МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ НАУК
Кафедра программной и системной инженерии

РЕКОМЕНДОВАНО К ЗАЩИТЕ
В ГЭК И ПРОВЕРЕНО НА ОБЪЕМ
ЗАИМСТВОВАНИЯ
Заведующий кафедрой
д.т.н., профессор
_____ *А.Г. Ивашко*
20.06. 2019 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

(магистерская диссертация)

Модуль для определения типа речевого акта отзыва и интернет-магазине

Прикладная информатика 09.04.03
Магистерская программа: Прикладная информатика в экономике

Выполнил работу
студент 2 курса
очной формы обучения
(Подпись)
Спрысков
Алексей
Алексеевич

Научный руководитель
к.фил.н., доцент
(Подпись)
Бидуля
Юлия
Владимировна

Рецензент
к.ф.-м.н., доцент
(Подпись)
Семихин
Дмитрий
Витальевич

г. Тюмень, 2019

Содержание

Введение

Интернет-магазином называют сайт в сети Интернет, предоставляющий посетителям возможность дистанционно совершать покупки. Поскольку в интернет-магазине, в отличие от магазина физического, нет контакта между представителем магазина (менеджером, администратором) и покупателем, предоставить покупателю рекомендации относительно определённого товара гораздо сложнее. Кроме того, мнение представителя магазина может быть ангажировано, поэтому не все покупатели изначально доверяют ему.

В то же время, сама природа интернет-магазина позволяет реализовать сбор и отображение отзывов покупателей о товарах. Такой вариант не требует затрат на оплату труда менеджеров, а также потенциально позволяет решить обе обозначенные выше проблемы. Согласно исследованию [1], 86% потребителей читают отзывы перед покупкой (в том числе 95% в возрастной группе 18-34), 78% доверяют отзывам в той же мере, что и личным рекомендациям, 89% читают и принимают во внимание ответы компании на отзывы. Эти цифры доказывают важность отзывов в интернет-магазине. Таким образом, отзывы действительно способны мотивировать колеблющегося посетителя совершить покупку.

Эти же цифры говорят и о важности работы компании с отзывами. Негативный эффект отрицательного отзыва можно значительно уменьшить, правильно ответив на него. В больших интернет-магазинах со значительным количеством посетителей обрабатывать поток входящих отзывов в ручном режиме может быть проблематично. Два этих фактора ведут к необходимости использования средств анализа имеющихся в ИМ отзывов.

Выявления речевого акта пользователя видится перспективным в первую очередь для ранжирования отзывов. Так, например, отзывы с более активными актами (например, вопросы) скорее всего требуют более скорого ответа, нежели менее активные (такие, как утверждения). Также данная характеристика может использоваться и для фильтрации отзывов – например, интернет-магазин может назначать менеджеров на работу с негативными и вопросительными отзывами,

тогда как отзывы-утверждения, выражающие удовлетворение товаром или услугой, не содержащие вопроса и не требующие содержательного ответа, можно обрабатывать автоматически или вручную при отсутствии загрузки по более приоритетным отзывам.

Отзывы с похожими утверждениями предлагается группировать. Это позволит выявлять настроения и мнения пользователей о товарах без необходимости чтения и анализа всех отзывов.

Глава 1. Постановка задачи

1.1 Цель и задачи

Цель данной работы – классификация отзывов и определение групп отзывов со схожими мнениями, а также численная оценка передаваемых каждой группой мнений в интернет-магазине путём выработки и последующего применения полученного метода классификации и группировки отзывов в интернет-магазине.

Данную цель можно разбить на несколько задач:

1. Выявление классов для классификации;
2. Выявление классификационных признаков;
3. Определение алгоритма классификации;
4. Выявление класса каждого отзыва;
5. Выявление групп отзывов для отзывов с выражением мнения;
6. Выявление слов, передающих мнение группы, для каждой группы;
7. Выявление численной оценки передаваемого мнения каждой группы.

Непосредственно информация об отзывах будет формироваться на этапах 4 и 5. Этапы 1-3 – подготовительные.

1.2 Входные данные

Анализируемыми данными будут выступать отзывы посетителей интернет-магазина на представленные в нём товары. Ожидается, что каждый отзыв можно отнести к одному из выделенных четырёх классов (утверждение, рекомендация, выражение мнения, вопрос). Распределение объектов между классами, предположительно, будет неравномерным, со значительным дисбалансом в сторону класса утверждений.

Отзыв представляется в виде текста. Длина текста не ограничивается, но можно предположить, что она будет составлять до 600 символов.

Текст отзыва может состоять из произвольного количества слов и предложений предложений. Он может содержать в себе, в том числе, смайлики, эмодзи и другие небуквенные символы.

Глава 2. Выбор алгоритма

2.1 Постановка задачи определения речевого акта

Для более формального определения речевого акта обратимся к работам [2] и [3].

В данных работах составлен список групп речевых актов, к одной из которых можно отнести отзыв:

1. Утверждение. К данной группе относятся тексты, явно подразумевающие правдивость передаваемой информации. «Простейшая проверка на утверждение: можно однозначно охарактеризовать его как верное или неверное»[4].

   В контексте задачи это отзывы, в которых описываются объективные характеристики товара.

2. Рекомендация. В эту группу попадают тексты, рекомендующие (либо осуждающие) некоторый объект, либо дающие совет относительно этого объекта.

   В контексте задачи это отзывы, в которых даются рекомендации по выбору и/или использованию товара.

3. Выражение мнения. Тексты данной группы выражают отношение и эмоции автора относительно чего-либо.

   В контексте задачи это отзывы, описывающие субъективные характеристики товара, а также отзывы с яркой эмоциональной окраской.

4. Вопрос. Тексты данной группы несут в себе вопросы по какой-либо области.

   В контексте задачи это отзывы, в которых пользователи задают вопросы относительно товара.

Примеры распределения отзывов по группам представлены в таблице 1.

Таблица 1. Примеры распределения отзывов по группам речевых актов.

Группа	Пример отзыва
Утверждение	Такой моделью пользуюсь уже 6 лет**,** часто использую гриль с вращающимся вертелем. Корочка поджаристая**,** равномерная. Готовлю также на других режимах. Есть еще режим - одновременное приготовление 2 блюд**,** тоже использовала. Отлично печет без нареканий**,** я очень довольна ей. Дверца с тремя стеклами**,** при готовке абсолютно не нагревается.Сейчас с переездом в другой дом встал вопрос о новой духовке**,** так по разумной цене с такими же 8 функциями не могу подобрать ничего взамен. Духовка на 5 с плюсом!
Рекомендация	Корпус мне понравился. Единственное обратите внимание на нижнее расположение БП. У меня из-за этого не хватило длины кабеля питания процессора 8pin. Пришлось наращивать.
Выражение мнения	Очень хороший смартфончик. Шустренький и простой в управлении, фотки получаются неплохие. Мне очень даже нравится эта модель))
Вопрос	Всем Здравствуйте!!! Ребята, можно ли заменить заднюю панель?(У меня она потрескалась и рассыпается. Нигде не могу найти замену этого стекла, она вообще возможна?????

Отзывы с типом речевого акта «Выражение мнения» должны группироваться по признаку схожести рассматриваемых аспектов объекта отзыва. Это позволит дифференцировать отзывы по обсуждаемым характеристикам товара. Число групп заранее неизвестно, группы могут иметь иерархию (так, группа отзывов, объединённая по признаку обсуждения конструктивных особенностей товара, может содержать в себе, подгруппы, объединённые по признаку обсуждения, в частности, размеров товара, а также формы товара.

Итого, входом для алгоритма служит набор отзывов. Входом первого этапа также служит набор отзывов. Выходом первого этапа служит набор типов речевых актов каждого входного отзыва. Входом второго этапа служит набор отзывов с типом речевого акта «Выражение мнения». Выходом второго этапа служит набор групп отзывов, сгруппированных по признаку близости высказываемого в них мнения. Выходом алгоритма в целом служит совокупность выходов первого и второго этапов.

2.2 Выбор признаков для определения принадлежности сообщения к группе

Опираясь на описанные в работе [6] признаки для определения принадлежности сообщения к группам речевых актов, определим следующие группы признаков:

Семантические

1.1. Слова, выражающие мнение. На основании подхода из работы [6] и словаря для русского языка[7] составлен список из 1017 слов, характерных для выражения ярких эмоций. Для каждого слова известна направленность эмоции (позитивная либо негативная). На основе этих данных сформированы девять бинарных признаков, отображающие:

1. Наличие хотя бы одного слова с яркой эмоциональной окраской;
2. Наличие хотя бы двух слов с яркой эмоциональной окраской;
3. Наличие не менее чем четырёх слов с яркой эмоциональной окраской;
4. Наличие хотя бы одного слова с позитивной эмоциональной окраской;
5. Наличие хотя бы двух слова с позитивной эмоциональной окраской;
6. Наличие хотя бы трёх слов с позитивной эмоциональной окраской;
7. Наличие хотя бы одного слова с негативной эмоциональной окраской;
8. Наличие хотя бы двух слова с негативной эмоциональной окраской;
9. Наличие хотя бы трёх слова с негативной эмоциональной окраской;

Семантический смысл данной группы признаков в том, что чаще всего такие слова есть в выражении мнения;

1.2. Вульгарные слова. Аналогично предыдущему пункту, на основании словаря и результатов работы морфологического анализатора Mystem[10] сформирован бинарный признак, отображающий наличие хотя бы одного слова из списка в сообщении. Семантический смысл данного признака в том, что чаще всего такие слова встречаются в выражении мнения и реже всего - в утверждениях;

1.3. Слова, выражающие рекомендацию. На основании анализа набора отзывов выделен список из пяти слов, характерных для выражения рекомендации. На его основе выделен один бинарный признак, отображающий наличие слов из списка в анализируемом тексте.

1.4. Смайлики. На основании словарей составлен список из 139 смайликов, а также 2699 эмодзи. Также алгоритмом определяется разница в количестве встречаемых открывающих и закрывающих скобок (то есть, «(» и «)»). Аналогично определяется количество вхождений в текст последовательностей из двух и более вышеприведённых символов. На основе вышеприведённого сформировано девять бинарных признаков, отображающих:

1. Наличие хотя бы одного элемента списка смайликов и эмодзи в сообщении;
2. Наличие хотя бы одного не парного символа «(»;
3. Наличие хотя бы трёх не парных символов «(»;
4. Наличие хотя бы одного не парного символа «)»;
5. Наличие хотя бы трёх не парных символов «)»;
6. Наличие хотя бы одной последовательности двух и более символов «(»;
7. Наличие хотя бы трёх последовательностей двух и более символов «(»;
8. Наличие хотя бы одной последовательности двух и более символов «)»;
9. Наличие хотя бы трёх последовательностей двух и более символов «)»;

Семантический смысл данного признака в том, что реже всего смайлики встречаются в утверждениях;

1.5. Глаголы, свойственные определенным речевым актам. На основе словаря[8] составлен список из 150 глаголов, свойственных определенным речевым актам. Соответственно, каждый из этих глаголов будет иметь значительно большую частотность в одном из выделенных групп речевых актов относительно всех остальных. На основании списка

составлено 150 бинарных признаков, отображающих наличие каждого из этих глаголов в сообщении.

1.6. n-граммы (унограммы, биграммы, триграммы). Аналогично элементам предыдущего пункта, некоторые сочетания слов могут выражать определенные речевые акты и таким образом их наличие в сообщении будет с некоторой долей вероятности сигнализировать о том, что сообщение принадлежит к определенной группе. Эти сочетания определяются эмпирически в автоматическом режиме на основе размеченных сообщений – алгоритмом выбираются те словосочетания, которые используются не менее чем в 2% сообщений, затем из них исключаются те, в которых есть имена собственные. Затем для каждого словосочетания определяется энтропия его распределения среди сообщений каждой группы. Затем значения энтропии для каждой группы суммируются в суммарную энтропию распределения. Чем это значение ниже, тем менее случайно рассматриваемая n-грамма распределена среди всех групп и, соответственно, тем чаще она встречается только в определенной группе речевых актов и тем вероятнее определить исходную группу сообщения по наличию факта включения в него рассматриваемой n-граммы. Затем суммарная энтропия распределения нормируется - делится на логарифм от количества сообщения, в которых встречается рассматриваемая n-грамма. Опытным путём и на основании анализа работы [6] установлено, что оптимальным следует считать извлечение тех n-грамм, полученное значение для которых не превышает 0,15. Количество выделяемых на данном этапе признаков равняется количеству полученных n-грамм, все признаки бинарные и отображают факт наличия n-граммы в сообщении;

2. Лексические

2.1. Знаки препинания. Такие знаки, как «!» и «?», встречаются далеко не в каждом сообщении. Следовательно, их наличие или отсутствие может быть признаком принадлежности или отсутствия принадлежности к определенной группе. Так, знак «?» будет присутствовать в большинстве сообщений, соответствующих группе «Вопрос». На данном этапе выделены шесть бинарных признаков, отображающих:

1. Наличие в тексте хотя бы одного знака «!»;

2. Наличие в тексте хотя бы трёх знаков «!»;

3. Наличие в тексте хотя бы одной последовательности из двух или более знаков «!»

4. Наличие в тексте хотя бы одного знака «?»;

5. Наличие в тексте хотя бы трёх знаков «?»;

6. Наличие в тексте хотя бы одной последовательности из двух или более знаков «?»

2.2. Сокращения. На основании словаря[9] создан список из 40 сокращений, часто используемых в интернет-среде. На его основании определен один бинарный признак для определения факта присутствия какого-либо сокращения в сообщении;

2.3. Части речи и словоформы. В сообщении на данном этапе определяется наличие прилагательных и междометий. Междометия зачастую используются для выражения эмоций и факт их наличия в сообщении может говорить о том, что сообщение выражает мнение. Прилагательные в свою очередь аналогично зачастую используются для выражения мнения или рекомендаций. Выделено два бинарных признака, отображающих наличие данных частей речи в сообщении.

Дополнительно анализируются следующие характеристики глаголов:

1. Нахождение анализируемого глагола в форме деепричастия;

2. Нахождение анализируемого глагола в начальной форме (инфинитив);

3. Нахождение анализируемого глагола в форме причастия;

4. Нахождение анализируемого глагола в форме изъявительного наклонения;

5. Нахождение анализируемого глагола в форме повелительного наклонения.Для каждой из этих характеристик выработано по одному бинарному признаку, отображающему наличие в тексте хотя бы одного соответствующего рассматриваемой характеристике глагола.

В сумме в данной группе выделено семь бинарных признаков.

Итого выделено 184 статичных признаков, плюс группа признаков, определяемых через анализ размеченных сообщений.

Входом данного этапа является набор отзывов. Выходом данного этапа является набор значений классификационных признаков отзыва.

2.3 Построение вектора признаков текста

Разберём на примере построение вектора признаков текста на основании вышеопределённых признаков. В качестве пробного текста возьмём следующий отзыв: «*Хороший планшет, со своими задачами справляется хоть для работы тяжелых приложений нужен clean master. Немного расстроил тест Antutu показал 13500 очков без режима экономии електроенергии и с помощью clean master, а так 9500 хоть читал что и так 13700.За такую цену отличный девайс советую для учебы, фильмов, серфинга и адекватных игр (по-моему Asphalt 8 без clean master это предел).*».

В процессе работы алгоритма в проверяемой конфигурации был построен следующий список n-грамм (слова каждого текста приводятся в виде лемм):

1. В целом
2. Советовать
3. Брать
4. Дешевый
5. Находить
6. Большой
7. Удобный
8. Мочь
9. Не мог
10. Мощный
11. Пожалеть
12. Не пожалеть
13. Свой деньги
14. Начинать
15. Рекомендовать
16. Делать
17. Пробовать
18. Ничто не

19. За такой деньги

20. Очень довольный

21. Данный

22. Высокий

23. Говорить

24. Становиться

25. Хотеться

Для сокращения объёма текста значение «Истина» признака будет заменяться значением «1», а значение «Ложь» - значением «0».

Соответствующая пункту 1.1 (Слова, выражающие мнение) часть вектора будет равна 100100000, поскольку в тексте есть одно слово с сильной позитивной эмоциональной окраской – «отличный».

Соответствующая пункту 1.2 (Вульгарные слова) часть вектора будет равна 0, поскольку слов обсценной лексики в данном тексте нет.

Соответствующая пункту 1.3 (Слова, выражающие рекомендацию) часть вектора будет равна 1, поскольку в тексте есть слово из списка слов рекомендаций – «советую».

Соответствующая пункту 1.4 (Смайлики) часть вектора будет равна 000000000, поскольку смайликов и эмодзи в тексте нет, а все имеющиеся скобки – парные, и не являются средствами передачи эмоций.

Соответствующая пункту 1.5 (Глаголы, свойственные определенным речевым актам) часть вектора будет равна 0, взятому 150 раз, поскольку слов из списка глаголов, связанных с речевыми актами, нет.

Соответствующая пункту 1.6 (n-граммы) часть вектора будет равна 01000000000000000000000000, поскольку из всех n-грамм в тексте встречается только слово «советую».

Соответствующая пункту 2.1 (Знаки препинания) часть вектора будет равна 000000, поскольку в тексте не встречаются символы «!» и «?».

Соответствующая пункту 2.2 (Сокращения) часть вектора будет равна 0, поскольку в тексте нет сокращений.

Соответствующая пункту 2.3 (Части речи и словоформы) часть вектора будет равна 0101000, поскольку в тексте встречается прилагательное (например, «хороший»), а также глагол в форме изъявительного наклонения («справляется»).

Для стемирования, получения исходной словоформы и морфологического анализа использовалось ПО MyStem[10] и библиотека pymystem3[11].

2.4 Определение принадлежности сообщения к классу

Задачу определения принадлежности объекта к классу решают алгоритмы классификации. В данной задаче классификация производится по нескольким заранее определенным классам. В качестве вариантов для реализации мультиклассовой классификации было выбрано четыре алгоритма - наивный Байесовский классификатор, дерево решений, логистическая регрессия и метод опорных векторов. Для всех четырёх алгоритмов использовалась их реализация в пакете Python Scikit-learn [12].

Входом данного этапа также служит набор классификационных признаков отзывов. Выходом данного этапа служит набор типов речевых актов каждого входного отзыва.

2.5 Выбор метода классификации

Для реализации метода опорных векторов использовалась стратегия “one versus one Коэффициент регуляризации (штраф при нормализации) C выбран равным 1. Веса классов балансируются, исходя из количества элементов в классе, что позволяет нивелировать диспропорцию в количестве элементов каждого класса. Критерию останова (минимальной близости имеющейся разделяющей гиперплоскости к оптимальной) выбрано значение, равное 0,5. Наилучшие результаты показало линейное ядро.

Для НБК, поскольку все признаки бинарные, использовалась реализация, модель которого использует распределение Бернулли. Сглаживающий параметр α, отвечающий за сглаживание результата функции правдоподобия, равен 0.

Для реализации логистической регрессии выбраны следующие параметры. Для регуляризации используется L1-регуляризатор. Метод работает по стратегии “one versus rests”. Таким образом, для N классов строится N бинарных классификаторов, каждый из которых оценивает принадлежность анализируемого объекта к своему классу. Затем в качестве результата выбирается тот класс, классификатор для которого показал наибольшую степень уверенности. Коэффициент регуляризации (штраф при нормализации) C выбран равным 0,6. Критерию останова (минимальной близости имеющейся разделяющей гиперплоскости к оптимальной) выбрано значение, равное 0,03. Веса классов балансируются, исходя из количества элементов в классе, что позволяет нивелировать диспропорцию в количестве элементов каждого класса.

Для реализации метода дерева решений выбраны следующие параметры. Максимальная глубина ограничивается 10 элементами. Минимальное количество элементов в листе, требуемое для того, чтобы алгоритм разделял данный лист, равно 10% от общего количества анализируемых элементов. Минимальное количество элементов, которое должно быть в листе, равно единице. Максимальное число признаков, рассматриваемых при разделении

листа, равно 90% от общего их числа. Критерий определения оптимального разделения – неопределенность Джини.

Для определения оптимального набора гиперпараметров для каждого классификатора использовались возможности пакета Scikit-learn[12] по перебору параметров и оценке точности работы классификатора при этих параметрах.

Для обучения классификаторов было вручную размечено 856 отзывов. Также для каждого сообщения программно были найдены значения всех признаков. Тестирование производилось с применением кросс-валидации по 5 блокам.

2.6 Сравнение качества работы методов классификации

Для оценки качества работы алгоритмов использовалась F1-мера, представляющая собой гармоническое среднее точности и полноты предсказания:

$$F1 = \frac{2 * prec * rec}{prec + rec};\ prec = \frac{TP}{TP + FP};\ rec = \frac{TP}{TP + FN}$$

, где $prec$ – точность, rec – полнота, TP, FP, FN – количество истинно-положительных, ложно-положительных и ложно-отрицательных решений соответственно.

Описание корпуса размеченных текстов и тестовой выборки, результаты измерений F1-меры для всех классификаторах представлены в таблицах 2-4. Всего было проанализировано 139 сообщений, что составляет 20% от общего числа корпуса размеченных текстов. Оставшиеся 80% составили обучающую выборку для классификатора. Все значения – средние от пяти измерений.

Таблица 2. Размер и состав корпуса размеченных текстов.

	Утверждение	Рекомендация	Выражение мнения	Вопрос	Всего
Количество элементов	425	90	160	20	695
Доля от общего числа, %	61	13	23	3	100

Таблица 3. Размер и состав тестовой выборки.

	Утверждение	Рекомендация	Выражение мнения	Вопрос	Всего
Количество элементов	85	18	32	4	139
Доля от общего числа, %	61	13	23	3	100

Таблица 4. Результаты измерения F1-меры для прогнозов, данных каждым алгоритмом по сообщениям каждой группы.

	Утв.	Рек.	Выраж. мн.	Вопрос	Среднее
Наивный байесовский классификатор	0.84	0.67	0.60	0.44	0.64
Дерево решений	0.89	0.77	0.73	0.57	0.74
Логистическая регрессия	0.90	0.76	0.76	0.75	0.79
Метод опорных векторов	0.89	0.79	0.69	0.57	0.74

При определении принадлежности отзывов к группе «Выражения мнения», а также при определении принадлежности ко всем группам в среднем лучший результат показал классификатор на основе логистической регрессии. Таким образом, на данном этапе работы выбрана и использовалась в дальнейшем классификация методом логистической регрессии. Результаты измерения качества работы выбранного метода классификации представлены в таблицах 5 и 6.

Таблица 5. Размер и состав тестовой выборки согласно предсказанию наилучшего классификатора.

	Утверждение	Рекомендация	Выражение мнения	Вопрос	Всего
Количество элементов	90	20	26	3	139
Доля от общего числа, %	65	14	19	2	100

Таблица 6. Результаты измерения точности и полноты выборки, полученной из предсказания наилучшего классификатора.

	Утверждение	Рекомендация	Выражение мнения	Вопрос
Точность, %	92	74	81	75
Полнота, %	89	80	71	75

2.7 Кластеризация сообщений с выражением мнения

Для решения задачи кластеризации сообщений с выражением мнения в рамках задачи поиска слухов наиболее оптимальным видится использование алгоритмов иерархической кластеризации, поскольку на результат кластеризации накладываются следующие ограничения:

1. Число кластеров неизвестно;
2. Кластеры могут иметь иерархию (поскольку в рамках нескольких слухов могут вестись спекуляции относительно одного аспекта обсуждаемого события, их можно как разделять на несколько слухов, так и сгруппировать в один).

Алгоритмы иерархической кластеризации работают по двум стратегиям:

1. «сверху вниз» (нисходящие, дивизивные), когда предварительно все объекты объединяются в один класс и затем на каждой итерации алгоритма новые кластеры образуются делением имеющихся;
2. «снизу вверх» (восходящие, агломеративные), когда предварительно каждый объект помещается в свой кластер и затем на каждой итерации алгоритма новые кластеры образуются слиянием имеющихся.

Нисходящие алгоритмы обладают экспоненциальной сложностью ($O(2^n)$, тогда как восходящие – полиномиальной $O(n^2)$. В связи с большим объёмом входных данных более целесообразной видится реализация кластеризации с восходящим алгоритмом.

В данной работе использовалась реализация иерархического агломеративного кластеризатора в пакете Python Scikit-learn[12].

Входом данного этапа является набор сообщений, отнесенных в группу «Выражение мнения». Выходом данного этапа является набор групп сообщений, сформированных по признаку близости передаваемой в них информации.

2.8 Выбор метода кластеризации

В соответствии с выводами работы [6] кластеризацию будем производить исходя из расстояний между всеми сообщениями. Расстояние между двумя сообщениями находится с помощью метрики TF-IDF и косинусной меры подобия.

Рассчитав метрику TF-IDF в тексте для набора слов, текст можно представить в виде вектора, пространством координат для которого будет являться рассматриваемый набор слов, а значениями по каждой оси – метрики TF-IDF соответствующих этой оси слов в данном документе. Представив в виде вектора два сообщения, в качестве набора слов для анализа используя совокупность всех слов, не являющихся служебными частями речи, в обоих документах, можно рассчитать косинусную меру подобия между ними, получив в итоге искомое расстояние между двумя сообщениями.

Метрика TF-IDF (Term Frequency – Inversed Document Frequency) представляет собой меру оценки важности слова в пределах текста[15]. Первая часть метрики (TF) представляет собой частоту встречаемости терма (слова). Вторая часть (IDF) представляет собой величину, обратную частоте встречаемости рассматриваемого терма во всех документах коллекции. Формула для расчёта метрики:

$$\mathrm{T}FIDF(w,d,D) = TF(w,d) * IDF(w,D);$$

$$TF(w,d) = 0.5 + \frac{0.5 * fr(w,d)}{\max\{fr(w_i,d) : w_i \in d\}};$$

$$IDF(w,D) = \log(\frac{|D|}{|\{d_i \in D \mid w \in d_i\}|})$$

, где w – слово, d – документ, D – коллекция документов, fr(w,d) – частота встречаемости слова w в документе d.

Косинусная мера сходства двух документов определяется как:

$$\cos(a, b) = \frac{a * b}{||a|| \times ||b||}$$

, где a,b – документы, представленные в виде векторов значений метрик TF-IDF слов обоих документов.

Для всех анализируемых текстов проводится их векторизация. Для построения вектора сообщения используется словарь слов, выражающих мнение[7]. Из него выбираются те слова, которые содержатся в любом из входных текстов. Затем из этого списка исключатся слова, не являющиеся прилагательным или его формой. Затем для каждого такого слова в каждом тексте определяется метрика TF-IDF. Вектором сообщения в таком случае является вектор в пространстве выявленных слов, значениями же являются найденные значения метрики TF-IDF выделенных слов в рассматриваемом сообщении.

Используемому классификатору на вход подаётся набор полученных векторных представлений сообщений. Классификатор рассчитывает матрицу косинусных расстояний между векторами, затем объединяет в кластеры те векторы, косинусное расстояние между которыми не больше порогового значения. В данном решении используется пороговое значение, равное 0,7.

Итоговые кластеры объединяют наборы сообщений, содержащие схожие мнения. Собирательно информация из полученного кластера сообщений представляет отдельное мнение о товаре.

После извлечения кластера производится его дальнейший разбор с целью определения численного представления передаваемого в кластере настроения, а также опорных слов кластера. Для этого используется упоминаемый выше словарь слов, выражающих мнение. Для каждого кластера определяется набор слов, встречаемых в каждом из текстов. Полученный набор слов считается списком опорных слов кластера. Затем из полученного набора слов исключатся те слова, которые не встречаются в словаре слов, выражающих мнение. Полученный набор слов считается списком опорных слов мнения кластера. Для

каждого слова этого набора из словаря берётся мера эмоциональности этого слова. Численное представление передаваемого в кластере настроения определяется как сумма мер эмоциональности опорных слов мнения кластера.

Таким образом, помимо собственно кластера извлекаются также набор ведущих слов (слов, общих для всех текстов кластера), набор передающих мнение ведущих слов и численная мера передаваемого текстами кластера настроения.

2.9 Построение кластера

Рассмотрим построение кластера на основании 26 сообщений, отмеченных классификатором как сообщения с типом речевого акта «Выражение мнения». Данные сообщения приводятся в Приложении 1.

Список слов, отобранных алгоритмом для построения вектора представления сообщения:

1. Хороший
2. Неплохой
3. Незаменимый
4. Положительный
5. Лишний
6. Печально
7. Великолепно
8. Пригодный
9. Разумный
10. Старый
11. Крутой
12. Полезный
13. Супер
14. Лёгкий
15. Разный
16. Удачный
17. Хлипкий
18. Шаткий
19. Плохой
20. Отличный
21. Жаль
22. Мощный
23. Довольный

24.Нормальный

25.Полностью

26.Адекватный

27.Хвалебный

28.Отлично

29.Вечный

30.Много

31.Красивый

Данные слова – слова выражения мнения, встречающиеся как минимум в одном тексте набора входных текстов, являющиеся прилагательным или его формой.

Рассмотрим сообщение №7. Вектор, представляющий это сообщение, приводится в таблице 7.

Таблица 7. Вектор, представляющий сообщение №7.

Позиция	1	2	3	4	5	6	7	8	9
Значение	0	0	0	0	0	0	0	0	0

Позиция	10	11	12	13	14	15	16	17	18
Значение	0	0	0	0,73	0	0	0	0	0

Позиция	19	20	21	22	23	24	25	26	27
Значение	0	0	0	0	0	0	0	0	0,48

Позиция	28	29	30	31
Значение	0	0	0	0

Индексом в данном векторе является порядковый номер слова из списка отобранных слов. Значением является значение метрики TF-IDF данного слова в текущем сообщении.

На основании данного вектора и векторов других сообщений данное сообщение было отнесено к кластеру №1, несмотря на наличие слова, являющегося ведущим для кластера №2.

Кластер 1. Опорное слово: супер.

Тексты кластера (здесь и далее в качестве текста приводится его номер из Приложения 1): 3, 7, 18

Кластер 2. Опорное слово: отлично.

Тексты кластера: 4, 12, 15, 24.

Кластер 3. Опорное слово: мощный.

Тексты кластера: 5, 26.

Кластер 4. Опорные слова: очень, довольный.

Опорные слова, передающие мнение: довольный

Тексты кластера: 17, 23.

Кластер 5. Опорное слово: жаль.

Тексты кластера: 13, 16.

2.10 Определение качества работы метода кластеризации

Для оценки результата работы кластеризатора тексты, которые он обрабатывал, предварительно были разделены на кластеры экспертом. Таким образом, для определения качества работы кластеризатора проводилось сравнение между сгенерированными кластеризатором группами текстов и группами, построенными экспертным методом.

Результаты измерений представлены в таблицах 8-12.

Таблица 8. Количество кластеров, выявленных экспертом и алгоритмом.

	Количество кластеров
Эксперт	5
Алгоритм	5

Таблица 9. Опорные слова для построенных экспертом и алгоритмом кластеров.

	Кластер 1	Кластер 2	Кластер 3	Кластер 4	Кластер 5
Эксперт	супер	отлично	мощный	доволен	жаль
Алгоритм	супер	отлично	мощный	доволен	жаль

Таблица 10. Количество ошибок первого рода - текстов, вошедших в кластер при данном методе построения кластеров, но не имеющихся в соответствующем кластере противоположного.

	Кластер 1	Кластер 2	Кластер 3	Кластер 4	Кластер 5
Алгоритм	0	1	0	0	0

Таблица 11. Количество ошибок второго рода - текстов, не вошедших в кластер при данном методе построения кластеров, но имеющихся в соответствующем кластере противоположного метода.

	Кластер 1	Кластер 2	Кластер 3	Кластер 4	Кластер 5
Алгоритм	0	3	0	2	0

Таблица 12. Общее количество текстов в каждом кластере.

	Кластер 1	Кластер 2	Кластер 3	Кластер 4	Кластер 5
Эксперт	3	6	2	4	2
Алгоритм	3	4	2	2	2

Общее количество анализируемых текстов составило 26 штук.

Таким образом, точность полученного классификатора равняется 92 процентам. Единственная имеющаяся ошибка первого рода не является грубой, поскольку текст, который был обозначен классификаторам как принадлежащий группе текстов, характеризуемых словом «отлично», содержит это слово, а экспертом был отнесён в группу «доволен» с сомнением. Полнота составляет 71 процент.

F1-мера точности кластеризации, исходя из полученных измерений, равна 0,8.

2.11 Программная реализация

Разработанный алгоритм был реализован на ЯП Python с использованием пакета scikit-learn[12] .

Класс app является точкой входа приложения. В нём происходит обращение к классу featuresExtractor для получения вектора признаков отзыва.

В данном классе происходят обращения к методам классов lexicFeatures, semanticFeatures и nGramsFeaturesBuilder. Каждый из этих классов отвечает за расчёт значений определенного набора классификационных признаков для входных сообщений. Класс nGramsFeaturesBuilder также отвечает за построение n-грамм, происходящее посредством обучения на корпусе размеченных текстов.

Также класс featuresExtractor взаимодействует с классом textPreps, отвечающим за предобработку текстов. Помимо предподготовки текстов, данный класс отвечает за синтаксический разбор текстов, а именно лемматизацию, выявление части речи слова, его формы и наклонения, определение обсценности для всех слов входных текстов. Помимо этого, описываемый класс взаимодействует с классом cacher. Данный класс реализует кеширование данных с помощью NoSQL-базы данных Redis. В классе featuresExtractor он используется для кеширования рассчитываемых векторов значений признаков текстов. Благодаря этому, для повторно разбираемого текста расчёт производиться не будет, а вместо этого вектор будет извлечён из кеша. Также кешируется список n-грамм, благодаря чему нет необходимости каждый раз строить его заново.

После обращения к классу featuresExtractor класс app обращается к классу classifier. Данный класс отвечает за непосредственно классификацию текстов, поступающих в виде векторов значений классификационных признаков. Методы этого класс позволяют производить классификацию по любому методу из четырёх имеющихся (реализации хранятся в подклассе models, что позволяет использовать обращение вида classifierObject.models.SVM), по всем методам

сразу, а также производить оценку качества работы модели, сохранить и загрузить её и подобрать оптимальные параметры для каждой модели.

После обращения к классу featuresExtractor класс app обращается к классу clusterizer. Этот класс отвечает за кластеризацию поступающих текстов.

Диаграмма классов для программой реализации алгоритма представлена на Рисунке 1.

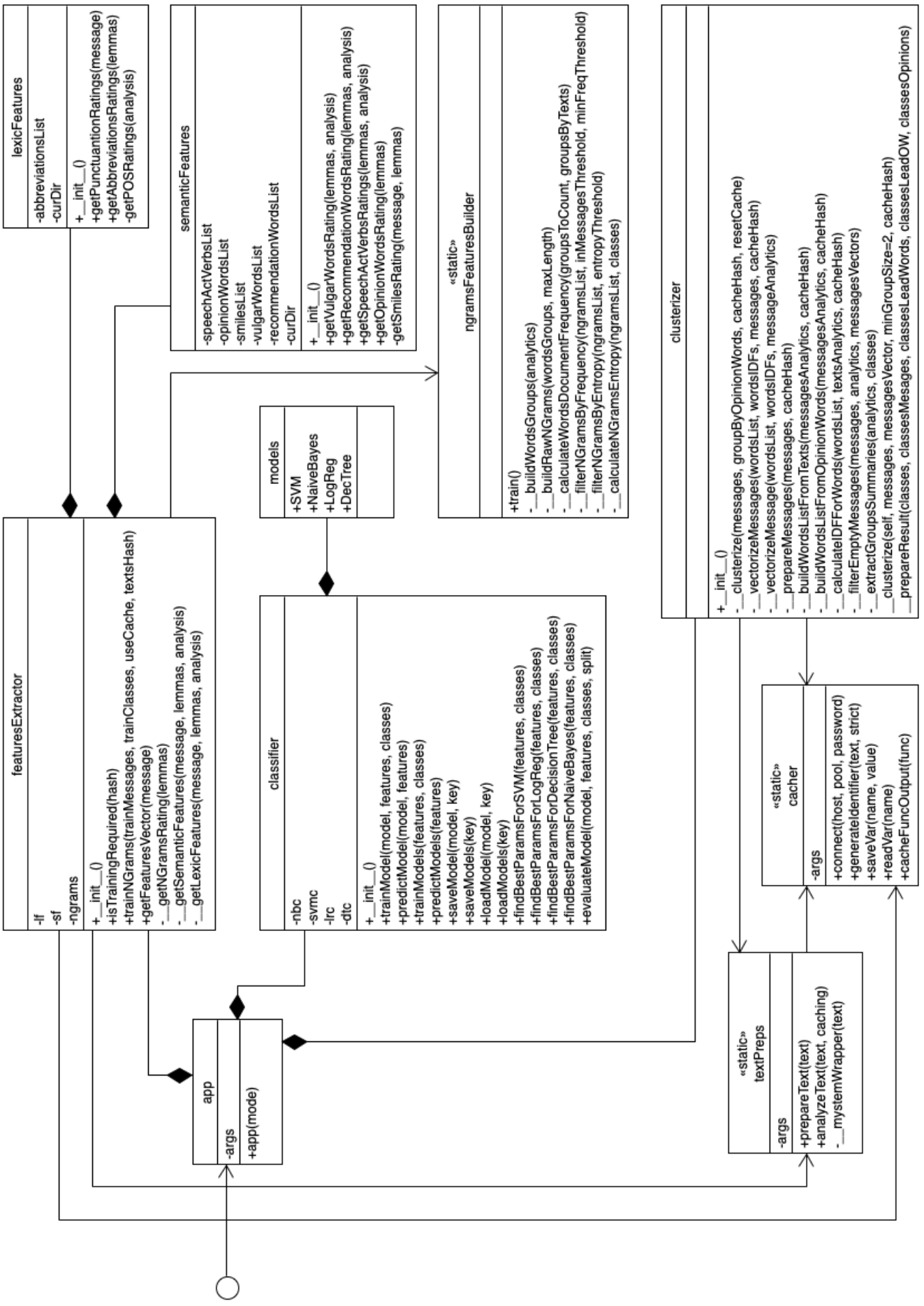

Рисунок 1. Диаграмма программных классов

На рисунках 2, 3 представлен внешний вид веб-интерфейса к разработанному алгоритму в системе управления контентом “Bitrix”. На рисунке 2 представлена фильтрация отзывов по типу речевого акта.

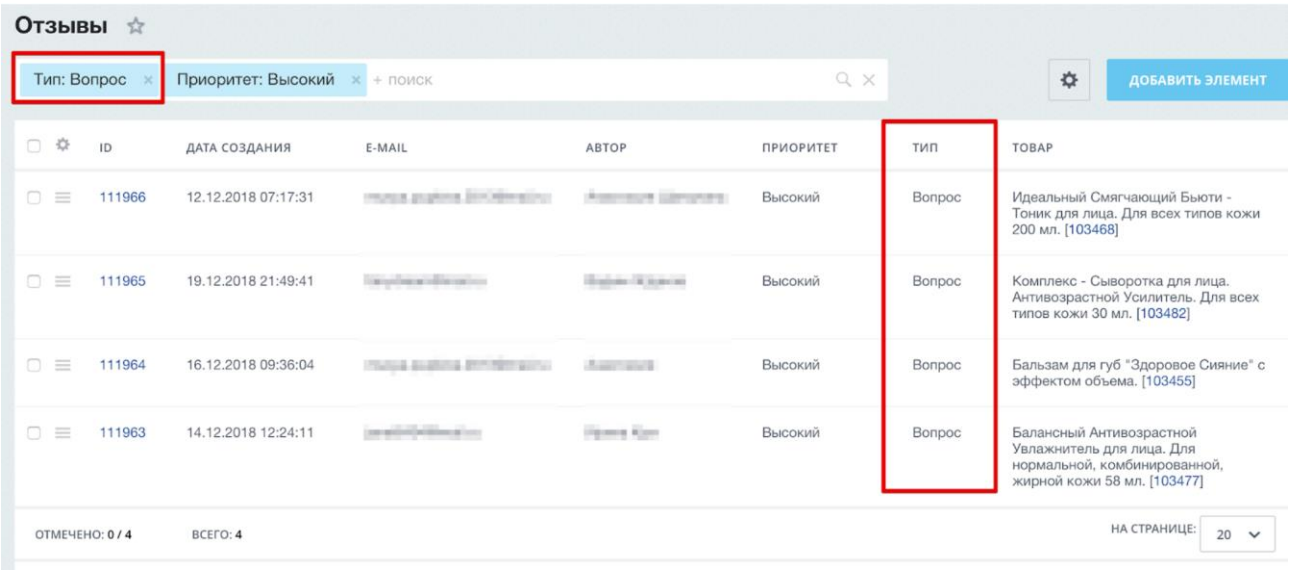

Рисунок 2. внешний вид веб-интерфейса к разработанному алгоритму в системе управления контентом “Bitrix”.

На рисунке 3 представлен вывод конкретной группы (кластера) отзывов.

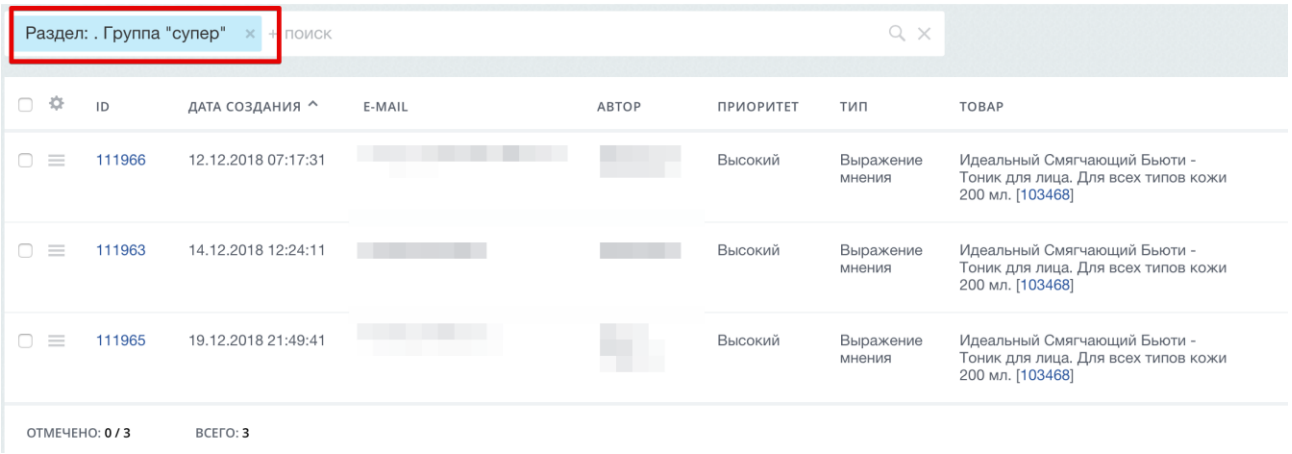

Рисунок 3. внешний вид веб-интерфейса к разработанному алгоритму в системе управления контентом “Bitrix”.

Заключение

В результате работы был разработан алгоритм, позволяющий определять тип речевого акта отзыва в интернет-магазине, а также группировать отзывы по принципу схожести обсуждаемых аспектов товара. Для групп отзывов определяется набор ключевых слов, а также численная оценка передаваемых в отзывах группы эмоций.

Были проведены измерения качества работы алгоритма, позволяющие говорить о его состоятельности.

Результаты работы были представлены и опубликованы в сборнике «Математика и междисциплинарные исследования – 2019» по материалам Всероссийской научно-практической конференции молодых ученых с международным участием.

Также результаты работы были представлены на 70-й ежегодной научной конференции ТюмГУ в секции «Разработка системных и программных решений» и были опубликованы в сборнике данной конференции.

Частично данная работа является продолжением работы “The summarization of search results” [16], представленной на международной конференции «IEEE 11th International Conference on Application of Information and Communication Technologies».

Список источников

1. Bright Local Consumer Review Survey 2018. https://www.brightlocal.com/learn/local-consumer-review-survey/
2. J. R. Searle. *A taxonomy of illocutionary acts*. Linguistic Agency University of Trier, 1976.
3. R. Zhang, D. Gao, and W. Li. *What are tweeters doing: Recognizing speech acts in twitter*. Analyzing Microtext, 2011.
4. J. R. Searle. *Expression and meaning: Studies in the theory of speech acts*. Cam- bridge University Press, 1985.
5. Soroush Vosoughi. *Automatic Detection and Verification of Rumors on Twitter*. Massachusetts Institute of Technology, 2015
6. P. Stone, D. C. Dunphy, M. S. Smith, and D. Ogilvie. *The general inquirer: A computer approach to content analysis.* Journal of Regional Science. 1968.
7. Проект лаборатории Интернет-исследований ВШЭ Linis Crowd. http://linis-crowd.org
8. Ермолаева И.А. *Семантическая классификация глаголов речи в русском языке.* Вестник СПбГУ. Язык и литература. 2017. Т.14. Вып. 3
9. D. Crystal. *Language and the internet*. Cambridge University Press, 2007.
10. Морфологический анализатор MyStem. https://tech.yandex.ru/mystem/
11. A Python wrapper of the Yandex Mystem 3.1 morphological analyzer. https://github.com/nlpub/pymystem3
12. Pedregosa,G.Varoquaux,A.Gramfort,V.Michel,B.Thirion,O.Grisel,M.Blon- del, P. Prettenhofer, R. Weiss, V. Dubourg, et al. *Scikit-learn: Machine learning in python.* The Journal of Machine Learning Research, 12:2825–2830, 2011.
13. A. Rajaraman and J. D. Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

14. Exhaustive search over specified parameter values for an estimator. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
15. A. Rajaraman and J. D. Ullman. *Mining of massive datasets.* Cambridge University Press, 2011.
16. Y.Bidulya, A.Spryskov. *The summarization of search results.* IEEE 11th International Conference on Application of Information and Communication Technologies (AICT) Conference Proceedings, 2017.

Приложение 1. Список текстов, отобранных классификатором в класс текстов с типом речевого акта «Выражение мнения».

1. жалею что купил гениуса, т.к. может сигнал у него и стабильней, но кнопки, колесо, это просто жуть. да и легкая очень в руке не лежит хорошо. оставлю прозапас))))
2. Достоинств никаких не заметила, вообще, не гладит, по одежде ездит. , как по наждачной бумаге. Хотела купить треф аль, продавец в магазине уговорила купить этот Bork 1510. Завтра пойду сдавать обратно. Кофточку так и не смогла погладить и это за 5990 тыс. Руб. Сейчас до стану свой старенький Тефаль. Просто ужас! Не покупайте! У меня bosh такой же новый в шкафу стоит, купила этот, думала порадует!
3. пылесос просто супер мы в восторге от покупки) долго выбирали но не ошиблись)
4. плита отлично работает – варит и жарит, газ перекрывает, если вдруг зальешь водою, купила в основном за внешний вид, не знаю как потом будет, может надоест и поменяю, но пока нравится!
5. долгие два дня выбирал себе ноутбук для жены чтобы она научилась работать в фотошопе, итог следующий: фотошопом никто не пользуется, зато я играю в готику 3, ворд оф тэнкс и вов))) мощный ноут. Купил за 35т, стоит своих денег)
6. Хорошая моделька для своей цены. Неплохо снимает видео, и фото получаются тоже неплохие, правда не с первого раза, ну что же, не беда. Аппаратом доволен!
7. Пользуемся год!супер!маленькая что для нас важно!стирает тихо и отлично отстирывает!все программы есть и чистка баробана!до этого

10 лет был самсунг-через года 3 начала прыгать.По поводу этой сказали что новые технологии и прыгать не будет-нет ремня у нее!это просто не может не радовать!советую всем!!!

8. Перу дней назад не включился, перезагрузил комп - не помогло, во время загрузки заметил мерцания, монитор маргнет и гаснет на долю секунды. В общем покупка была изначально не удачной))) попробую реанимировать, не получится пойдет на мусорку.
9. Купила, этот красивый пылесос год назад, пару раз попылесосила, не сразу заметила, а у меня в месте куда ставиться пластмассовый контенер для моющей жидкости, образовались трещины и пылесос перестал подавать воду((Очень печально!
10. Не ожидал, что это старьё будет превосходно работать на новейшей Windows 8.1. Материнке скоро будет 5 лет, а актуальности не теряет. Все старые игры идут великолепно. Встроенная видеокарта тянет игры своего времени на ура!!! Мини игры так вообще только в путь!
11. Перешел с Айфона, ни разу не пожалел, все нравится, особенно цена))
12. Купил примерно месяц назад, откатал по пересеченке уже почти 500км.

    За это время ничего с велосипедом не произошло, все отлично работает, ничего не скрипит, нет проблем с эксплуатацией (изначально все настроил, после этого только пару раз цепь смазал) Вообщем полностью доволен!
    Для кроскантри то что нужно, особенно для начинающих)
13. Не то Сони(((а жаль (
14. в общем не плох)пока что держится)и для игр пригоден)тот же планшет))
15. все замечательно, для меня ios это очень простая и понятная система, в ней нет ничего лишнего, как это обычно в андройде. Летает по сравнению с любым андройдом, никаких глюков, все сделанно

качественно и хорошо. Как писали что через айтюнс проблемно сузыку кидать, я опровергну, это такой бред честно :) скачал айтюнс и скачал треки, все еще легче чем на андройд :) просто отличная модель) советую, тем более за такую цену :)

16. вроде бы вполне адекватный кнопочник.... был... на протяжении 8 месяцев....но потом стал самопроизвольно выключаться. за сегодня уже 11 раз выключился - реально бесит! сдам в ремонт, пока на гарантии. но тут уже писали, что это не лечится. жаль! к остальному вполне привыкнуть можно, тем более, что используется в основном как звонилка и СМС-илка.
17. Я очень доволен этим смартфоном, всем советую!
18. Для поездок просто супер! Купила себе такой планшет, сын отобрал в тот же день. Опять купила точно такой же не раздумывая. Аналог только iPad. Но Supra удивит ещё и ценой!
19. мой незаменимый помощник на кухне! Мультиварка Scarlett маленькая, удобная, но при этом готовит на отлично. Да, чашу хотелось бы побольше. Раньше я вообще практически в мультиварке не готовила, но теперь готовлю постоянно и поучается очень вкусно, поэтому хочется продлить удовольствие подольше)) Хотя в принципе 4 литра для мультиварки – это уже хорошо)) Ну, конечно, хочется отметить дизайн – эту мультиварку ни с какой не спутаешь, ну очень не обычно и не стандартно выглядит!
20. Вот не знала моя мама куда выбросить деньги! Воет как самолет, практически ничего не сосоет, с пола может какой мусор и подбирает-чего-то в мусоросборник попадает, но с ковра..... постоянно нагибаешься чтоб собрать руками мусоринку. В трубку мусор с руки не засасывается-иду в мусорку выбрасывать.Волосы с ковра не подбирает ваще. Щетка и трубка сплошное недоразумение-все хлипкое и шаткое. впрочем как и сам пылесос. Честно удивлена

хвалебным отзывам.. заказные они что ли? После месяца мучений выкидываю его с чувством полного облегчения-нервы знаете ли дороже.

21. Чайник, к сожалению, разочаровал. Покупала его по рекламе, как самый тихий и ещё клюнула на дизайн. Хотя перед покупкой читала отзывы и видела, что люди писали, что "он ревёт как самолет", и есть проблемы с крышкой, но это меня не остановило, и будучи оптимисткой, решила поверить положительным отзывам по данному прибору, как оказалось, зря! Он, конечно, функции свои выполняет и по поддержанию температуры, и по нагреву в разных режимах, и ни разу не сломался за длительный срок работы, в этом, надо отдать ему должное. Все предыдущие чайники, в основном это были Tefal начинали течь через года полтора эксплуатации. Но вот уровень шума у него абсолютно не соответствует заявляемому, он стоит у нас на кухне, там же периодически смотрим и телевизор, так вот при включённом чайнике, звук в телевизоре приходится делать значительно громче, что бы уловить хоть что-то из интересующей передачи! В общем смотрите сами, может вам с конкретным экземпляром повезёт больше чем мне, но риск того что нет тоже присутствует, и не малый.
Я считаю, что он совершенно не стоит того что бы рисковать ~ 9 000 рублей, опробовала на себе.
22. 'Недостатки могу перечислять до бесконечности. Этот телефон - сплошной глюк и вечный тормоз. Да к тому же недавно уронила его и дисплей сдох, еще и хиленький оказался, буду покупать люмку теперь, надеюсь, не разачаруюсь))))'
23. Такой моделью пользуюсь уже 6 лет, часто использую гриль с вращающимся вертелем. Корочка поджаристая, равномерная.Готовлю также на других режимах. Есть еще режим -

одновременное приготовление 2 блюд, тоже использовала. Отлично печет без нареканий, я очень довольна ей. Дверца с тремя стеклами, при готовке абсолютно не нагревается.Сейчас с переездом в другой дом встал вопрос о новой духовке, так по разумной цене с такими же 8 функциями не могу подобрать ничего взамен. Духовка на 5 с плюсом!

24. мне очень понравилась ! Низкая цена ! Можно болтать как по гарнитуре !) отличный дизайн !) маленький размер)) блин она крутая) кто хочет маленькую колоночку и не дорогую зайдите в магазин послушайте может она и вас зацепит)
25. Покупал как полезный подарок для родителей. Специально с пультом, чтобы лишний раз не бегали. Пульт удобный, большой, не теряется. Мама говорит, что отец храпеть перестал. Очень доволен, подарок делал не зря.
26. Брал взамен Philips 732, после него аппарат впечатляет. Европеец 2 ГБ Большой, мощный, хороший звук, как в динамиках, так и в наушниках (наушники cx-400 shenhaiser). Батарейки честно хватает на 3-4 дня - много звонков, немного музыки, интернет 3g и wi-fi. Камера огорчила, синит, мылит, но я ей не почти пользуюсь. Плохая связь. С обновлениями прошивок становится незначительно, но получше. За эти деньги нормальный аппарат.