ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

на тему:

Исследование оценочной лексики потребительских отзывов в системе Яндекс.Маркет

основная образовательная программа магистратуры по направлению подготовки

45.04.02 «Лингвистика»

Исполнитель:

Обучающийся 2 курса
Образовательной программы
«Прикладная и экспериментальная
лингвистика»,
Профиль «Современные технологии
языкового воздействия»

очной формы обучения
Чечнева Надежда Сергеевна

Научный руководитель:
проф., д.ф.н. Мартыненко Г.Я.

Рецензент:
проф., д.ф.н. Наумов В.В.

Санкт–Петербург
2018

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В последние годы стремительно развивается интернет, в том числе его русскоязычный сегмент. В интернете и повседневной жизни мы ежечасно сталкиваемся с оценками: прежде чем что-нибудь купить, мы знакомимся с отзывами, ставим «лайки», сочиняем комментарии, оставляем записи в блогах. Нас окружает мир оценок, рейтинговый мир. Это явление приобрело такую тотальную массовость, что возникает потребность в ее внимательном изучении, обратившись к языковым проявлениям всех форм оценки.

В современном языкознании стала активно развиваться область исследований, которая занимается анализом мнений, чувств, эмоций, оценок людей по отношению к различным объектам. Эта область называется оценкой тональности. Наше исследование вписывается в эту область.

Основные подходы к изучению тональности текста можно разделить на две большие группы. Подходы первой группы основаны на использовании словарей и правил, вторая группа использует методы машинного обучения.

В данной работе предлагается подход к составлению словаря оценочной лексики для заданной предметной области.

Актуальность выбранной темы обусловлена необходимостью разработки новых методов автоматического анализа оценочной лексики.

Цель работы – представить систему оценок объекта в виде словаря-тезауруса, основанного на иерархическом принципе.

В качестве **базы для исследования** использовались отзывы на товары, размещенные на портале Яндекс.Маркет.

В соответствии с данной целью необходимо решить следующие **задачи**:

- дать понятие оценки и рассмотреть, как развивались философские взгляды на определение оценок;
- выделить отличительные характеристики оценок как предмета изучения лингвистики;

- проанализировать различные существующие классификации оценок;
- дать характеристику проблеме анализа тональности;
- проанализировать существующие словари оценочной лексики для целей анализа тональности;
- создать словарь-тезаурус оценочной лексики потребительских отзывов.

Практическая значимость работы заключается в том, что её результаты могут быть использованы для автоматического анализа тональности текстов.

Апробация исследования: основные положения исследования и полученные результаты были представлены в докладе на XIX Международной научной конференции молодых филологов, проходившей в период с 15 до 17 февраля 2018 года в Таллине.

Структура квалификационной работы: работа состоит из введения, трёх глав, заключения, списка использованной литературы и приложений.

В первой главе рассматривается сущность оценок с философской точки зрения, а также особенности категории оценки как предмета изучения лингвистики. В конце главы приводятся различные подходы к классификации оценок.

Во второй главе дается характеристика проблемы анализа тональности. Затем приводится обзор существующих словарей оценочной лексики для задач анализа тональности.

В третьей главе описывается материал исследования, этапы составления тезауруса оценочных слов и словосочетаний. В заключении главы дается характеристика полученного тезауруса, приводится экспериментальная проверка применения тезауруса для анализа тональности.

ГЛАВА 1. ФИЛОСОФСКИЙ И ЛИНГВИСТИЧЕСКИЙ АСПЕКТЫ КАТЕГОРИИ ОЦЕНКИ

1.1. Философский аспект оценок

В любом языке существуют понятия, выражающие ценностное отношение человека. В русском языке это, прежде всего, понятия «ценность», «оценка», «достоинство».

Этимология слов «ценность» и «оценка» связана со словом «цена», которое в современном русском языке выражает как стоимость в денежных единицах, так и употребляется в более широком смысле, например, в выражениях «любой ценой», «бесценный». Этимология слова «достоинство» связана со словом «достой», что означает приличие, соответствие, соразмерность [Столович, 1994, 8-9 с.].

Понятие «ценность», как правило, объединяет в себе три значения:

- вещественно-предметное – характеристика внешних свойств вещей, предметов, явлений, являющихся объектом ценностного отношения;
- психологическое – психологические качества человека, выступающего субъектом этого отношения;
- социальное – отношения между людьми, благодаря которым ценности приобретают общую значимость.

С понятием «ценность» тесно связано понятие «оценка». Ценности реализуются, проявляется в оценках. Ивин называет оценками высказывания о ценностях [Ивин, 1970]. Оценка, как и ценность, является одним из важнейших аксиологическим понятий. Существует множество трактовок данного понятия. В толковом словаре русского языка значение оценки приведено в широком смысле, как «мнение о ценности, уровне или значении кого-чего-н» [Ожегов, 2011, 445 с.]. В некоторых философских работах оценка часто рассматривается как сам умственный акт [Тугаринов, 1968; Брожик, 1982; Неновски, 1987]. Также встречается мнение, когда под

оценкой понимается результат процесса оценивания, который проявляется в форме «представления, понятия, суждения о значении» [Анисимов, 1988, 40 с.].

Теория ценности как отдельная отрасль философии возникла во второй половине XIX века. Это связано с возникновением «философии ценности», однако вопрос о ценностном отношении (в том числе вопрос о том, что такое хорошо и что такое плохо) и стремление к его решению возникли задолго до этого [Столович, 1994, 8 с].

Понятие ценности возникло вместе с формированием философии, а именно как часть направления, которое занималось изучением морали, нравственных компонентов, т.е. этики. Для того чтобы лучше понять природу оценок, кратко рассмотрим, как развивались философские представления об оценках и ценностях.

Одним из центральных вопросов в теории ценности является вопрос онтологии оценок и ценностей. Существует большое количество разнообразных и порой противоречивых взглядов на эту проблему [Батурин, 1998].

Представления древнегреческих и средневековых философов о ценностях можно охарактеризовать как «натуралистический» или «природный» подход к ценности. В основе данного подхода лежит представление о том, что наряду с миром людей и вещей существует отдельный, объективно существующий мир ценностей. Под ценностями при этом понимаются объективные свойства предметов. А эмоции или оценки выступают в роли анализатора, позволяющего отобразить объективные качества вещей и явлений. Оценивая предметы, человек познает, в чем состоят ценности объективного мира.

Данный подход подвергался критике за то, что объективное существование ценностей невозможно доказать никакими научными способами. Помимо этого, с точки зрения психологии неизвестно, какими

рецепторами психических процессов человек может воспринимать объективные ценности предметов и явлений.

Кроме того, с помощью объективного подхода сложно объяснить, почему разные люди в разных ситуациях могут оценивать один и тот же предмет по-разному.

Когда философы заметили, что оценки зависят от человека и от ситуации, они направили свое внимание на субъект ценностного отношения, т.е. на человека, дающего оценки. Это привело к возникновению «субъективистского» подхода. Этот подход понимает ценности не как объективные свойства, а как изобретение человека. Только человек может наделить предметы или явления ценностным смыслом.

Данный подход также не избежал критики. Так, очевидным становилось противоречие: если оценки и ценности не являются объективными свойствами, то почему все же в отношении большого количества предметов и явлений многие люди солидарны в своих оценках. Стало быть, все-таки в самом предмете есть нечто такое, что приводит к признанию данного предмета ценностью разными людьми.

В поисках разрешения данного противоречия некоторые ученые пришли к тому, что ценности и оценки – объективно существующие категории, но только через человека они приобретают свой ценностный характер. Идею о том, что «только для человека и только через человека действительность приобретает ценностный характер» В. Брожик считает основным положением теории ценности [Брожик, 1968].

Такое положение вещей влечет за собой введение понятия «отношение». Отношение связывает человека и предметы, выражает связь человека и предметов действительности, мира, общества.

Все вышеназванные подходы находят отражение в концепциях философов разных периодов времени.

Уже в античности наблюдается разнообразие взглядов на теорию ценности [Столович, 1994]. Так, среди античных философов есть сторонники

идеи объективности ценностей. К ним можно отнести пифагорейцев, Гераклита, Демокрита, Сократа, Платона, Аристотеля. В то же время уже в античности появляются приверженцы субъективистской концепции. К их числу можно отнести софистов и скептиков от Пиррона до Секста Эмпирика.

Кроме того, выделяется позиция киников, которые отрицают не сами ценности, а их почитание, основанное на ложных принципах.

Древняя философия ценности закладывает общечеловеческие основы ценностной концепции. В ней можно обнаружить модель всего последующего развития аксиологии.

В Средние века возникает главенство религиозного сознания, которое проявляется в том, что в Боге воплощается высшее совершенство и благо, он олицетворяет триединство истины, добра, красоты [Анисимов, 2001]. Наряду с таким объективным идеализмом в Средние века многие мыслители вполне осознают субъективность ценностного отношения. Так, Августин разрабатывает психологическую теорию восприятия, которая включает в себя идею способности судить на основе чувства удовольствия или неудовольствия.

Некоторые исследователи [Столович, 1994] отмечают символичность сознания, характерную для средневековья. Поскольку ценности и ценностные отношения выражаются в знако-символьной форме, то можно говорить о том, что в раннем средневековье появляется своеобразная семиотика, которая содержит важные аксиологические идеи.

В эпоху Возрождения высшей ценностью признается человек. Особо важной становится проблема истины, добра, красоты в связи с проблемой человека. Это приводит к новому этапу разработки понятия, которое объединяло бы все вышеупомянутые понятия – понятия блага, ценности.

В эпоху Возрождения аксиологические концепты разрабатываются с точки зрения человеческой субъективности, а понятие ценности – с точки зрения общечеловеческого значения.

В XVII веке ценностная концепция получает рационалистическую интерпретацию в связи с борьбой различных политических сил и с развитием общественных отношений, в том числе товарно-денежных [Анисимов, 2001].

К представителям рационалистического подхода этого периода можно отнести Декарта, Паскаля, Спинозу. Они признают субъективность и относительность оценочной деятельности, в основе которой, по их мнению, находится мышление.

Некоторые другие философы XVII века особое внимание обращают на психологическую природу человека [Арутюнова, 1988]. Они пытаются выделить ее простейшие составляющие. Гоббс выделяет шесть страстей в эмоциональной сфере: желание, любовь, отвращение, ненависть, радость и горе. Через эти понятия он определяет добро и зло: «Все вещи, являющиеся предметом влечения, обозначаются нами в виду этого обстоятельства общим именем добро, или благо; все же вещи, которых мы избегаем, обозначаются как зло». Таким образом, в его концепции хорошее приравнивается к желаемому, а плохое – к нежелаемому. Из этого следует, что хорошее и плохое – относительные, субъективные концепты, и они могут отличаться у разных людей.

Дж. Локк считает хорошее и плохое категориями сознания, но через отсылку к чувственному опыту, т.е. хорошее – это то, что осознается как вызывающее удовольствие, а плохое – то, что осознается как вызывающее страдание. У него, как и у Гоббса, добро связано с удовольствием.

Таким образом, если средневековые философы связывали относительность и сложность описания аксиологических понятий с их приложением к разным категориям объектов, то Гоббс – с различием субъективных мнений человека о хорошем и плохом, а Дж. Локк же – с разными вызываемыми ими эмоциональными ощущениями.

Категоричный отход от гедонизма в определениях оценок и ценностей осуществляет И. Кант. Он заменяет принцип ощущений, чувствований принципом разума, желание – долженствованием, эмпирическое –

априорным, относительное – абсолютным, добро как средство – добром как цель. По Канту, «человек обладает некоторым достоинством (некоей абсолютной внутренней ценностью)» [Кант, 1994]. Он рассматривает нравственно-моральную ценность как эталон ценности.

В XIX веке формируется отдельная отрасль философского знания, получившая название «философия ценности» [Столович, 1994]. Философы XIX века продолжают разрабатывать основные положения теории ценностных отношений, появившиеся в предыдущем столетии. Рассмотрение ценности как философской категории принято связывать с исследованиями Лотце, он утверждает необходимость признания отдельного мира ценностей, который четко отделен от мира обликов и фактов. В его работах наблюдается тенденция субъективизации ценностей.

Такие представители неокантианства как Виндельбанд, Риккерт считают, что ценности составляют особый идеальный мир, не доступный научному познанию.

В XX веке появляется термин «аксиология». Большую роль начинает играть семиотический подход к анализу ценностей, особое внимание уделяющий значению слов, выражающих ценности.

Мур в своей работе «Принципы этики» [Мур, 1984] критикует «натуралистический» подход, называя «натуралистической ошибкой» установление фактических свойств объектов, якобы определяющие эти объекты как ценности.

Позднее идеи Мура развивает Хэар. Он подчеркивает невозможность сведения оценок к каким-то естественным свойствам. Оценочные суждения могут быть дополнены разным количеством фактической (дескриптивной) информации, но оценочное значение первично по сравнению с дескриптивным. Следовательно, то, что Мур называет «натуралистическим», Хэар характеризует как «дескриптивное».

Интересными и актуальными являются взгляды итальянского философа-неогегелянца Бенедетто Кроче. Он становится своеобразным

предшественником современного семиотического подхода к пониманию ценности [Столович, 1994].

По Кроче, деление мира на мир ценностей и мир фактов соответствует делению на дух и природу, бытие и долженствование. Ценность – это идеальное образование.

Кроче рассматривает ценностные суждения как выражение чувств. Однако, выражение ценности – не есть создание ценности. Словосочетание «красивая вещь» по Кроче является оксюмороном, поскольку красота не является физическим фактом и относится не к сфере вещей, а к деятельности человека, к сфере его духовной энергии.

В своем труде «Философия как наука о духе» (1902-1916) Кроче выстраивает систему категорий бытия, которое мыслится ему как «дух». Он рассматривает два вида духа: теоретический (познание) и практический (действие). Теоретическому духу присущи две основные формы: интуитивная и логическая. Первой присущи человеческая способность воображения и наука эстетика. Второй – разум и наука логика. Практический дух проявляется в экономической и этической форме. Экономическая форма выражается в способности человека желать и науке экономике. Этическая – в воле и науке этике.

Каждой составляющей духа соответствует ценностная категория: красота, истина, польза, добро. Таким образом, можно выделить эстетические, интеллектуальные, экономические и этические ценности и противоценности. Границы между ценностями нежесткие (не замкнуты). Например, красота может быть интеллектуальной, красотой деятельности, моральной красотой.

В своих последующих исследованиях разные философы, вплоть до настоящего времени, продолжают искать ответы на такие вопросы как соотношение оценок и естественных свойств объектов, соотношение оценок и норм и т.д.

Таким образом, как можно увидеть из всего вышесказанного, вопрос оценок интересовал ученых-философов с давних времен, но до сих пор остается открытым и актуальным.

1.2. Оценки как предмет изучения лингвистики

Категория оценок является предметом изучения широкого круга наук: от философии и аксиологии до психологии, политологии и лингвистики. Лингвистика изучает средства и способы выражения оценок в тексте и речи на всех уровнях языка: фонетическом, морфологическом, лексическом, синтаксическом [Кочеткова, 2004].

На фонетическом уровне оценки могут быть выражены звукоподражаниями, ритмоподражаниями, фонетическими каламбурами, а также аллитерацией или ассонансом [Усминский, 1997]. То есть определенные сочетания гласных и согласных звуков, находясь во взаимодействии с другими языковыми средствами, создают нужный психоэмоциональный фон. Фонетические единицы языка прямую оценку не выражают, а лишь косвенно влияют на восприятие высказывания.

На уровне морфологии положительная или отрицательная оценка может быть представлена суффиксами субъективной оценки, которые придают словам различные оттенки (ласкательное, сочувствия, пренебрежения, презрения, уничижения, иронии, также реального уменьшения или увеличения).

На синтаксическом уровне оценки явно не выражаются. Синтаксические конструкции могут только усиливать то или иное (положительное или отрицательное) восприятие адресантом высказывания.

Наиболее ярко и полно оценки могут быть выражены на лексическом уровне [Кочеткова, 2004]. Лексические единицы языка с оценочным компонентом значения эксплицитно могут выражать положительную или

отрицательную оценку в тексте или речи, а выбор того или иного слова напрямую оказывает воздействие на восприятие адресантом высказывания.

В повседневной речи встречается огромное количество слов со значением оценки. По мнению Т.В. Маркеловой это связано с тем, что «ценностное отношение к миру в его языковой семантической интерпретации пронизывает кровеносными сосудами всю систему языка» [Маркелова, 1993].

Одним из центральных вопросов в лингвистическом изучении аксиологических значений является вопрос определения статуса слов, выражающих оценки.

Л.В. Щерба рассматривает оценочные предикаты в числе предикатов категории состояния [Щерба, 1974].

Золотова утверждает, что в русском языке можно выделить наличие особого лексико-грамматического класса слов со статусом категории оценки [Золотова, 1980].

М.А. Богданова [Богданова, 2014] выделяет оценочные слова в отдельную категорию. Основанием для этого по ее мнению служит наличие форм времени (в отличие от наречий), которые синтаксически выражаются связками, и сочетание с инфинитивом. Слова категории оценки обозначают оценку действия, выраженного инфинитивом, а не состояния, которое подразумевает безличный предикатив.

Большой вклад в изучение лингвистического аспекта категории оценок в русском языке внесли Е.М. Вольф и Н.Д. Арутюнова.

Е.М. Вольф исследует связь оценочного компонента языковых единиц с субъектом речи. Согласно ее взглядам, смысл оценки может зависеть от картины мира, пресуппозиций участников речевой коммуникации, контекста, ситуации общения. [Вольф, 1979]. В своих работах она определяет оценку как «отношение субъекта к объектам действительности, которое является социально устоявшимся и закрепленным в семантике единиц языка, оно может быть положительным или отрицательным, эксплицитным или имплицитным». [Вольф, 1985, 1986].

Н.Д. Арутюнова изучает оценку как логическую категорию и применяет методы коммуникативного и логического анализа [Арутюнова, 1988]. Она выделяет два типа непредметных объектов: процессы (состояния, свойства, события) и факты. Первый тип включает в себя всё то, что относится к области погружения людей в мир, а второй связан с тем, что составляет результат погружения мира в сознание людей [Арутюнова, 1985]. Из этого следует, что оценки в любом случае связаны с человеком. Также она отмечает, что оценки неразрывно связаны со сравнением [Арутюнова, 1988]. Оценка всегда предполагает существование какого-то эталона, который может быть представлен как другим объектом, так и какой-то нормой или идеалом.

Н.Д. Арутюнова противопоставляет дескриптивному значению оценочное, поскольку последнее не содержит каких-либо объективных признаков предметов.

Похожим образом Н.В.Телия [Телия, 1986] рассматривает языковые значения, содержащие информацию о мире, и значения, содержащие информацию об оценочном отношении к миру.

Кроме того, ролью оценки в образовании лексического значения слова может быть участие в коннотации. Так, А. В. Вестфальская [Вестфальская, 2015] считает, что оценка образует основу коннотации. Положительный или отрицательный оценочный компонент является определяющим элементом коннотации и обусловлен культурными, нравственными, политическими, особенностями каждого народа.

Научные исследования в области оценок доказывают, что лингвистический аспект категории оценок активно изучается в настоящее время. Растущий интерес к изучению оценочной лексики можно объяснить следующими причинами:

Во-первых, возможность использования оценочной лексики в практических целях. В области маркетинга анализ оценочной лексики в отзывах на товары и услуги позволяет производителям узнать, как

потребители относятся к их товарам. Изучая лексику с оценочным значением в текстах СМИ и социальных медиа, можно определить, как преподносятся разные социально значимые темы, и, соответственно, как они воспринимаются людьми. Психологи, анализируя комментарии человека в социальных сетях, могут составить его психологический портрет.

Во-вторых, по данным некоторых исследований, в русском языке доля лексических единиц, обладающих оценочным компонентом значения, достигает 40%, но при этом словарей оценочной лексики для русского языка существует немного [Тихонова, 2015].

Сложность составления словарей оценочной лексики заключается, прежде всего, в том, что вопрос о том, какие слова считать оценочными, не однозначен, поскольку формальных оценочных помет нет.

В работе Г.Я. Мартыненко [Мартыненко, 2015] описывается создание одного из первых тезаурусов оценочной лексики. Одним из основоположников тезаурусного подхода к лексикографии оценок можно считать Корнея Ивановича Чуковского. В середине XX века этот подход, включающий в себя метод ключевых слов, стал активно применяться в области разработки лингвистического обеспечения информационно-поисковых систем.

В своей книге «Леонид Андреев – большой и маленький» (1908 г.) Корней Иванович собрал коллекцию критических статей, характеризующих творчество Леонида Андреева. Корней Иванович преследовал цель – показать, как оценивали творчество писателя критики-современники. А также, он хотел представить эти оценки в виде системы, которая была бы простой и наглядной.

Для этих целей он собрал критические статьи, т.е. сформировал корпус, каждой единице которого была приписана соответствующая библиографическая информация: автор статьи, название издания, номер, страница. В первую очередь, статьи относились к произведениям писателя,

получившим скандальную и одиозную известность. К их числу относились: «В тумане» (1902), «Бездна» (1903), «Тьма» (1907).

На основе корпуса этих статей он создал словарь-тезаурус оценочных слов, а точнее, словарь ругательных наименований, которыми отзывались о нем критики.

Для этого Корней Иванович выделил ключевые слова из всех статей, которыми критики оценивали творчество Андреева. Затем все ключевые слова (с указанием источника, откуда они взяты) были упорядочены в алфавитном порядке.

В результате Чуковский получил словарь-тезаурус, который дает, с одной стороны, обобщенную картину оценки творчества Леонида Андреева в русской литературной критике начала XX века, а с другой – представление о лексиконе критических статей, посвященных творчеству одного из самых выдающихся писателей XX века. Фрагмент этого словаря (буквы «А», «Б». «В») приведен в Таблице 1.

Таблица 1. *Фрагмент словаря К.И.Чуковского, составленного по критическим статьям, оценивающим творчество Л. Андреева*

А	***Абракадабр***	«Голос правды» 907, VI
Б	***Балаган отвратительный***	Д. В.Философов, «Речь» 908, IV
	Барабанная трескотня	Измайлов, «Р. Слово» 906, X
	Бахвал	А.Басаргин, «Московские ведомости»
	Бедламовщина	«Голос правды», 907, VI
	Безграмотность	Л. Т-цкий, «Север» (Вологда), 907, VI
	Бездарность	«Развлечение», 903, №17
	Бездарность	«Рус. Знамя», 907, VI
	Бездарная и неумная вещь	З. Гиппиус, «Весы», 907, V
	Беспримерная ерунда	А.Е., «Харьк. Вед.», 902
	Белиберда	Буренин, «Новое время», 902, IX
	Бесчестит язык, как женщину	Мережковский, «Русская мысль», 908, I
	Безобразие	А.Е., «Харьковские ведомости», 902
	Богомерзкая книжка	«Русское Знамя», 907, VII
	Босяк	«Вече», 907, № 63
	Бешенство притязательного безмыслия	Буренин, «Новое время», 902, IX
В	***Вранье***	В.Розанов, «Новое время», 907, VII
	Выкрученная и наглая чепуха	Буренин, «Новое время», 902
	Вызывает тошноту	А.Ефимов, «Живописное обозрение», 902, X

Поскольку тезаурус построен на основе анализа критических статей, то в нем превалирует оценочная лексика. При этом эта оценочная лексика чаще имеет явную негативную окраску.

Каждая лексическая единица словаря сопровождается указанием источника: именем автора, названием периодического издания и выходными данными. В результате чего, данный тезаурус позволяет осуществлять документальный поиск (поиск документов по ключевым словам), авторский поиск (поиск документов по имени автора), семантико-авторский поиск (формирование коллекции слов-оценок, фигурирующих в работах конкретного критика) и др.

Таким образом, исследование языковых и неязыковых данных показывает, что категория оценки является универсальной, характерной для любого языка. В процессе своей жизнедеятельности мы оцениваем объекты и явления, и это находит отражение в языке.

Однако стоит отметить, что проблеме лексикографирования оценочной лексики русского языка до настоящего времени не уделялось должное внимание в силу объективных сложностей: неоднозначность критериев оценочности, отсутствие оценочных помет.

Создание словарей оценочной лексики позволило бы, с одной стороны, упорядочить разноуровневые средства выражения семантики оценки, а с другой – расширить возможности практического изучения важнейшей тенденции в современном обществе, а именно – аксиологизации сознания.

1.3. Типология оценок

Существует большое разнообразие классификаций оценок. Они отличаются тем, какой критерий классификации положен в их основу.

В самом общем виде, с точки зрения аксиологической интерпретации, все оценки можно разделить на два вида: положительные и отрицательные

[Фомина, 2007]. С помощью этой классификации можно ответить на вопрос, положительно или отрицательно относится субъект оценки к объекту.

Одна из самых известных классификаций оценок предложена Арутюновой [Арутюнова, 1988]. В самом общем плане она выделяет положительную и отрицательную оценки – это два общеоценочных значения.

Кроме них она рассматривает частнооценочные значения, такие как: сенсорно-вкусовые (вкусный, приятный), психологические (интересный, глупый, грустный), эстетические (красивый), этические (аморальный), утилитарные (полезный, вредный), нормативные (правильный) и телеологические (удачный, целесообразный).

Классификация, разработанная Арутюновой, в основном, ориентирована на оценку объекта, представленного предметом. Она не рассматривает в качестве объекта оценки ситуацию. В статье [Сердобольская, 2005] авторы обращают свое внимание на ситуации в качестве объектов оценок. Так, например, они рассматривают утилитарные оценки применительно не к предметам («полезное приспособление»), а к ситуациям («поражение было для него полезным). Кроме того, они выделяют следующие классы оценок:

- Интеллектуальные оценки. К этой группе авторы относят такие предикаты, как интересный, которые Н. Д. Арутюнова характеризует как психологические, и предикаты сенсорно-вкусовой оценки (вкусный, приятный);
- Истинностные оценки. В классификации Н. Д. Арутюновой это нормативные оценки (правильный, верный). Эти оценки могут употребляться в двух случаях – оценка с точки зрения некой нормы или оценка по признаку «истина/ложь»;
- Универсальные оценки. Сюда авторы относят предикаты общей оценки (хорошо, плохо);
- Психологические оценки. К этой группе авторы относят телеологические (удачный, целесообразный) и психологические оценки

(глупый) из классификации Арутюновой, а также предикаты «важно», «уместно» и т.д.;

- Дедуктивные оценки. К таким оценкам они относят предикаты логического вывода (ясно, понятно, видно, очевидно);
- Вероятностные оценки (возможно, вероятно). Кроме собственно вероятностной оценки они включают в эту группу предикаты «странно», «удивительно», т.к. в некоторых ситуациях они могут обозначать оценку вероятности наступления события;
- Временные оценки (пора, рано);
- Количественные оценки (мало, много).

В зависимости от наличия эмотивного компонента, можно выделить два типа оценок: рациональные и эмоциональные [Кабирова, 2011]. Рациональные оценки, как правило, выражают осмысленное сравнение объекта с неким установленным стандартом или нормой. Эмоциональные оценки проявляются в виде определенной реакции человека на объекты и явления окружающей действительности, которые затрагивают личность субъекта, его мировоззрение, представление о поведении, которое он воспринимает как важное для себя.

С точки зрения коммуникативной цели высказывания Е.М. Вольф, основываясь на работах Сёрля, выделяет пять типов оценочных слов: ассертивы, директивы, комиссивы, экспрессивы, декларативы. Все они в разной степени могут содержать и рациональные, и эмоциональные компоненты. Так, Е.М. Вольф считает, что, например, в ассертивах преобладает рациональный компонент, который заключается в том, что говорящий рассчитывает на то, что адресат согласится с его оценкой. Экспрессивы же могут быть как эмоциональными, так и рациональными [Вольф, 1985].

Финский логик Георг Хенрик фон-Вригт выделяет следующие типы оценок [Фон-Вригт, 1986]:

1. Инструментальные (например, хороший нож);

2. Технические – оценки мастерства (например, хороший специалист);

3. Оценки благоприятствования (например, полезный для здоровья);

4. Утилитарные (например, хороший план);

5. Медицинские – оценки, характеризующие физические органы и психическое состояние (например, плохая память);

6. Гедонистические (например, хороший вкус).

Как разновидность оценки благоприятствования он рассматривал этическую оценку (например, хороший поступок).

Первые два вида оценок (инструментальная и техническая) основаны на принципе функциональности, поэтому фон-Вригт рассматривает их неразрывно друг от друга.

Функции, выполняемые объектами, объединяют эти объекты в классы. Поэтому оценки функционального типа характеризуют объекты как члены классов: например, хороший нож хорош как нож, хороший генерал должен удовлетворять требованиям, предъявляемым к полководцам.

Таким образом, можно сказать, что пресуппозицией функциональной оценки является суждение о пригодности объекта к выполнению некоторой задачи. Отрицательная функциональная оценка означает, что объект плохо справляется со своей задачей.

Оценки благоприятствования связаны с инструментальным оценкам, поскольку они выражают, например, что объект является хорошим для чего-либо. В свою очередь, благоприятствование может рассматриваться как вид утилитарных оценок. Их объединяет направленность на получение положительного эффекта. Утилитарные оценки и оценки благоприятствования, в отличие от инструментальных, выражаются безотносительно к классу объектов.

Медицинские оценки относятся к органам и некоторым психологическим и ментальным процессам. Медицинские оценки характеризуют основные функции организма. Маркированной является

отрицательная медицинская оценка, поскольку положительная оценка является нормой и, следовательно, вторичной.

Медицинские оценки могут быть связаны с удовольствием, которое выражают гедонистические оценки. Эти оценки относятся к самому ощущению, независимо от того, чем оно было вызвано. Гедонистические оценки не подлежат верификации.

Классификацию оценок, несколько отличную от предыдущих, предложила Миронова Н.Н. [Миронова, 1997]. Она анализирует структуру оценочного дискурса и выделяет следующие виды оценок:

1) аксиологические оценки (этические, эстетические, утилитарные, политические/идеологические, религиозные, эмоциональные);

2) модальные (необходимость, долженствование, возможность);

3) экзистенциальные;

4) временные;

5) оценки величин;

6) пространственные оценки.

Кроме этого, Н.Н. Миронова, довольно традиционно делит оценки на общие и частные, а последние, в свою очередь – на рациональные и эмоциональные оценки.

Большое разнообразие классификаций оценок связано со сложностью структурирования оценочного значения. В связи с этим, важной проблемой является систематизация средств выражения оценочных значений в русском языке. Для практического достижения этой цели требуется всесторонний анализ функционирования языковых средств на уровне семантики, грамматики и прагматики.

Для анализа оценочных высказываний не представляется возможным создание какой-то одной универсальной классификации оценок. Для каждой конкретной ситуации становится необходимым разработка своей классификации оценок, учитывающей как средства выражения оценочного значения, так и цель систематизации.

ГЛАВА 2. СЛОВАРИ ОЦЕНОЧНОЙ ЛЕКСИКИ ДЛЯ ЦЕЛЕЙ АНАЛИЗА ТОНАЛЬНОСТИ

2.1. Анализ тональности

2.1.1. Понятие анализа тональности

В последние годы происходит бурное развитие интернета, в том числе его русскоязычного сегмента. Все большую популярность приобретают социальные сети, блоги и форумы. Это привело к возрастанию интереса к изучению оценок и мнений пользователей, как со стороны коммерческих организаций, так и со стороны научного сообщества. Проблему автоматического анализа оценок, мнений, эмоций и чувств позволяет решить анализ тональности [Клековкина, 2012].

Анализ тональности является одним из примеров практического использования оценочной лексики. Анализ тональности (англ. Sentiment Analysis) – класс методов, позволяющих провести автоматизированный анализ эмоциональной окраски текстов. Другими словами, это одна из областей компьютерной лингвистики, которая занимается анализом оценок, мнений, чувств и эмоций людей по отношению к различным объектам. В качестве объектов могут выступать продукты, услуги, организации, личности, проблемы, события, темы и т.д.

Анализ тональности позволяет узнать, не что говорят о каком-то объекте, а насколько эмоционально о нем говорят [Хохлова, 2016].

Эта область не предполагает работу с фактической информацией, она изучает только степень эмоциональной окраски текстов.

Тональность текста представляет собой эмоциональную оценку, выраженную в тексте по отношению к некоторому объекту, которая определяется тональностью составляющих его лексических единиц и правилами их сочетания. В самом простом случае классификация текстов по тональности осуществляется на два класса – положительный и

отрицательный, но даже эта задача является сложной. При большем числе классов задача, соответственно, становится еще сложнее [Клековкина, 2012].

Несмотря на то, что лингвистика в целом и обработка естественного языка в частности имеют большую историю, исследования в области анализа тональности до двухтысячного года практически не проводились. Но в настоящее время данная область стала активно развиваться. Это можно объяснить несколькими причинами [Liu, 2012]:

1. Широкое распространение коммерческих приложений по оценке тональности. Это обеспечивает сильную мотивацию для исследований;
2. Данная область содержит много сложных проблем, которые ранее не решались вовсе;
3. В современном мире с развитием информационных технологий мы сталкиваемся с огромными объемами информации, в том числе содержащей оценки и мнения.

Автоматическое определение тональности текста предполагает выделение тех частей текста, которые имеют позитивную или негативную оценку по отношению к объекту тональности. Объект оценки может быть как один для всего текста (с учетом его синонимических и анафорических вариантов употребления), так их может быть и несколько [Пазельская, 2011].

Как правило, оценка тональности проводится на трех уровнях [Pang, Lee, 2008]:

- Уровень документа;
- Уровень предложения;
- Уровень объекта.

Задача анализа тональности на уровне документа заключается в том, чтобы определить, какую тональность имеет документ в целом: положительную или отрицательную.

Например, имеется отзыв на товар. Система определяет, какую оценку выражает отзыв в целом на этот продукт: положительную или отрицательную.

Этот уровень анализа предполагает, что документ выражает оценку только по отношению к одному объекту (например, к одному продукту). В связи с этим, данного уровня анализа недостаточно для тех случаев, когда мнения высказываются сразу по отношению к нескольким объектам.

Задача анализа тональности на уровне предложения заключается в установлении, какую тональность имеет отдельное предложение, содержащее оценку или мнение: положительную или отрицательную.

На этом уровне важно отделить предложения, содержащие фактическую информацию (объективные предложения) от предложений, содержащих оценочные взгляды и мнения (субъективные предложения). Но нужно отметить, что часто бывают случаи, когда объективные предложения могут иметь оценочный компонент. Например, предложение может содержать только фактическую информацию, но оно будет выражать оценку, поскольку положение вещей, описанное в нем, не совпадает с ожидаемым.

Ни уровень документа, ни уровень предложения не позволяют узнать, что именно субъекту нравится или не нравится в объекте. Более детально выполнить оценку тональности позволяет анализ на уровне объекта. Он позволяет рассмотреть непосредственно саму оценку или мнение. Анализ на уровне объекта основывается на идее о том, что любая оценка или мнение состоит из двух частей: тональность (положительная или отрицательная) и объект (то, что оценивается).

Осознание важности объектов оценок и мнений помогает лучше понять проблему анализа тональности. Например, несмотря на то, что предложение «Я люблю этот ресторан, хотя обслуживание в нем не очень» имеет положительную тональность, нельзя сказать, что оно полностью положительное. В действительности, оно выражает положительную оценку ресторана в целом и отрицательную – обслуживания в нем. В вышеуказанном

примере ресторан будет являться объектом, а обслуживание в нем – аспектом объекта. Таким образом, цель на данном уровне анализа – выявить оценки и мнения по отношению к объектам и/или аспектам этих объектов.

В отличие от фактической информации, мнения, оценки и эмоции имеют важную характеристику – они субъективны. Из этого следует, что анализ тональности требует распознавания смысла (субъективной оценки), заложенного автором в текст. В работе [Ермаков, 2005] авторы приводят следующие характеристики субъективности содержания текста с лингвистической точки зрения:

- использование лексико-грамматических средств, выражающих модальные характеристики ситуации, модусные смыслы и явное отношение автора к описываемой ситуации, включая выбор слова с эмоциональной окраской вместо стилистически нейтрального слова;
- расстановка акцентов, ракурс представления ситуации, трансформация «обычной» структуры предложения (например, изменение порядка слов, осложнения, изменение залога и т.д.).

Факторы первой группы активно используются для автоматического определения тональности текста. Факторы второй группы использовать для анализа тональности текста крайне сложно.

Остальная информация в тексте является объективной с точки зрения лингвистики – это некоторая совокупность семантических отношений между объектами в описанном фрагменте внеязыковой действительности, которые автор решил выразить.

При этом, анализ тональности не ставит задачу определить, является ли описанная ситуация истинной или искаженной, хотя в этих случаях содержание перестает быть объективным. Задача установления истинности – это отдельная задача, выходящая за рамки анализа тональности, которая должна использовать экстралингвистические факторы.

Для анализа тональности релевантными являются только оценочные суждения. В работе [Ермаков, 2012] авторы представляют оценочное

суждение в виде тройки (субъект высказывания, объект и валентность). Субъектом является автор текста или суждения. Под объектом понимают предмет, лицо или явление, в отношении которого производится эмоциональное высказывание, а под валентностью – эмоциональное отношение субъекта к объекту.

В работе [Liu, 2012] оценочное суждение представлено в виде совокупности четырех компонентов (субъект высказывания, объект, эмоции по поводу объекта и время, в которое было сделано высказывание). При этом под объектом мнения понимается продукт, услуга, тема, проблема, человек, организация, событие. Он, в свою очередь, может состоять из частей, иметь какие-то признаки. Составляющие объекта основаны на отношениях часть-целое. Корневой узел – название объекта. Все другие узлы являются частями. Оценочное суждение может быть высказано по поводу любой части объекта и любого признака любой части.

Объект как иерархия неограниченного числа уровней требует вложенных отношений для такого представления, которые оказывается слишком сложными для применения. Для упрощения часто сокращают иерархию до двух уровней и используют термин «аспект» как по отношению к частям, так и по отношению к признакам.

Учитывая вышесказанное, понятие «оценочное суждение» можно модифицировать, представив его в виде совокупности пяти компонентов (субъект высказывания, объект, аспект объекта, эмоции по поводу объекта и время, в которое было сделано высказывание) [Liu, 2012].

Эмоции по поводу объекта могут быть положительными, отрицательными или нейтральными. При этом они могут быть выражены с разной силой (интенсивностью).

Необходимо подчеркнуть, что все пять компонентов должны взаимодействовать друг с другом. Мнение должно быть выражено субъектом по поводу объекта или аспекта объекта в определенное время. Все пять компонентов существенны. Потеря любого из них является проблематичной

в целом. Например, если отсутствует компонент «время», невозможно проанализировать, как изменяются оценки во временном интервале, что является очень важным на практике, поскольку мнение, выраженное, например, два года назад, и мнение, выраженное вчера могут сильно отличаться.

Цель анализа тональности в данном случае – выявить все пять составляющих оценочного суждения в некотором тексте.

2.1.2. Подходы к определению тональности текстов

Основные подходы к автоматическому определению тональности текста можно разделить на две большие группы [Хохлова, 2016]. Подходы первой группы основаны на применении методов машинного обучения, вторая группа методов использует словари оценочной лексики и правила.

В первом подходе можно выделить два класса методов: машинное обучение без учителя (unsupervised learning) и машинное обучение с учителем (supervised learning) [Клековкина, 2012]. Машинное обучение без учителя предполагает, что наибольший вес в тексте имеют слова, которые чаще других встречаются в этом тексте и в то же время присутствуют в небольшом количестве текстов всей коллекции. Такие слова извлекаются, затем определяется их тональность, и на основе этого определяется тональность всего текста. Машинное обучение с учителем требует наличия обучающей коллекции размеченных текстов, на основе которой строится статистический или вероятностный классификатор (например, байесовский).

Второй подход может быть основан на использовании правил и шаблонов (rule-based with patterns). В таком случае происходит генерация правил, с помощью которых определяется тональность текста. Для этой цели текст разбивается на слова или последовательности слов (N-граммы). Затем с помощью этих данных происходит определение часто встречающихся шаблонов, которым присваивается числовое значение положительной или

отрицательной полярности. Полученные шаблоны применяются при разработке логических условий вида «Если *условие*, То з*аключение*».

Что касается словарей, то они могут быть представлены, например, в виде списка слов с указанием вероятности быть оценочными. Чаще всего словари используются в сочетании с правилами. Более подробно подходы к составлению словарей оценочной лексики будут рассмотрены в следующем параграфе.

Кроме этого, бывают случаи, когда классификаторы работают на основе применения нескольких методов в определенной последовательности, это так называемый гибридный подход.

У каждого из вышеперечисленных подходов есть свои достоинства и недостатки.

Подходы первой группы требуют наличия размеченных коллекций текстов для тренировки моделей машинного обучения. Кроме того, многие существующие программы доступны только для английского языка. Но зато они не требуют больших затрат сил и времени.

Подходы второй группы не требуют размеченных коллекций текстов и показывают хорошие показатели точности. Но зато они весьма трудоемки и сильно зависят от предметной области.

Для русского языка чаще всего используется второй подход, поскольку для русского языка существует немного доступных размеченных корпусов и программных средств [Хохлова, 2016].

Все вышесказанное делает в настоящее время задачу создания словарей оценочной лексики очень актуальной.

Таким образом, за последние годы интерес к области автоматического анализа тональности текстов сильно возрос. Но стоит отметить, что на данном этапе развития в этой области компьютерной лингвистики все еще остается много нерешенных проблем. Анализ тональности затруднителен не только в связи с проблемой выделения компонентов оценочных суждений, но

и из-за неоднозначности эмоциональной составляющей лексических единиц [Ермаков, 2012].

2.1.3. Системы анализа тональности текстов на русском языке

В завершение параграфа рассмотрим несколько существующих систем анализа тональности текстов на русском языке.

Система SentiStrength [Thelwall, Buckley, 2012] изначально разрабатывалась для анализа коротких неструктурированных текстов на английском языке. Затем появилась возможность сконфигурировать ее для работы с текстами на ряде других языков, в том числе и на русском языке. Результатом работы системы является оценка положительной эмоциональной окраски текста, а также оценка негативной эмоциональной окраски текста. Оценки выдаются по шкале от 1 до 5. Также система может выполнить бинарную классификацию текста: положительный или отрицательный. Кроме того, может быть выполнена классификация на три класса: положительный, отрицательный, нейтральный. И последний вариант – оценка по единой шкале от -4 до +4.

В основе алгоритма системы лежит поиск слов с максимальной негативной оценкой и слов с максимальной положительной оценкой. Алгоритм учитывает простейшие взаимодействия слов (например, слова-операторы с усиливающим значением тональности. При работе с русским языком в системе могут возникнуть сложности с русской морфологией: могут отсутствовать некоторые словоформы слов. Данная система позволяет узнать лишь общую тональность текста, не выделяя субъекты и объекты тональности.

В составе системы «Аналитический курьер» также реализована функция анализа тональности текста [Анализ тональности текста, Ай-Теко]. Она были разработана компанией «Ай-Теко». Данная система использует словари и правила для определения тональности текста. Классификация

текстов по тональности осуществляется на три группы (позитивные, негативные, нейтральные). Система основана на глубоком лингвистическом анализе входящего текста и работает в несколько этапов. На первом этапе происходит предварительная обработка текста, извлекаются и классифицируются оценочные слова. Затем найденные слова объединяются в связанные друг с другом цепочки. На последнем этапе выделяются объекты тональности. Данная система развивается долгое время, и в настоящее время уровень точности определения тональности в оценочном высказывании составляет 75-85%. При этом покрытие системы находится на уровне 30-40%

Другая система оценки тональности – система «Ваал» [Проект ВААЛ]. Эта система примечательна тем, что в фокусе ее внимания находится фонетическая структура текста. Цель системы – определить степень воздействия фонетической структуры текста и отдельных слов на подсознание человека. Для этой цели система составляет на основе текста частотный словарь и относит некоторые слова к определенным психолингвистическим категориям. В результате анализа пользователь получает набор оценок по нескольким критериям, относящихся к данному тексту или слову. Программа не занимается анализом семантической составляющей текста, что ведет к сильной ограниченности применимости продукта.

Функция анализа тональности реализована в системе RCO Fact Extractor, разработанной компанией RCO [RCO Fact Extractor SDK]. В этой системе используется подход, основанный на правилах. Эта система учитывает синтаксическую структуру текста и взаимодействие слов. Работа системы состоит из нескольких этапов. На первом этапе извлекаются все упоминания объекта во всех формах, включая полные, краткие и другие. На втором этапе выполняется синтаксический разбор предложений, в которых встречается целевой объект. На третьем этапе выделяются и классифицируются фрагменты с ярко выраженной тональностью или эмоционально-оценочными коннотациями. На заключительном этапе

оценивается общая тональность текста на основе тональностей всех входящих в него элементов. Эта система не позволяет осуществить количественную оценку тональности.

Вышеуказанные системы анализа тональности основаны на различных подходах и предназначены для использования в различных целях. Каждая система имеет свои достоинства и недостатки.

Несмотря на актуальность и перспективность задача анализа тональности на данный момент полностью не решена, в том числе и для русского языка.

Безусловно, одним из важных компонентов универсальных и качественных систем анализа тональности являются словари оценочных слов. Все вышесказанное делает задачу составления словарей оценочной лексики еще более актуальной.

2.2. Словари оценочной лексики

Разработке словарей оценочной лексики, как важному компоненту систем анализа тональности, посвящено много исследований. Такие словари могут быть, например, представлены в виде списка слов с указанием рассчитанной вероятности их быть оценочными. Бóльшая часть существующих словарей оценочной лексики была создана для английского языка.

Одним из примеров такого словаря может служить словарь оценочной лексики английского языка MPQA [Wilson, 2005]. В словаре представлено более 8000 оценочных слов. Каждое слово в словаре имеет указание полярности (положительная, отрицательная или нейтральная), а также силы оценочного содержания (сильное или слабое).

Другим примером словаря оценочной лексики является англоязычный словарь AFINN, который был создан вручную [Nielsen, 2012]. Данный словарь был дополнен нецензурными и сленговыми выражениями с целью

получения лучшего результата при автоматическом анализе сообщений в социальных сетях. Его объем – около 2400 слов. Каждому слову приписано числовое значение полярности. Так, слова с самой высокой положительной оценкой имеют значение полярности, равное +5, а слова с резко отрицательным значением имеют значение полярности, равное -5.

Словарь SentiWordNet [Baccianella, 2010] был создан на основе автоматической разметки синонимических рядов тезауруса английского языка WordNet: каждому ряду были приписаны три числа (доля положительной, отрицательной и нейтральной оценки слов из данного синонимического ряда). Следовательно, многозначные слова могут иметь разные оценки тональности.

Многие исследователи пытались решить проблему создания словарей оценочной лексики для нескольких языков.

В работе [Steinberger, 2012] описывается опыт создания словарей для нескольких языков, основанный на методе триангуляции. Материалом для исследования была коллекция новостных текстов. На первом этапе работы авторы создали словари оценочной лексики высокого уровня для двух языков (английский и испанский), затем переводили их на третий язык (например, французский). Перевод выполнялся автоматически с помощью сервиса Google translate. Те слова, которые встречались в обоих списках слов, переведенных с двух языков, были использованы в качестве основы для словаря на третьем языке. Затем полученный перечень подвергался корректировке: нерелевантные слова исключались вручную, а описание словарных статей расширялось за счет введения морфологической парадигмы. Такой параллельный словарь был подготовлен для восьми языков: английского, испанского, арабского, чешского, французского, немецкого, итальянского и русского.

Работа [Volkova, 2013] также посвящена описанию подхода к созданию словарей оценочной лексики для нескольких языков. Основой метода послужила техника бутстрэппинг. Авторы анализировали публикации в

социальных медиа на примере Твиттера. Для своего эксперимента авторы использовали две тысячи размеченных твитов для запуска процедуры бутстрэппинга, две тысячи размеченных твитов для оценки работы алгоритма и один миллион неразмеченных для анализа. Каждый твит размечался пятью независимыми экспертами с ресурса Amazon Mechanical Turk. В итоге, твиту приписывалась та оценка, которую поставило большинство. Для создания исходного словаря оценочной лексики они использовали разработанный ранее словарь MPQA и небольшой набор размеченных вручную твитов. Затем выполнялась процедура бутстрэппинга: на вход алгоритма поступал неразмеченный твит. Если данный твит содержал хотя бы одно слово из первоначального словаря, то твит считался оценочным. Затем для каждого слова в данном твите вычислялась его вероятность быть оценочным. Все новые оценочные слова с их удельными весами добавлялись в исходный словарь. Следующая итерация бутстрэппинга выполнялась уже на основе нового расширенного словаря. И процедура повторялась. Таким образом, был подготовлен словарь для трех языков: английского, испанского и русского.

Для русского языка известны словари оценочной лексики РуСентиЛекс [Лукашевич, Левчик, 2016], ProductSentiRus [Chetviorkin, Loukachevitch, 2012] и улучшенная версия последнего ProductSentiRus+ [Chetviorkin, Loukachevitch, 2015].

Словарь ProductSentiRus был создан автоматически с использованием лингвистических и статистических признаков, позволяющих находить оценочные слова в тексте. Такие признаки комбинировались с помощью методов машинного обучения. Модель извлечения оценочной лексики создавалась для предметной области фильмов, а затем переносилась на другие предметные области. Затем качество полученных словарей оценочных слов оценивалось с помощью ручной разметки. В итоге авторы составили общий словарь оценочных слов из отдельных словарей в нескольких предметных областях.

Полученный словарь представляет собой список из пяти тысяч слов, упорядоченных по мере снижения вычисленной вероятности их быть оценочными. Полярность оценок в данном словаре не указана.

В результате улучшения этого словаря был получен словарь ProductSentiRus+. Создание данного словаря происходило в несколько этапов.

На первом этапе извлекались оценочные слова из текстов предметной области с помощью методов машинного обучения. Для извлечения оценочных слов использовались три тестовые коллекции:

- отзывы, собранные с рекомендательного сервиса Имхонет (imhonet.ru) и с сайта Яндекс.Маркет (market.yandex.ru) для пяти предметных областей: фильмы, книги, игры, цифровые камеры, мобильные телефоны;
- тексты-описания объектов отзывов (описания сюжетов фильмов и книг, описания технического устройства);
- коллекция новостей.

На основе этих корпусов для каждого знаменательного слова (существительного, прилагательного, глагола, наречия), встречающегося в корпусе отзывов, вычислялись признаки, описывающие «поведение» слова в каждой из вышеупомянутых коллекций.

Сначала вычислялись частотные признаки, включающие в себя такие показатели как

- общая частота встречаемости слова в разных коллекциях;
- документная частота, т.е. количество документов, в которых встретилось данное слово;
- частота употребления слов с заглавной буквы;
- частота встречаемости слова в сочетании со словами-операторами, которые усиливают или меняют тональность слова на противоположную.

Затем вычислялись контрастные признаки. Это признаки, которые вычислялись с целью показать, что доля оценочных слов в коллекции отзывов выше по сравнению с другими коллекциями.

Кроме этого, авторы вычисляли признаки, учитывающие оценки, которые ставили пользователи за товар в отзыве. К числу таких признаков относятся отклонение от средней оценки (вычисляется как разница между средней оценкой отзывов, в которых упоминалось слово, и средней оценкой отзыва) и дисперсия оценки слова. Дисперсия показывает, насколько сильно отличаются оценки отзывов, в которых упоминается слово. Предполагалось, что если слово упоминается в различных отзывах с похожими оценками, то, скорее всего, это оценочное слово.

Затем вычислялись лингвистические признаки, включающие признаки части речи, многозначности части речи для слова, признак наличия приставки из некого списка приставок, часто встречающихся в составе оценочных слов.

С помощью этих признаков происходило обучение алгоритма. В результате автоматически был сформирован список слов, упорядоченный по оценочным весам (вероятности слов быть оценочными).

Затем обученная модель была перенесена на другие четыре области отзывов (книги, игры, цифровые камеры и мобильные телефоны).

На втором этапе полученный перечень оценочных слов уточнялся на основе информации о лексических отношениях между словами, представленных в тезаурусе русского языка РуТез.

На заключительном этапе словари оценочной лексики разных предметных областей объединялись в один словарь. В результате был получен новый лексикон оценочных слов ProductSentiRus+.

Другой пример словаря оценочной лексики для русского языка – лексикон РуСентиЛекс. Он был создан на основе сочетания автоматических методов извлечения оценочной лексики из текстов и последующего их ручного просмотра и описания. В данном словаре содержится более десяти

тысяч русских слов и выражений, имеющих некоторое оценочное значение. Для слов, имеющих несколько значений с разной полярностью, сделаны ссылки на соответствующие понятия тезауруса русского языка РуТез.

РуСентиЛекс содержит следующие типы слов, значения которых связаны с тональностью:

- слова (словосочетания) литературного русского языка, значение которых связано с оценочностью из тезауруса русского языка РуТез;
- слова (словосочетания) из корпуса новостей, напрямую не выражающие оценки, но имеющие оценочную коннотацию;
- сленговые слова из Твиттера.

Для всех словарных единиц в РуСентиЛекс указана полярность слова (позитивная, негативная или нейтральная), а также источник тональности (оценка, эмоция или коннотация). Кроме этого, представлены тональные различия между значениями многозначного слова. Если все значения многозначного слова имеют одну и ту же тональность во всех значениях, то указана просто тональность слова.

Многие из рассмотренных выше словарей оценочной лексики позволяют осуществлять классификацию текстов по тональности лишь по двум полярным категориям – «положительная» и «отрицательная».

Для более глубокого анализа тональности текстов требуется создание более сложных словарей. В своем исследовании мы поставили задачу получить такой словарь, который позволял бы определить не только нравится или не нравится субъекту какой-то объект, но и что именно в этом объекте ему нравится, а что нет. Такую задачу можно решить только при использовании словарей-тезаурусов с семантическими связями между единицами словаря, разработанных для конкретных предметных областей.

ГЛАВА 3. СОЗДАНИЕ ТЕЗАУРУСА ОЦЕНОЧНОЙ ЛЕКСИКИ

3.1. Материал исследования

В качестве базы для исследования использовались отзывы на товары, размещенные на портале Яндекс.Маркет. Отзывы как материал для исследования были выбраны потому, что они представляют собой тексты с высокой концентрацией оценочной лексики. В качестве группы товаров были выбраны кофемашины, т.к. они представляют собой изделие средней технической сложности, которым может пользоваться широкий круг потребителей – и мужчины и женщины (нет привязки к гендеру).

На портале Яндекс.Маркет посредством отзывов пользователи могут делиться своими впечатлениями о приобретенном товаре и выбранном магазине.

Отзыв может содержать оценку пользователя, а также текстовое описание опыта использования товара или взаимодействия с конкретным магазином, выявленных преимуществ или отрицательных качеств.

Магазины и модели на Яндекс.Маркете имеют рейтинг – оценку потребительских свойств, которая показывает отношение к товару или магазину пользователей. Рейтинг рассчитывается в основном исходя из оценок и отзывов пользователей [API Маркета].

Отзывы, размещенные на портале Яндекс.Маркет, как правило, имеют следующую структуру:

- Субъект оценки (автор отзыва);
- Оценка по пятибалльной шкале;
- Опыт использования (количество месяцев);
- Достоинства;
- Недостатки;
- Комментарий;
- Дата, город.

Пример отзыва представлен в Таблице 2.

Таблица 2. *Структура отзыва на кофемашину с портала Яндекс.маркет*

Субъект оценки (автор отзыва):	Лахтионов Виктор. Автор 2 отзывов
Оценка по пятибалльной шкале:	5 (Отличная модель)
Опыт использования:	более года
Достоинства:	*Надежный агрегат. Отработала 9 лет в коллективе 25 человек.*
Недостатки:	*Если придраться - резервуар для воды маленький.*
Комментарий:	*Покупали на работу 9 лет назад, коллектив 25 человек. Нагрузка такая, что думал года не протянет.. Закрепили два человека, которые смотрели за машиной, проводили профилактику. Ничего не смазывали и не разбирали за все время.. Каждый день по несколько десятков чашек, как автомат Калашникова. Неделю назад перестала готовить кофе, разобрали.. какой то пластиковый переходник лопнул пополам. С учетом амортизации в целом решили купить новую. Пока искал, увидел такую же на маркете:) Решил пару строк написать. Однозначно рекомендую. Живучий и надежный аппарат!*
Дата, город:	22 января 2018, Москва

Для автоматического извлечения отзывов была разработана программа на python, использующая контентный API Яндекс.Маркета.

Программа работала по следующей схеме:

1. Через обращение к API Яндекс.Маркета [API Маркета] извлекался перечень id моделей для товарной категории «кофемашины»;
2. В цикле перебирались id моделей, и для каждой извлекались отзывы.

Отзывы извлекались только на самые популярные модели, т.е. на модели, которые имели большое количество отзывов. С полным кодом программы можно ознакомиться в Приложении А.

Поскольку базовый доступ имеет ограничение по допустимому числу обращений к API – 100 в сутки, то информация извлекалась в несколько этапов.

Отзывы извлекались в формате json, метаинформация сохранялась для дальнейшего анализа, а для составления словаря использовался непосредственно текст отзывов.

Таким образом, был получен корпус общим объемом 3850 отзывов на кофемашины (368 193 словоупотребления).

Кроме этого, были составлены подкорпусы мужских и женских отзывов, каждый объемом 500 отзывов. Целью анализа являлась проверка возможности классификации текстов по гендерному признаку автора отзыва (мужские и женские) при помощи статистических и лингвистических показателей.

Для этой цели была написана программа, которая извлекала из общего корпуса сведения об авторе отзыва. Использовались только те отзывы, в которых у автора было указано и имя и фамилия. Отбирались самые часто встречающиеся женские и мужские имена, по которым затем составлялись корпусы. Для каждого из корпусов считались такие характеристики как средняя длина отзыва и распределение слов по частям речи.

Результаты представлены в Таблице 3.

Таблица 3. *Показатели, рассчитанные по подкорпусам мужских и женских отзывов*

Пол автора	Ср. длина отзыва (в словах)	% от общего числа слов в отзыве					
		Имя сущ.	Предлог	Наречие	Местоимение	Имя прил.	Глагол
мужской	108,11	31,35	11,54	7,08	2,85	12,54	11,31
женский	92,712	29,15	10,62	7,79	3,55	13,09	11,65

Разница данных показателей проверялась на значимость. Поскольку распределения выборок отзывов не являются нормальными, то для проверки значимости использовался критерий Бейли, который можно использовать для

сравнения малых выборок при полной неизвестности структуры генеральных совокупностей [Плохинский, 1980, с. 106].

Формула для расчета критерия Бейли выглядит следующим образом:

$$t_d = \frac{M_2 - M_1}{\sqrt{m_1^2 + m_2^2}} \geq t_{st} \left\{ \nu_d = \frac{\nu_1 \cdot \nu_2}{\nu_2 a_1 \cdot \nu_1 a_2} \right\} \begin{Bmatrix} \beta_1 = & 0{,}9 \\ \beta_2 = & 0{,}95 \\ \beta_3 = & 0{,}99 \end{Bmatrix}$$

Расчет критерия Бейли для показателя «средняя длина отзыва» приведен ниже.

$n_{\text{м}} = 500;\ n_{\text{ж}} = 500$ – объемы выборок,

$\text{M}_{\text{м}} = 108{,}1;\ \text{M}_{\text{ж}} = 92{,}7$ – сравниваемые средние,

$\sigma_{\text{м}} = 86{,}9;\ \sigma_{\text{ж}} = 71{,}7$ – среднее квадратичное отклонение,

$m_{\text{м}}^2 = \frac{86{,}9^2}{500} = 15{,}1;\ m_{\text{ж}}^2 = \frac{71{,}7^2}{500} = 10{,}3$ – квадраты ошибок репрезентативности,

$\text{a}_{\text{м}} = \left(\frac{15{,}1}{25{,}4}\right)^2 = 0{,}35;\ \text{a}_{\text{ж}} = \left(\frac{10{,}3}{25{,}4}\right)^2 = 0{,}16$ – дополнительные величины для вычисления числа степеней свободы(ν_d), рассчитываемые по формулам:

$\text{a}_{\text{м}} = \left(\frac{m_{\text{м}}^2}{m_d^2}\right)^2$ и $\text{a}_{\text{ж}} = \left(\frac{m_{\text{ж}}^2}{m_d^2}\right)^2;\ \left(m^2 = \frac{\sigma^2}{n}\right)$

$\nu_{\text{м}} = 500 - 1 = 499,\ \nu_{\text{ж}} = 500 - 1 = 499$ – объемы выборок, уменьшенные на единицу,

$m_d^2 = 15{,}1 + 10{,}3 = 25{,}4$ – квадрат ошибки выборочной разницы,

$m_d = \sqrt{25{,}4} = 5$ – ошибка выборочной разницы.

Для проверки, достоверна ли разность, значение критерия Бейли $(t_{d)}$ сравнивается со стандартными значениями критерия Стьюдента из таблицы критериев Стьюдента по числу степеней свободы (ν_d) для трех порогов вероятности:

$\nu_d = \frac{499 \cdot 499}{500 \cdot 0{,}35 + 500 \cdot 0{,}16} = 963 \rightarrow t_{st}\{1{,}65; 1{,}96; 2{,}58\}$ – стандартные значения критерия Стьюдента из таблицы критериев Стьюдента по числу степеней свободы ($\nu_d = 963$) для трех порогов вероятности,

$t_d = \frac{15}{5} = 3 -$ критерий Бейли,

$t_d = 3 \geq t_{st}\{1{,}65; 1{,}96; 2{,}58\}$

Поскольку полученное значение 3 больше значения 2,58 из таблицы критериев Стьюдента, то разница между сравниваемыми величинами достоверна. Это означает, что разница средних длин отзывов в мужской и женской коллекции является статистически значимой.

Проверка разницы распределения слов по частям речи показала, что данные отличия не являются значимыми.

Таким образом, задача классификации текстов по гендерному признаку авторов отзывов – отдельная и нетривиальная задача. Она требует разработки системы более сложных показателей. В данном исследовании не ставилась цель ее решить.

Для дальнейшего анализа использовался общий корпус отзывов.

3.2. Этапы разработки тезауруса оценочной лексики

3.2.1. Извлечение оценочных слов и словосочетаний и группировка в семантические категории

Создание словаря происходило в несколько этапов. На первом этапе вручную извлекались оценочные слова или словосочетания и относящиеся к ним оцениваемые параметры. Для этого вручную была проанализирована часть отзывов по следующей схеме: отдельно выписывалось оценочное слово или словосочетание и отдельно – оцениваемый параметр, т.е. то, к чему эта оценка относится. Кроме того, у оценочного слова сразу же определялась полярность – положительная или отрицательная. Например:

(1) Вкусный кофе (+)

Оценочное слово *вкусный* относится к оцениваемому параметру *кофе* и обладает положительной полярностью *(+)*.

(2) Интуитивно понятные настройки (+)

Оценочное словосочетание *интуитивно понятные* относится к оцениваемому параметру *настройки* и обладает положительной полярностью (+).

(3) Изумительная пенка (+)

Оценочное слово *изумительная* относится к оцениваемому параметру *пенка* и обладает положительной полярностью *(+)*.

(4) Страшный шум (-)

Оценочное слово *страшный* относится к оцениваемому параметру *шум* и обладает отрицательной полярностью *(-)*.

На втором этапе полученные оцениваемые параметры были сгруппированы во взаимосвязанные семантические категории, упорядоченные по иерархическому принципу (Рис.1).

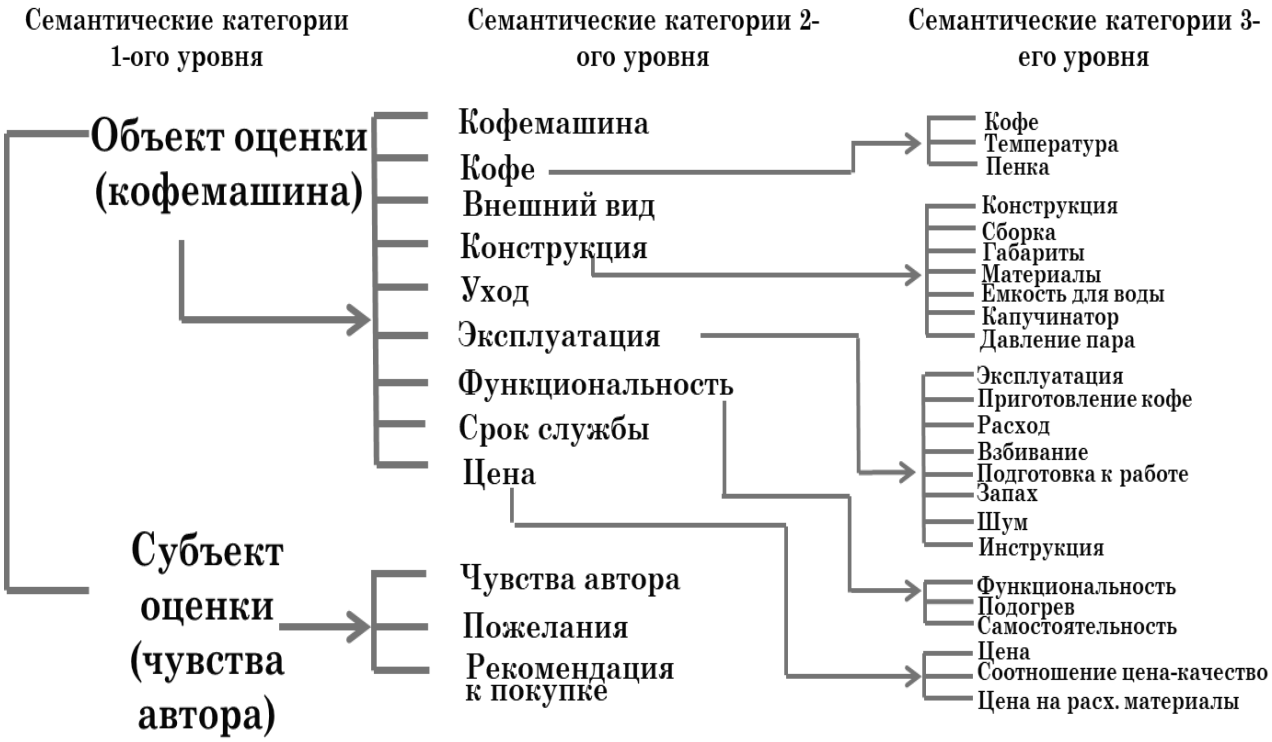

Рис. 1. Группировка оцениваемых параметров в обобщенные семантические категории, упорядоченные по иерархическому принципу

На первом уровне иерархии все оцениваемые параметры были разделены на две большие группы, в зависимости от того, к чему относится оценка: объект оценки (кофемашина) и сам субъект оценки (чувства автора отзыва).

При этом, категория «объект оценки (кофемашина)» включает в себя такие семантические категории второго уровня, как «кофе», «внешний вид», «конструкция», «уход», «эксплуатация», «функциональность», «срок службы», «цена» и оценки относительно изделия в целом. В категорию «субъект оценки (чувства автора)» входят «пожелания», «рекомендация к покупке» и оценки, выражающие чувства автора в целом.

Семантические категории второго уровня, в свою очередь, могут включать в себя категории третьего уровня. Например, в категории «кофе» выделяются категории: «температура», «пенка» и прочие оценки, выражающие отношение к кофе в целом. Категории «эксплуатация» соответствуют такие категории третьего уровня как «приготовление кофе», «подготовка к работе», «шум», «взбивание», «расход (воды и кофе)», «инструкция», «запах» и прочие оценки эксплуатации, не попадающие ни в одну из ранее названных категорий.

Основанием для выделения данных семантических категорий служила частота встречаемости описываемых параметров. Так, если, например, было замечено, что в отзывах часто встречаются оценки издаваемых кофемашиной звуков при приготовлении кофе, то данные оценки выделялись в отдельную категорию – «шум». Или, оценки, описывающие сложность или легкость промывки, частоту требуемых чисток, были отнесены к категории «уход». Затем, эти полученные категории «шум» и «уход» были сгруппированы в категорию «эксплуатация». Таким образом, категории выделялись от частных – к более общим. В результате получилась иерархическая структура оценок.

В результате, на первом этапе был получен тезаурус, в котором каждой семантической категории соответствуют перечни оценочных слов или

словосочетаний с указанием полярности (положительной или отрицательной).

Объем словаря после первого этапа составил 1500 словарных единиц.

3.2.2. Расширение тезауруса с помощью правил

На втором этапе полученный словарь был дополнен автоматически на основании набора правил. Пример правила для автоматического извлечения словосочетаний представлен в Приложении Б.

Во-первых, были использованы правила, учитывающие сочетания лексем. Поиску подлежали такие слова, которые довольно часто сочетаются с оценочными лексемами, т.е. объекты и аспекты, которые оцениваются в отзывах. После установления таких слов извлеклись словосочетания с их участием. Добавление таких словосочетаний было сделано автоматически. Для этой проверки все слова были приведены к словарной форме, их частеречная принадлежность была определена с помощью морфологического анализатора pymorphy2 [Korobov, M. 2015].

Таким образом, автоматически извлекались словосочетания при выполнении правил:

1. ***Если*** *Часть речи слова «X» = Прил.* ***И*** *за этим словом следует слово из списка существительных, представленных в Таблице 4.*

Таблица 4*. Существительные для автоматического извлечения словосочетаний с их участием*

№	Объект/аспект (существительное)	Примеры словосочетаний
1	кофе	добротный кофе (+), мерзкий кофе (-)
2	пенка	эффектная пенка (+), махонькая пенка (-)
3	цена	разумная цена (+), дикая цена (-)
4	дизайн	премиальный дизайн (+), убогий дизайн (-)
5	модель	удачная модель (+), замороченная модель (-)

6	машина	бесконфликтная машина (+), плохая машина (-)
7	кофемашина	умная кофемашина (+), капризная кофемашина (-)
8	шум	незначительный шум (+), страшный шум (-)

2. ***Если*** *Часть речи слова «X» = Нареч.* ***И*** *до/после этого слова следует слово из списка глаголов, представленных в Таблице 5.*

Таблица 5*. Глаголы для автоматического извлечения словосочетаний с их участием*

№	Действие (глагол)	Примеры словосочетаний
1	варить	превосходно варит (+), долго варит (-)
2	готовить	шикарно готовит (+), шумно готовит (-)
3	мыть	удобно мыть (+), часто мыть (-)
4	ухаживать	легко ухаживать (+), трудно ухаживать (-)
5	чистить	легко чистить (+), неудобно чистить (-)
6	ломаться	редко ломается (+), постоянно ломается (-)
7	делать	вкусно делает (+), ужасно делает (-)

Объем словаря после применения двух вышеуказанных правил составил 1810 словарных единиц, т.е. прирост составил 310 словарных единиц (21%).

Кроме этого, использовались правила, основанные на поиске слов-операторов, которые усиливают или ослабляют оценочное значение. Такие слова-операторы чаще всего являются наречиями, и они практически всегда используются в сочетаниях с оценочными словами. Поэтому автоматическое извлечение словосочетаний с их участием могло бы пополнить словарь. Таким образом, извлекались словосочетания при выполнении правила:

3. ***Если*** *«X» – слово из списка, представленного в Таблице 6,* ***И*** *за ним следует словосочетание «X» (прилагательное) + «Y» (существительное)*

***Таблица 6**. Слова-операторы для автоматического извлечения словосочетаний с их участием*

№	Объект/аспект (существительное)	Примеры словосочетаний
1	очень	(очень) быстрое образование пара (+)
2	чрезвычайно	(чрезвычайно) стойкая пенка (+)
3	безумно	(безумно) вкусный кофе (+)
4	излишне	(излишне) нежная техника (-)
5	чуть	(чуть) теплый кофе (-)
6	недостаточно	(недостаточно) густая пена (+)
7	почти	есть (почти) все функции (+)
8	слишком	(слишком) мелкие порции (-)
9	невероятно	(невероятно) громкая очистка (-)
10	действительно	(действительно) стоящий кофе (+)
11	весьма	(весьма) достойные характеристики за такую цену (+)
12	самый	(самый) оптимальный выбор (+)
13	до ужаса	(до ужаса) слабый кофе (-)
14	ужасно	(ужасно) шумная машина (-)

Сами слова-операторы в словарь не включались, они использовались только для поиска словосочетаний, в которых данные слова употребляются.

Объем словаря после применения правила, учитывающего слова-операторы, составил 2098 словарных единиц, т.е. прирост составил 288 словарных единиц (19%).

Словосочетания, извлеченные с помощью трех вышеназванных правил, просматривались вручную, поскольку не все такие словосочетания являются оценочными. Оценочные словосочетания добавлялись в словарь, а словосочетания, не являющиеся оценочными, исключались из дальнейшего рассмотрения.

Кроме этого, было замечено, что для многих объектов, к которым относятся оценки, существует более одного наименования. Для пополнения словаря использовалось выделение контекстных синонимов для основных

классов объектов, которых было бы достаточно для обеспечения точности распознавания оцениваемых объектов. Например, кофемашину в отзывах часто могут называть машиной, моделью и т.д.

Не только объекты могут иметь синонимичные наименования, но и действия. Например, очень часто встречаются синонимичные наименования процесса приготовления кофе: варит кофе, готовит кофе, делает кофе.

Таким образом, мы выделили синонимичные наименования объектов и действий, которые встречаются в отзывах и автоматически добавили различные комбинации оценочных словосочетаний с их участием в словарь.

Например, у нас было в словаре оценочное сочетание «большой бак для воды». Но было замечено, что в отзывах «бак для воды» авторы часто называют «резервуаром». Поэтому мы автоматически сгенерировали словосочетание «большой резервуар для воды» и добавили его в наш словарь. Аналогичным образом были проработаны все синонимичные варианты для других слов.

Автоматическое добавление словосочетаний с синонимичными наименованиями каких-либо объектов или действий позволило увеличить полноту охвата словаря.

Контекстные синонимичные слова и выражения представлены в Таблице 7.

Таблица 7. *Контекстные синонимичные наименования объектов и действий*

Объект/действие	Синонимичные наименования объекта/действия
кофемашина	машина
	аппарат
	вещь
	устройство
	модель
	прибор
	машинка
	агрегат
внешний вид	дизайн
шум	звук
	гул

размеры	габариты
инструкция	руководство
емкость (для воды)	резервуар (для воды)
	бак (для воды)
	отсек (для воды)
цена	стоимость
выглядит	смотрится
готовит	варит
взбивает	вспенивает

Объем словаря после добавления словосочетаний с синонимичными наименованиями объектов и действий составил 2900 словарных единиц, т.е. прирост составил 802 словарных единицы (53%).

Таким образом, объем словаря до применения правил составлял – 1500 словарных единиц, после применения всех правил – 2900 словарных единиц. То есть объем словаря увеличился на 1400 словарных единиц (на 93%).

Поэтапное расширение словаря с указанием прироста после каждого этапа представлено в Таблице 8.

Таблица 8. *Прирост словаря после применения правил*

Вариант словаря	Объем словаря, сл.ед.	Прирост, сл.ед.	Прирост, %
Исходный	1500		
После применения правил 1-2 (сущ+прил, нареч+гл)	1810	310	21%
После применения правила 3 (учитывающего слова-операторы)	2098	288	19%
После добавления контекстных синонимичных наименований объектов и действий	2900	802	53%
Итого, прирост:		1400	93%

3.2.3. Характеристика полученного тезауруса

Таким образом, был подготовлен тезаурус оценочных слов и словосочетаний, упорядоченных по семантическим категориям трех уровней.

Тезаурус доступен в сети Интернет по адресу: https://github.com/Chechneva/CoffeeThesaurus.

В полученном тезаурусе словарной единицей является слово или словосочетание. Слово или словосочетание может быть представлено либо лексемой, выражающей оценку, либо лексемой, выражающей оценку, в сочетании с лексемой, обозначающей объект или аспект объекта. Слова не приводились к начальной форме, в словаре они представлены в той форме, в которой они чаще всего встречаются в отзывах.

Отрывок из полученного тезауруса представлен в Таблице 9.

Таблица 9. *Отрывок из тезауруса оценочных слов и словосочетаний*

Семантические категории 1-ого уровня	Семантические категории 2-ого уровня	Семантические категории 3-его уровня	Оценочное слово или словосочетание	Полярность
объект оценки (кофемашина)	внешний вид	внешний вид	великолепный дизайн	+
объект оценки (кофемашина)	внешний вид	внешний вид	неброский дизайн	-
объект оценки (кофемашина)	конструкция	габариты	компактность	+
объект оценки (кофемашина)	конструкция	габариты	великоваты габариты	-
объект оценки (кофемашина)	конструкция	давление пара	высокое давление	+
объект оценки (кофемашина)	конструкция	давление пара	давление пара маленькое	-
объект оценки (кофемашина)	конструкция	капучинатор	уникальный капучинатор	+
объект оценки (кофемашина)	конструкция	капучинатор	хлипкий капучинатор	-
объект оценки (кофемашина)	конструкция	материалы	приятен на ощупь	+
объект оценки (кофемашина)	конструкция	материалы	дешевый пластик	-
объект оценки (кофемашина)	конструкция	сборка	надёжная сборка	+
объект оценки (кофемашина)	конструкция	сборка	отвратная сборка	-
объект оценки (кофемашина)	кофе	кофе	благородный кофе	+

объект оценки (кофемашина)	кофе	кофе	бурда	-
объект оценки (кофемашина)	кофе	пенка	восхитительная пенка	+
объект оценки (кофемашина)	кофе	пенка	жалкое подобие пенки	-
объект оценки (кофемашина)	кофемашина	кофемашина	великолепный агрегат	+
объект оценки (кофемашина)	кофемашина	кофемашина	капризная машина	-
объект оценки (кофемашина)	срок службы	срок службы	неубиваемая	+
объект оценки (кофемашина)	срок службы	срок службы	быстро сломалась	-
объект оценки (кофемашина)	уход	уход	легко мыть	+
объект оценки (кофемашина)	уход	уход	промывка неудобная	-
объект оценки (кофемашина)	функциональность	функциональность	богатый функционал	+
объект оценки (кофемашина)	функциональность	функциональность	малая функциональность	-
объект оценки (кофемашина)	цена	соотношение цена - качество	за свои деньги отлично	+
объект оценки (кофемашина)	цена	соотношение цена - качество	деньги на ветер	-
объект оценки (кофемашина)	цена	цена	доступная цена	+
объект оценки (кофемашина)	цена	цена	ощутимая цена	-
объект оценки (кофемашина)	эксплуатация	взбивание	отлично взбивает	+
объект оценки (кофемашина)	эксплуатация	взбивание	сложно взбивать	-
объект оценки (кофемашина)	эксплуатация	запах	запаха пластмассы в напитке нет	+
объект оценки (кофемашина)	эксплуатация	запах	воняет	-
объект оценки (кофемашина)	эксплуатация	инструкция	грамотная инструкция	+
объект оценки (кофемашина)	эксплуатация	инструкция	корявое руководство	-
объект оценки (кофемашина)	эксплуатация	подготовка к работе	быстрый старт	+
объект оценки (кофемашина)	эксплуатация	подготовка к работе	долгий прогрев	-
объект оценки (кофемашина)	эксплуатация	расход	экономно расходует	+
объект оценки (кофемашина)	эксплуатация	расход	большой расход	-
объект оценки (кофемашина)	эксплуатация	шум	малошумная	+
объект оценки (кофемашина)	эксплуатация	шум	свистит	-
объект оценки (кофемашина)	эксплуатация	эксплуатация	легка в управлении	+
объект оценки (кофемашина)	эксплуатация	эксплуатация	дребезжит	-
субъект оценки (чувства автора отзыва)	чувства автора	рекомендация	рекомендую	+
субъект оценки (чувства автора отзыва)	чувства автора	чувства автора	доволен	+

Распределение словарных единиц по семантическим категориям представлено в Таблице 10.

Таблица 10. *Распределение словарных единиц по семантическим категориям*

Семантические категории 1-ого уровня	Семантические категории 2-ого уровня	Семантические категории 3-его уровня	Кол-во словарных единиц (сл.ед.)
Всего слов			2900
объект оценки (кофемашина)	кофемашина	кофемашина	651
	кофемашина Итог		***651***
объект оценки (кофемашина)	эксплуатация	эксплуатация	156
объект оценки (кофемашина)	эксплуатация	приготовление кофе	147
объект оценки (кофемашина)	эксплуатация	шум	125
объект оценки (кофемашина)	эксплуатация	инструкция	46
объект оценки (кофемашина)	эксплуатация	взбивание	22
объект оценки (кофемашина)	эксплуатация	подготовка к работе	20
объект оценки (кофемашина)	эксплуатация	расход	19
объект оценки (кофемашина)	эксплуатация	запах	12
	эксплуатация Итог		***547***
объект оценки (кофемашина)	кофе	пенка	213
объект оценки (кофемашина)	кофе	кофе	210
объект оценки (кофемашина)	кофе	температура кофе	17
	кофе Итог		***440***
объект оценки (кофемашина)	цена	соотношение цена - качество	196
объект оценки (кофемашина)	цена	цена	109
объект оценки (кофемашина)	цена	цена на расходные материалы	13
	цена Итог		***318***
объект оценки (кофемашина)	конструкция	конструкция	100

объект оценки (кофемашина)	конструкция	габариты	52
объект оценки (кофемашина)	конструкция	емкость для воды	46
объект оценки (кофемашина)	конструкция	материалы	32
объект оценки (кофемашина)	конструкция	капучинатор	30
объект оценки (кофемашина)	конструкция	сборка	28
объект оценки (кофемашина)	конструкция	давление пара	11
	конструкция Итог		***299***
объект оценки (кофемашина)	внешний вид	внешний вид	246
	внешний вид Итог		***246***
объект оценки (кофемашина)	функциональность	функциональность	125
объект оценки (кофемашина)	функциональность	подогрев	15
объект оценки (кофемашина)	функциональность	самостоятельность кофемашины	9
	функциональность Итог		***149***
объект оценки (кофемашина)	срок службы	срок службы	97
	срок службы Итог		***97***
объект оценки (кофемашина)	уход	уход	74
	уход Итог		***74***
субъект оценки (чувства автора)	субъект оценки (чувства автора)	чувства автора	55
субъект оценки (чувства автора)	субъект оценки (чувства автора)	рекомендация	16
субъект оценки (чувства автора)	субъект оценки (чувства автора)	пожелания автора	8
	субъект оценки (чувства автора отзыва)		***79***

Как видно из Таблицы 10, на втором уровне иерархии больше всего словарных единиц содержится в категории «кофемашина» (651 сл.ед.) – т.е. это оценки, описывающие изделие целом (например, хорошая машина, удачный вариант). Это можно объяснить тем, что, действительно, в отзывах встречается много оценок в отношении непосредственно самого объекта, т.е. кофемашины. А с другой стороны, в словаре много синонимичных

наименований кофемашины, для каждого из которых повторяются все оценочные выражения.

На втором месте по количеству словарных единиц находится категория «эксплуатация» (547 сл.ед.). Авторы отзывов уделяют много внимания эксплуатационным характеристикам изделия, включая процесс приготовления кофе, взбивание, шум при работе, подготовку к работе, расход воды и т.д.

На третьем месте находится продукт – категория «кофе». Поскольку коемашины покупают, прежде всего, для того, чтобы получить результат их работы – кофе, то в этой категории содержится большое разнообразие оценок (440 сл.ед.). Примечательно, что примерно половину от этого числа (213 сл.ед.) составляют оценки не самого кофе, а его части, а именно – пенки.

Четвертое место занимают оценки категории «цена» (318 сл.ед.). Оценивается как непосредственно сама цена изделия, так и то, насколько она соотносится с его качеством, т.е. стоит ли товар своих денег.

Значительную долю составляют оценки конструкции изделия (299 сл.ед.). В отзывах часто встречаются описания конструкционных характеристик – габаритов, сборки; уделяется внимание отдельным составным частям изделия – капучинатору и емкости для воды.

Несмотря на то, что кофемашины покупают, прежде всего, для того, чтобы готовить кофе, а не из-за эстетических потребностей, их внешнему виду – оценкам дизайна, оценкам того, насколько удачно кофемашины вписываются в интерьер – в отзывах уделяется много внимания. Категория «внешний вид» (246 сл.ед.) находится на пятом месте.

На шестом месте находятся оценки того, насколько реализованы те или иные функции в изделии, сгруппированные в категорию «функциональность» (149 сл.ед.).

На седьмом месте находится категория «срок службы» (97 сл.ед.). Сюда относятся описания различных поломок изделия, сколько оно прослужило, оценки сервисного обслуживания.

Отличные от всех оценки, выражающие чувства автора, находятся на восьмом месте (79 сл.ед.). Сюда относятся оценки, которые относятся не к объекту отзыва, а к субъекту. Это эмоции людей: радость, довольствование чем-то, пожелания и т.д.

И на последнем месте находится категория «уход» (74 сл.ед.). Сюда относятся оценки того, насколько легко ухаживать за изделием, в том числе осуществлять чистку, промывку.

Анализируя структуру семантических категорий, можно заметить, что выделенные категории соотносятся с различными классификациями оценок, представленными в Параграфе 1.3.

Так, традиционное деление оценок на общие и частные находит отражение в словаре. Например, оценки, характеризующие изделие (кофемашину) в целом чаще носят общий характер (хорошая модель, отличная машина). Оценки, характеризующие аспекты чаще носят частный характер (стойкая пенка, высокое давление пара).

Если рассматривать классификацию оценок Арутюновой, то можно заметить, что полученные нами семантические категории во многом совпадают с ее группами.

Таблица 11. *Сопоставление семантических категорий тезауруса оценочных слов и классификации оценок Арутюновой*

Группа оценок из классификации Арутюновой	Семантические категории тезауруса оценочных слов (и примеры)
сенсорно-вкусовые	«кофе», «конструкция» (вкусный кофе, приятен на ощупь)
психологические	«субъект оценки (чувства автора отзыва)» (безумно рад, жалею о покупке)
эстетические	«внешний вид» (выглядит великолепно, замечательный дизайн)
этические	«цена», «срок службы» (заламывают цены, обманули при покупке)

утилитарные	«функциональность» (дисплей - полезная опция, подставка для прогрева кружек абсолютно бесполезна)
нормативные	«эксплуатация» (работает как надо, грамотная инструкция)
телеологические	«цена», «кофемашина» (стоит своих денег, удачное приобретение)

Структура категорий полученного нами тезауруса оценочных слов и словосочетаний во многом находит отражение в типологии оценок, которую предложил логик Георг Хенрик фон-Вригт (Таблица 12).

Таблица 12. *Сопоставление семантических категорий тезауруса оценочных слов и классификации оценок Георга Хенрика фон-Вригта*

Группа оценок из классификации Георга Хенрика фон-Вригта	Семантические категории тезауруса оценочных слов (и примеры)
Инструментальные	«конструкция» (отличный капучинатор)
Технические	«конструкция», «эксплуатация» (мощная подача пара, вибрирует)
Оценки благоприятствования	«кофе», «субъект оценки (чувства автора отзыва)» (красивый кофе, приятно удивлен)
Утилитарные	«кофемашина» (бесполезный прибор, практичная вещь)
Медицинские	«конструкция» (опасный для здоровья из-за коррозии)
Гедонистические	«кофе», «субъект оценки (чувства автора отзыва)» (вкус кофе не передать, наслаждаться, полный восторг)

Сопоставление семантических категорий тезауруса оценочных слов с классификацией оценок, предложенной Н.Н. Мироновой, представлено в Таблице 13.

Таблица 13. *Сопоставление семантических категорий тезауруса оценочных слов и классификации оценок Н.Н. Мироновой*

Группа оценок из классификации Н.Н. Мироновой	Семантические категории тезауруса оценочных слов (и примеры)
аксиологические оценки (этические, эстетические, утилитарные, эмоциональные)	«цена», «кофе», «внешний вид» (хорошая цена, изумительная пенка, лаконичный дизайн)
модальные (необходимость, долженствование, возможность)	«уход», «эксплуатация» (нужна частая чистка, можно все отрегулировать на свой вкус)
экзистенциальные	«функциональность» (смысла в подогревателе чашек не узрел)
временные	«эксплуатация», «срок службы» (быстро готовит, долгий прогрев, быстро сломалась)
оценки величин	«конструкция», «цена» (громоздкая, большой резервуар для воды, высокая цена)
пространственные оценки	«конструкция» (легко помещается практически везде, прекрасно вмещается)

Как видно из Таблиц 11-13, структура оценок такой узкой предметной области как отзывы на кофемашины во многом носит универсальный характер.

В связи с этим, можно предположить, что другие классы товаров (фотоаппараты, ноутбуки, телефоны) будут иметь очень схожие семантические категории оценок. В отзывах на них, вероятней всего, также будут встречаться оценки как самих изделий в целом, так и их эксплуатационных свойств, конструкционных характеристик, внешнего вида, цены, ухода, результатов их работы и т.д. Авторы отзывов, с большой долей вероятности, будут описывать свои личные эмоции от опыта использования данных изделий.

Таким образом, полученная нами структура семантических категорий оценок может носить в некоторой степени универсальный характер применительно к отзывам на разные группы товаров.

Для более детального анализа структуры оценочных словосочетаний были составлены частотные словари отдельных слов из словаря по частям речи. Результаты представлены в Таблице 14.

Таблица 14. *Распределение словарных единиц по семантическим категориям*

Имя прилагательное			
№	Слово	Часто	Кол-во
1	вкусный	1315	4
2	хороший	1216	23
3	отличный	988	20
4	большой	669	13
5	компактный	480	1
6	удобный	471	8
7	горячий	415	2
8	простой	297	9
9	небольшой	297	11
10	маленький	294	8

Наречие			
№	Слово	Часто	Кол-во
1	быстро	651	9
2	много	534	7
3	легко	486	9
4	удобно	470	5
5	хорошо	393	18
6	долго	366	4
7	отлично	274	15
8	вкусно	151	2
9	правильно	93	3
10	сложно	85	5

Глагол			
1	нравиться	486	10
2	рекомендов	386	2
3	советовать	238	2
4	сломаться	216	2
5	шуметь	161	1
6	пожалеть	136	2
7	радовать	130	7
8	устраивать	115	5
9	жалеть	85	3
10	справляться	79	4

Имя существительное			
1	хлипкость	5	2
2	шок	5	6
3	продуманнос	5	4
4	недоразумен	4	2
5	малютка	4	2
6	лидер	4	3
7	украшение	3	1
8	соратник	3	1
9	ржавчина	3	2
10	пойло	3	1

Как видно из Таблицы 14, некоторые слова встречаются в большом количестве категорий:

Хороший (23 категории)

Отлично (15 категорий)

Другие же слова представлены только в одной из категорий:

Компактный (1 категория)

Украшение (1 категория)

Это связано с тем, что существуют «универсальные» оценочные слова, при помощи которых можно описать практически любой объект или аспект, а также есть более специальные оценочные лексемы, которые отражают специфические свойства объекта.

Также необходимо отметить, что среди единиц словаря есть как явные (эксплицитные) оценки:

Ущербный дизайн (-)

Приятная цена (+)

Так и дескриптивные (имплицитные) оценки:

Дала течь (-)

Справится даже ребенок (+)

Если в явных оценках оценочный компонент выражен прямо, то в дескриптивных он не столь нагляден, однако описанное положение вещей может однозначно расцениваться как хорошее (желаемое) или плохое (нежелаемое).

3.3. Экспериментальная проверка

Разработанный тезаурус мы применили для автоматического анализа тональности. Были проанализированы новые отзывы, не включенные в первоначальный корпус. Результат применения словаря для первых 20 отзывов представлен в Таблице 15.

Таблица 15. *Результаты применения тезауруса оценочной лексики*

ID отзыва	внешний вид	конструкция	кофе	кофемашина	срок службы	уход	Функц-ть	цена	чувства автор	эксплуатация	Общий итог
17955112	2(2;0)		3(3;0)	1(1;0)				1(1;0)	1(1;0)	1(1;0)	9(9;0)
21552269	1(1;0)			1(1;0)					1(1;0)		3(3;0)
26262233			2(2;0)	1(1;0)					3(3;0)	2(2;0)	8(8;0)
29922395				1(1;0)				1(1;0)	1(1;0)	1(1;0)	4(4;0)
29966280			1(1;0)	3(3;0)				1(1;0)	2(2;0)	-1(1;-2)	6(8;-2)
31949589		-1(0;-1)	1(1;0)	2(3;-1)						0(1;-1)	2(5;-3)
32461420			-2(0;-2)	1(1;0)			2(2;0)			2(2;0)	3(5;-2)
34561124	1(1;0)		1(1;0)						2(2;0)	1(1;0)	5(5;0)
34579963				1(1;0)				-1(0;-1)		-1(0;-1)	-1(1;-2)
34633602	1(1;0)		-1(0;-1)	1(1;0)	-1(0;-1)					1(1;0)	1(3;-2)
35054530		1(1;0)	-2(0;-2)							-1(0;-1)	-2(1;-3)
35150814		3(3;0)		1(1;0)				-1(0;-1)		2(2;0)	5(6;-1)
36340921	2(2;0)		1(2;-1)								3(4;-1)
36671881	1(1;0)	-1(0;-1)	3(3;0)					-1(0;-1)	1(1;0)	1(3;-2)	4(8;-4)
36789441	-1(0;-1)			1(1;0)		1(1;0)		1(1;0)	4(1;0)	1(4;0)	7(8;-1)
67923453		3(3;0)	2(2;0)	3(3;0)				-1(0;-1)	1(1;0)	1(1;0)	9(10;-1)
67924720		2(2;0)	1(1;0)				2(2;0)		1(1;0)	2(2;0)	8(8;0)
68247582	2(2;0)		2(2;0)	5(5;0)		-1(0;-1)			1(1;0)	2(2;0)	11(12;-1)
68436661	2(2;0)		0(1;-1)	2(2;0)		2(2;0)				2(3;-1)	8(9;-1)
68574151	1(1;0)	1(1;0)	2(2;0)								4(4;0)

Результат работы представлен в виде матрицы ID отзыва - семантическая категория 2-ого уровня, в которой строкам соответствуют ID отзыва, а столбцам – встречающиеся в них семантические категории. Ячейка таблицы содержит следующие данные:

- Первое число в скобках - количество встречаемых словарных единиц из тезауруса с положительной полярностью.
- Второе число в скобках - количество встречаемых словарных единиц из тезауруса с отрицательной полярностью.
- Число перед скобками - суммарное значение по каждой категории (т.е. сумма двух вышеназванных чисел).

Кроме этого, была проведена первичная оценка результатов для задачи классификации отзывов по общей полярности. В качестве метрик были выбраны точность (precision) и полнота (recall). Точность в пределах класса – это доля документов, действительно принадлежащих данному классу, относительно всех документов, причисленных к этому классу. Полнота – отношение числа найденных документов, принадлежащих классу, к числу всех документов этого класса.

На точность и полноту тестировалась только правильность определения общей полярности отзывов. Точность и полнота для задачи определения семантических категорий не оценивалась из-за отсутствия размеченного корпуса с указанием семантических категорий. Для тестирования точности и полноты результаты классификации, выполненной при помощи словаря, сравнивались с классификацией, выполненной вручную. Проверка осуществлялась на коллекции из 120 отзывов. Оценка качества результата приведена в Таблице 16.

Таблица 16. *Оценка качества результата*

Класс	Точность	Полнота
положительные	92%	93%
отрицательные	78%	65%

Таким образом, применение словаря дает неплохие показатели точности и полноты. Данные показатели получились ниже для отрицательных отзывов. Это можно объяснить тем, что для определения отрицательной полярности помимо самого словаря использовалось лишь одно правило, учитывающее сочетание частицы «не» со словарными единицами.

Для улучшения показателей точности и полноты для отрицательных отзывов требуется разработка новых правил и настройка системы весов для оценочных слов.

ЗАКЛЮЧЕНИЕ

Таким образом, данная работа посвящена проблеме разработки тезауруса оценочных слов для заданной предметной области.

В работе были проанализированы особенности оценок как предмета изучения философии и лингвистики. Было показано, что оценочная лексика является неотъемлемым компонентом систем автоматического анализа тональности. В работе были рассмотрены подходы к анализу тональности, особое внимание было уделено подходу с использованием словарей оценочной лексики, дана характеристика некоторым существующим словарям оценочных слов.

В работе был описан собственный подход к созданию тезауруса оценочной лексики для заданной предметной области. На основе коллекции из 3850 отзывов на кофемашины был составлен тезаурус оценочных слов и словосочетаний, упорядоченных по семантическим категориям трех уровней, который затем был автоматически расширен с помощью правил. Общий объем словаря составил 2900 словарных единиц.

Полученная нами структура семантических категорий оценок может носить в некоторой степени универсальный характер применительно к отзывам на разные группы товаров.

Применение разработанного тезауруса может служить основой для глубокого анализа тональности, позволяющего определять, не только, как пользователь оценивает объект в целом – положительно или отрицательно, но и выявить, что именно в объекте ему нравится, а что нет.

Первичная оценка результатов показала неплохие значения точности и полноты для задачи классификации отзывов по общей полярности.

В дальнейшем планируется расширить спектр правил и настроить систему весов оценочных слов для улучшения производительности. Кроме этого, планируется более тщательно оценить эффективность тезауруса и попытаться адаптировать тезаурус к другим предметным областям.

СПИСОК ЛИТЕРАТУРЫ

1. Анисимов С. Ф. Духовные ценности: производство и потребление //М.: Мысль. – 1988. 253 с.
2. Анисимов С.Ф. Введение в аксиологию. Учебное пособие для изучающих философию. – М.: Современные тетради, 2001. 128 с.
3. Арутюнова Н.Д. Об объекте общей оценки // Вопросы языкознания. - 1985, № 3. - С.13-24.
4. Арутюнова Н.Д. Типы языковых значений. Оценка. Событие. Факт. – М.: Наука, 1988. 346 с.
5. Бабаева Е. В. Культурно-языковые характеристики отношения к собственности (на материале немецкого и русского языков: диссертация на соискание ученой степени кандидата филологических наук. Волгоград, 1997.
6. Батурин Н.А. Оценочная функция психики: диссертация на соискание ученой степени доктора психологических наук. Санкт-Петербург, 1998.
7. Богданова М. А. Хорошо – слово категории оценки // Вестник Московского государственного областного университета. Серия: Русская филология. 2014. №. 3. С. 26-30.
8. Брожик В. Марксистская теория оценки. – М.: Прогресс, 1982. 264 с.
9. Вестфальская А. В. Оценка и коннотация: современные подходы // Язык и текст. 2015. Т. 2. №. 3. С. 3-11.
10. Вольф Е.М. Варьирование в оценочных структурах // Семантическое и формальное варьирование. – М.: Наука, 1979. С.273-294.
11. Вольф Е.М. Оценочное значение и соотношение признаков "хорошо/плохо" // Вопросы языкознания. 1986. № 5. С.98-106.
12. Вольф Е.М. Функциональная семантика оценки. – М.: Наука, 1985. 228 с.

13. Воркачев С.Г. Оценка и ценность в языке. Монография. – Волгоград: Парадигма, 2006. 186 с.

14. Воронина И. Е., Гончаров В. А. Анализ эмоциональной окраски сообщений в социальных сетях (на примере сети «вконтакте») // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2015. № 4. С. 151-158.

15. Ермаков А. Е., Киселев С. Л. Лингвистическая модель для компьютерного анализа тональности публикаций СМИ // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог. 2005. С. 282-285.

16. Ермаков С. А., Ермакова Л. М. Методы оценки эмоциональной окраски текста. // Вестник Пермского университета. №. 1. 2012. С. 85-89

17. Золотова Г.А. Коммуникативные аспекты русского синтаксиса. Монография. – М.: Наука, 1982. 368 с.

18. Ивин А.А. Основания логики оценок. – М.-Берлин: Диркет-Медиа, 2015. 337 с.

19. Кабирова Г. У. Оценка как языковой концепт // Актуальные вопросы филологических наук: материалы Междунар. науч. конф. (г. Чита, ноябрь 2011 г.). – Чита: Издательство Молодой ученый, 2011. С. 85-87. URL https://moluch.ru/conf/phil/archive/25/1253/
(дата обращения: 08.03.2018).

20. Кант И. Собрание сочинений в 8-ми томах. Том 6. – М.: Чоро, 1994. 613 с.

21. Клековкина М. В., Котельников Е. В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции RCDL-2012», Переславль-Залесский, Россия, 15-18 октября 2012 г.

22. Кочеткова Е. В. Языковые средства выражения негативной оценки мира и человека в поэзии Игоря Северянина: диссертация на соискание ученой степени кандидата филологических наук. – Хабаровск: Дальневосточный государственный университет, 2004.

23. Лукашевич Н. В., Левчик А. В. Создание лексикона оценочных слов русского языка РуСентиЛекс // Труды VI международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем OSTIS-2016», Минск, Белоруссия, 18-20 февраля 2016 г.

24. Лукашевич Н. В., Четвёркин И. И. Комбинирование тезаурусных и корпусных знаний для извлечения оценочных слов // Системы и средства информатики. 2015. Т. 25. № 1. С. 20-33.

25. Маркелова Т.В. Семантика оценки и средства ее выражения в русском языке. Учеб. пособие по спецкурсу. – М.: Изд-во МПУ, 1993. 125 с.

26. Мартыненко Г. Я. Об истоках тезаурусного подхода: к 130-летию со дня рождения К.И. Чуковского // Структурная и прикладная лингвистика. 2012 . № 9. С. 65-07.

27. Меньшиков И. Л., Кудрявцев А. Г. Обзор систем анализа тональности текста на русском языке // Молодой ученый. 2012. №12. С. 140-143.
URL https://moluch.ru/archive/47/5951/
(дата обращения: 10.02.2018).

28. Миронова Н. Н. Оценочный дискурс: проблемы семантического анализа // Известия Российской академии наук. Серия литературы и языка. 1997. Т. 56, № 4. С. 52–59.

29. Мур Дж. Принципы этики. – М.: Прогресс, 1984. 326 с.

30. Неновски Н. Право и ценности. Пер. с болг. – М.: Прогресс, 1987. 248 с.

31. Ожегов С.И. Толковый словарь русского языка. – М.: Мир и Образование, Оникс, 2011. 736 с.

32. Пазельская А. Г., Соловьев А. Н. Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог». – М.: Изд-во РГГУ. 2011. №. 10. С. 17.

33. Плохинский Н.А. Алгоритмы биометрии. – М.: Изд-во Моск. гос. ун-та, 1980. 150 с.

34. Сердобольская Н. В., Толдова С. Ю. Оценочные предикаты: тип оценки и синтаксис конструкции // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог. 2005. С. 436-443.

35. Столович Л.Н. Красота. Добро. Истина: Очерк истории эстетической аксиологии. М.: Республика, 1994, 464 с.

36. Телия В.Н. Коннотативный аспект семантики номинативных единиц. – М.: Наука, 1986. 144 с.

37. Тихонова М. А. Оценочная лексика русского языка: проблемы лексикографирования // Вестник Московского государственного университета печати. 2015. №. 2. С. 352-358

38. Тугаринов В. П. Теория ценностей в марксизме. Л.: – Изд. Ленингр. ун-та, 1968. 124 с.

39. Усминский О. И. Сенсорно-прагматические и типологические аспекты русских тропов: автореф. дис. д-ра филол. наук, Екатеринбург. 1997.

40. Фомина Ю. А. Аспекты изучения языковой оценки // Вестник Челябинского государственного университета. 2007. № 20.

41. Фон В. Г. Х. Логико-философские исследования: избранные труды. – М.: Прогресс. 1986. 600 с.

42. Хохлова М. В. Глава 5. Анализ тональности // Прикладная и компьютерная лингвистика. – М.: Ленанд, 2016. С. 245-258.

43. Щерба Л. В. О частях речи в русском языке // Языковая система и речевая деятельность. 1974. С. 77-100

44. Wilson T., Wiebe J., Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis // Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005. P. 347-354.

45. Andreevskaia A., Bergler S. Mining WordNet for a Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses // EACL. 2006. Vol. 6. P. 209-216

46. Chetviorkin I., Loukachevitch N. Extraction of domain-specific opinion words for similar domains // Proceedings of the Workshop on Information Extraction and Knowledge Acquisition held in conjunction with RANLP 2011. 2011. P. 7-12.

47. Chetviorkin I., Loukachevitch N. Extraction of Russian sentiment lexicon for product meta-domain // Proceedings of COLING 2012. 2012. P. 593-610.

48. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages // International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham, 2015. P. 320-332.

49. Liu B. Sentiment analysis and opinion mining // Synthesis lectures on human language technologies. 2012. Vol. 5. №. 1. P. 1-167.

50. Pang B., Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts // Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004. P. 271.

51. Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and Trends® in Information Retrieval. 2008. Vol. 2. №. 1–2. P. 1-135.

52. Qiu G. et al. Expanding domain sentiment lexicon through double propagation // IJCAI. – 2009. Vol. 9. P. 1199-1204.

53. Steinberger J. et al. Creating sentiment dictionaries via triangulation //Decision Support Systems. 2012. Vol. №. 4. P. 689-694.

54. Thelwall, M., Buckley, K., Paltoglou, G. Sentiment strength detection for the social Web // Journal of the American Society for Information Science and Technology, 63(1), 2012 P. 163-173.

55. Volkova S., Wilson T., Yarowsky D. Exploring Sentiment in Social Media: Bootstrapping Subjectivity Clues from Multilingual Twitter Streams // ACL (2). 2013. P. 505-510.

56. Ру Тез [Электронный ресурс]:
URL: http://www.labinform.ru/pub/ruthes/index.htm
(дата обращения: 18.03.2018).

57. Анализ тональности текста (на примере русского и английского языков) [Электронный ресурс]:
URL:http://www.i-teco.ru/solutions/business_intelligence_products/analiz_tonalnosti_teksta/
(дата обращения: 16.03.2018).

58. Проект ВААЛ [Электронный ресурс]:
URL: http://www.vaal.ru/
(дата обращения: 08.03.2018).

59. RCO Fact Extractor SDK [Электронный ресурс]:
URL: http://www.rco.ru/product.asp?ob_no=5047
(дата обращения: 08.03.2018).

60. API Маркета [Электронный ресурс]:
URL:https://tech.yandex.ru/market/
(дата обращения: 10.04.2018).

ПРИЛОЖЕНИЕ А. КОД ПРОГРАММЫ ДЛЯ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ ОТЗЫВОВ С РЕСУРСА ЯНДЕКС.МАРКЕТ

```
import http.client
import json
import time
headers = {"Host": "api.content.market.yandex.ru", "Accept": "*/*",
"Authorization": "IPjAnHXl5xpdNdJcUfl8q8hT8CnrSJ"}
conn = http.client.HTTPSConnection("api.content.market.yandex.ru")
for j in range(1,11):
    conn.request("GET",'/v1/category/90589/models.json?geo_id=10174
&count=30&page='+str(j),headers=headers)
    result = conn.getresponse().read().decode("utf-8")
    b=open("bazapredv.txt",'ab')
    b.write(result.encode('utf-8'))
    b.close()
    js=json.loads(result)
    for i in js['models']['items']:
        conn =
http.client.HTTPSConnection("api.content.market.yandex.ru")
        conn.request("GET",
"/v1/model/"+str(i['id'])+"/opinion.json?sort=rank&count=30", headers=headers)
        st = conn.getresponse().read().decode("utf-8")
        time.sleep(1)
        f=open("bazaotz.txt",'ab')
        f.write(st.encode('utf-8'))
        f.close()
```

ПРИЛОЖЕНИЕ Б. ПРИМЕР ПРАВИЛА ДЛЯ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ СЛОВОСОЧЕТАНИЙ

```
mport json
import pymorphy2
morph=pymorphy2.MorphAnalyzer()
ids=[]
e=open("rule1sentenceverbs.txt",'w',encoding='utf-8')
for i in range(1,7):
    f=open('bazaotzkofe'+str(i)+'.txt','r',encoding='utf-8')
    st=f.read()
    f.close()
    js=json.loads(st)
    for model in js:
        if 'modelOpinions' in model['opinions']:
            for op in model['opinions']['modelOpinions']['opinion']:
                if not op["id"] in ids:
                    ids.append(op["id"])
                    data=''
                    if 'text' in op:
                        data=op['text']
                    if 'pro' in op:
                        data=data+'.'+op['pro']
                    if 'contra' in op:
                        data=data+'.'+op['contra']
                    for c in u'!?':
                        data=data.replace(c,'.')
                    sentences=data.split('.')
                    for sent in sentences:
                        s=sent
```

```
                    for c in u',.!?:;»«()"':
                        s=s.replace(c,' ')
                    s=s.lower()
                    words=s.split()
                    prevword=''
                    for w in words:
                        if morph.parse(w)[0].normal_form in [u'варить',
u'готовить', u'мыть', u'ухаживать', u'чистить', u'ломаться', u'делать', u'стоить']:
                            if morph.parse(prevword)[0].tag.POS=='ADVB':
                                print(prevword,w,file=e)
                        prevword=w
    e.close()
```