

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Базовая кафедра «Интеллектуальные системы управления»

УТВЕРЖДАЮ

Заведующий кафедрой

_____ Ю.Ю. Якунин
подпись инициалы, фамилия

«_____» _____ 2020 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Оценка релевантности текстовых отзывов сервиса анкетирования

09.04.04 Программная инженерия

09.04.04.02 Технологии индустриального производства
программного обеспечения интеллектуальных систем управления

Научный

руководитель

_____ доцент, канд. техн. наук
подпись, дата

А.А. Даничев

Выпускник

подпись, дата

Е.И. Высотенко

Рецензент

_____ доцент, канд. физ.-мат. наук
подпись, дата

А.Л. Двинский

Красноярск 2020

РЕФЕРАТ

Выпускная квалификационная работа по теме «Оценка релевантности текстовых отзывов сервиса анкетирования» содержит 51 страницу текстового документа, 12 использованных источников, 20 иллюстраций, 7 формул, 21 таблицу.

КЛАССИФИКАЦИЯ, ОБРАЗОВАТЕЛЬНАЯ СРЕДА, МАШИННОЕ ОБУЧЕНИЕ, ТОНАЛЬНОСТЬ ТЕКСТОВ, НОРМАЛИЗАЦИЯ ТЕКСТА, СТЕММИНГ, ЛЕММАТИЗАЦИЯ.

Объект изучения – сервис анкетирования на базе личного кабинета «Института космических и информационных технологий».

Целью является разработка модуля расчета критериев оценки релевантности текстовых отзывов сервиса анкетирования студентов Института космических и информационных технологий.

В результате данной работы проведено исследование эффективности методов нормализации текстов, исследование эффективности моделей машинного обучения. На основе результатов исследований был создан модуль расчета критериев оценки релевантности текстовых отзывов.

В итоге был разработан модуль расчета критериев оценки релевантности текстовых отзывов сервиса анкетирования.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....	4
1 Теоретические основы расчета критериев оценки релевантности текстовых отзывов.....	6
1.1 Сервис анкетирования студентов СФУ.....	6
1.2 Критерии оценки релевантности текстовых отзывов.....	7
1.3 Методы нормализации текста.....	8
1.3.1 Стемминг.....	8
1.3.2 Лемматизация.....	9
1.4 Подходы к решению задачи автоматической классификации текстовых отзывов.....	11
1.4.1 Методы, основанные на машинном обучении.....	11
1.4.2 Метод, основанный на словаре тональностей.....	17
2 Исследование и выбор методов классификации текстов.....	20
2.1 Анализ и сравнение методов машинного обучения при расчете критерия принадлежности.....	24
2.1.1 Предобработка данных.....	24
2.1.2 Машинное обучение.....	25
2.2 Анализ и сравнение методов нормализации при расчете критерия тональной принадлежности словарным подходом.....	28
3 Программный модуль расчета критериев оценки релевантности текстовых отзывов.....	31
3.1 Описание программного модуля.....	31
3.2 Апробация программного модуля.....	35
ЗАКЛЮЧЕНИЕ.....	49
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ.....	50

ВВЕДЕНИЕ

Сервисы анкетирования на сегодняшний день находят широкое применение во многих областях деятельности, как с практической, так и с научной точки зрения. Рассматривая механизм анкетирования студентов на базе личного кабинета «Института космических и информационных технологий», можно сказать о том, что анкетирование является одним из непростых процессов механизма мониторинга образовательной среды. Данный факт связан с тем, что в опросе определяющую роль играет человеческий фактор. Именно в данном случае стоит более ответственно подходить к анализу полученных результатов.

В случае, когда система анкетирования предусматривает возможность дополнять числовую оценку текстовым комментарием, возникает задача анализа текстовой части отзыва. Регулярная обработка результатов практически всегда требует большое количество ресурсов, тем более если данные представлены в свободной форме. Так же, не все отзывы можно использовать как достоверную информацию. Проблема релевантности таких результатов связано с субъективным отношением человека и эмоциональным фактором. Анализ текстовой части отзыва позволяет наиболее продуктивно реализовывать обратную связь путем выделения числовых показателей релевантности текста, используя современные методы машинного обучения.

Объектом исследования является сервис анкетирования студентов на базе личного кабинета Института космических и информационных технологий.

Целью работы является разработка модуля расчета критериев оценки релевантности текстовых отзывов сервиса анкетирования студентов «ИКИТ».

Для достижения поставленной цели необходимо решить следующие задачи:

— исследовать технологии, необходимые для вычисления критериев оценки релевантности текстовых отзывов;

- выбрать метод нормализации, классификации, выделения эмоционального окраса текста;
- реализовать и проверить эффективность методов;
- на основе исследований, разработать модуль оценки релевантности текстовых отзывов.

Результаты теоретических исследований апробированы на международной конференции молодых ученых «Перспектив Свободный – 2020». В данный момент, публикация тезисов находится на этапе печати.

1 Теоретические основы расчета критериев оценки релевантности текстовых отзывов

1.1 Сервис анкетирования студентов СФУ

Одним из основных способов контроля качества в высших образовательных учреждениях является изучение мнения основных потребителей образовательных услуг – студентов. Данный механизм, как правило, называется анкетированием. Он позволяет объективно оценить слабые и сильные стороны функционирования образовательного процесса, или, обучающей кафедры, в частности. Как правило, мнение, предоставленное обучающимися, имеет немаловажное значение в принятии решений о корректировке образовательного процесса, например, улучшение качества преподавания и др. Но не смотря на существенное значение отзывов студентов, с теоретической точки зрения, мнение не всегда может быть объективным. Данная действительность связана с тем, что в опросе наиболее влияющую роль, как правило, играет человеческий фактор.

В Сибирском федеральном университете, так же, уделяется внимание изучению мнений обучающихся. Не является исключением и Институт космических и информационных технологий, на базе личного кабинета которого реализован сервис анкетирования. Опросы проходят каждый учебный семестр. Основные темы анкетирования:

- отношения к работе выпускающей кафедры;
- работы учебно-организационного отдела (деканата);
- характеристики преподавания предметов и обеспечивающих их материалов.

Ответы студентов предоставляются в формате оценки по бальной шкале, опционально, числовые оценки дополняются текстовыми комментариями.

Пользовательский интерфейс сервиса анкетирования представлен на рисунке 1.

Рисунок 1 – Пользовательский интерфейс сервиса анкетирования СФУ

1.2 Критерии оценки релевантности текстовых отзывов

Главным направлением обработки результатов анкетирования является анализ бальных оценок, и проведение параллели с результатами итоговой аттестации студентов. При данном подходе текстовая часть отзывов используется крайне ограничено.

В большинстве случаев, автоматизированный анализ текстовых ответов позволит наиболее эффективно осуществить поддержку обратной связи на отзывы студентов.

В процессе автоматизации процесса анализа отзывов, требуется выделять числовые оценки текста – критерии оценки релевантности. Предлагается выделять следующие показатели [1]:

- степень принадлежности отзыва определенному вопросу;
- сумма тональных показателей слов отзыва.

Степень принадлежности позволит определить процент соответствия текста отзыва словарю вопроса (принадлежность вопросу). Сумма тональных показателей, в свою очередь, позволит узнать соответствие эмоциональных оттенков отзыва числовой оценке.

1.3 Методы нормализации текста

1.3.1 Стемминг

Суть стемминга заключается в отсечении окончаний суффиксов слов так, чтобы оставшаяся часть была одинаковой для всех грамматических форм слова. Получившийся по итогу обработки результат называют «стеммом». На текущий момент существует множество вариантов реализаций алгоритмов стемминга, например стемминг Портера. Возможные ошибки при использовании стемминга:

- усечение слова до слишком короткой формы чревато совпадением получившегося «стема» со словами другого смыслового значения;
- при недостаточном усечении слова «стем» может не охватить все те слова, подходящие по грамматическому смыслу;
- невозможность правильно построить «стем» ввиду изменения букв в корне слова.

В таблице 1 показана нормализация метода стемминга.

Таблица 1 – Пример работы стемминга

Форма	Суффикс	Стем
Ваза	-а	Ваз
Главный	-ый	Главн

В контексте исследований, используется стеммер «SnowballStemmer». Snowball – реализация алгоритма, разработанного Мартином Портером в 1979

году для английского языка. Основная идея стеммера Портера заключается в том, что существует ограниченное количество словообразующих суффиксов, и стемминг слова происходит без использования каких-либо баз основ: только множество существующих суффиксов (при этом сложные составные суффиксы разбиваются на простые) и вручную заданные правила. Алгоритм состоит из пяти шагов. На каждом шаге отсекается словообразующий суффикс, и оставшаяся часть проверяется на соответствие правилам (например, для русских слов основа должна содержать не менее одной гласной). Если полученное слово удовлетворяет правилам, происходит переход на следующий шаг. Если нет – алгоритм выбирает другой суффикс для отсечения. Согласно официальному сайту проекта, на первом шаге отсекается максимальный формообразующий суффикс, на втором – буква «и», на третьем – словообразующий суффикс, на четвертом – суффиксы превосходных форм, «ь» и одна из двух «н». То, что стеммер Портера не использует никаких словарей и баз основ, является плюсом для быстрого действия и спектра применения (он неплохо справляется с несуществующими словами), и одновременно минусом с точки зрения точности выделения стеммы. Алгоритм часто обрезает слово больше необходимого, что затрудняет синтез нормальной формы по получающейся стемме: кровать -> кровя (при этом реально неизменяемая часть 8 слова – кроват, но стеммер обрезает наиболее длинную морфему) и не справляется с выпадающими гласными в корне: кошек -> кошек, кошками -> кошк. Кроме того, к минусам стеммера Портера часто относят человеческий фактор: то, что правила для проверки задаются вручную и иногда связаны с грамматическими особенностями языка, увеличивает вероятность ошибки.

1.3.2 Лемматизация

В отличие от стемминга, лемматизация представляет из себя инструмент, который использует морфологический анализ слов и словарь. Данный процесс

заключается в удалении флективных окончаний слов и возвращении словарной или базовой формы слова, которая определяется как лемма. Процесс лемматизации представлен в таблице 2.

Таблица 2 – Пример работы лемматизации

Форма	Морфологические данные	Лемма
Кошку	Сущ., вин. падеж, жен. род, ед. ч., одуш., нач. форма «кошка»	Кошка
Преподавателей	Сущ., род. падеж, муж. род, мн.ч., одуш., нач. форма «преподаватель»	Преподаватель

Процесс, представленный в таблице, рассматривает лишь одно слово, не учитывая контекст. Данный процесс является одним из видов лемматизации, и называется наивной лемматизацией. Одной из важных проблем данного вида нормализации – частые неоднозначности при определении частей речи.

Неоднозначность наивной лемматизации решается привлечением морфологического анализатора. Как правило, такая модель анализатора является вероятностной, и основана на машинном обучении. Такая модель вероятностной морфологии, как правило, обучается по специально размеченному корпусу данных. Увеличение модели точности, как правило, полностью определяется процессом обучения, либо, ограничениями самой модели.

```
#lemmatize
'Мы ели суп, а вдоль аллеи стояли раскидистые ели.'
['я', 'есть', 'суп', ',', 'а', 'вдоль', 'аллея', 'стоять', 'раскидистый', 'ель', '.']
```

Рисунок 2 – Пример работы лемматизатора с учетом морфологического контекста

На рисунке 2 представлен итог работы лемматизатора с учетом морфологического контекста. При анализе предложения, данный лемматизатор делает вывод, в каком контексте используется слово, например, слово «ели». В данном случае слово встречается дважды: как глагол, отражающий потребление пищи – «есть», и как существительное «ель» – дерево рода хвойных вечнозелёных.

1.4 Подходы к решению задачи автоматической классификации текстовых отзывов

Существуют два основных подхода к задаче автоматической классификации текстов – подход, основанный на использовании словарей и подход, основанный на машинном обучении.

1.4.1 Методы, основанные на машинном обучении

Обучение на размеченных данных или обучение с учителем – это наиболее распространенный класс задач машинного обучения. К нему относятся те задачи, где нужно научиться предсказывать некоторую величину для любого объекта, имея конечное число примеров [2].

В задаче автоматической классификации текстов с помощью машинного обучения используются заранее размеченные корпуса данных, на которых происходит обучение модели, которая в дальнейшем используется для классификации.

1.4.1.1 Наивный байесовский классификатор

Наивный байесовский классификатор — простой вероятностный классификатор, основанный на применении теоремы Байеса со строгими

(наивными) предположениями о независимости. Формула теоремы байеса представлена на формуле 1 [6].

$$P(H|E) = (P(E|H) * P(H)) / P(E) \quad (1)$$

где,

$P(H)$ – априорная вероятность события H ;

$P(E|H)$ – вероятность наступления E при истинности H ;

$P(E)$ – полная вероятность события E ;

$P(H|E)$ – вероятность события H при наступлении E .

Преимущества метода:

- высокая скорость работы;
- поддержка инкрементного обучения;
- относительно простая программная реализация алгоритма;
- легкая интерпретируемость результатов работы алгоритма.

Недостаток: относительно низкая качество классификации и неспособность учитывать зависимость результат классификации от сочетания признаков.

1.4.1.2 Метод опорных векторов

Основная идея метода опорных векторов состоит в оптимальной разделяющей гиперплоскости в пространстве признаков высокой размерности. Под оптимальностью понимается минимизация верхней оценки вероятности ошибки классификации [7].

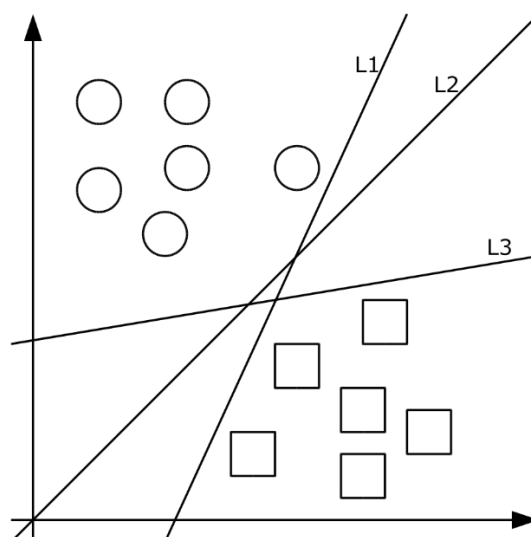


Рисунок 3 – Классифицирующие разделяющие прямые метода «SVM»

В качестве преимуществ метода можно выделить возможность работы с небольшим набором обучающих данных.

Преимущества метода:

- один из наиболее качественных методов;
- возможность работы с небольшим набором данных для обучения;
- сводимость к задаче выпуклой оптимизации, имеющей единственное решение.

Недостатки метода: сложная интерпретируемость параметров алгоритма и неустойчивость по отношению к выбросам в исходных данных.

1.4.1.3 Метод k-средних

Метод k-средних основан на разделении множества наблюдений X на k кластеров, которые локально минимизированы относительно расстояния между информационной точкой и центроидом кластера. Целевая функция алгоритма представлена следующей формулой [8]:

$$J = \sum_{h=1}^k \sum_{x_i \in x_h} \|x_i - \mu_h\|^2 \quad (2)$$

где,

μ_h - значение центроида.

Целевая функция является локально минимизированной для каждого кластера, что означает, что каждая точка из набора данных находится на минимальном расстоянии от центроида кластера, к которому она относится.

Преимущества метода:

- возможность обновления обучающей выборки без переобучения классификатора;
- устойчивость алгоритма к аномальным выбросам в исходных данных;
- относительно простая программная реализация алгоритма;
- легкая интерпретируемость результатов работы алгоритма;
- хорошее обучение в случае с линейно неразделимыми выборками.

Недостатки метода:

- репрезентативность набора данных, используемого для алгоритма;
- высокая зависимость результатов классификации от выбранной метрики;
- большая длительность работы из-за необходимости полного перебора обучающей выборки;
- невозможность решения задач большой размерности по количеству классов и документов.

1.4.1.4 Логистическая регрессия

В методе логистической регрессии для задачи классификации подбирается вектор коэффициентов $b = (w_1, w_2, \dots, w_n)^T$, который используется для осуществления классификации [10]:

$$f(x) = \frac{1}{1 + \exp(-(x, b))} = \sigma((x, b)) \quad (3)$$

Преимущества метода:

- является одним из наиболее качественных методов;
- поддерживает инкрементное обучение;
- имеет относительно простую программную реализацию алгоритма.

Недостатки метода: сложная интерпретируемость параметров алгоритма и неустойчивость по отношению к выбросам в исходных данных.

1.4.1.5 Дерево решений

Деревом решений называют ациклический граф, по которому производится классификация объектов (в нашем случае текстовых документов), описанных набором признаков. Каждый узел дерева содержит условие ветвления по одному из признаков [9]. У каждого узла столько ветвлений, сколько значений имеет выбранный признак. В процессе классификации осуществляются последовательные переходы от одного узла к другому в соответствии со значениями признаков объекта. Классификация считается завершенной, когда достигнут один из листьев (конечных узлов) дерева. Значение этого листа определит класс, которому принадлежит рассматриваемый объект. На практике обычно используют бинарные деревья решений, в которых принятие решения перехода по ребрам осуществляется простой проверкой наличия признака в документе. Если значение признака

меньше определенного значения, выбирается одна ветвь, если больше или равно, другая. В отличие от остальных подходов, представленных ранее, подход, использующий деревья решений, относится к символьным (то есть нечисловым) алгоритмам. Алгоритм построения бинарного дерева решений состоит из следующих шагов. Создается первый узел дерева, в который входят все документы, представленные всеми имеющимися признаками. Размер вектора признаков для каждого документа равен n , так как $d = (t_1, \dots, t_n)$. Для текущего узла дерева выбираются наиболее подходящий признак t_k и его наилучшее пограничное значение v_k . На основе пограничного значения выбранного признака производится разделение обучающей выборки на две части. Далее выбранный признак не включается в описание фрагментов в этих частях, то есть фрагменты в частях представляются вектором с размерностью $n - 1$. Образовавшиеся подмножества обрабатываются аналогично до тех пор, пока в каждом из них не останутся документы только одного класса или признаки для различения документов. Когда говорят о выборе наиболее подходящего признака, как правило, подразумевают частотный признак, то есть любой признак текста, допускающий возможность нахождения частоты его появления в тексте. Лучшим для разделения является признак, дающий максимальную на данном шаге информацию о категориях. Таким признаком для текста может являться, например, ключевое слово. С этой точки зрения любой частотный признак можно считать переменной. Тогда выбор между двумя наиболее подходящими признаками сводится к оценке степени связанности двух переменных. Поэтому для выбора подходящего признака на практике применяют различные критерии проверки гипотез, то есть критерии количественной оценки степени связанности двух переменных, поставленных во взаимное соответствие, где 0 соответствует полной независимости переменных, а 1 – их максимальной зависимости. Для исследования связи между двумя переменными удобно использовать представление совместного распределения этих переменных в виде таблицы сопряженности (факторной таблицы, или матрицы частот появления признаков). Она является наиболее

универсальным средством изучения статистических связей, так как в ней могут быть представлены переменные с любым уровнем измерения. Таблицы сопряженности часто используются для проверки гипотезы о наличии связи между двумя признаками при помощи различных статистических критериев: критерия Фишера (точного теста Фишера), критерия согласия Пирсона (критерия хи-квадрат), критерия Крамера, критерия Стьюдента (t-критерия Стьюдента) и пр.

Преимущества метода:

- относительно простая программная реализация алгоритма;
- легкая интерпретируемость результатов работы алгоритма.

Недостатки метода: неустойчивость алгоритма по отношению к выбросам в исходных данных и большой объем данных для получения точных результатов.

1.4.2 Метод, основанный на словаре тональностей

Существует три основных подхода к составлению словарей оценочной лексики:

- экспертный;
- на основе текстовых коллекций;
- на основе тезаурусов (словарей).

При экспертном подходе словарь составляется непосредственно экспертами. Данный подход отличается от остальных трудоемкостью, ввиду присутствия человеческого фактора, и высокой вероятностью отсутствия специфических для предметной области слов. Из плюсов данного подхода можно выделить высокое качество словаря относительно присвоенной тональности.

В подходе, основанном на текстовых коллекциях для компоновки словарей, применяется статистический анализ размеченных текстов. Как

правило, текста принадлежат предметной области, в которой составляются словари. При таком подходе нивелируется вероятность отсутствия специфических слов, принадлежащих рассматриваемой области. Впрочем, качество составленного словаря напрямую зависит от качества размеченных текстов.

При подходе на основе текстовых тезаурусов, имеющийся список слов расширяется с привлечением различных словарей, например с помощью словаря синонимов или антонимов. Из минусов данного подхода можно выделить связи формируемых словарей с предметной областью.

Глубоко аннотированный корпус русского языка («СинТагРус») – один из первых аннотированных корпусов русского языка, который разрабатывается с 1998 года [11]. На текущий момент в корпусе присутствуют следующие жанры:

- биографии;
- публицистика;
- журнальные и газетные статьи периода с 1960 по современное время;
- современная научно-популярная литература;
- художественная проза 20 века;
- различные новостные ленты.

Особенность корпуса «СинТагРус» заключается в нескольких уровнях аннотаций разной глубины. Данные из уровней извлекаются независимо, и в теории, количество таких уровней не ограничено. Язык разметки – XML.

```
<S CLASS="LF" COREF="(2:его,-2;3:Юлия);" ID="23" MICROSYNТ="(в присутствии,{6:в...7:присутствии});">
<W DOM="3" FEAT="ADV" ID="1" KSNAME="СНАЧАЛА" LEMMA="СНАЧАЛА" LINK="обст">Сначала</W>
<W DOM="3" FEAT="S ЕД МУЖ ВИН ОД" ID="2" KSNAME="ОН" LEMMA="ОН" LINK="1-компл">его</W>
<W DOM="_root" FEAT="V НЕСОВ ИЗЪЯВ ПРОШ МН" ID="3" KSNAME="ПОИТЬ" LEMMA="ПОИТЬ">поили</W>
<W DOM="3" FEAT="S ЕД МУЖ ТВОР НЕОД" ID="4" KSNAME="ЧАЙ1" LEMMA="ЧАЙ" LINK="2-компл">чаем</W>,
<W DOM="6" FEAT="ADV" ID="5" KSNAME="НЕПРЕМЕННО" LEMMA="НЕПРЕМЕННО" LINK="огранич">непременно</W>
<W DOM="3" FEAT="PR" ID="6" KSNAME="В2" LEMMA="В" LINK="обст">в</W>
<W DOM="6" FEAT="S ЕД СРЕД ПР НЕОД" ID="7" KSNAME="ПРИСУТСТВИЕ" LEMMA="ПРИСУТСТВИЕ" LINK="предл">присутствии</W>
<W DOM="7" FEAT="S ЕД ЖЕН РОД ОД" ID="8" KSNAME="ПАЦИЕНТКА" LEMMA="ПАЦИЕНТКА" LINK="квaziагент">пациентки</W>.
<LF LFARG="7" LFFUNC="_ADV2-UN" LFVAL="6"/>
</S>
```

Рисунок 4 – Разметка экземпляра корпуса «СинТагРус»

Каждый экземпляр корпуса разбит на предложения. Каждое предложение, в свою очередь, разбито на отдельные слова. Пример представлен на рисунке 4.

В корпусе содержит следующие типы разметок:

— анафорическая разметка – для каждого анафорического местоимения указан антецедент;

— микросинтаксическая разметка – разметка фразеологизмов с синтаксической спецификой;

— лексико-функциональная разметка – разметка путем указаний словосочетаний, интерпретируемых в терминах лексических функций;

— лексико-семантическая разметка – указание значений соответствующих статей толково-комбинаторного словаря;

— синтаксическая разметка – разметка в рамках грамматики зависимостей;

— морфологическая разметка – разметка частей речи и морфологических характеристик.

Так же, из особенностей корпуса «СинТагРус» можно выделить то, что опущенные фрагменты эллиптированных предложений восстанавливаются явно.

В контексте исследования, применяется подход, основанный на текстовых коллекциях. Данный метод был описан Peter D. Turney в 2002 году [28]. Идея заключается в следующем: каждое слово текста рассматривается на наличие в словаре, как слово, несущее положительную или отрицательную тональность. Если слово встречается в словаре, как слово с положительным весом, то счетчик для слов, несущих положительную тональность, увеличивается. Аналогично счетчик для слов, несущих отрицательную тональность, увеличивается, если слово имеет отрицательный вес.

Тональность текста определяется большим количеством того или иного счетчика.

2 Исследование и выбор методов классификации текстов

Исследование выполнено на множестве отзывов студентов Института космических и информационных технологий (ИКИТ) Сибирского федерального университета, собираемых и хранимых в персональной образовательной среде института [6]. Анкетирование студентов проводится регулярно, начиная с 2015 года после каждой промежуточной аттестации. Исходный корпус состоит из 4 900 отзывов студентов. Всего вопросов 16, и они сгруппированы по трем аспектам образовательной среды:

- отношение студентов к работе выпускающей кафедры;
- отношение студентов к работе учебно-организационного отдела (деканата);
- отношение студентов к преподаванию дисциплин и обеспечивающих их материалов.

Как правило студенты оставляют отрицательные текстовые отзывы, а для положительных оценок ограничиваются только балами.

Но не все оценки студентов и ответы на вопросы анкет можно использовать как достоверную информацию. Проблема релевантности оценок тесно связана с эмоциональным фактором, субъективным отношением.



Рисунок 5 – Соотношение отзывов сервиса анкетирования в тыс.

На рисунке 5 представлено соотношение отзывов сервиса анкетирования. Процент отзывов с комментарием составляет 7,25 % от общего числа отзывов. На следующем рисунке (рис. N) представлено процентное соотношение оставленных отзывов, с учетом числовой оценки.

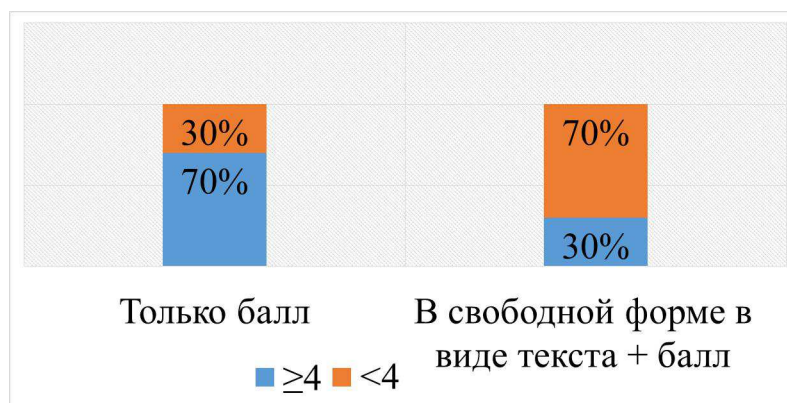


Рисунок 6 – Процентное соотношение текстовых отзывов с учетом числовой оценки

На основе рисунка 6 можно сделать вывод о том, что при отрицательной оценке обучающиеся чаще оставляют текстовый комментарий.

Анализ текстовых ответов позволяет наиболее эффективно осуществлять обратную связь на отзыв студента. Автоматический анализ текстов позволит применить дополнительные критерии оценки релевантность ответов студентов наравне с статистическим анализом балльных оценок.

Для автоматизации анализа отзывов необходимо получать численные оценки текста. Для текстовых данных предлагается два критерия оценки релевантности ответа:

- соответствие текста отзыва словарю вопроса;
- соответствие эмоциональных оттенков отзыва числовой оценке.

Оценку соответствия предлагается выполнять с помощью классификатора. Предполагаемый алгоритм расчета данного показателя предлагается организовать следующим способом: при поступлении нового

ответа из него выделяются значимые наборы слов и вычисляются степени соответствия этих наборов слов перечню вопросов. Пример указан в таблице 3.

Таблица 3 – Критерий соответствия текста отзыва словарю вопроса

№	Вопрос	Отзыв	Прогноз	Кр.
1	Предложения по улучшению преподавания	Если ты болел и принес справку, она не учитывается и все пропуски приходится отрабатывать	Уровень сервиса по работе с заявлениями студентов, выдачи справок	0,55
2	дисциплины	Хотелось бы еще и заниматься говорением, а не только работать в эк	Полнота и качество электронного образовательного ресурса	0,29

Для определения показателя тональности, в свою очередь, предлагается сопоставлять наборы слов с частотными тональными словарями аспектов.

По итогу, в результате автоматически возможно выделить случаи, когда набор слов ответа студента не соответствует типичным словосочетаниям для данного вопроса или балльная оценка ответа не соответствует эмоциональному окрасу текста.

Для машинного обучения необходимо сформировать обучающие выборки (словари) с информацией о соответствии слов и вопросов, а также указать для ответов тональность (положительную, отрицательную, нейтральную).

При формировании словарей, как правило, применяется стемминг. Стемминг не требователен к ресурсам, но может совершенно разные по смыслу слова привести к одной форме. Сами словари в этом случае представляют собой наборы букв и не позволяют интерпретировать результаты. Альтернативой является лемматизация текста (приведение

различных словоформ к базовому виду с использованием морфологических словарей и алгоритмов контекстного анализа). В общем случае процесс лемматизации крайне ресурсоемкий. Так как все ответы рассматриваются в одном контексте (в сфере образования), то для данной задачи удалось составить единый словарь, который почти безошибочно позволяет определять базовую форму слов.

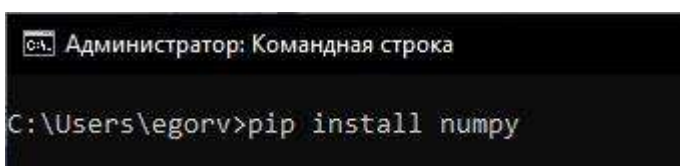
В качестве лемматизатора, в контексте исследований, используется библиотека «TreeTagerWrapper», которая официально зарегистрирована в PyPI – каталоге программного обеспечения Python [12].

Исследования были проведены, используя объектно-ориентированный язык программирования Python 3.8. Так же использовались подключаемые библиотеки, список которых представлен ниже:

- csv – модуль для структурного анализа и обработки файлов в формате .csv;
- openpyxl – модуль для создания и работы с рабочими книгами xlsx/xlsm/xltx/xltn;
- pickle – модуль сериализации и десериализации объектов Python;
- sklearn – модуль машинного обучения;
- os – модуль для работы с операционной системой;
- sys – модуль для взаимодействия с интерпретатором;
- datetime – модуль для обработки времени и даты различными способами;
- nltk – пакет библиотек для символьной и статистической обработки естественного языка;
- xml – стандартный модуль для парсинга файлов XML формата;
- re – стандартный модуль для работы с регулярными выражениями;
- treetaggerwrapper – модуль для взаимодействия со словарями лемматизации;

- `numpy` – модуль, добавляющий поддержку больших многомерных массивов и матриц;
- `statistics` – модуль описательной статистики;
- `pylab` – модуль для визуализации двухмерных и трехмерных данных;
- `random` – модуль для генерации случайных чисел.

Каждая из библиотек (python-пакетов) подключается посредством системы управления пакетами `pip` и командной консоли Windows 10. Пример подключения представлен на рисунке 7.



```
Администратор: Командная строка
C:\Users\egorv>pip install numpy
```

Рисунок 7 – Пример установки python-пакета «Numpy»

2.1 Анализ и сравнение методов машинного обучения при расчете критерия принадлежности

2.1.1 Предобработка данных

Так как исходные данные представляют собой необработанный текст, была проведена предварительная обработка данных.

Прежде всего, используя библиотеку регулярных выражений `re`, из текста были исключены все знаки пунктуации, хештеги, гиперссылки.

Далее, данные выборки были очищены от стоп-слов и знаков пунктуации, а также переведены в нижний регистр. Для очистки была использована библиотека для обработки естественного языка - NLTK, которая содержит 151 стоп-слов [4].

После очистки текста, необходимо привести слова в исходном тексте к некой канонической форме. Для этого были проанализированы и использованы два алгоритма – стемминг и лемматизация.

В качестве алгоритма стеммера был взят SnowballStemmer все той же библиотеки NLTK [nltk.stem.SnowballStemmer('russian')]. Для работы с текстом создается экземпляр класса SnowballStemmer [4].

Алгоритм стемминга содержит следующие принципы:

— Поиск и удаление окончаний, свойственных деепричастиям («в, вши, вшись, ыв, ывши»). Если такие не найдены, производится поиск возвратного постфикса («ся, съ») и его удаление, затем таким же образом ищутся и удаляются окончания, свойственные причастиям, прилагательным или существительным.

— если слово заканчивается на букву «и», удалить её;

— если окончание слова относится к превосходной степени («ейш, ейше»), оно отсекается;

— замена двойного «н» на одинарное;

— удаление «ь», если он стоит в конце слова.

В качестве лемматизатора, в свою очередь, был рассмотрен инструмент TreeTaggerWrapper. TreeTagger применим к множеству языков, в том числе и к русскому. Перед подключением библиотеки к проекту установлен пакет TreeTagger, к нему подключены необходимые скрипты и файл параметров языка. Лемматизация слов осуществляется с помощью объекта класса TreeTaggerWrapper, одним из основных методов работы с которым является процедура tag_text() [3].

2.1.2 Машинное обучение

Для точного определения работы моделей машинного обучения, было принято решение о тестировании путем постепенного увеличения среднего

количества слов в предложениях. Всего было произведено 50 итерации тестирования для 20 выборок с разными показателями среднего количества слов в предложениях.

Результаты сравнения эффективности работы трех классификаторов представлены таблице 4, на рисунке 8 (предварительный стемминг текста) и рисунке 9 (предварительная лемматизация текста).

Таблица 4 – Основные показатели классификации текстовых отзывов

Метод классификации	Доля правильных решений		
	Минимум	Максимум	Среднее
Стемминг			
Наивный байесовский классификатор	0.34	0.7	0.52
Метод к-средних	0.01	0.18	0.09
Метод опорных векторов	0.68	0.95	0.77
Логистическая регрессия	0.55	0.85	0.72
Дерево решений	0.54	0.93	0.70
Лемматизация			
Наивный байесовский классификатор	0.34	0.68	0.52
Метод к-средних	0.01	0.12	0.05
Метод опорных векторов	0.66	0.96	0.77
Логистическая регрессия	0.57	0.87	0.72
Дерево решений	0.54	0.92	0.70

Метод опорных векторов проявил себя одинаково эффективно как для стемминга, так и для наивной лемматизации.

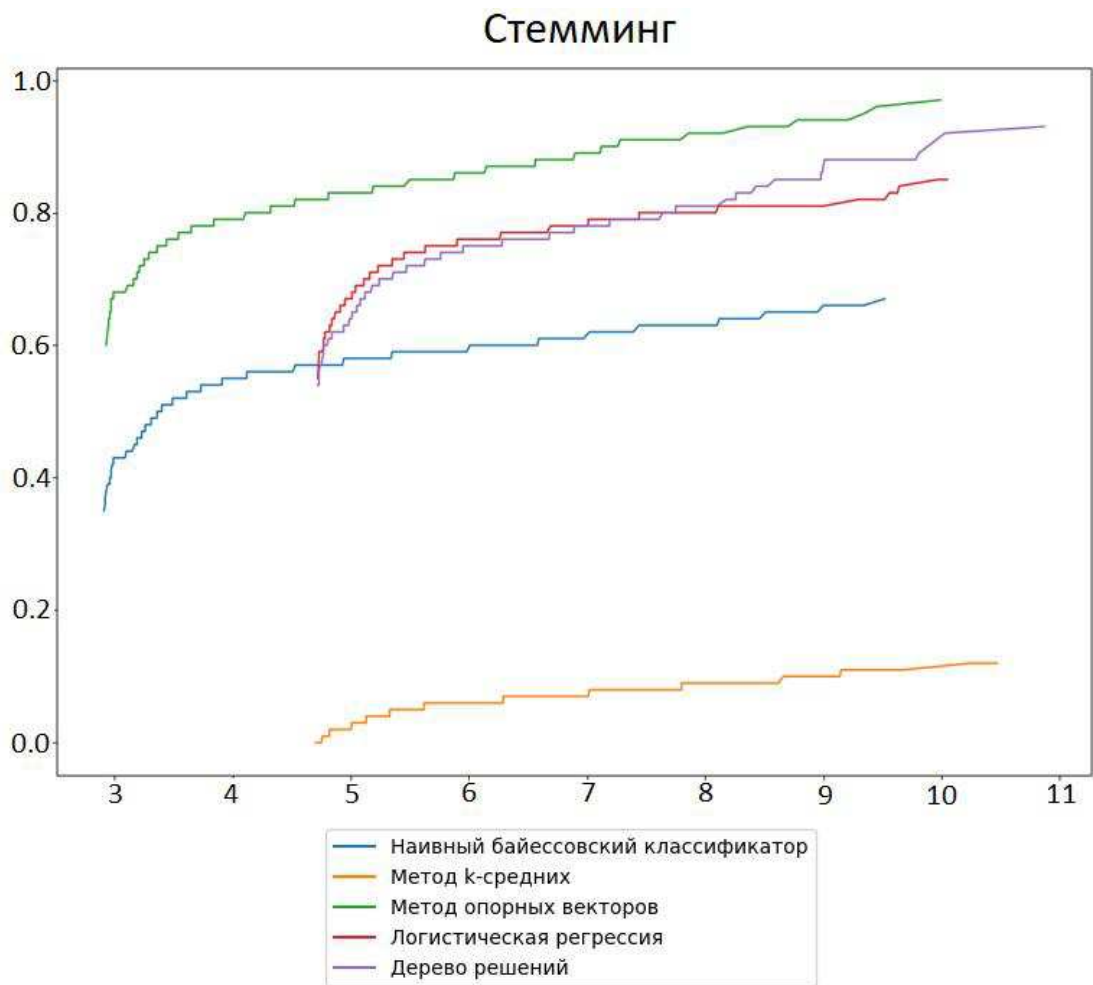


Рисунок 8 – Результаты классификаторов при предварительном стемминге текста

Лемматизация

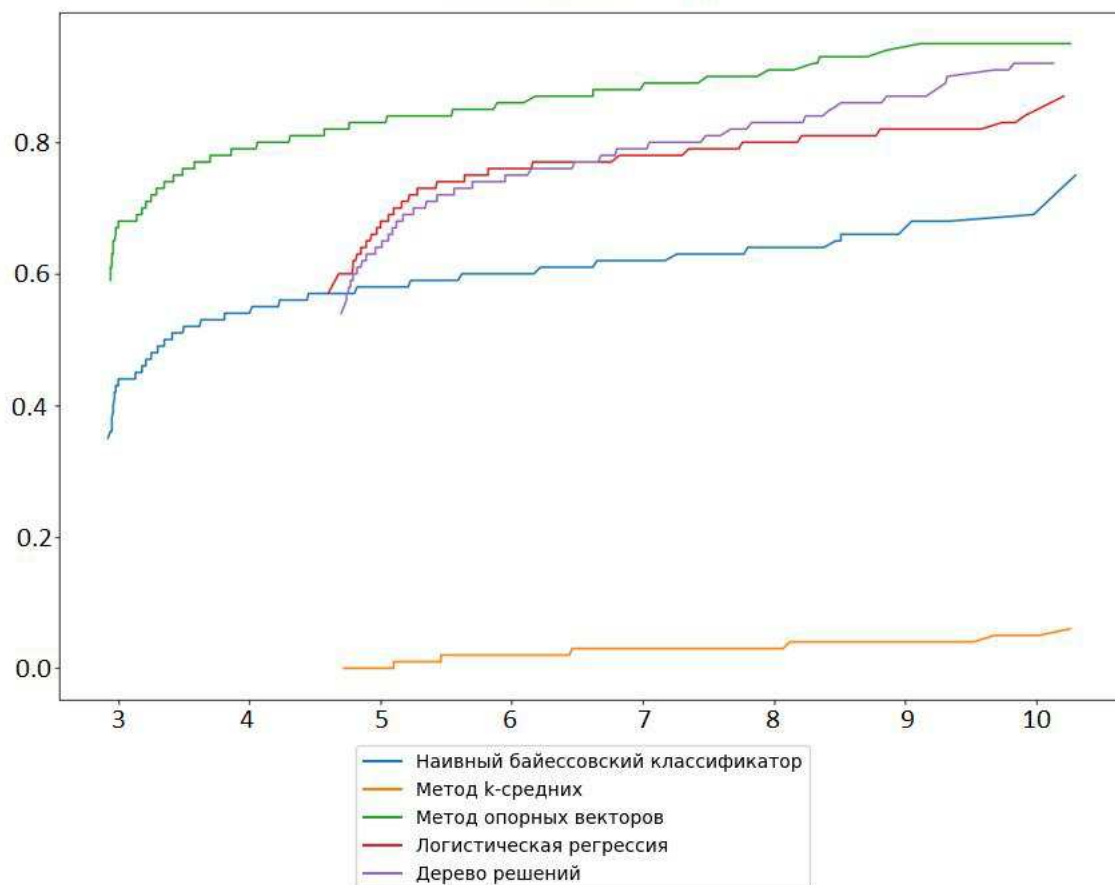


Рисунок 9 – Результаты классификаторов при предварительной лемматизации текста

2.2 Анализ и сравнение методов нормализации при расчете критерия тональной принадлежности словарным подходом

В качестве тонального словаря был взят открытый словарь русского языка, состоящий из 28197 слов [5]. Структура словаря представляет собой размеченные на положительные, нейтральные и отрицательные значения эмоционального окраса слов в диапазоне от 3 до -3. Так, же для удобства присутствует соответствующий числовому значения тэг: PSTV (позитивное), NEUT (нейтральное), NGTV (негативное). Наглядная структура словаря представлена в таблице 5.

Таблица 5 – Структура тонального словаря

Term	Tag	Value
счастливая	PSTV	3.0
заученный	NEUT	0.519
напиваться	NGTV	-3.0

Для оценки эффективности определения эмоционального окраса тестовых данных с предварительным стеммингом, был создан еще один тональный словарь, который был обработан теми же средствами, что и при нормализации данных при тестировании.

Тест применения тональных словарей русского языка показал значительное преимущество наивной лемматизации (рисунок 10). Нормализация текста посредством стемминга приводит к частым ошибочным в словаре.

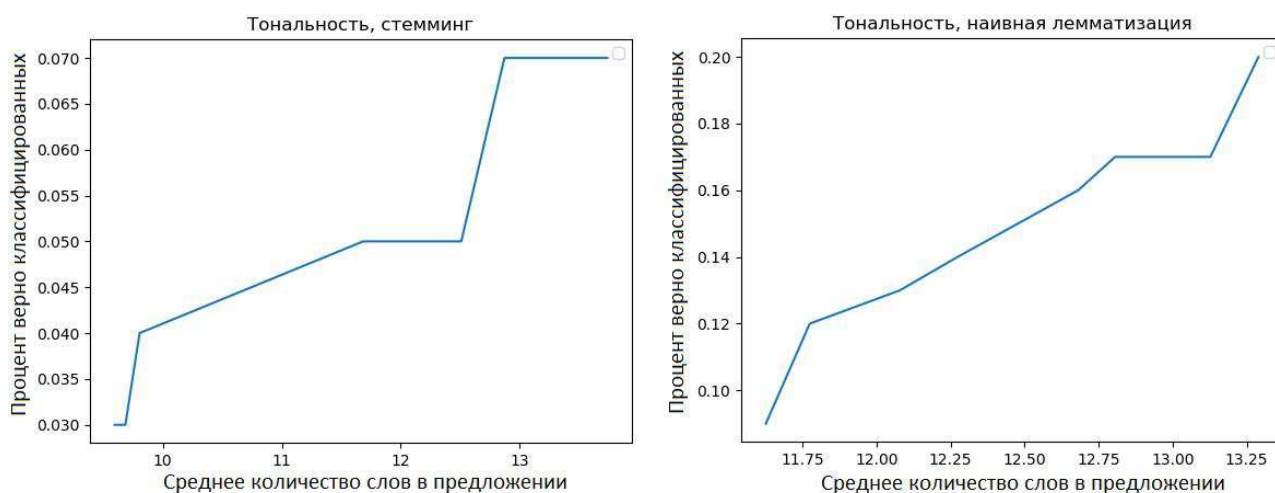


Рисунок 10 – Результаты определения тональности текстовых отзывов

Исходя из графиков, представленных на рисунке 4, можно сделать вывод о том, что наиболее эффективным методом нормализации является лемматизация. Точность определения эмоциональной окраски текста прямо пропорционально зависит от количества слов в предложении.

По итогу, результаты проведенного исследования позволяют сделать вывод о том, какой набор технологий для классификации текстовых отзывов и расчете эмоциональной окраски являются оптимальными. В задаче определения критерия принадлежности наиболее эффективно показывает себя совокупность метода опорных векторов с лемматизацией. В задаче определения тональности метод основанный на тональных словарях с предварительной лемматизацией текста, так же, показал себя весьма успешно.

3 Программный модуль расчета критериев оценки релевантности текстовых отзывов

3.1 Описание программного модуля

В качестве данных была взята коллекция размеченных данных, разбитых по источникам. В каждой из тем присутствует разбитый на предложения текст. Каждое из предложений, в свою очередь, токенизировано на слова и знаки пунктуации.

Коллекция представлена в виде xml-кода, отрывок которого представлен на рисунке 11:

```
<sentence id="31274">
<source>Ах, дурни, прости господи!</source>
<tokens>
<token id="559691" text="Ах"><tfr rev_id="1392773" t="Ах"><v><l id="20493" t="ах"><g v="INTJ"/></l></v></tfr></token>
<token id="559692" text=","><tfr rev_id="1392774" t=","><v><l id="0" t=","><g v="PNCT"/></l></v></tfr></token>
<token id="559693" text="дурни"><tfr rev_id="1392775" t="дурни"><v><l id="93159" t="дурень"><g v="NOUN"/><g v="anim"/><g v="masc"/><g v="plur"/><g v="nomn"/></l></v></tfr></token>
<token id="559694" text=","><tfr rev_id="1392776" t=","><v><l id="0" t=","><g v="PNCT"/></l></v></tfr></token>
<token id="559695" text="прости"><tfr rev_id="1392777" t="прости"><v><l id="285009" t="простил"><g v="VERB"/><g v="perf"/><g v="tran"/><g v="sing"/><g v="impr"/><g v="excl"/></l></v></tfr></token>
<token id="559696" text="господи"><tfr rev_id="1392778" t="господи"><v><l id="74056" t="господь"><g v="NOUN"/><g v="anim"/><g v="masc"/><g v="Sgtm"/><g v="sing"/><g v="voct"/></l></v></tfr></token>
<token id="559697" text="!"><tfr rev_id="1392779" t="!"><v><l id="0" t="!"><g v="PNCT"/></l></v></tfr></token>
</tokens>
</sentence>
<sentence id="31276">
```

Рис. 11 – Фрагмент XML-кода исходных данных

После предварительного анализа корпуса данных, было определено что корпус содержит 4070 источников, 110129 предложений, 1992129 слов.

Модуль бы написан с использованием высокоуровневого языка программирования Python, версии 3.8. Библиотеки, которые используются для работы модуля:

- pickle – модуль сериализации и десериализации объектов Python;
- sklearn – модуль машинного обучения;
- os – модуль для работы с операционной системой;
- sys – модуль для взаимодействия с интерпретатором;

- `datetime` – модуль для обработки времени и даты различными способами;
- `nltk` – пакет библиотек для символьной и статистической обработки естественного языка;
- `xml` – стандартный модуль для парсинга файлов XML формата;
- `re` – стандартный модуль для работы с регулярными выражениями.

Для удобства сопровождения программного модуля, программа была спроектирована модульным способом. Структура программы представлена на рисунке 12.

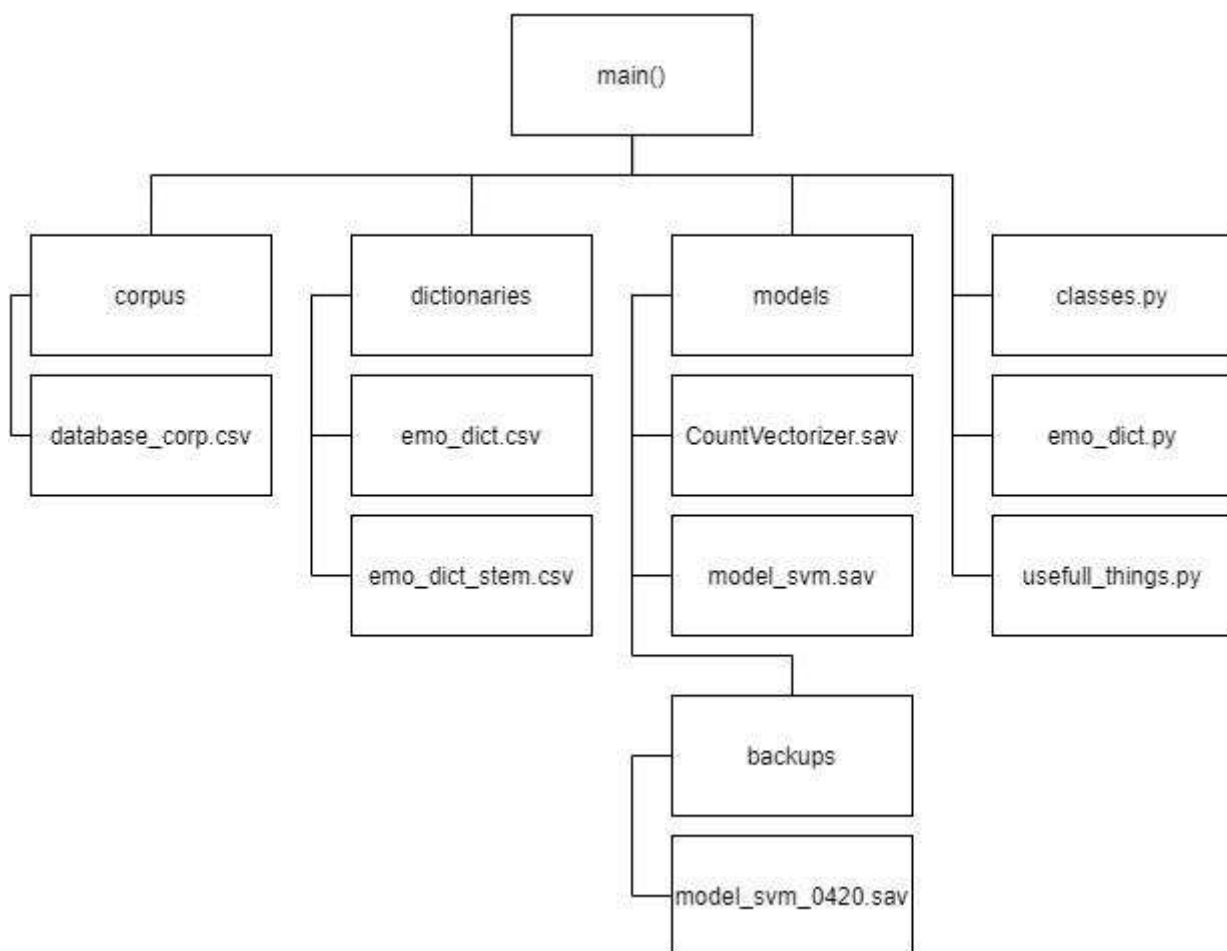


Рисунок 12 – Архитектура модуля

Модуль `main()` содержит в себе основной код, который связывает все остальные модули. Модуль `classes.py` содержит в себе класс `machine_model`,

экземпляр которого создается при расчете критерия соответствия вопросу или при переобучении модели. Функции класса `machine_model` способны сохранять или читать модель из директории `models`, переобучать текущую модель и делать резервные копии в директорию `models\backups`.

Сегмент `emo_dict.py`, в свою очередь, взаимодействует со словарями в директории `dictionaries`, и отвечает за расчет эмоционального окраса текстового отзыва.

Часть `usefull_things.py` отвечает за нормализацию исходных данных. Данный модуль содержит функции, очищающие текст от знаков пунктуации и стоп слов, а также функции для стемминга и лемматизации.

Помимо лемматизации и стемминга, функция нормализации содержит список регулярных выражений, посредством которых осуществляется очистка данных. Список выражений представлен на рисунке 13:

```
text = text.lower()
text = re.sub( r'https?://[\S]+', ' ', text)
text = re.sub( r'[\w\./]+\.[a-z]+', ' ', text)
text = re.sub( r'\d+[-/\.\]\d+[-/\.\]\d+', ' ', text)
text = re.sub( r'\d+ ?rr?', ' ', text)
text = re.sub( r'\d+:\d+(\:\d+)?', ' ', text)
text = re.sub( r'@\w+', ' |', text )
text = re.sub( r'#\w+', ' ', text )
text = re.sub( r'<[^>]*>', ' ', text)
text = re.sub( r'[\W]+', ' ', text )
```

Рисунок 13 – регулярные выражения для очистки текста

Каждое из выражений отвечает за определенные символы в тексте, например, выражение `r'#\w+'` позволяет убрать из данных все хештеги.

В целом, общий принцип функционирования программы представлен на блок-схеме N.

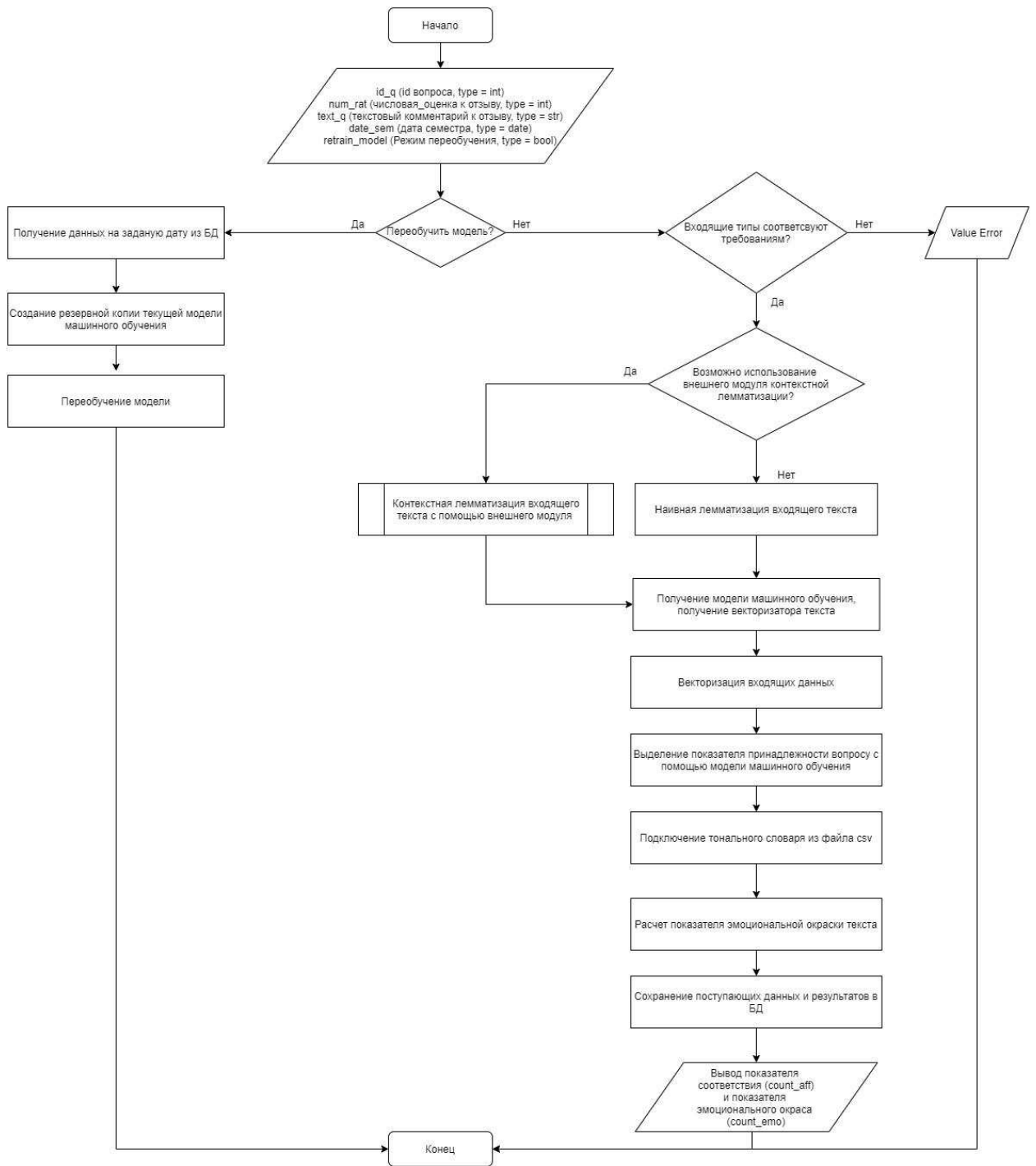


Рисунок 14 – Блок-схема модуля оценки релевантности текстовых отзывов

3.2 Апробация программного модуля

Для оценки качества построения алгоритма, необходимо определить численную оценку его качества.

Для этого, было принято решение воспользоваться следующими метриками: точность и полнота.

Точность (precision) – часть текстовых отзывов, которые действительно принадлежат вопросу (классу), которая модель отнесла к этому вопросу.

Полнота (recall), в свой черед, это часть найденных классификатором текстовых отзывов, принадлежащих вопросу относительно всех текстовых отзывов рассматриваемого вопроса в тестовой выборке.

Значения можно рассчитать, воспользовавшись таблицей контингентности:

Таблица N – Таблица контингентности

Категория i		Экспертная оценка	
		положительная	отрицательная
Оценка системы	положительная	TP	FP
	отрицательная	FN	TN

Формулы точности и полноты выглядят следующим образом:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

где,

TP – истинно-положительное значение;

FP – ложно-положительное значение;

FN – ложно-отрицательное значение;

TN – истинно-отрицательное значение.

Высокая точность и полнота отражают факт качественного построения модели.

Следующая не менее важная метрика – f-мера. Данный показатель объединяет данные о точности и полноте алгоритма, и представляет собой гармоническое среднее между точностью и полнотой. Значение метрики прямо пропорционально зависит от значений полноты и точности, например, если полнота и точность стремятся к нулю, то и значение f-меры стремится к нулю. Формула, вычисляющая f-меру представлена следующим образом:

$$F = \frac{2*Precision*Recall}{Precision+Recall} \quad (6)$$

где,

Precision – значение точности модели;

Recall – значение полноты модели.

Для каждого критерия из показателей оценки релевантности было принято решение о проверке всех методов нормализации, то есть стемминга, наивной лемматизации, контекстной лемматизации. Для более точного похода к проверке, выборка была сгруппирована по среднему количеству слов в предложениях, и случайно составлена из 1000 случайных отзывов. Всего таких выборок получилось 13.

Таблица 6 – Результаты расчета критерия соответствия при стемминге текста при среднем количестве слов от двух слов до семи

ср. слов	2	3	4	5	6	7
precision	0,6410	0,6920	0,7030	0,7240	0,7650	0,7890
recall	0,6284	0,6784	0,6892	0,7098	0,7500	0,7735
f-мера	0,6347	0,6851	0,6960	0,7168	0,7574	0,7812

Таблица 7 – Результаты расчета критерия соответствия при стемминге текста при среднем количестве слов от восьми слов до тринадцати

ср. слов	8	9	10	11	12	13
precision	0,8700	0,8900	0,8990	0,9040	0,9050	0,9500
recall	0,8529	0,8725	0,8814	0,8863	0,8873	0,9314
f-мера	0,8614	0,8812	0,8901	0,8950	0,8960	0,9406

Таблица 8 – Результаты расчета критерия соответствия при наивной лемматизации текста при среднем количестве слов от двух слов до семи

ср. слов	2	3	4	5	6	7
precision	0,6660	0,6830	0,7010	0,7500	0,7680	0,7400
recall	0,6529	0,6696	0,6873	0,7353	0,7529	0,7255
f-мера	0,6594	0,6762	0,6941	0,7426	0,7604	0,7327

Таблица 9 – Результаты расчета критерия соответствия при наивной лемматизации текста при среднем количестве слов от восьми слов до тринадцати

ср. слов	8	9	10	11	12	13
precision	0,8770	0,8910	0,8990	0,9040	0,9680	0,9870
recall	0,8598	0,8735	0,8814	0,8863	0,9490	0,9676
f-мера	0,8683	0,8822	0,8901	0,8950	0,9584	0,9772

Таблица 10 – Результаты расчета критерия соответствия при контекстной лемматизации текста при среднем количестве слов от двух слов до семи

ср. слов	2	3	4	5	6	7
precision	0,5680	0,5770	0,6040	0,6480	0,7020	0,7400
recall	0,5569	0,5657	0,5922	0,6353	0,6882	0,7255
f-мера	0,5624	0,5713	0,5980	0,6416	0,6950	0,7327

Таблица 11 – Результаты расчета критерия соответствия при контекстной лемматизации текста при среднем количестве слов от восьми слов до тринадцати

ср. слов	8	9	10	11	12	13
precision	0,9300	0,9590	0,9870	0,9960	0,9960	0,9990
recall	0,9118	0,9402	0,9676	0,9765	0,9765	0,9794
f-мера	0,9208	0,9495	0,9772	0,9861	0,9861	0,9891

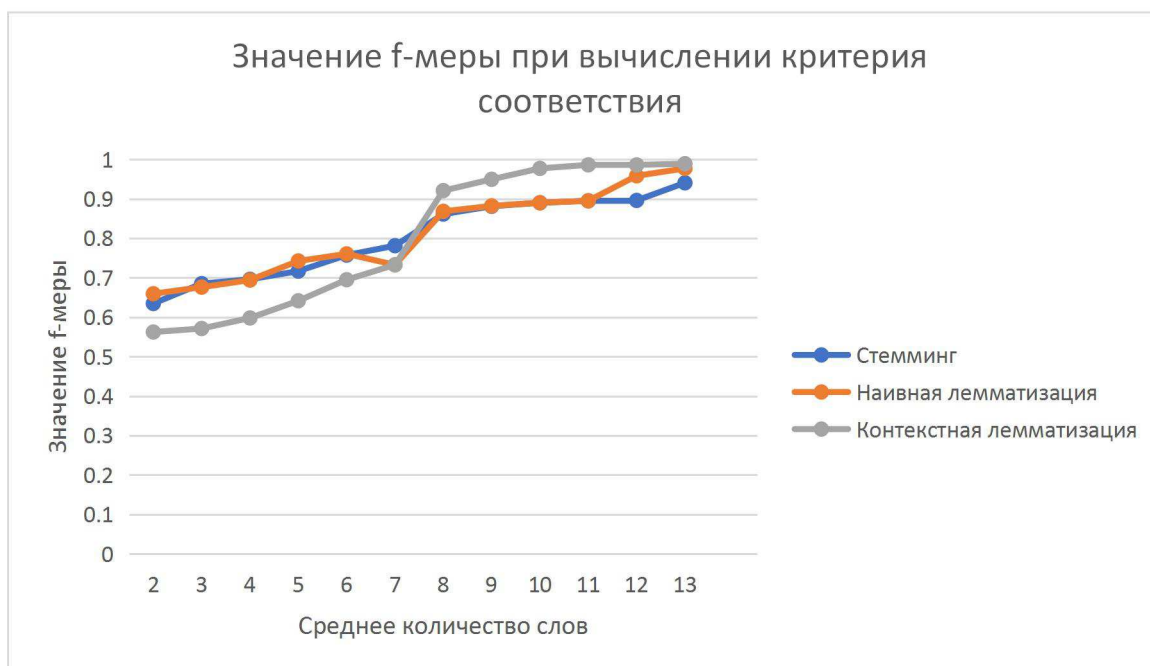


Рисунок 15 – Значение f-меры при вычислении критерия соответствия

На основе вышеперечисленных таблиц был построен график (рис. 15) зависимости значения показателя f-меры от среднего количества слов в экземпляре выборки. На основе графика можно понять, что самый эффективный способ нормализации текста – контекстная лемматизация. Эффект от нормализации данным способом достигается лишь при тех условиях, когда количество слов в анализируемом тексте превышает семи. Данный факт связан с тем, что при анализе контекста, во многих случаях, важен именно объем текста. Исходя из этого, можно сделать вывод о том, что при нормализации текста, преобладает вариант наивной лемматизации.

Следующий этап апробации программы, проверка качества расчета эмоционального окраса текстовых отзывов. Для определения эффективности, так же, воспользуемся методом вычисления f-меры. Таблицы показателей для каждого из трех методов нормализации представлены ниже.

Таблица 12 – Результаты расчета эмоционального окраса отзыва при стемминге текста при среднем количестве слов от двух слов до семи

ср. слов	2	3	4	5	6	7
precision	0,2410	0,3420	0,3590	0,4020	0,3710	0,5310
recall	0,2363	0,3353	0,3520	0,3941	0,3637	0,5206
f-мера	0,2386	0,3386	0,3554	0,3980	0,3673	0,5257

Таблица 13 – Результаты расчета эмоционального окраса отзыва при стемминге текста при среднем количестве слов от восьми слов до тринадцати

ср. слов	8	9	10	11	12	13
precision	0,6030	0,7210	0,8020	0,7230	0,8020	0,8510
recall	0,5912	0,7069	0,7863	0,7088	0,7863	0,8343
f-мера	0,5970	0,7139	0,7941	0,7158	0,7941	0,8426

Таблица 14 – Результаты расчета эмоционального окраса отзыва при наивной лемматизации текста при среднем количестве слов от двух слов до семи

ср. слов	2	3	4	5	6	7
precision	0,3040	0,4050	0,4220	0,4650	0,4340	0,5940
recall	0,2980	0,3971	0,4137	0,4559	0,4255	0,5824
f-мера	0,3010	0,4010	0,4178	0,4604	0,4297	0,5881

Таблица 15 – Результаты расчета эмоционального окраса отзыва при наивной лемматизации текста при среднем количестве слов от восьми слов до тринадцати

ср. слов	8	9	10	11	12	13
precision	0,6660	0,7840	0,8650	0,7860	0,8650	0,8990
recall	0,6529	0,7686	0,8480	0,7706	0,8480	0,8814
f-мера	0,6594	0,7762	0,8564	0,7782	0,8564	0,8901

Таблица 16 – Результаты расчета эмоционального окраса отзыва при контекстной лемматизации текста при среднем количестве слов от двух слов до семи

ср. слов	2	3	4	5	6	7
precision	0,2080	0,4980	0,5150	0,5580	0,6400	0,6870
recall	0,2039	0,4882	0,5049	0,5471	0,6275	0,6735
f-мера	0,2059	0,4931	0,5099	0,5525	0,6337	0,6802

Таблица 17 – Результаты расчета эмоционального окраса отзыва при контекстной лемматизации текста при среднем количестве слов от восьми слов до тринадцати

ср. слов	8	9	10	11	12	13
precision	0,8090	0,8770	0,9580	0,9790	0,9980	0,9590
recall	0,7931	0,8598	0,9392	0,9598	0,9784	0,9402
f-мера	0,8010	0,8683	0,9485	0,9693	0,9881	0,9495

На основе полученных результатов был построен график, тек же, отражающий зависимость f-меры от количества слов в тексте отзыва.

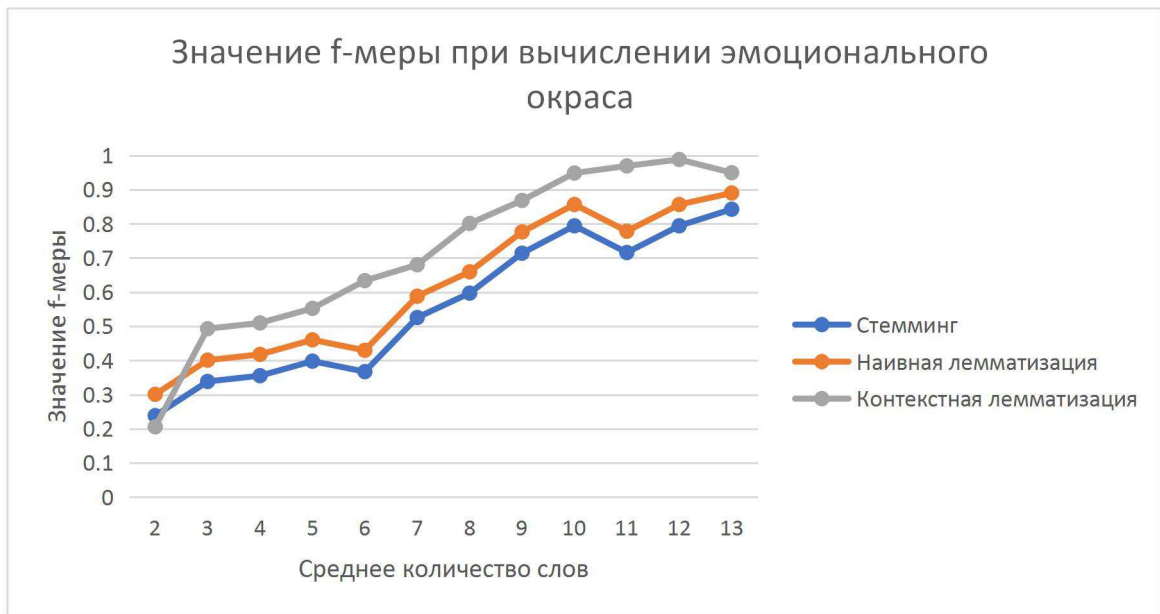


Рисунок 16 – Значение f-меры при вычислении эмоционального окраса

На основе данных графика (рис. 16) можно сделать вывод о том, что, так же, наиболее эффективным методом оказалась наивная лемматизация. По графику отслеживается, что эффективность расчета эмоционального окраса текста в большей степени зависит от количества слов, нежели при расчете критерия соответствия.

На следующем этапе апробации программного модуля необходимо проверить эффективность построенного алгоритма, основываясь на реальных отзывах студентов. Было принято решение о том, чтобы в качестве обучающей и тестовой выборки взять данные за разные семестры. Структура данных построена следующим образом:

Таблица 18 – Структура данных отзывов

text_id	text	class_fk	score
86523	Не знаю о ней	5	3
87409	Не интересовался	5	2
88790	Нет куратора	10	1

Каждый текстовый ответ сопровождается уникальным идентификатором – text_id, вопрос, к которому был оставлен текстовый комментарий – class_fk, и числовая оценка – score.

Как условность, определим названия выборкам. Данные для обучения – данные за первый семестр, соответственно, данные для проверки – данные за второй семестр.

В таблице 19 представлены обучающие данные, сгруппированные по вопросам.

Таблица 19 – Обучающие данные

Вопрос	Количество
Доброжелательность и корректность сотрудников отдела	24
Доступность изложения теоретического материала на лекционных занятиях	118
Качество проведения практических занятий	82
На сколько Вы удовлетворены организацией и проведением практик (учебной, производственной и др.)	36
Насколько хорошо Вы знакомы с деятельностью заведующего кафедрой, на которой обучаетесь?	75
Оцените Ваше участие в деятельности кафедры, на которой обучаетесь	81
Оцените работу Вашего куратора	69
Полнота и качество электронного образовательного ресурса	148
Удобство графика работы со студентами	41
Уровень ведения разъяснительной работы (консультирования) сотрудниками отдела	21
Уровень сервиса по работе с заявлениями студентов, выдачи справок	21

Всего	716
--------------	------------

Данные по тестовым отзывам имеют то же количество вопросов (классов).

Таблица 20 – Обучающие данные

Вопрос	Количество
Доброжелательность и корректность сотрудников отдела	9
Доступность изложения теоретического материала на лекционных занятиях	94
Качество проведения практических занятий	67
На сколько Вы удовлетворены организацией и проведением практик (учебной, производственной и др.)	18
Насколько хорошо Вы знакомы с деятельностью заведующего кафедрой, на которой обучаетесь?	36
Оцените Ваше участие в деятельности кафедры, на которой обучаетесь	48
Оцените работу Вашего куратора	62
Полнота и качество электронного образовательного ресурса	100
Удобство графика работы со студентами	30
Уровень ведения разъяснительной работы (консультирования) сотрудниками отдела	12
Уровень сервиса по работе с заявлениями студентов, выдачи справок	16
Всего	492

Данное подразделение на классы необходимо для проверки модуля на правильность определения критерия принадлежности.

Так же, все отзывы за первый и второй семестр имеют отрицательный окрас относительно числовой оценки, то есть числовая оценка не превышает трех баллов.

При описании исходных данных, важно учитывать среднее количество слов в обучающей и тестовых выборках. Данные показатели составляют 6,35 и 6,62 слов соответственно.

В качестве численной оценки качества каждой из комбинаций алгоритма, предлагается выделять показатель точности (формула 7).

$$\text{Accuracy} = \frac{P}{N} \quad (7)$$

где,

P – количество верно классифицированных отзывов;

N – общий размер текстовых данных.

На рисунке 17 представлена гистограмма, отражающая количество верно и ошибочно классифицированных данных, с предварительным стеммингом текста.

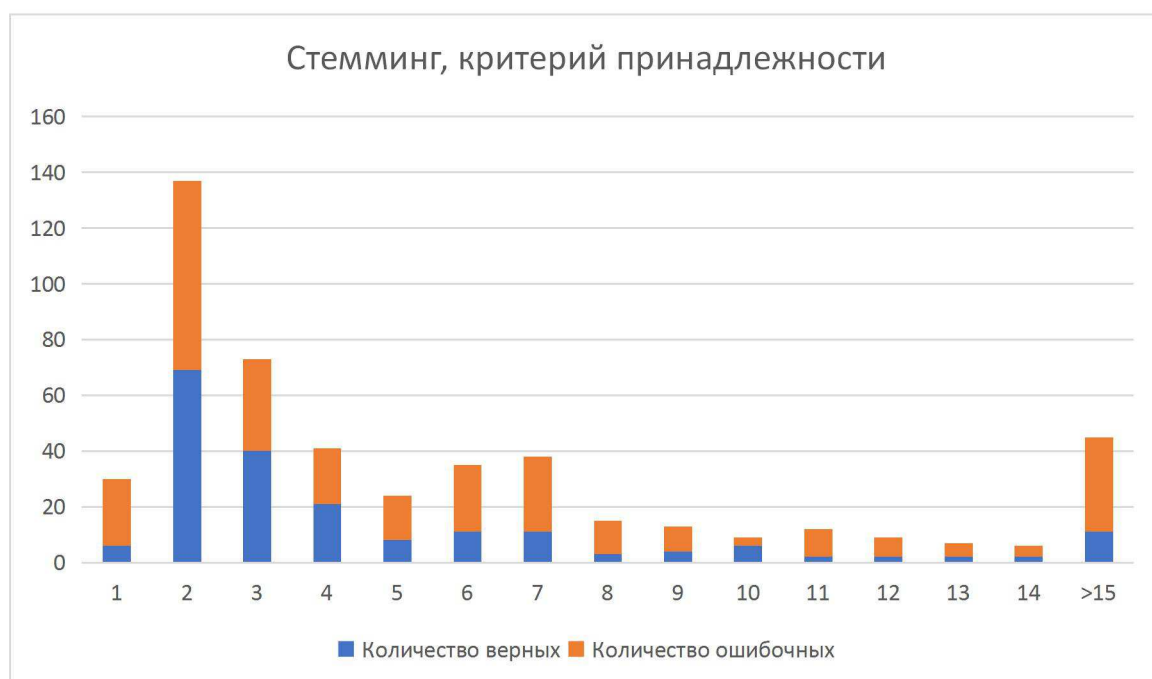


Рисунок 17 – Результаты классификации с предварительным стеммингом

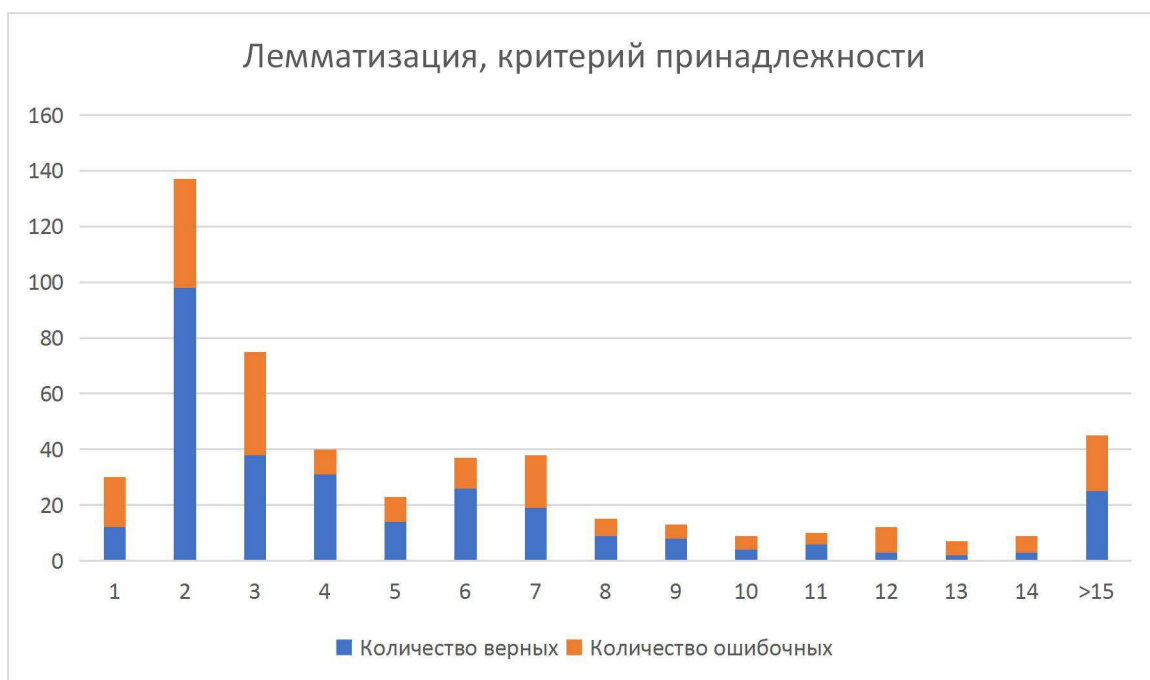


Рисунок 18 - Результаты классификации с предварительной лемматизацией

На основе рисунка 17 и рисунка 18 можно сделать вывод о том, что в данном наборе данных лемматизация, показывает лучший результат как на маленьком количестве слов в тексте отзыва, так и на большом. Показатель точности для стемминга составил 0,605, для лемматизации 0,831.

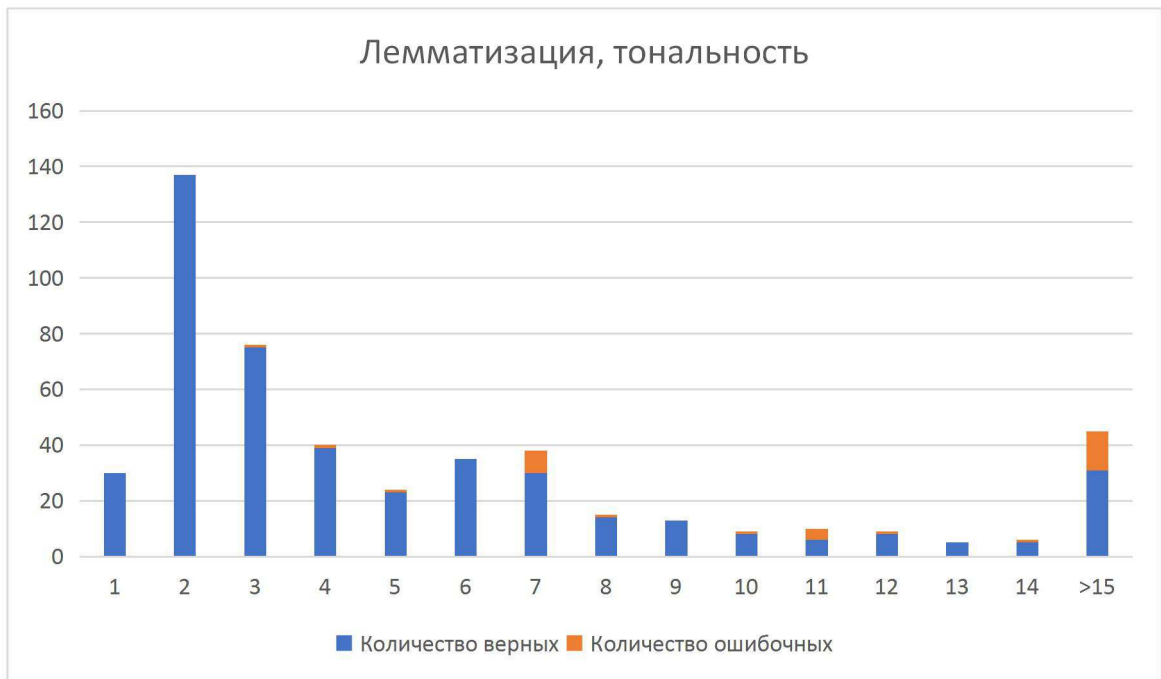


Рисунок 19 - Результаты определения тональности с предварительной лемматизацией

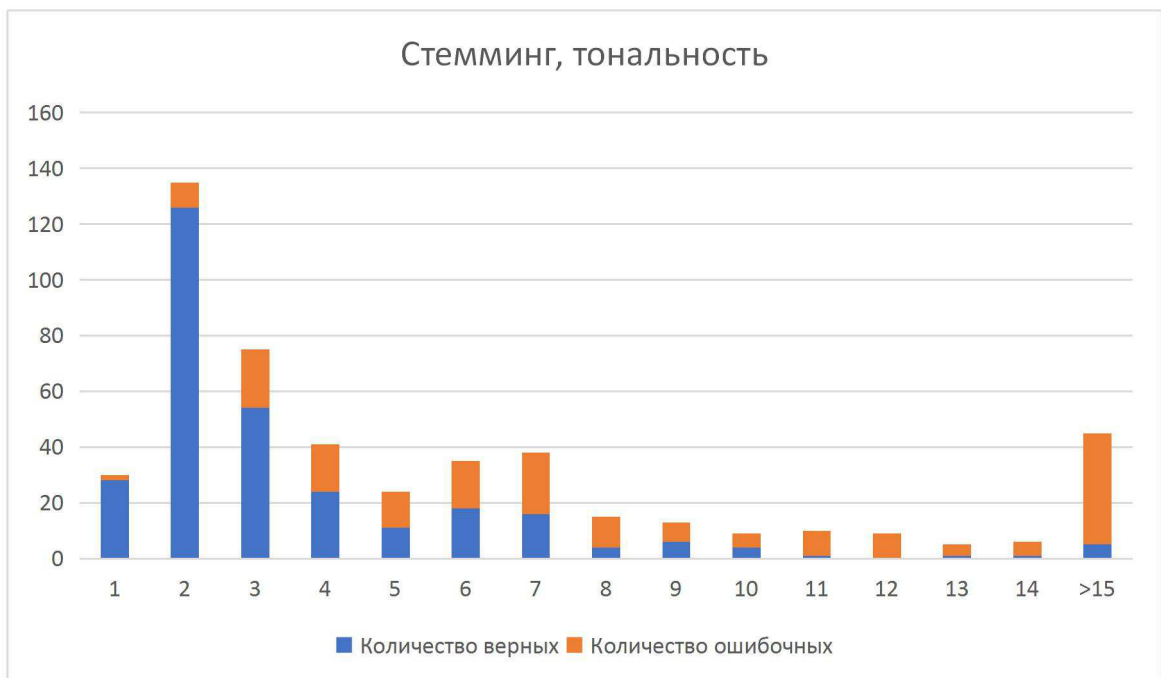


Рисунок 20 - Результаты определения тональности с предварительным стеммингом

На основе результатов определения тональности (рис. 19 и рис. 20) можно сказать о том, что лемматизация, пусть даже без учета контекста, показывает хороший результат в определении тональности текстовых отзывов. Особенно данный факт отражается на объеме отзыва от одного до шести слов. Несмотря на это, точность падает в зависимости от увеличения количества слов. Показатели точности для расчета критерия эмоционального окраса используя стемминг – 0,607, используя лемматизацию 0,93.

Как правило, основываясь на показателе среднего количества слов в выборке равному 6,62 слова, можно сделать вывод о том, что наивная лемматизация успешно справляется со своей задачей на небольшом количестве слов в отзыве.

В результате апробации программного модуля используя данный набор данных можно выделить показатели точности, полноты и f-меры. Результаты представлены в таблице 21.

Таблица 21 – Результаты апробации на реальных данных отзывов студентов

Критерии	Критерий соответствия	Критерий эмоционального окраса
Стемминг		
precision	0,6988	0,7006
recall	0,8227	0,8239
precision	0,7557	0,7572
Лемматизация		
precision	0,8816	0,9526
recall	0,9371	0,9757
f-мера	0,9085	0,9640

В результате апробации программного модуля расчета критериев оценки релевантности на данных студентов можно сделать вывод о том, что, в целом, все комбинации программных средств показали высокие результаты.

Особенно высокие результаты показала лемматизация при расчете тонального критерия, с использованием тональных словарей.

Таким образом метод опорных векторов, взятый за основу расчета критерия соответствия, показывает большую эффективность при нормализации не только методом контекстной лемматизации, но и наивной. Несмотря на высокие затраты ресурсов при вычислении контекста текстового отзыва, высокие показатели эффективности модели компенсируют этот факт.

Так же, при проверке расчета критерия эмоционального окраса преобладает метод контекстной лемматизации. Метод, основанный на тональных словарях, показывает не плохие результаты, но есть возможность для дальнейшей модернизации программы, путем исследования других методов, например, основанных на машинном обучении.

ЗАКЛЮЧЕНИЕ

Не смотря на развитие современных IT-технологий, на текущий момент нет совершенного алгоритма автоматического анализа текстов. Как правило, любая задача из области классифицирования текстов отталкивается от конечного результата, и успешность решения таких задач состоит из правильного подбора технологий, например, технологий нормализации данных перед машинным обучением и т.п.

В ходе работы по теме диссертации были проведены исследования по выбору оптимальных технологий для автоматического анализа текстовых отзывов сервиса анкетирования Сибирского федерального университета. Прежде всего, были изучены аспекты нормализации данных. На практике, рассмотрены стемминг и лемматизация текстов. Далее, были проанализированы два метода автоматической классификации текстовых отзывов: метод, основанный на машинном обучении и метод, основанный на применении тональных словарей. В результате изучения метода с применением машинного обучения, были рассмотрены пять математических моделей.

По итогу работы, основываясь на результатах исследований, был разработан модуль расчета критериев оценки релевантности текстовых отзывов. По результатам апробации разработанного алгоритма был сделан вывод о том, что выбранные технологии анализа являются оптимальными в контексте решаемой задачи.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Даничев А.А., Якунин Ю.Ю. Аспектный анализ тональности отзывов в образовательной среде // Информатизация образования и методика электронного обучения Материалы III Международной научной конференции. В двух частях. Сибирский федеральный университет, Институт космических и информационных технологий. 2019. С. 61-65.
2. Обзор методов классификации в машинном обучении [Электронный ресурс]. IT-портал «TProger.ru» – Режим доступа: <https://tproger.ru/translations/scikit-learn-in-python/> (дата обращения 13.04.2020).
3. Документация модуля «Tree Tagger» [Электронный ресурс]. Режим доступа: <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>
4. Документация библиотеки «nltk», описание стеммера «SnowballStemmer» [Электронный ресурс]. IT-ресурс «kite» – Режим доступа: <https://kite.com/python/docs/nltk.SnowballStemmer>
5. Онлайн-тезаурус русского языка «Карта слов» [Электронный ресурс]. – URL: <https://github.com/dkulagin/kartaslov> (дата обращения: 12.04.2020).
6. Описание наивного байесовского алгоритма для «Python» [Электронный ресурс]: IT-ресурс «Stack Abuse» – Режим доступа: <https://stackabuse.com/the-naive-bayes-algorithm-in-python-with-scikit-learn/>
7. Описание метода опорных векторов для «Python» [Электронный ресурс]: IT-ресурс «Stack Abuse» – Режим доступа: <https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/>
8. Описание метода k-средних [Электронный ресурс]: IT-портал «data science» – Режим доступа: <https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/>

9. Описание классификатора дерева решений для «Python» [Электронный ресурс]: IT-ресурс «Stack Abuse» – Режим доступа: <https://stackabuse.com/decision-trees-in-python-with-scikit-learn/>

10. Описание метода логистической регрессии [Электронный ресурс]: «Википедия» – Режим доступа: https://ru.wikipedia.org/wiki/Логистическая_регрессия

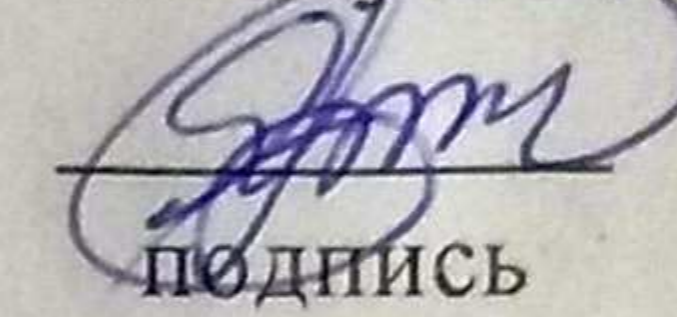
11. Глубоко аннотированный корпус русского языка [Электронный ресурс]: «Википедия» – Режим доступа: https://ru.wikipedia.org/wiki/Глубоко_аннотированный_корпус_русского_языка

12. Документация модуля TreeTager для «Python» [Электронный ресурс]: TreeTagger Python Wrapper's documentation – Режим доступа: <https://stackabuse.com/decision-trees-in-python-with-scikit-learn/>

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Базовая кафедра «Интеллектуальные системы управления»

УТВЕРЖДАЮ

Заведующий кафедрой



Ю.Ю. Якунин

подпись инициалы, фамилия

«25» июня 2020 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

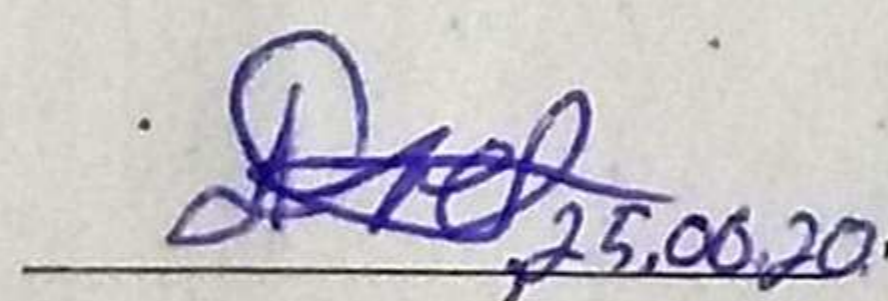
Оценка релевантности текстовых отзывов сервиса анкетирования

09.04.04 Программная инженерия

09.04.04.02 Технологии индустриального производства

программного обеспечения интеллектуальных систем управления

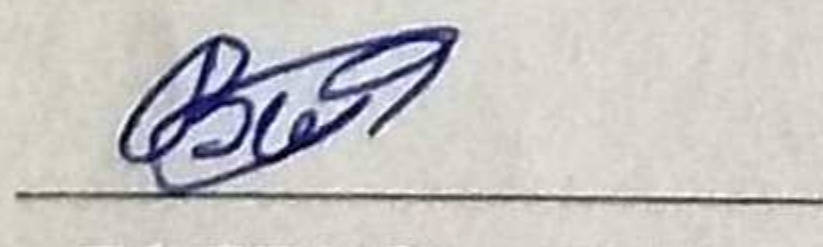
Научный
руководитель


25.06.20
подпись, дата

доцент, канд. техн. наук

А.А. Даничев

Выпускник


подпись, дата

Е.И. Высотенко

Рецензент


подпись, дата

канд. физ.-мат. наук

А.Л. Двинский

Красноярск 2020