первый подход: для каждого документа проверялось наличие/отсутствие эмоционально окрашенных слов. Если таковых не было обнаружено, то документ считался нейтральным. При данной стратегии $M2=0.64$.

Второй подход: теперь положительно и негативно окрашенные именованные сущности разделяются. Проверяется гипотеза, что сентименты разной эмоциональной окраски могут друг друга компенсировать. Тогда документ считается нейтральным, если такие слова отсутствуют или если их одинаковое количество на каждой полярности. В таком случае $M2=0.65$.

5 Headings: first level

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur 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 5.

5.1 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 consectetur.

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

5.1.1 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, 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.

6 Examples of citations, figures, tables, references

6.1 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...

6.2 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 consectetur odio sem sed wisi. See Figure 8. 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, consectetur 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.

6.3 Tables

See awesome Table 1.

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

<https://www.ctan.org/pkg/booktabs>

6.4 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

Список литературы

¹Sample of the first footnote.



Рис. 8: Sample figure caption.

Таблица 1: Sample table title

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

Список литературы

- [1] Е.К.Миллота. Языковые модели для обнаружения поляризации общественного мнения в новостном потоке, 2022.
- [2] Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных, 2017.
- [3] Д.Г.Фельдман. Комбинирование фактов, семантических ролей и тональных слов в генеративной модели для поиска мнений, 2020.
- [4] Kajal Yadav. Text Clustering using K-means, 2021.
- [5] Francis Ndiritu. How to Build an NLP Based Emotion Detection Model using Neattext and Scikit-learn, 2021.
- [6] Yasir Ali Solangi, Zulfiqar Ali Solangi. Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis, 2018.
- [7] Federico Pascual. Getting Started with Sentiment Analysis using Python, 2022.
- [8] CY Yam. Emotion Detection and Recognition from Text Using Deep Learning, 2015.
- [9] Enrique Amigo, Julio Gonzalo, Javier Artiles, Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints, 2008.
- [10] Estela Saquete, David Tomás, Paloma Moreda, Patricio Martínez-Barco, Manuel Palomar. Fighting post-truth using natural language processing: A review and open challenges, 2019.
- [11] Ajitesh Kumar. Cohen Kappa Score Python Example: Machine Learning, 2022.