САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

КАФЕДРА ТЕХНОЛОГИИ ПРОГРАММИРОВАНИЯ

Паздникова Мария Валерьевна

Магистерская диссертация

Автоматическое реферирование отзывов на основе аспектно-ориентированного тонального анализа

Направление 02.04.02

«Фундаментальные информатика и информационные технологии»

Магистерская программа «Технологии баз данных»

Научный руководитель,

ст. преп.

Малинина М.А.

Санкт-Петербург

2017

Оглавление

Введение

Последние годы ознаменовались бурным ростом рынка интернет-коммерции. Для русского сегмента всемирной сети эта тенденция наметилась еще в 2008-2009 гг. — и сохраняется до сих пор [1]. Множество торговых площадок предлагают самые разнообразные товары. Принимая решение о покупке, пользователи часто учитывают не только характеристики товаров, но и отзывы, оставленные о них другими покупателями.

Сложившаяся ситуация привела к появлению большого количества площадок для размещения отзывов. А с увеличением числа отзывов, оставляемых в Интернете, растет и потребность в их автоматической обработке. В частности, для данной области актуальной остается задача распознавания эмоциональной окраски текстов, позволяющая на основе текстов отзывов оценить мнение автора (субъекта) по отношению к некоторому товару (объекту), т.е. определить некоторое оценочное отношение. Описанная задача относится к задачам автоматической обработки естественного языка.

Данная задача представляет большой интерес, как с исследовательской точки зрения, так и с практической. Она предполагает решение множества нетривиальных задач из области компьютерной лингвистики и машинного обучения, а ее решение позволит эффективно отслеживать отношение целевой аудитории к продуктам и брендам, своевременно устранять выявленные недостатки и тем самым получать большую прибыль.

Анализ тональности можно условно разделить на объектно-ориентированный и аспектно-ориентированный. К первому относится определение общего оценочного отношения к одному объекту в целом, ко второму — определение оценочного отношения к набору характеристик некоторого объекта. Разницу между ними можно продемонстрировать на примере следующего фрагмента оценочного отзыва:

«*Посредственный* **ресторан**. *Неудобное* **расположение**, **зал** *маленький* и *душный*. А вот сама **еда** была *вкусной*.»

В предложенном фрагменте объектом тональности выступает ресторан. Жирным шрифтом выделены его отдельные характеристики (или аспекты). Заметим, что «ресторан» в данном случае является аспектом категории RESTAURANT#GENERAL. Выделенные курсивом слова выражают тональность. В случае объектно-ориентированного анализа аспекты, как правило, игнорируются, и происходит подсчет оценочных выражений. В данном случае к ним относятся четыре негативных («посредственный», «неудобное», «маленький», «душный») и одно позитивное («вкусной»). Таким образом, общая тональность будет негативной. В то время как при проведении аспектно-ориентированного анализа оценочная информация будет представлена пользователю вместе с аспектами, один из которых в рассматриваемом случае получит положительную оценку (еда). Подобное разделение позволяет получать релевантные результаты согласно интересующим критериям.

Об актуальности данной задачи свидетельствует и интерес к ней научного сообщества. В частности, на ежегодном международном семинаре SemEval, посвящённом решению различных задач компьютерной лингвистики, задача аспектно-ориентированного тонального анализа была представлена с 2014 по 2016 годы [2–4]. Русское научное сообщество не осталось в стороне: в 2015 году в рамках конференции Диалог-2015 было проведено соревнование SentiRuEval [5]. В разделе, посвященном анализу тональности, пять из шести работ относились к тональному анализу относительно аспектов.

В рамках задачи аспектно-ориентированного тонального анализа обычно выделяют следующие этапы [6]:

1. извлечение аспектов;
2. анализ тональности относительно аспектов;
3. реферирование мнения.

Первый этап направлен на выявление во входных данных множества аспектов (чаще всего возможные аспекты заданы заранее). В ряде работ аспекты классифицируется на следующие типы: (i) явный аспект; (ii) неявный аспект; (iii) тональный факт [5–7; 21].

К явным аспектам относятся термины, содержащие упоминание объекта, для которого написан отзыв, или его аспекта. Например, «ресторан», «облуживание», «еда». Неявные аспектные термины включают оценочные слова, специфичные для какого-то аспекта: например, «вкусный» представляет позитивную оценку аспекта «еда». Тональными фактами называются термины, которые представляют объективную информацию, но неявно выражают мнение автора отзыва: например, «была вежлива» представляет позитивную оценку аспекта «обслуживание».

На втором этапе происходит непосредственно определение тональности относительно выделенных аспектов по определенной шкале (чаще всего по шкале «позитивное-нейтральное-негативное»).

Что касается третьего этапа, то, как правило, и выявление аспектов, и тональный анализ осуществляются на уровне предложений. Для того чтобы получить сведения об аспектах для целого отзыва, необходимо некоторым образом реферировать полученные для отдельных высказываний результаты. Однако эта задача не сводится к простому обобщению: каждый аспект может упоминаться в отзыве несколько раз с различной эмоциональной окраской. К примеру:

«Зал показался достаточно уютным. <…> Мягкие на вид диванчики оказались ужасно жесткими и неудобными, а лампы просто не давали достаточно света — мы пользовались фонариками, чтобы разобрать меню»

В таком случае необходимо определить основную тональность (в указанном выше примере для аспекта «обстановка» она будет отрицательной). Однако в отдельных случаях это может быть невозможно. К примеру, для отзыва ниже:

«Лапша с курицей была очень вкусной. Но они так залили все маслом, что есть было почти невозможно — в следующий раз попрошу без масла.»

Для данного случая необходимо ввести некоторую специальную метку (например, «конфликт»).

В данной работе рассматриваются все три этапа задачи аспектно-ориентированного тонального анализа. Для решения подзадач каждого из первых двух этапов используются два различных подхода: метод, основанный на семантической близости слов, описанный в статье [8], а так же предложенный в статье [9] подход к классификации текстов с использованием сверточных нейронных сетей. Путем проведения вычислительных экспериментов, результат работы каждого из методов сравнивается с результатом, полученным при использовании контрольного метода классификации текстов в применении к аспектно-ориентированному тональному анализу на том же корпусе [4]. Подбираются оптимальные параметры работы перечисленных методов классификации. Предлагается способ обобщения результатов, полученных на первых двух этапах.

Основной целью данной работы можно считать апробацию двух указанных методов с целью оценить их результативность для решения задачи аспектно-ориентированного тонального анализа отзывов на русском языке.

Постановка задачи

Задачу аспектно-ориентированного тонального анализа отзывов можно описать следующим образом: для каждого отзыва d_i из имеющегося набора отзывов $D = \{d_1, d_2, \ldots, d_n\}$ необходимо найти подмножество $A_i = \{a_{i1}, a_{i2}, \ldots, a_{i|A_i|}\}$ аспектов из $A = \{a_1, a_2, \ldots, a_k\}$, которые упоминаются в данном отзыве, и для каждого $a_{ij} \in A_i$ определить тональность из множества $Y = \{-1, 0, 1\}$: «отрицательная», «нейтральная» и «положительная» соответственно. Для случая, когда отзыв содержит противоречащие друг другу суждения по одному аспекту, для него указывается специальная метка C — конфликт. Т.е. для каждого аспекта необходимо выбрать метку из множества $Y \cup C$.

Как уже было сказано выше, данная задача может быть разбита на три отдельные подзадачи: извлечение аспектов, определение тональности и реферирование мнения для отзыва.

Запишем формальную постановку каждой из подзадач.

Подзадача 1. Извлечение аспектов

Данная подзадача может быть рассмотрена как задача классификация объектов (отзывов) на пересекающиеся классы.

$S = \{s_1, s_2, \ldots, s_p\}$ — множество предложений отзыва $d \in D$.

$A = \{a_1, a_2, \ldots, a_k\}$ — конечное множество известных для данной предметной области аспектов.

$A^* = \{0,1\}^k$ — множество допустимых ответов классификатора.

$a^*: S \to A^*$ — неизвестная целевая зависимость, значения которой известны только на объектах конечной обучающей выборки $S_m = \{(s_1, A_1^*), \ldots, (s_m, A_m^*)\}$.

Требуется построить алгоритм $a: D \to A^*$, способный классифицировать произвольный объект $s_i \in S$.

Подзадача 2. Выявление тональности относительно аспектов

Может быть рассмотрена как задача классификации на непересекающиеся классы.

$S = \{s_1, s_2, \ldots, s_p\}$ — множество предложений некоторого отзыва $d \in D$, причем для каждого $s_i \in S$ определено $A_i^* = \{a_{i1}^*, a_{i2}^*, \ldots, a_{ir_i}^*\}, r_i \leq k$ — множество упомянутых в этом документе аспектов.

$Y = \{-1, 0, 1\}$ — множество тональных меток, которые соответствуют шкале «отрицательный» — «нейтральный» — «положительный».

$y^*: S \to Y^p$ — неизвестная целевая зависимость, значения которой известны только на объектах конечной обучающей выборки $S_m = \{(s_1, a_{11}^*, y_{11}^*), \ldots, (s_m, a_{mr_m}^*, y_{mr_m}^*)\}$.

Требуется построить алгоритм $y: S \to Y^p$, способный классифицировать произвольный объект $s_i \in S$.

Подзадача 3. Реферирование мнения для отзыва

$D = \{d_1, d_2, \ldots, d_n\}$ — множество текстовых документов (отзывов), причем каждый отзыв d_i состоит из множества предложений $S_i = \{s_{i1}, s_{i2}, \ldots, s_{i|S_i|}\}$.

Для каждого предложения $s_{ij}, i \in [1, n], j \in [1, |Si|]$ известно множество пар $\{(a_{ijl}^*, y_{ijl}^*)\}_{l=1}^{L_{ij}}$, где a_{ijl}^* — l-й аспект предложения s_{ij}, y_{ijl}^* — тональность аспекта a_{ijl}^*.

Требуется построить алгоритм, способный для каждого $d_i \in D$ указать множество пар (a_{ih}, y_{ih}), таких, что: $\forall s_{ij} \in d_i, \forall (a_{ijl}^*, y_{ijl}^*), l \in [1, L_{ij}] \; \exists (a_{ih}, y_{ih}): a_{ih} = a_{ijl}^*, y_{ih} \in Y^* \cup C$.

$Y^* = \{y_{ijl}^*, y_{qwe}^*, C\}, (q, w, e): \exists (a_{qwe}^*, y_{qwe}^*): a_{qwe}^* = a_{ijl}^*$

C — метка конфликта.

$h \in [1, H_i]$, где H_i — суммарное количество аспектов, встретившихся в тексте $d_i \in D$, без учета повторяющихся аспектов, $H_i = \sum_{j=1}^{|S_i|} L_{ij} - L_i^*$.

Обзор литературы

Большинство работ, связанных с тональным анализом, посвящено определению тональности на уровне объектов и гораздо меньше связано с аспектно-ориентированным анализом. Однако с проведением таких мероприятий как SemEval [2–4] и SentiRuEval [5], был опубликован ряд работ на эту тему, в том числе и на русском языке.

В данном обзоре отдельно будут рассмотрены важнейшие публикации и методы решения основных подзадач аспектно-ориентированного тонального анализа: извлечение аспектов и определение тональности по аспектам.

Сначала обратимся к задаче извлечения аспектов. Можно выделить несколько наиболее часто используемых при решении данной задачи подходов:

- частотный подход [10–12],
- машинное обучение [13–22].

Первые попытки решения задачи извлечения аспектов были основаны на классическом подходе к задаче извлечения информации, основанном на использовании наиболее часто встречающихся слов. К примеру, как в работе [10], общая идея которой сводится к поиску существительных либо словосочетаний с ними и выбору наиболее часто встречающихся в качестве аспектов. В дальнейшем был предложен ряд модификаций такого метода, направленных на сокращение итогового набора лексических единиц. Например, с помощью статистических критериев [11, 12].

Основанные на частотном подходе методы хорошо работают для выделения аспектов, которые тесно связаны с одним конкретным термином. Однако они не подходят для случаев, когда аспект представлен множеством низкочастотных терминов (к примеру, для предметной области «рестораны» аспект «еда» может быть описан множеством разных блюд), или для

выявления абстрактных аспектов (таких как «атмосфера»), которые могут быть описаны вообще без использования конкретных существительных.

Проблема извлечения аспектов решается так же методами машинного обучения с учителем. Однако специфика задачи подразумевает необходимость учитывать взаимосвязи между словами, что достаточно сложно при использовании традиционных методов классификации. По этой причине для решения данной задачи чаще всего применяются обобщения существующих методов, которые еще называют методами сегментации и разметки последовательностей.

Одним из представителей этой группы методов является метод на основе скрытых марковских моделей (СММ), который позволяет учитывать предыдущие состояния. В работе [13] описана одна из модификаций СММ.

К методам сегментации относится так же метод условных случайный полей (англ. Conditional Random Fields, CRF), который, используя входную последовательность терминов, моделирует условное вероятностное распределение для последовательности меток. Для извлечения явных аспектных терминов модель CRF с различным набором признаков применялась в работах [14; 15]. Качество решения задачи во многом зависело от выбора признаков для описания данных.

В настоящее время наилучшие результаты для английского языка демонстрируют именно методы разметки последовательностей, однако они обладают определенными недостатками. Во-первых, из-за необходимости наличия специально размеченных данных для обучения их достаточно сложно перенастроить на различные предметные области. Во-вторых, для языков со свободным порядком слов в предложении (к таким относится и русский язык), эти методы будут показывать достаточно скромные результаты. Это связано с тем, что из-за свободного порядка слов уменьшается вероятность встретить размеченную цепочку слов в обучающей выборе, что негативно влияет на процесс обучения.

Для извлечения аспектов применяются так же методы тематического моделирования, которые относятся к машинному обучению без учителя. Они осуществляют анализ слов в больших наборах текстовых документов и на его основе выявляют некие тематические составляющие, прослеживаемые в коллекции. К методам тематического моделирования относится метод скрытого размещения Дирихле (LDA). Однако применением стандартной модели LDA к задаче выделения аспектов сложно добиться хороших результатов, поскольку она имеет тенденцию захватывать только глобальные темы, но не конкретные узкие аспекты в каждом тексте. Для устранения этого недостатка модель LDA можно разделить на глобальную и локальную копии (MG-LDA), как это сделано в работе [16].

В последние годы для решения данной задачи так же используются нейронные сети, в том числе: глубокие нейронные сети [17], рекуррентные нейронные сети [18] и свёрточные нейронные сети [19], — которые демонстрируют очень неплохие результаты, наряду с методами разметки последовательностей.

Среди исследований на тему извлечения аспектов в русском языке интерес представляют несколько исследований, опубликованных в рамках SentiRuEval. В работе [8] предложен метод решения задачи извлечения аспектов, основанный на использовании семантической близости между словами для определения аспектных терминов и набора правил для определения многословных аспектных терминов. Этот подход показал один из лучших результатов, наряду с использованием рекуррентных нейронных сетей в [20] и методом классификации последовательностей, использующим метод опорных векторов на наборе морфологических, синтаксических и семантических признаков [21].

Теперь рассмотрим задачу определения тональности по аспектам. Для нее так же можно выделить несколько наиболее часто используемых при решении подходов:

- на основе знаний или правил [22, 23],

- машинное обучение [24–29].

Первый подход связан с созданием экспертом в выбранной предметной области некоторого набора правил или шаблонов, с помощью которого можно судить о содержании в тексте определенной тональности.

Для автоматизации процесса построения правил во многих системах, разработанных в рамках этого подхода, используется синтаксический анализатор [22, 23]. В таком случае не требуется привлечение экспертов для составления правил, однако для некоторых языков, в том числе и для русского, не существует открытых программных модулей, предназначенных для синтаксического анализа текстов (отсюда и отсутствие исследований для русского языка по этой тематике). Кроме того, в случае низкого качества входного текста, качество разбора и, следовательно, качество определения тональности также будут невысокими. Так, к текстам низкого качества часто относятся опубликованные в интернете отзывы, что негативно влияет на качество определения их тональности [23].

Для решения задачи тонального анализа текста часто применяют методы машинного обучения.

Здесь, как и для предыдущего набора методов, важную роль играют оценочные словари. В рамках данного подхода они используются для составления признаковых описаний выбранных фрагментов текста. В работе [24] было показано, что наиболее эффективным с точки зрения повышения качества классификации является использование словарей тональности. Как и в случае с синтаксическими анализаторами, такие словари часто бывают недоступны для некоторых языков. Большая часть существующих на данный момент в открытом доступе тональных словарей была разработана для английского языка; для русского языка их существует куда меньше.

Для составления тональных словарей пользуются в том числе и автоматическими методами. К примеру, в работе [25] для вычисления значения тональности использовалась взаимная информация между рассматриваемой лексической единицей и словами «poor» (плохой) и

«excellent» (отличный). Стоит помнить, что автоматические методы составления словарей уступают экспертным аналогам, хоть и дают значительный прирост по времени.

Для повышения полноты получаемых словарей часто используются методы их автоматического пополнения, за счет добавления синонимов и антонимов слов. Такой приём реализован в работе [26]. Кроме того, словарь можно пополнять и вручную. К примеру, авторы [27] расширили словарь специфичными терминами предметной области, выражающими тональность, а так же провели дополнительную фильтрацию выражений, которые отрицательно влияли на качество решения задачи.

С использованием составленного словаря формируется некоторое признаковое описание рассматриваемого фрагмента текста. Например, в работе [26] использовались такие признаки, как количество позитивных и негативных слов в контексте аспектного термина и суммарная оценка тональности этих терминов. Помимо этого часто используются признаки, связанные с частями речи и синтаксическим анализом.

Наконец, применяются некоторые алгоритмы машинного обучения. Для определения тональности неплохие результаты обычно демонстрируют метод опорных векторов [24, 27] и метод логистической регресии [26, 28].

Для русского языка интерес представляют работы, основанные на использовании подхода Word2Vec [8, 29] для векторного представления слов и применение глубокого машинного обучения [17] для определения тональности относительно аспектов.

Глава 1. Свёрточные нейронные сети

С увеличением объемов данных и вычислительных возможностей все большее распространение получают нейронные сети; в частности — свёрточные нейронные сети, архитектура которых была предложена в 1998 году Яном Лекуном [37]. Их отличает использование операции свёртки, когда каждый фрагмент данных умножается на матрицу (ядро) свёртки поэлементно, после чего результат суммируется и записывается в аналогичную позицию выходных данных.

Изначально свёрточные нейронные сети использовались для эффективного распознавания изображений. Они получили особое внимание после конкурса ImageNet в 2012 году [38]: использование свёрточной нейронной сети для распознавания образов позволило победителю значительно превзойти остальных участников.

Этот успех привел к попыткам применить свёрточные нейронные сети к другим областям. В частности, их стали использовать для решения задач, связанных с классификацией текстов.

1.1 Нейронные сети

Нейронные сети представляют собой семейство моделей, используемых для решения практических задач: управления, прогнозирования, классификации, кластеризации и др. В основу работы искусственных нейронных сетей заложен принцип работы их биологических прототипов — сетей нервных клеток живого организма.

Вычислительной единицей нейронной сети является нейрон. Модель искусственного нейрона была предложена еще в 1943 году У. Маккалоком и У. Питтсом в статье [35].

На рисунке 1.1 приведен пример искусственного нейрона.

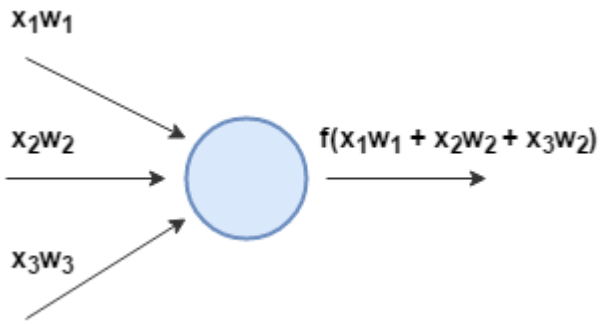

Рис. 1.1: Искусственный нейрон. Здесь x_i — входной сигнал, w_i — вес входного сигнала, $f(\cdot)$ — функция активации.

Нейрон получает некоторую информацию, производит над ней простые вычисления и передает ее дальше. Информация может поступать как в виде исходного входного сигнала, так и от других нейронов. Она поступает в нейрон через несколько входных каналов. Каждый входной сигнал проходит через соединение, имеющее определенный вес, и для каждого нейрона вычисляется взвешенная сумма таких входов, которая затем преобразуется с помощью функции активации. Так получается выходной сигнал нейрона, который передается следующему нейрону или является результатом работы сети.

В зависимости от расположения в сети нейроны делятся на входные, скрытые и выходные. Кроме того, исходя из принципов своего функционирования, нейроны группируются в слои.

1.1.1 Метод обратного распространения ошибки

Одним из методов обучения нейронной сети является метод обратного распространения ошибки [36]. Его можно разбить на четыре отдельных этапа: прямое распространение, функцию потерь, обратное распространение и обновление веса.

Во время прямого распространения исходные данные пропускаются через всю сеть. Каждый нейрон получает некоторые данные, вычисляет значение своей активационной функции и передает это значение дальше. При этом все веса или значения фильтров в сети взяты произвольным образом.

Очевидно, что сеть со случайно заданными весами не сможет правильно классифицировать объект. Тогда с помощью функции потерь L можно посчитать отличие полученного результата от ожидаемого. На первых этапах значения функции будут высокими, и задача обучения сети сводится к задаче оптимизации: подобрать веса таким образом, чтобы значение функции потерь стало минимальным.

Чтобы добиться этого, необходимо выполнить обратное распространение ошибки. Этот этап позволит определить, какие веса сильнее всего повлияли на потери, чтобы найти способы их преобразования для уменьшения потерь. Наконец, значения весов обновляются, и весь процесс повторяется снова — он будет повторяться фиксированное количество раз для каждого входного объекта. После завершения обновления параметров на последнем тренировочном образце обучение сети считается законченным.

1.1.2 Функция активации

Функцией активации нейронной сети называется функция, вычисляющая выходной сигнал нейрона сети. Ниже перечислены используемые в работе функции активации:

- Сигмоида:

$$f(s) = \frac{1}{1 + e^{-s}}$$

- Линейная:

$$f(s) = s$$

- Усеченное линейное преобразование (англ. rectified linear unit, ReLU):

$$f(s) = \max(0, s)$$

- Мягкий максимум (англ. softmax, софтмакс):

$$f(s) = \frac{e^{s_j}}{\sum_{k=1}^{K} e^{s_k}}, j = 1, \dots, K$$

1.1.3 Функция потерь

Функция потерь характеризует отличие значений, полученных на выходе нейронной сети, от ожидаемых значений, которые содержатся в тренировочных данных.

Введем следующие обозначения: X — множество описаний объектов, Y — множество допустимых ответов. Предполагается, что существует неизвестная целевая зависимость — отображение $y^*: X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X_m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.

Можно ввести функцию потерь $L(y, y')$, которая характеризует отличие ответа y от правильного ответа $y' = y^*(x)$ на произвольном объекте $x \in X$.

Тогда эмпирический риск — это функционал качества, характеризующий среднюю ошибку на обучающей выборке:

$$Q(a, X_m) = \frac{1}{m}\sum_{i=1}^{m} L(y_i, y^*(x_i))$$

Целью процесса обучения нейронной сети является минимизация эмпирического риска.

Для задач мультиклассовой классификации в качестве функции потерь часто используется перекрестная энтропия:

$$L(y_i, y^*(x_i)) = -\sum_{j=1}^{K} y_{ij}^* log y_{ij},$$

где K — количество меток классов в задаче.

1.1.4 Регуляризация

Поскольку обучение нейронных сетей обычно производят стохастическим градиентным спуском со случайным выбором некоторых объектов из выборки, они подвержены проблеме переобучения. Для предотвращения переобучения нейронной сети применяется операция отсева (англ. dropout) [39]. Она заключается в следующем: на каждой итерации

обучения любые нейроны могут быть исключены из сети с некоторой вероятностью p. По оставшейся сети проводится обучение, для оставшихся весов делается градиентный шаг, после чего все исключенные нейроны возвращаются в сеть.

Так, для значений $z = [\hat{c}_1, \dots, \hat{c}_m]$, где m — количество фильтров, на выходе вместо

$$y = w \cdot z + b$$

для прямого распространения используется

$$y = w \cdot (z \circ r) + b,$$

где $\circ$ — посимвольное умножение, r — вектор, состоящий из нулей и единиц, причем вероятность появления единицы равна p.

Таким образом, на каждом шаге стохастического градиента происходит настройка одной из возможных 2^N структур связей между нейронами, где N — общее количество нейронов.

При тестировании нейронной сети нейроны уже не исключаются, однако выход каждого нейрона домножается на p, что превращает его в математическое ожидание ответа этого нейрона по всем 2^N возможным структурам. Т.е. полученную в итоге сеть можно рассматривать, как усреднение 2^N сетей.

1.2 Архитектура свёрточных нейронных сетей

Чаще всего свёрточная нейронная сеть состоит из трёх видов слоев: свёрточные, субдискретизирующие и полносвязные, — которые могут чередоваться в произвольном порядке [37].

Типовая структура свёрточной нейронной сети изображена на Рис. 1.2 и представляет собой следующее: после начального слоя данные проходят серию чередующихся свёрточных и субдискретизирующих слоев. Чередование слоев позволяет составить карты свойств, отвечающие за определенные признаки данных. На каждом следующем слое каждая

отдельная карта уменьшается в размере, но их становится всё больше. На практике это означает способность распознавания всё более сложных характеристик. Наконец, после завершения серии чередующихся слоев дополнительно устанавливают несколько слоев полносвязной нейронной сети, на вход которым подаются сформированные карты признаков.

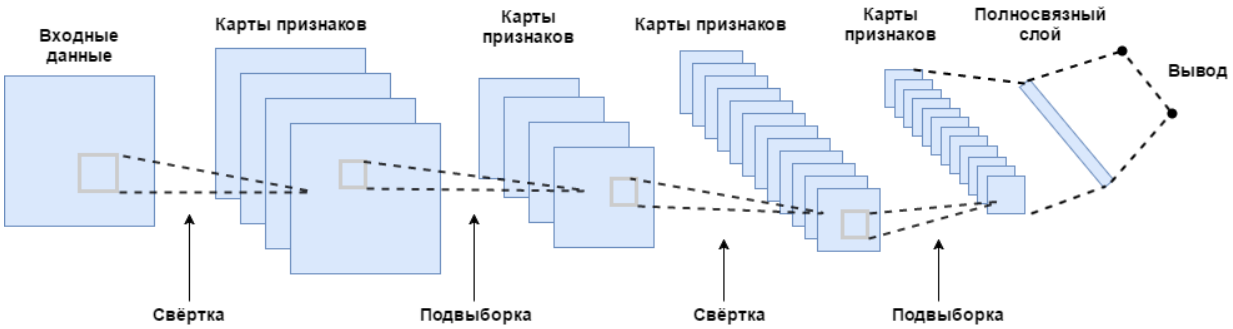

Рис. 1.2: Архитектура сверточной нейронной сети

Свёрточный слой представляет собой основной элемент свёрточной нейронной сети. С использованием фильтра (ядра свёртки), каждый предыдущий слой обрабатывается по фрагментам; результаты матричного произведения для каждого фрагмента суммируются. При этом весовые коэффициенты ядра свёртки устанавливаются в процессе обучения.

На Рис. 1.3 показан пример свёрточного слоя с фильтром размера 3 × 3.

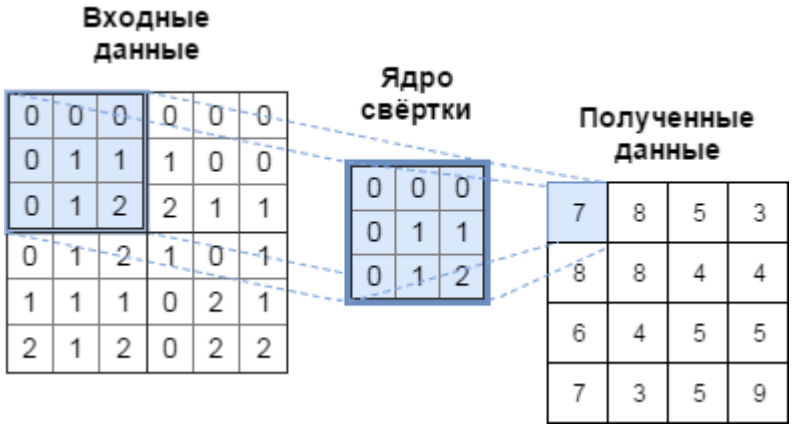

Рис. 1.3: Свёрточный слой

Особенностью свёрточного слоя является небольшое, по сравнению с полносвязными нейронными сетями, количество параметров, определяемое при обучении. Кроме того, за счет краевых эффектов размер исходных данных уменьшается.

Субдискретизирующий слой, который еще называют слоем подвыборки, обычно располагается следом за свёрточным слоем и служит для значительного сокращения размерности без потерь наиболее важной информации. Чаще всего подвыборка осуществляется с помощью выбора максимального элемента — каждая карта признака разбивается на ячейки, из которых выбирается та, что имеет максимальное значение.

Можно представить каждую карту признака, как отражение некоторого свойства предложения, например, наличия в нем фразы «не самый интересный». Если эта фраза встретилась где-то в предложении, то в карте признака значения в соответствующей ячейке будет высоким. После выделения максимального элемента, информация о наличии данной фразы в предложении сохранится, но ее точное местоположение будет утеряно.

Пример субдискретизирующего слоя с выбором максимального элемента представлен на Рис. 1.4.

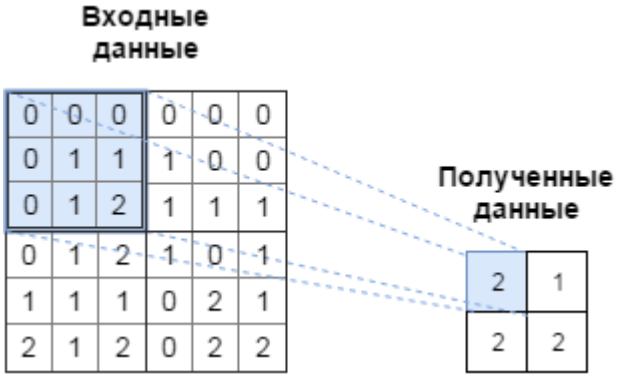

Рис. 1.4: Субдискретизирующий слой

В **полносвязном слое** каждый нейрон соединен со всеми нейронами на предыдущем уровне, и каждая такая связь имеет свой весовой коэффициент.

Обычно полносвязные слои размещают в конце свёрточной сети: они, используя входные данные, возвращают N-пространственный вектор, где N — число классов, а каждое возвращаемое значение описывает вероятность принадлежности к соответствующему классу.

Определение указанной вероятности осуществляется с помощью данных, полученных на предыдущих слоях. В полносвязном слое происходит

определение карт свойств, которые сильнее связаны с определенным классом. Так, если предложение содержит уже упомянутую выше фразу «не самый интересный», то в карте свойств, представленной этой фразой, будет высокое значения, и можно будет с некоторой вероятностью сделать вывод о наличии в предложении отрицательной тональности.

На Рис. 1.5 показан пример полносвязного слоя.

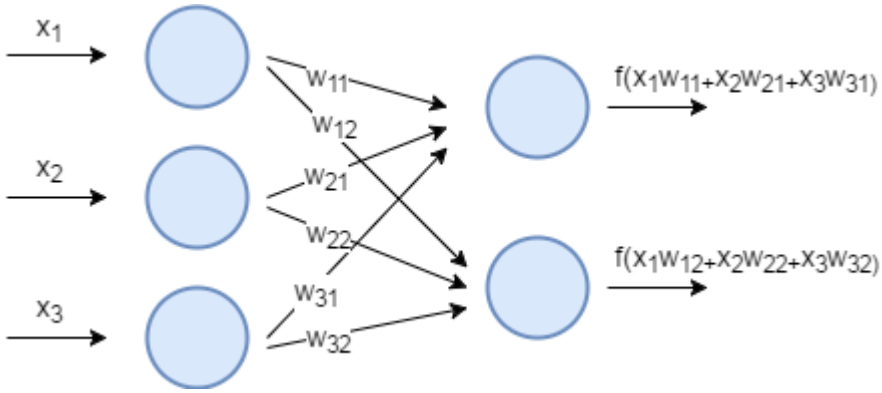

Рис. 1.5: Полносвязный слой

1.3 Использование свёрточной нейронной сети для работы с текстами

Рассматриваемая в данной работе модель свёрточной нейронной сети для классификации текстов на естественном языке основана на подходе с использованием кодирования слов, описанном в статье [9].

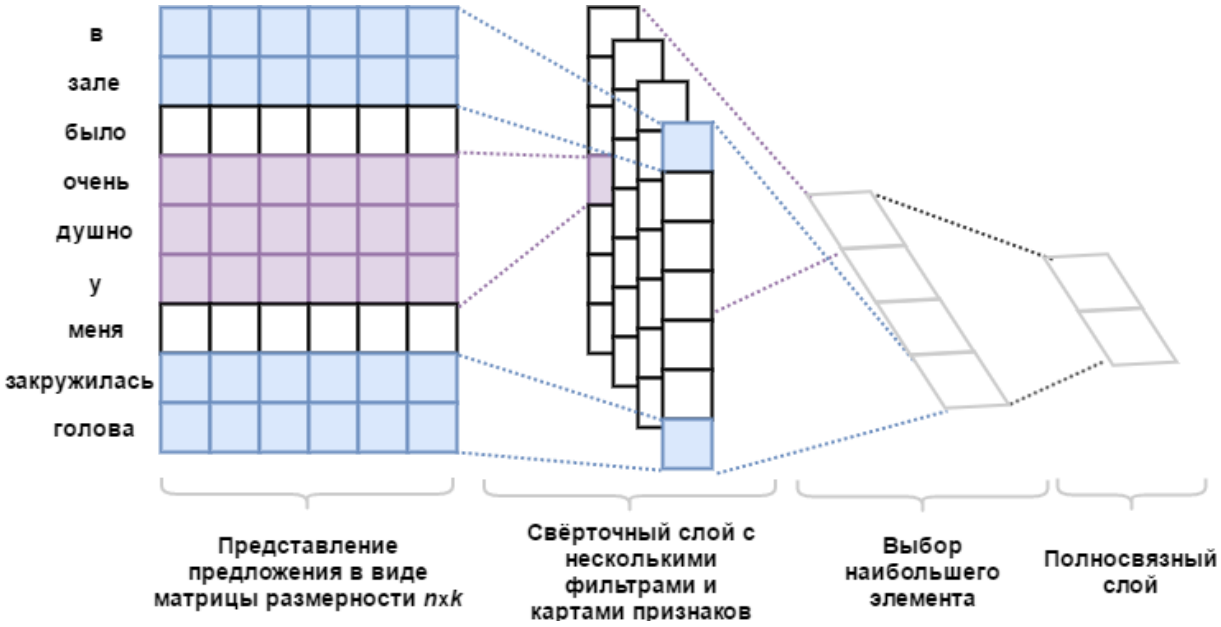

Рис. 1.6: Архитектура свёрточной нейронной сети для работы с текстами

На Рис. 1.6 приведена общая архитектура рассматриваемой модели.

Опишем формально данный подход.

Пусть $x_i \in R^k$ — вектор размерности k, соответствующий i-тому слову в предложении.

Тогда предложение длины n (если необходимо, дополненное специальными метками, заменяющими недостающие слова) можно представить как:

$$x_{1:n} = x_1 \oplus x_2 \oplus \dots \oplus x_n,$$

где $\oplus$ — операция конкатенации векторов.

В целом, через $x_{i:i+j}$ будем обозначать конкатенацию слов $x_i, x_{i+1}, \dots, x_{i+j}$. Операция свёртки использует фильтр $w \in R^{\mathrm{hk}}$, после применения которого к последовательности из h слов формируется новый признак. К примеру, для признака c_i, полученного из последовательности слова $x_{i:i+h-1}$:

$$c_i = f(wx_{i:i+h-1} + b),$$

где f — функция активации свёрточной нейронной сети, b — некоторая константа.

Указанный фильтр применяется к каждой возможной последовательности слов в предложении $\{x_{1:h}, x_{2:h+1}, \dots, x_{n-h+1:n}\}$. В результате получается следующая карта свойств:

$$c = [c_1, c_2, \dots, c_{n-h+1}], c \in R^{\mathrm{n-h+1}}$$

Затем производится операция выбора наибольшего элемента:

$$\hat{c} = max\{c\}$$

Полученные применением всех значения передаются в полносвязный слой с функцией мягкого максимума.

1.3.1 Векторное представление слов

Для перевода слова в вектор фиксированной длины существует множество различных способов. В данной работе для экспериментов с

применением свёрточных нейронных сетей будет использован метод, основанный на алгоритме word2vec, который подробно описан в разделе 2.1 настоящей работы.

1.3.2 Применение в задаче аспектно-ориентированного тонального анализа

Как было показано выше, свёрточные нейронные сети можно использовать для решения задач классификации текстов, к которым относится и аспектно-ориентированный тональный анализ.

Отдельные предложения могут быть рассмотрены в виде конкатенации векторных представлений входящих в них слов. Предложения с заранее определенными метками аспектов/тональности будут использованы для обучения нейронной сети.

При этом задача извлечения аспектов, рассматриваемая как задача мультиклассовой классификации на пересекающиеся классы, может быть сведена к нескольким задачам бинарной классификации, как описано в разделе 2.2 данной работы. В таком случае необходимо обучить по одной нейронной сети для каждого аспекта.

Следуя тому же принципу, для тонального анализа предложения можно обучить две нейронные сети, первая из которых будет классифицировать предложение по принципу «нейтральный» — «эмоциональный», а вторая определять тональность эмоциональных предложений.

Глава 2. Семантическое сходство

Наиболее распространенные подходы к автоматической обработке текстов на естественном языке, в частности, основанные на векторном представлении текстов, не учитывают наличие семантических связей между различными терминами.

Согласно [30], под мерой семантической близости будем понимать численную меру, отражающую степень схожести двух лексем (например, существительных или многословных выражений) и предназначенную для количественной оценки семантической схожести этих лексем. Такая метрика показывает высокие значения для пар слов, находящихся в семантических отношениях (синонимов, гипонимов, ассоциаций), и низкие или нулевые значения для всех остальных пар.

Необходимость учитывать лексико-семантические особенности языка на практике объясняется тем, что семантически близкие понятия могут быть выражены совершенно разными способами. Так, «солянка», «суп» и «первое» в одном тексте могут обозначать один и тот же объект. Однако без явного указания соответствий (т.е. семантических отношений) между этими словами, вычислительной системе непросто будет прийти к выводу об эквивалентности их значения.

Одним из подходов к разрешению проблемы неучтенных семантических связей между словами является использование описанной в статье [31] модели векторного представления терминов на основе распределённых представлений слов.

Эксперименты показывают, что такая модель способна кластеризовать семантически схожие слова [31]. Такое свойство оказывается полезным при решении подзадач аспектно-ориентированного тонального анализа [8].

2.1 Word2Vec

Word2Vec — это технология, разработанная группой исследователей компании Google [33] и предназначенная для расчета векторных представлений слов. Ее работа основана на предположении, что слова, находящиеся в похожих контекстах, обычно значат похожие вещи, т.е. являются семантически близкими.

Для обучения модели Word2Vec необходим достаточно большой корпус. Используя этот корпус, Word2Vec собирает статистику о частоте появлении слов в данных, удаляет наиболее редкие и наиболее частые слова, после чего, используя нейронную сеть, решает задачу снижения размерности, чтобы получить компактные векторные представления слов заранее определенной длины.

Более формально, решаемая Word2Vec задача может быть сформулирована как максимизация косинусной близости между векторами слов, которые упоминаются в близких контекстах, и минимизация косинусной близости между векторами слов, которые появляются рядом очень редко или никогда.

Косинусное сходство $cos(\theta)$ между векторами A и B — это мера сходства между ними, определяемая формулой (1) [32].

$$cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

Одной из существующих моделей обучения векторных представлений является skip-gram [31].

2.1.1 Skip-gram

Идея данной модели заключается в том, чтобы по единственному входному слову попытаться определить его наиболее вероятный контекст. Текст просматривается окном ширины $2k + 1$, и для каждого окна по

центральному слову w_t предсказывается его контекст $w_{t-k}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+k}$.

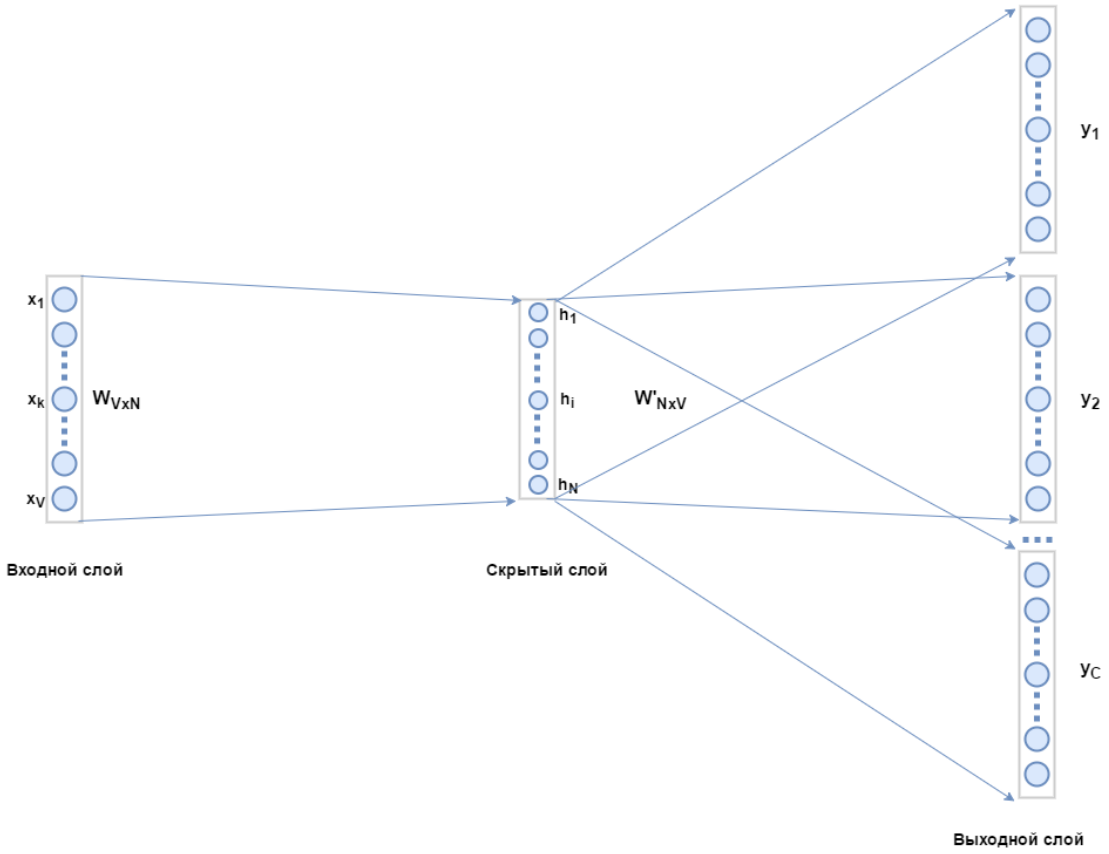

Рис. 1.7: Архитектура нейронной сети skip-gram

Модель skip-gram представляет собой нейронную сеть. Ее архитектура представлена на Рис. 1.2. Все смежные слои являются полносвязными. Входной слой содержит V нейронов, где V — объем словаря (количество различных слов в обучающих данных). Скрытый слой содержит N нейронов, а выход представляет собой набор из C векторов размерности V, где C — количество слов контекста.

Для обработки подается вектор размерности V: $\{x_1, \cdots, x_V\}$, у которого все элементы будут равны 0, кроме элемента $x_k = 1$, соответствующего рассматриваемому слову.

Обозначим матрицу весов между входным и скрытым слоем через $W_{V\times N} = \{w_{ki}\}$, а матрицу весов между скрытым и выходным слоем — через $W'_{N\times V} = \{w'_{ij}\}$. Тогда выход нейронов на скрытом слое представляет собой k-тую строку w_k матрицы W:

$$h = x^T W = w_k$$

Это означает, что скрытый слой имеет линейную функцию активации.

Вход каждого из C×V выходных нейронов вычисляется по формуле

$$u_{cj} = u_j = w'_j h, \quad \mathrm{c} = 1,2,\dots,\mathrm{C}$$

для j-го нейрона c-го слова на выходе, с учетом, что выходной слой для каждого слова на выходе имеет одинаковые веса.

Наконец, используя функцию активации софтмакс, можно вычислить выход j-го нейрона для c-го слова:

$$y_{cj} = \frac{\exp(u_{cj})}{\sum_{j'=1}^{V} \exp(u_{cj'})} = p\left(w_{cj} = w_{Oc} \middle| w_I\right),$$

где w_{cj} — полученное слово, w_{Oc} — настоящее с-тое слово полученного контекста, w_I — исследуемое слово.

Обучаясь, нейронная сеть максимизирует y_{j^*}, где j^* — номера предсказываемых слов в словаре (вектор длины C), или, что то же самое в контексте данной задачи, максимизирует вероятность получения выходных слов $w_{O1}, w_{O2}, \dots w_{Oc}$ при условии, что известно входное слово w_I.

$$max\ y_{j^*} = max\ log\ y_{j^*}$$

$$= \sum_{c=1}^{C} u_{j_c^*} - C \log \sum_{j'=1}^{V} \exp\left(u_{j'}\right) = \max p(w_{OI}, w_{O2}, \dots, w_{OC} | w_I)$$

Наконец, из формулы Байеса

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

следует, что увеличение $p(w_{OI}, w_{O2}, \dots, w_{OC}|w_I)$ будет способствовать увеличению $p(w_I|w_{OI}, w_{O2}, \dots, w_{OC})$.

Обучение по модели skip-gram по сравнению с другими моделями является более затратным, однако, согласно [33], данная модель обычно дает лучшие результаты.

2.1.2 Иерархический мягкий максимум

В описанной выше модели время вычисления функции softmax линейно зависит от объема словаря V, который на практике очень велик (порядка $10^5 - 10^7$ термов) [31]. Данный подход не является эффективным, поэтому в реализации Word2Vec часто используются аппроксимирующий метод для быстрого вычисления функции активации: иерархический мягкий максимум (англ. hierarchical softmax) [34].

В данном методе по всем словам в словаре строится бинарное дерево, в котором V листьев, а узлы содержат относительную вероятность выбора каждого из его дочерних узлов. Так определяется случайное прохождение по дереву, которое присваивает вероятности словам.

Формально, условная вероятность вычисляется следующим образом:

$$p(v|w) = \prod_{n \in path(v)} \sigma(ch(n)w'_n h),$$

где $\sigma(\cdot)$ − сигмоидная функция активации, $path(v)$ − путь от вершины до корня, а

$$ch(n) = \begin{cases} 1, \text{если вершина } n - \text{правый потомок} \\ 0, \text{если вершина } n - \text{левый потомок} \end{cases}$$

2.2 Задача извлечения аспектов как множество задач бинарной классификации

Задача извлечения аспектов, рассматриваемая как задача мультиклассовой классификации на пересекающиеся классы, может быть сведена к нескольким задачам бинарной классификации. Подобно описанному в статье [8] подходу, решено было воспользоваться стратегий «один против всех»: для каждого из заявленных для предметной области аспектов строится отдельный классификатор, обученный на данных, относящихся к одному аспекту, против данных для всех остальных аспектов.

Формально, если $D = \{d_1, d_2, \ldots, d_n\}$ — множество текстовых документов (отзывов), причем каждый текст $d_i \in D$ состоит из множества

предложений: $d_i = \{s_1, s_2, \ldots, s_n\}$, которые подлежат классификации, $A = \{a_1, a_2, \ldots, a_k\}$ — конечное множество аспектов, $p^* = \{0,1\}^k$ — множество допустимых ответов, $S = \{(s_1, p_1^*), \ldots, (s_m, p_m^*)\}$ — конечная обучающая выборка, то для каждого аспекта $a_i \in A$:

$z_i = \{z_{i1}, z_{i2}, \ldots, z_{im}\}$ — новый вектор меток, причем $z_{ij} = 1$, если $p_{ji}^* = 1$, иначе $z_{ij} = 0$.

$S_i = \{(s_1, z_{i1}), \ldots, (s_m, z_{im})\}$ — новая обучающая выборка.

Получим набор из k обучающих выборок, по одной для каждого аспекта. Тогда $c = \{c_1, c_2, \ldots, c_k\}$ — набор из классификаторов, где $c_i, i \in [1; k]$ был обучен на соответствующей выборке S_i.

2.3 Метод извлечения аспектов

Для каждого рассматриваемого аспекта из размеченной обучающей коллекции могут быть выбраны эталонные термины: существительные, глаголы, прилагательные и наречия, которые являются характерными при описании аспекта. Так, можно предположить, что для аспекта «еда» предметной области «рестораны» в состав эталонных терминов войдут различные наименования блюд.

Заметим, что в оригинальной работе [8] в качестве эталонных терминов рассматривались исключительно существительные и глаголы. Такой отбор усложняет определение неявных аспектов:

«Все было очень вкусно!»

Здесь наречие «вкусно» указывает на наличие аспекта «еда» без каких-либо уточняющих явных терминов, характерных для данного аспекта.

После составления словаря эталонных терминов, каждый новый проверяемый термин, имеющий распределенное представление $\vec{a} = (a_1, \ldots, a_n)$, может быть сопоставлен с конкретным аспектом A^* одним из двух способов:

- поэлементным сравнением с каждым эталонным термином $\overrightarrow{b_\iota} \in B_{A^*}$ аспекта A^*,
- вычислением суммарного сходства с аспектом A^*.

B_{A^*} — множество эталонных терминов аспекта A^*. Каждый $\overrightarrow{b_\iota} \in B_{A^*}$ имеет распределенное векторное представление $\overrightarrow{b_\iota} = (b_1, \dots, b_n)$. В качестве меры близости между векторами в обоих случаях используется косинусное сходство [32].

Для первого способа:

$$sim_1(\vec{a}, A^*) = \max_{i=1,\dots,k} \frac{\vec{a} \cdot \overrightarrow{b_\iota}}{\|\vec{a}\| \cdot \|\overrightarrow{b_\iota}\|},$$

$\overrightarrow{b_\iota} \in B_{A^*}$, $k = |B_{A^*}|$ — количество эталонных терминов.

Для второго способа:

$$sim_2(\vec{a}, A^*) = \sum_{i=1}^{k} \frac{\vec{a} \cdot \overrightarrow{b_\iota}}{\|\vec{a}\| \cdot \|\overrightarrow{b_\iota}\|},$$

$\overrightarrow{b_\iota} \in B_{A^*}$, $k = |B_{A^*}|$ — количество эталонных терминов.

Если полученное значение близости превышает некоторый порог, то проверяемый термин считается аспектным. Пороговое значение в каждом из случаев можно определить экспериментальным путем.

2.4 Метод определения тональности аспектных терминов

Тональность для каждого из найденных аспектов можно определить по его контексту. Для того чтобы дать интерпретацию контекста в числовом формате используется тональный словарь для рассматриваемой предметной области. Далее с его помощью для каждого оцениваемого аспекта генерируется набор признаков, который используется в качестве входных данных для классификатора на основе решающих деревьев (Gradient Boosting Classifier) [41].

2.4.1 Составление тонального словаря

Для составления словаря из рассматриваемых текстов выбираются кандидаты в эмоциональные выражения: все существительные, прилагательные, глаголы и наречия, а также отдельные фрагменты текста.

Далее необходимо численно определить эмоциональную окраску каждого термина. Для этого вновь используется семантическое сходство.

$$sim^{+}(\vec{a}, B^{+}) = \sum_{i=1}^{k} \frac{\vec{a} \cdot \overrightarrow{b_i}}{\|\vec{a}\| \cdot \|\overrightarrow{b_i}\|}$$

$$sim^{-}(\vec{a}, B^{-}) = \sum_{i=1}^{k} \frac{\vec{a} \cdot \overrightarrow{b_i}}{\|\vec{a}\| \cdot \|\overrightarrow{b_i}\|}$$

Здесь B^{+} и B^{-} — множества эталонных эмоциональных терминов положительной и отрицательной тональностей соответственно. Состав этих множеств определяется экспертом, каждое из них содержит по 20 выражений для позитивной и негативной тональностей. Каждый элемент этих множеств имеет распределенное представление $\overrightarrow{b_i} = (b_1, \dots, b_n)$.

$\vec{a} = (a_1, \dots, a_n)$ — распределенное представление рассматриваемого термина.

Наконец, $sim^{+}(\vec{a}, B^{+})$ и $sim^{-}(\vec{a}, B^{-})$ — значения суммарных сходств.

В качестве тональности рассматриваемого слова выбирается тональность, чье суммарное сходство больше по модулю.

После определения эмоциональной окраски каждого выражения получается законченный тональный словарь, сопоставляющий каждому лексическому выражению оценку тональности на основе семантического сходства.

2.4.2 Признаковое описание аспекта

При определении тональности для каждого аспектного термина необходимо учитывать, что в предложении может одновременно упоминаться несколько аспектов. Это означает, что необходимо учитывать

как близкий контекст для определения случаев, когда аспекты находятся в различных частях составного предложения, так и дальний — для случаев перечисления контекстов, когда выражающий тональность термин может быть значительно удален от аспектного термина. Кроме того, не следует забывать, что и сам аспектный термин может содержать эмоциональную окраску.

Исходя из этого, было предложено сформировать вектор признаков из следующих составляющих:

- ближайший контекст (три термина слева и справа),
- дальний контекст (шесть терминов слева и справа)
- непосредственно сам термин.

Каждый термин с помощью составленного тонального словаря можно представить его тональным значением. Чтобы учесть также термины с инвертированной тональностью, для фрагментов текста, соответствующих шаблону <не> + <термин>, значение из тонального словаря бралось с противоположным знаком.

2.4.3 Случай присутствия нескольких аспектов в предложении

Определение тональности отдельно для каждого аспекта осложняет тот факт, что в предложении их, вообще говоря, может быть несколько.

Изучение отзывов имеющейся коллекции показало, что можно выделить три типичных случая расположения нескольких аспектов в одном предложении:

1. В разных частях сложносочиненного предложения с соединительным союзом. Пример: «В зале было очень уютно, и официант был замечательно вежлив». Оба аспекта: «обстановка» и «обслуживание», — имеют одинаковую тональность.
2. В разных частях сложносочиненного предложения с противительным союзом. Пример: «Все заказанное мне очень понравилось, но музыка в ресторане играла просто

отвратительная». Здесь для аспекта «еда» должна быть определена положительная тональность, а для аспекта «обстановка» — отрицательная.

3. В одной части предложения. Пример: «И интерьер, и обслуживание, и сама еда, — все было на высшем уровне!». Для всех аспектов здесь должна быть указана одинаковая тональность.

Для первого и третьего случаев никаких дополнительных действий при определении тональности предпринимать не следует, поскольку тональность одного из аспектов совпадает с тональностью остальных. Во втором случае, когда аспекты выражают различную тональность, каждая часть составного предложения должна быть рассмотрена отдельно. Для этого необходимо разбить каждое сложносочиненное предложение с противительным союзом на две части — и далее считать каждую из этих частей отдельным предложением.

Поскольку в открытом доступе отсутствуют мощные инструменты для синтаксического разбора предложений с указанием их составных частей в русском языке, а разработка такого инструмента не является целью данной работы, необходимо использовать упрощенный вариант поиска и разбиения таких предложений.

Для решения указанной задачи предлагается следующий подход: для каждого предложения проверять в нем наличие одного из противительных союзов: *но, однако, а, зато* [40]. Если такой союз был найден, то содержащее его предложение разбивается на две части: до союза и после. Дополнительным условием является наличие запятой перед союзом для исключения срабатывания метода для тех противительных союзов, которые используются не для разделения частей сложносочиненных предложений.

Исходя из имеющихся в коллекции примеров, предложенный подход позволит верно определять предложения с несколькими разнотональными аспектами в подавляющем большинстве случаев.

Глава 3. Реферирование отзывов

В главах 1 и 2 данной работы были рассмотрены способы извлечения аспектов и определения тональности по этим аспектам для отдельных предложений отзыва. Исходная же задача предусматривает обобщение полученной информации для отзывов, которым принадлежат рассматриваемые предложения.

Делать это предполагается в два этапа: сначала найти все аспекты, упомянутые в отзыве, затем — для каждого из этих аспектов определить суммарную тональность.

3.1 Обобщение аспектов для отзыва

Перечисление аспектов для отзыва, если известны аспекты, упомянутые в его предложениях, может быть осуществлено простой операцией объединения.

Будем считать, что аспект $a_k \in A$ считается упомянутым в тексте $d_i \in D$, где $d_i = \{s_{i1}, s_{i2}, \dots, s_{i|d_i|}\}$, если он упомянут хотя бы в одном его предложении s_{ij}. Тогда A_i^* — множество упомянутых в тексте $d_i \in D$ аспектов можно получить следующим образом:

$$A_i^* = \bigcup_{j=1}^{|d_i|} a_{ij}^* \cup a_{gen},$$

где $a_{ij}^* = \{a_{ij1}^*, a_{ij2}^*, \dots a_{ijL_{ij}}^*\}$ — множество аспектов, упомянутых в предложении $s_{ij} \in d_i$, a_{gen} — аспект «GENERAL», характеризующий рассматриваемый объект в целом.

3.2 Обобщение тональности по аспектам отзыва

Как уже было сказано во введении данной работы, задача выявления общей тональности аспектов отзыва не сводится к простому объединению

полученных результатов: каждый аспект может упоминаться в отзыве несколько раз с различной эмоциональной окраской.

В случае наличия в предложении множества разнотональных аспектов предлагается применять следующий набор правил для обобщения тональных меток аспектов:

- В случае, если аспект встречается в отзыве ровно один раз, то тональность определяется его тональностью в предложении.
- Если аспект встречается в отзыве несколько раз с одинаковой тональностью, то в качестве итоговой выбирается эта единственная тональность.
- Если аспект встречается в отзыве несколько раз с различными значениями тональности, среди которых нет одновременно положительной и отрицательной, то итоговой считается наиболее эмоциональная тональность. Т.е. для набора {«отрицательный», «нейтральный»} итоговой тональностью будет отрицательная. Аналогично для набора {«положительный», «нейтральный»} будет выбрана положительная тональность.
- Если аспект встречается в отзыве несколько раз с различными значениями тональности, среди которых встречается как положительная, так и отрицательная, то аспект в качестве тональности аспект получает метку C — конфликт.

Отдельно необходимо сказать об аспекте a_{gen}, выражающем общее мнение о рассматриваемом объекте. Возможны два случая:

1. Аспект a_{gen} был упомянут в одном из предложении отзыва наравне с остальными аспектами. В таком случае его тональность определяется по приведенным выше правилам.
2. Аспект a_{gen} был искусственно добавлен на шаге обобщения аспектов (см. раздел 3.1 данной работы). Тогда его тональность определяется по следующим правилам:

- Если среди множества тональностей других аспектов нет метки конфликта или положительной и отрицательной тональностей одновременно, тональность аспекта a_{gen} приравнивается к наиболее часто встречающейся среди остальных аспектов тональности. Если же различных тональностей равное количество, a_{gen} получает метку тональности C — конфликт.
- Во всех остальных случаях аспект a_{gen} получает метку тональности C — конфликт.

Глава 4. Вычислительные эксперименты

В данной главе приводится описание и результаты решения задачи автоматического реферирования отзывов на основе аспектно-ориентированного тонального анализа с использованием семантического сходства, а также свёрточных нейронных сетей. Для оценки качества используемых методов, результаты их работы будут сравниваться с контрольными результатами, полученными на тех же данных организаторами семинара SemEval-2016 и используемыми в качестве проходного порога для участников семинара.

4.1 Данные

Для проведения экспериментов использовались данные, предоставленные организаторами семинара SemEval-2016 в рамках задания по аспектно-ориентированному тональному анализу [4]. Данные представляют собой набор из вручную размеченных русскоязычных отзывов предметной области «рестораны»: 300 отзывов (3548 предложений) для обучающей выборки и 103 отзыва (1209 предложений) для тестовой.

Разметка имеющихся данных включает аспекты и их тональность как на уровне предложений, так и на уровне целых отзывов. Всего для данной предметной области характерно 12 различных аспектов: AMBIENCE#GENERAL, FOOD#STYLE_OPTIONS, FOOD#PRICES, FOOD#QUALITY, SERVICE#GENERAL, RESTAURANT#GENERAL, DRINKS#QUALITY, RESTAURANT#PRICES, RESTAURANT#MISCELLANEOUS, DRINKS#STYLE_OPTIONS, LOCATION#GENERAL, DRINKS#PRICES. Каждый аспект помечен одной из трех тональностей: positive, negative, neutral.

В таблицах 4.1 и 4.2 приведена статистическая информация об используемых данных.

	Обучающая выборка	**Тестовая выборка**
Количество предложений, содержащих аспекты	2733	908
Количество предложений без аспектов	815	301

Таблица 4.1: Статистика наличия аспектов в предложениях

	Обучающая выборка	**Тестовая выборка**
Положительная тональность	3103	870
Отрицательная тональность	709	321
Нейтральная тональность	276	103
Всего	4088	1294

Таблица 4.2: Количество содержащихся в предложениях аспектов

4.2 Контрольный алгоритм

Для извлечения аспектов организаторами SemEval-2016 был использован метод опорных векторов (англ. Support Vector Machine, SVM) с линейной функцией ядра. Для каждого предложения формировался вектор признаков, по которому классификатор должен был определить вероятность наличия в предложении определенного аспекта. Далее аспекты, вероятность которых преодолела некоторый порог (организаторы использовали порог 0.2), приписывались к рассматриваемому предложению. В качестве признаков использовались 1000 наиболее часто встречающихся в обучающей

выборке униграмм, за исключением стоп-слов. Для получения аспектов на уровне отзыва суммировались все аспекты на уровне предложений с удалением дубликатов.

Для определения тональности также был обучен классификатор SVM с линейной функцией ядра. Для признакового описания предложений, как и в предыдущем случае, использовались униграммы, однако для этой подзадачи к ним также была добавлена метка найденного в предложении аспекта. Для получения тональности на уровне отзыва, из полученных на уровне предложения тональностей для каждого аспекта выбиралась наиболее часто встречаемая.

С более подробным описанием предложенных алгоритмов можно ознакомиться в оригинальной публикации [4].

В таблице 4.3 приведены результаты их работы. Для численной оценки качества используются: F-мера (англ. F-measure, F-1) для извлечения аспектов и точность (англ. Accuracy, Acc) для определения тональности.

Задача	**Для предложений**	**Для отзыва целиком**
Извлечение аспектов, F-1	0.55882	0.84792
Определение тональности, Acc.	0.71	0.706

Таблица 4.3: Результаты работы контрольных алгоритмов

4.3 Свёрточные нейронные сети

С использованием библиотеки TensorFlow [42] была реализована свёрточная нейронная сеть с архитектурой, описанной в разделе 1.3 данной работы.

Первоначально параметры работы нейронной сети были подобраны в соответствие с рекомендациями в работах [9, 19]. В частности, были использованы фильтры размеров 3, 4, 5, их количество составляло 100 для

каждой размерности, а вероятность отсева нейронов на этапе регуляризации была равна 0.5. Для обучения нейронной сети использовался пакетный градиентный спуск с размером пакета 64; было проведено 5 тренировочных эпох.

Для векторного представления слов на основе алгоритма word2vec была использована предоставленная сайтом RusVectōrēs [44] дистрибутивная семантическая модель для русского языка, обученная на веб-корпусе, состоящем из 900 миллионов слов. Для выяснения эффективности алгоритма word2vec для векторного представления слов был также использован встроенный метод векторизации слов с установкой весов в процессе обучения.

Для извлечения аспектов было обучено двенадцать нейронных сетей, по одной для каждого из аспектов, как описано в разделе 1.3.2 данной работы.

Таблица 4.4 описывает первоначальные результаты работы. Для численной оценки качества используются значения точности (англ. Precision), полноты (англ. Recall) и F-меры с микро-усреднением.

Векторное представление слов	**Precision**	**Recall**	**F-1**
Определение весов «на лету»	0.36850	0.41971	0.39240
word2vec	0.47776	0.69383	0.56587

Таблица 4.4: Результаты работы набора свёрточных нейронных сетей с первоначальными параметрами для задачи извлечения аспектов

Заметим, что алгоритм word2vec ожидаемо показал более высокий результат, тогда как при использовании алгоритма с встроенным определением весов не удалось достичь планки, установленной контрольным алгоритмом. В связи с этим дальнейшая оптимизация параметров будет

приведена только для набора свёрточных нейронных сетей с использованием алгоритма word2vec для векторного представления слов.

Рисунки 4.1–4.5 демонстрируют влияние различных параметров свёрточной нейронной сети на результаты работы.

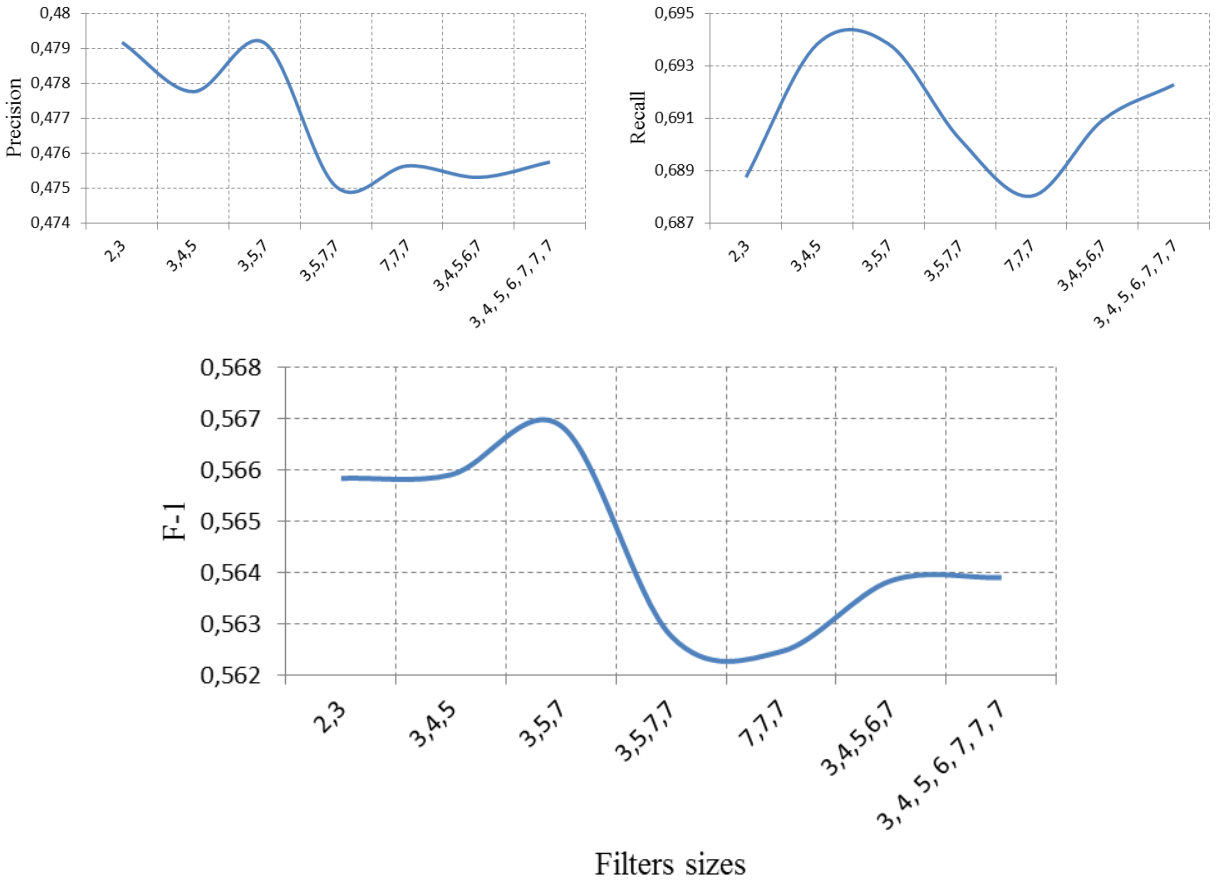

Рис. 4.1: Влияние выбора различных комбинаций фильтров на результаты работы свёрточных нейронных сетей в применении к задаче извлечения аспектов

Были проведены эксперименты для различных комбинаций фильтров (рис. 4.1). Размеры фильтров выбраны исходя из результатов предварительных экспериментов на данных малой размерности и результатов, полученных в работе [43]. Наилучшего результата удалось достичь с комбинацией фильтров размеров 3, 5, 7. Это можно объяснить необходимостью учитывать как ближний, так и дальний контекст слова, не перегружая при этом сеть дополнительной информацией в виде дублирующихся или близких по размеру фильтров.

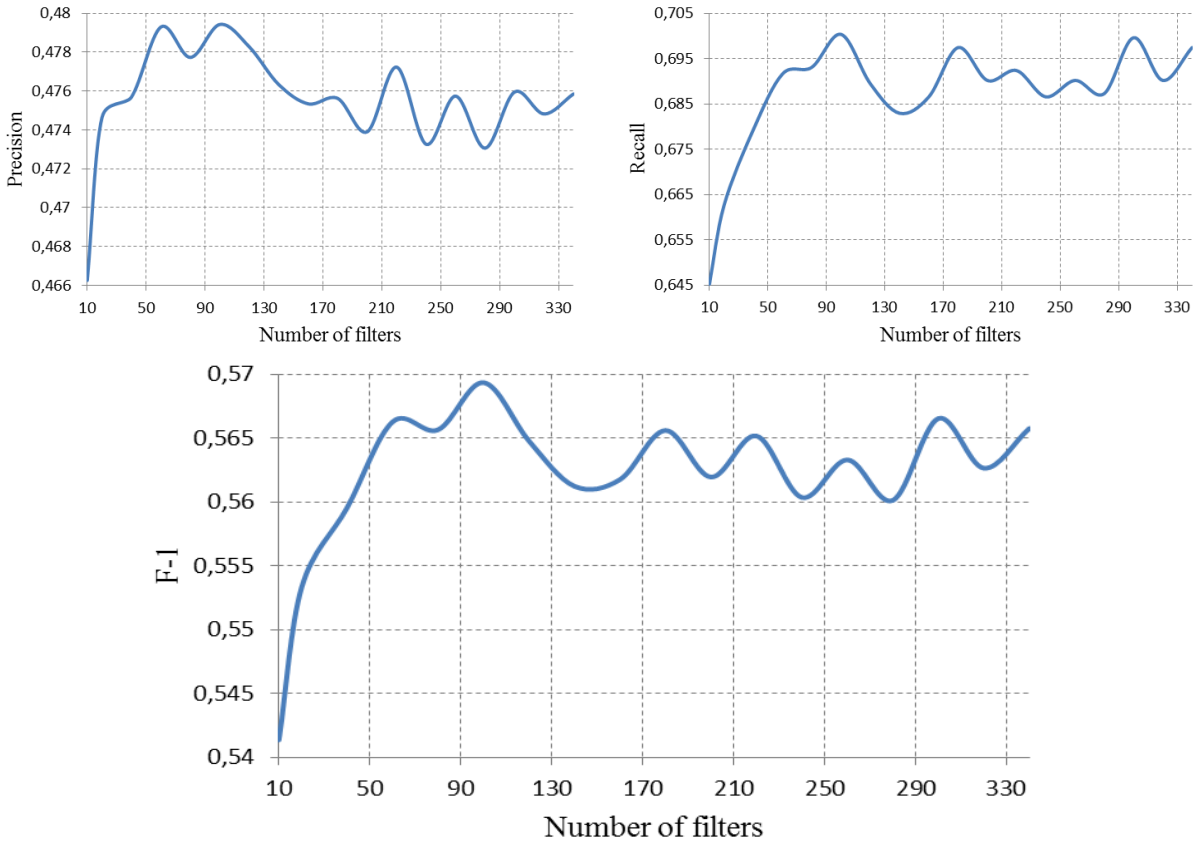

Рис. 4.2: Влияние количества фильтров на результаты работы свёрточных нейронных сетей в применении к задаче извлечения аспектов

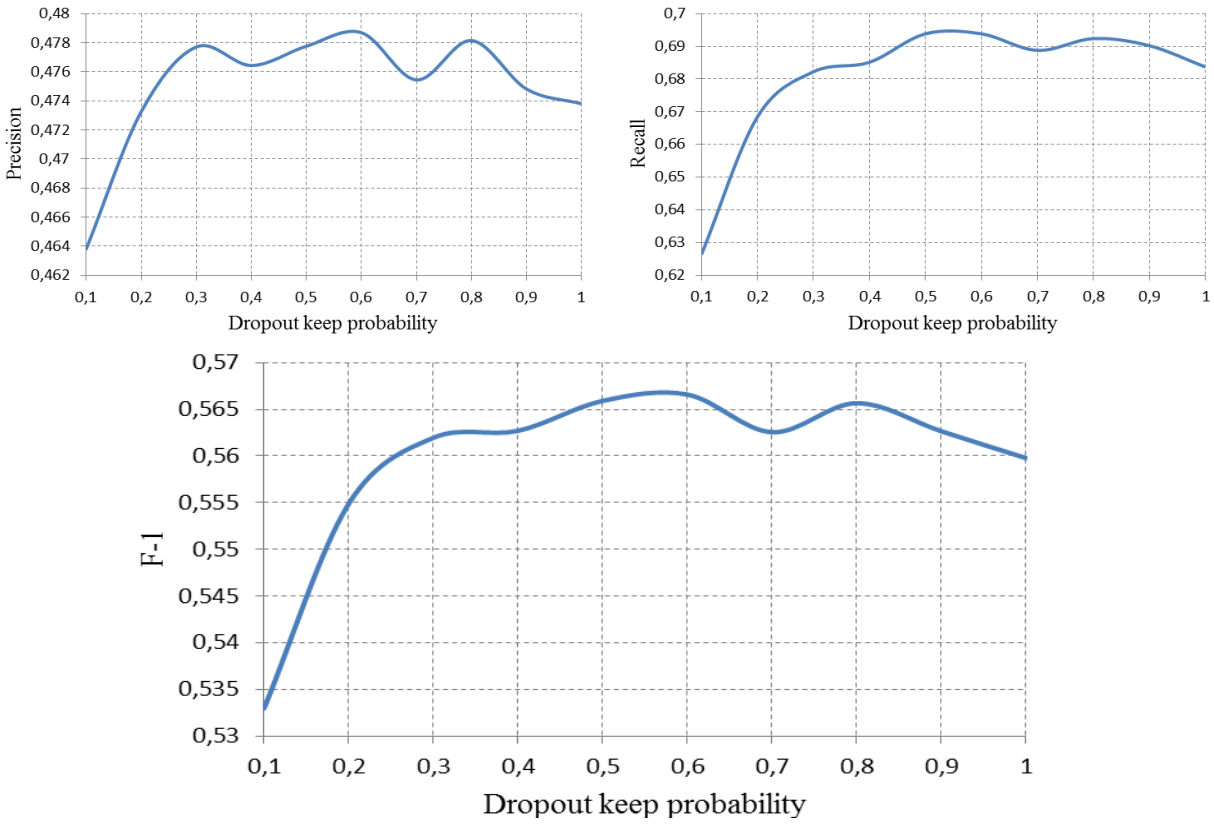

Рис. 4.3: Влияние вероятности исключения нейронов на этапе отсева на результаты работы свёрточных нейронных сетей в применении к задаче извлечения аспектов

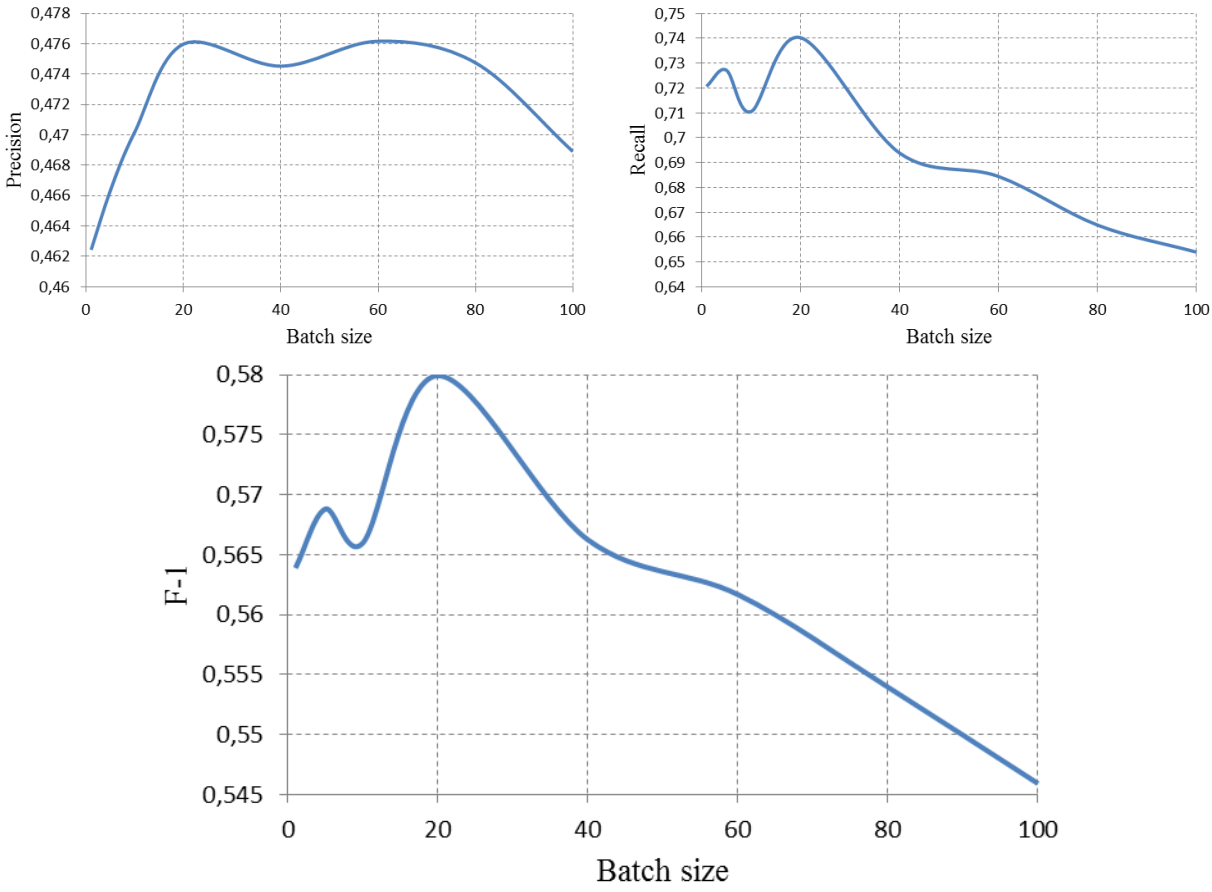

Рис. 4.4: Влияние размера пакета на результаты работы свёрточных нейронных сетей в применении к задаче извлечения аспектов

Увеличение количества фильтров (рис. 4.2) позволяет незначительно повлиять на результат, однако после достижения определенного порога рост точности и полноты прекращается.

Наилучшие результаты достигаются при сохранении вероятности отсева нейронов в районе 0.5 (рис. 4.3), поскольку в таком случае остается достаточно нейронов для классификации, и, одновременно, регуляризация сети позволяет избежать проблемы переобучения.

Пакетный градиентный спуск считается более быстрой и стабильной реализацией метода обратного распространения ошибки, в сравнении со стохастическим градиентным спуском, однако он имеет тенденцию останавливаться и застревать в локальных минимумах, особенно на высоком размере пакетов, что демонстрируют полученные результаты (рис. 4.4). В качестве компромисса можно использовать пакеты небольших размеров.

Увеличение количества эпох обучения оказывает отрицательное влияние на результаты (рис. 4.5), поскольку на небольшой обучающей выборке нейронная сеть становится более чувствительна к проблеме переобучения.

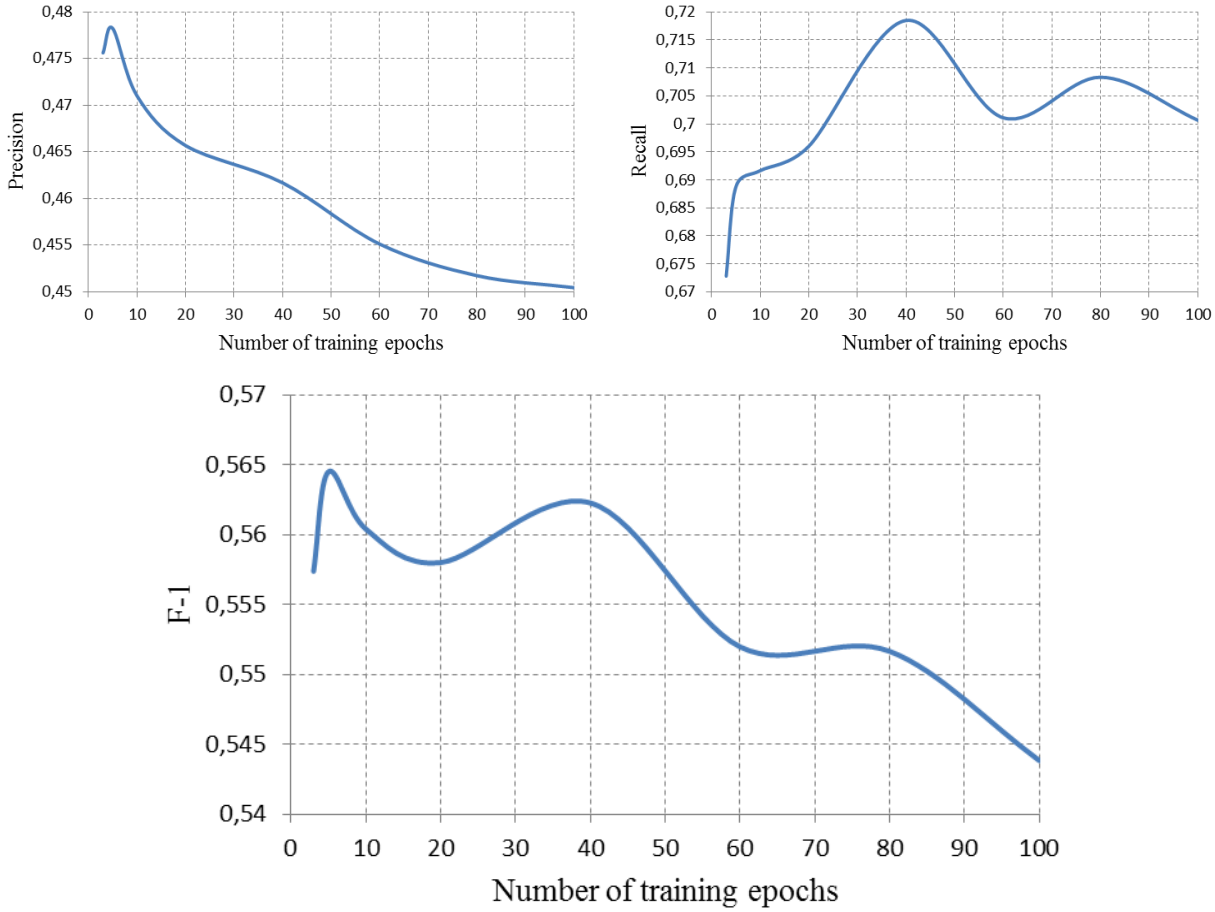

Рис. 4.5: Влияние количества тренировочных эпох на результаты работы свёрточных нейронных сетей в применении к задаче извлечения аспектов

Уровень извлечения аспектов	**Precision**	**Recall**	**F-1**
Для предложения	0.47506	0.72429	0.57384
Для отзыва	0.94394	0.83	0.88332

Таблица 4.5: Результаты работы метода на основе свёрточных нейронных сетей для извлечения аспектов на уровне предложений

Суммируя все вышесказанное, наилучшие результаты для извлечения аспектов были достигнуты с использованием фильтров размерности 3,5,7,

количеством 100 каждый, с вероятностью отсева нейронов равной 0.55, размером пакета 20, на 5 тренировочных эпохах. Полученные на этих параметрах результаты содержатся в таблице 4.5.

Использование свёрточных нейронных сетей для задачи извлечения аспектов позволило превзойти результаты контрольного алгоритма как на уровне предложений, так и на уровне отзыва.

Теперь обратимся к задаче тонального анализа. В первую очередь оценим, как влияет на результаты классификации использование двух свёрточных сетей вместо одной, как описано в разделе 1.3.2. В таблице 4.6 приведены результаты работы свёрточной нейронной сети с первоначальными параметрами (описанными ранее для случая извлечения аспектов). Для численной оценки качества приводится точность, как отношение правильно угаданных меток тональностей к общему числу меток.

Здесь и далее будут использованы следующие обозначения рассматриваемых свёрточных нейронных сетей: All — единая нейронная сеть, Neut — нейронная сеть, классифицирующая аспекты по классам «нейтральный» — «эмоциональный», Emo — нейронная сеть, классифицирующая эмоциональные аспекты по классам «положительный» — «отрицательный», {Neut; Emo} — последовательное применение нейронных сетей Neut и Emo.

Свёрточная нейронная сеть	**Acc.**
All	0.72411
Neut	0.85317
Emo	0.78673
{Neut; Emo}	0.66847

Таблица 4.6: Результаты работы свёрточных нейронных сетей с первоначальными параметрами для задачи определения тональности

Заметим, что последовательное применение двух свёрточных нейронных сетей показывает худшие результаты по сравнению с единой сетью, однако по отдельности нейронные сети из набора классифицируют тональность на порядок лучше. Это связано с превалированием одной тональности (положительной) над остальными — так, высокие значения точности достигаются за счет отнесения большей части аспектов к наиболее распространенной тональности. Нейронная сеть, отделяющая эмоциональную тональность от нейтральной, показала лучшие результаты, поскольку аспекты с нейтральной тональностью встречаются очень редко.

Далее проведем уточнение параметров свёрточной нейронной сети, подобно тому, как это было сделано для задачи извлечения аспектов. Графики, демонстрирующие влияние различных параметров на результаты, приведены на рисунках 4.6–4.10.

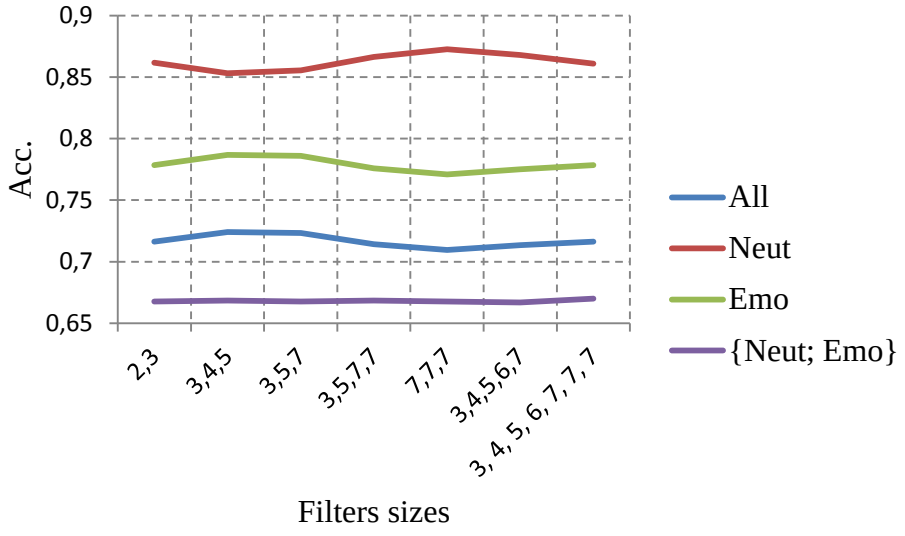

Рис. 4.6: Влияние выбора различных комбинаций фильтров на результаты работы свёрточных нейронных сетей в применении к задаче тонального анализа

Заметим, что наибольшее влияние на полученные результаты оказывает изменение вероятности отсева нейронов (рис. 4.8). Как и в случае предыдущей задачи, наилучшие результаты оказались достигнуты на вероятности в окрестности 0.5. Это верно для всех рассматриваемых нейронных сетей, кроме сети Neut, которая достигает наибольших показателей точности при максимальной рассмотренной вероятности отсева:

0.9. Это вновь можно объяснить превалированием положительной тональности над остальными. При отсеве соотношение нейтральных и эмоциональных предложений выравнивается. Однако при этом ухудшается точность определения конкретной тональности аспекта, что сказывается на результатах остальных нейронных сетей.

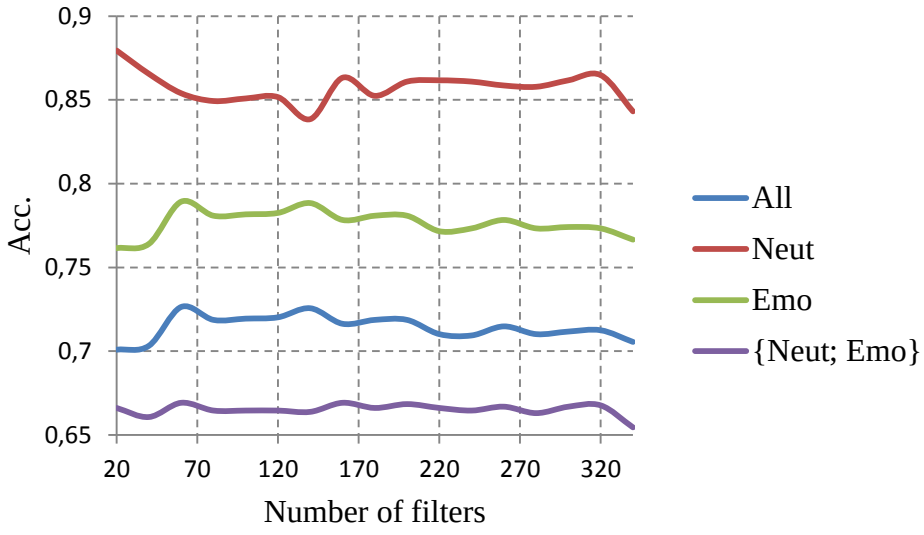

Рис. 4.7: Влияние количества фильтров на результаты работы свёрточных нейронных сетей в применении к задаче тонального анализа

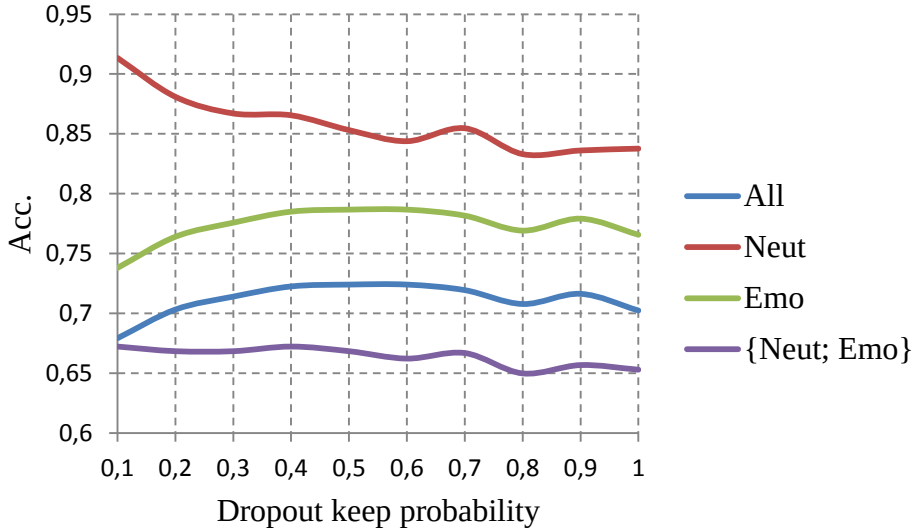

Рис. 4.8: Влияние вероятности исключения нейронов на этапе отсева на результаты работы свёрточных нейронных сетей в применении к задаче тонального анализа

Увеличение количества эпох (рис. 4.10) не способствует улучшению обучения: напротив, точность продолжает падать. Это происходит из-за переобучения на ограниченном наборе обучающих данных: для обучения

свёрточных сетей обычно используют наборы данных большей размерности. Однако для задач классификации текстов достаточно сложно подобрать и разметить выборку необходимого объема.

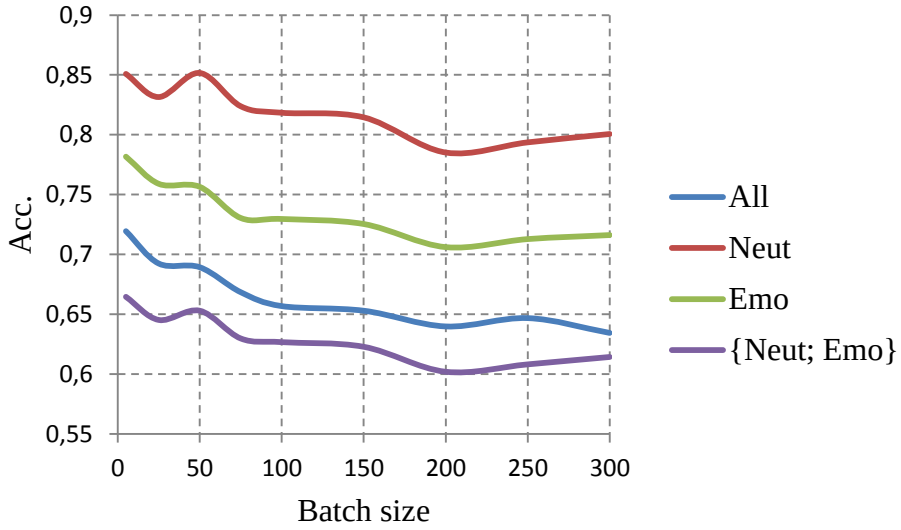

Рис. 4.9: Влияние размера пакета на результаты работы свёрточных нейронных сетей в применении к задаче тонального анализа

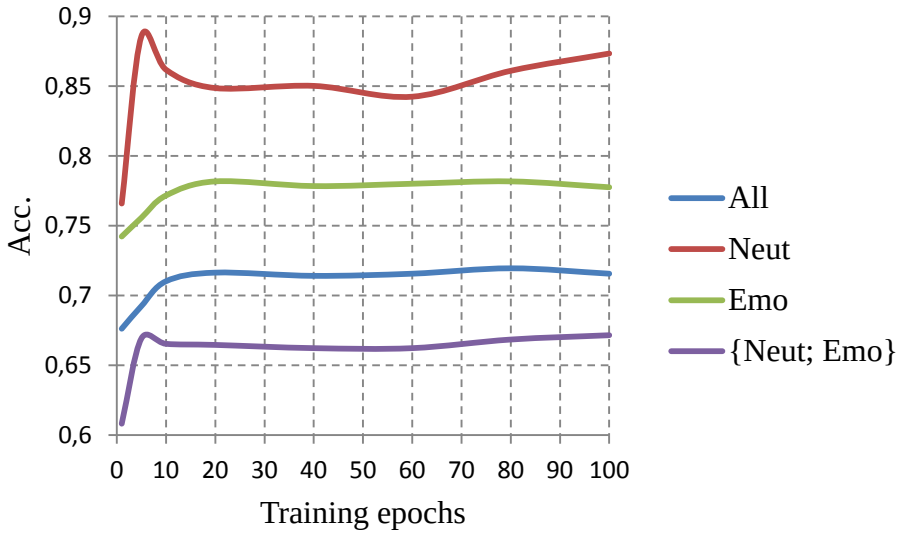

Рис. 4.10: Влияние количества тренировочных эпох на результаты работы свёрточных нейронных сетей в применении к задаче тонального анализа

Прочие параметры практически не оказывают влияния на результаты классификации.

Наилучший полученный результат, а также результат реферирования обнаруженных аспектов для отзывов приведены в таблице 4.7.

Реферирование осуществлялось по алгоритму, описанному в разделе 3.2 данной работы.

Свёрточная нейронная сеть	**Для предложений**	**Для отзыва**
All	0.72947	0.72104
Neut	0.89099	–
Emo	0.78169	–
{Neut; Emo}	0.67156	0.65939

Таблица 4.7: Результаты работы свёрточных нейронных сетей для задачи определения тональности, Acc

Итак, используя свёрточные нейронные сети, удалось превзойти показатели контрольного алгоритма в задаче определения тональности. При этом единая свёрточная нейронная сеть, решающая задачу мультиклассовой классификации, показала лучшие по точности результаты, чем набор из двух сетей, обученных для бинарной классификации.

Однако, учитывая разбалансированность данных обучающей выборки, необходимо также оценить другие метрики классификации. Для выравнивания баланса классов обучающей выборки был использован метод SMOTE (Synthetic Minority Over-sampling Technique) [46], реализованный в библиотеке imbalanced-learn [47]. Это одна из стратегий увеличения количества примеров миноритарного класса (англ. over-sampling), которая заключается в искусственной генерации примеров, которые были бы сходны с примерами миноритарного класса, но не повторяли их. Данный алгоритм использует алгоритм *k* ближайших соседей для формирования похожих примеров.

Результаты в таблице 4.8 демонстрируют, что использование методов over-sampling позволяет повысить значение F-меры за счет снижения

точности вычислений. Вместо того, чтобы классифицировать подавляющее большинство аспектов как положительные, нейронная сеть пытается более равномерно отнести их к другим классам.

Свёрточная нейронная сеть	**Для предложений**		**Для отзывов**	
	Acc.	**F-1**	**Acc.**	**F-1**
All	0.72947	0.46073	0.72104	0.41012
All + SMOTE	0.63210	0.55984	0.62379	0.49834

Таблица 4.8: Результаты работы свёрточных нейронных сетей для задачи определения тональности с использованием алгоритма SMOTE

К сожалению, авторы работы [4], откуда были взяты результаты работы контрольного алгоритма, не предоставили значения F-меры. Однако изучение работ, посвященных аспектно-ориентированному тональному анализу в той же предметной области, в частности, публикации [8], позволяет судить о том, что демонстрируемое свёрточной нейронной сетью значение F-меры находится на уровне текущих разработок в данной области.

4.4 Семантическое сходство

Для решения задачи извлечения аспектов были вручную сформированы двенадцать — по числу аспектов — словарей с эталонными аспектными терминами. Термины для словарей выбирались исключительно из слов, предложенных в обучающей выборке, чтобы исключить подбор специфических для тестовой выборки терминов.

Термины в словаре прошли первоначальную обработку: отсев слов, содержащих ошибки, приведение в нормальную форму, добавление части речи. Последнее было сделано из-за особенностей используемой модели word2vec.

Далее на языке python была написана программа, реализующая описанные в разделах 2.2 и 3.1 данной работы алгоритмы для извлечения

аспектов на уровне предложения и обобщения полученных результатов на уровень отзыва. Результаты работы программы представлены в таблице 4.9.

Метод сопоставления слова с аспектом	**Для предложений**			**Для отзыва**		
	Prec.	**Recall**	**F-1**	**Prec.**	**Recall**	**F-1**
Суммарное сходство	0.34806	0.39518	0.37002	0.69159	0.45485	0.54423
Поэлементное сравнение	0.46736	0.72257	0.56769	0.91544	0.82673	0.86897

Таблица 4.9: Результаты работы метода на основе семантического сходства для задачи извлечения аспектов

Качество работы рассмотренных подходов сильно зависело от составленных экспертных словарей. Проведенные эксперименты показали, что лучше всего они работают на словарях низкой размерности (5–15 терминов), содержащих ограниченное число специфичных для аспекта терминов.

При использовании поэлементного сравнения с аспектными терминами были получены более высокие результаты. Объясняется это тем, что за счет определения отдельных специфичных для аспекта терминов данный подход позволил успешно определять даже редкие аспекты и реже совершать ошибки первого рода.

Однако для данного подхода характерны ложные срабатывания для аспектов одной направленности, к примеру, таких как FOOD#PRICES и DRINKS#PRICES. Относящиеся к таким аспектам термины часто совпадают или являются семантически схожими. Отсюда и низкие значения точности. Впрочем, при обобщении результатов на уровень отзывов, и точность, и полнота находятся на достаточно высоком уровне. Это можно объяснить тем, что каждый отзыв имеет достаточно обширный набор аспектов, поэтому

даже для случаев ошибки второго рода на уровне предложений существует вероятность, что такой аспект в отзыве действительно существует.

Несмотря на то, что подход, основанный на вычислении суммарного сходства, показал равномерные результаты по точности и полноте, их значения оказались существенно ниже результатов, показанных контрольным алгоритмом.

Обратимся к задаче тонального анализа по аспектам. Для ее решения были составлены словари эталонных эмоциональных терминов и написана программа, реализующая алгоритм, описанный в разделе 2.4 данной работы. Для реализации классификатора на основе решающих деревьев была использована библиотека scikit-learn [45] языка python. Для выравнивания баланса классов обучающей выборки был вновь использован метод SMOTE.

Были рассмотрены два различных подхода: с использованием единого классификатора (All), и двух отдельных классификаторов, которые решали задачи бинарной классификации между классами «нейтральный» — «эмоциональный» (Neut) и «положительный» — «отрицательный» (Emo).

Классификатор	**Для предложений**		**Для отзыва**	
	Acc.	**F-1**	**Acc.**	**F-1**
All	0.67626	0.3668	0.66616	0.31775
All + SMOTE	0.53777	0.50428	0.50747	0.45185
Neut	0.89915	0.48122	–	–
Emo	0.76366	0.53375	–	–
{Neut; Emo}	0.68345	0.27698	0.70833	0.20758
{Neut; Emo} + SMOTE	0.67626	0.27518	0.71710	0.27201

Таблица 4.10: Результаты работы метода на основе семантического сходства для задачи определения тональности, Acc

Исходя из полученных результатов (таблица 4.10) можно выделить несколько особенностей использования предложенного метода. Во-первых, заметим, что использование набора из двух классификаторов демонстрирует более высокие значения точности, однако отстает по F-мере. Такой подход хуже справляется с разбалансированностью данных, и даже применением метода SMOTE не удается справиться с этой проблемой. Высокие значения точности на уровне отзыва обусловлены тем, что классификатор отнёс подавляющее число примеров к одному доминирующему классу.

В случае единого классификатора применение SMOTE, напротив, дало прирост значения F-меры. С учетом высокой разбалансированности классов необходимо ориентироваться именно на это значение, поэтому можно сделать вывод, что в рамках рассматриваемого метода лучшие результаты показал единый классификатор с методом SMOTE для выравнивания баланса классов при обучении.

4.5 Выводы

На полученных результатов можно сделать следующие выводы:

1. Свёрточные нейронные сети являются эффективным методом для решения задач, связанных с аспектно-ориентированным тональным анализом. В связке с векторным представлением слов с помощью word2vec они позволили решить поставленные задачи с результатами, превышающими результаты контрольного алгоритма по точности и F-мере.
2. Варьируемые параметры свёрточных нейронных сетей позволяют незначительно влиять на результаты их работы, однако по-настоящему сильного воздействия на них не оказывают. Использование же различных подходов к векторному представлению слов позволило значительно повысить эффективность работы сети. Из этого можно сделать вывод, что в задачах, связанных с классификацией текстов, для

улучшения результатов необходимо сконцентрироваться не на подборе параметров, а на предварительной обработке и числовом представлении текстов. Впрочем, для уточнения этого вывода требуется проведение дальнейших экспериментов.

3. Для обучения свёрточной нейронной сети необходим большой корпус документов, создание которого может представлять определенную сложность, особенно, если данные в корпусе должны быть определенным образом размечены.
4. Метод, основанный на семантическом сходстве, на каждой из задач показывает результаты, сравнимые с результатами контрольного алгоритма. При этом для сравнения слова с аспектом лучше подходит метод, основанный на поэлементном сравнении слова с аспектными терминами. А для определения тональности — единый классификатор для определения всех трех классов.
5. На данный момент использование метода на основе семантического сходства при решении практических задач для выявления мнений пользователей по конкретным аспектам представляется достаточно затруднительным по ряду причин:
    a. высокая зависимость качества работы метода от экспертных словарей. Как следствие, узкая направленность метода: подходит только для одной предметной области,
    b. трудоемкость процесса формирования словарей и субъективность экспертных мнений,
    c. необходимость наличия обширной сбалансированной обучающей выборки.
6. Применение алгоритма SMOTE для выравнивания количества примеров различных классов в данных позволяет выровнять отношение точность–полнота для полученных результатов, что очень полезно при использовании обучающей выборки, разбалансированной по числу классов.

Заключение

В данной работе производилось исследование свёрточных нейронных сетей и метода, основанного на семантическом сходстве, в применении к задаче аспектно-ориентированного тонального анализа. Результаты работы этих методов сравнивались с контрольным алгоритмом, основанном на классических методах представления и классификации текстов.

Задача аспектно-ориентированного тонального анализа была разбита на несколько подзадач: извлечение аспектов для отдельных предложений, определение тональности для отдельных предложений и реферирование результатов для отзыва.

В качестве данных для вычислительных экспериментов использовались размеченные отзывы предметной области «рестораны» на русском языке.

Была проверена гипотеза о том, что свёрточные нейронные сети, которые зарекомендовали себя, как мощный алгоритм для классификации изображений, также пригодны для решения задач классификации текстов. Была реализована свёрточная нейронная сеть; показано, что она справляется с поставленными задачами на уровне или лучше, чем контрольный метод. Исследовано влияние различных способов векторного представления слов на качество работы сети; проведены испытания на различных наборах параметров, сделаны выводы об их влиянии на результаты.

Был предложен метод использования семантического сходства для решения поставленных задач, написана программа, реализующая этот метод. Показано, что данный метод показывает результаты на уровне контрольного, сделаны выводы о его применимости к задаче аспектно-ориентированного тонального анализа.

Наконец, был предложен метод для решения задачи реферирования аспектов и тональных меток для отзыва. Его применение позволило получить результаты, превышающие результаты контрольного алгоритма.

Список литературы

1. Показатели развития информационного общества в Российской Федерации. http://www.gks.ru/free_doc/new_site/business/it/monitor_rf.xls
2. Pontiki M., Galanis D., Pavlopoulos I., Papageorgiou H., Androutsopoulos I., Manandhar S. SemEval 2014 Task 4: Aspect Based Sentiment Analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) at (COLING 2014), 2014. P. 27–35.
3. Pontiki M., Galanis D., Papageorgiou H., Androutsopoulos I., Manandhar S. SemEval-2015 Task 12: Aspect Based Sentiment Analysis // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), at the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015), 2015. P. 486–495.
4. Pontiki M. [et al.] SemEval-2016 Task 5: Aspect Based Sentiment Analysis // Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), 2016. P. 19–30.
5. Loukachevitch N. [et al.] SentiRuEval: testing object-oriented sentiment analysis systems in russian // Proceedings of International Conference Dialog, 2015. P. 12–24.
6. Тутубалина Е. В. Методы извлечения и резюмирования критических отзывов пользователей о продукции. http://www.ispras.ru/dcouncil/docs/diss/2016/tutubalina/dissertacija-tutubalina.pdf
7. Poria S. [et al.] A rule-based approach to aspect extraction from product reviews / Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP), 2014. P. 28–37.
8. Блинов П.Д., Котельников Е.В. Семантическое сходство в задаче аспектно-эмоционального анализа // Российский научный электронный журнал, 2015. Т. 18, № 3–4. С. 120–137.

9. Kim, Y. Convolutional neural networks for sentence classification / IEMNLP, 2014. P. 1746–1751.
10. Hu M., Liu B. Mining and summarizing customer reviews // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004. P. 168–177.
11. Popescu A., Etzioni O. Extracting product features and opinions from reviews // Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005. P. 339–346.
12. Schouten K., Frasincar F., Jong F. COMMIT-P1WP3: A Co-occurrence based approach to aspect-level sentiment analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014. P. 203–207.
13. Jin W. A Novel Lexicalized HMM-based Learning Framework for Web Opinion mining // Proceedings of the 26th Annual International Conference on Machine Learning, 2009. P. 465–472.
14. Chernyshevich M. IHS R&D Belarus: Cross-domain extraction of product features using CRF // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014. P. 309–313.
15. Toh Z., Su J. NLANGP: Supervised Machine Learning System for Aspect Category Classification and Opinion Target Extraction / Z. Toh, J. Su // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015. P. 568–573.
16. Titov, I. Modeling Online Reviews with Multi-grain Topic Models // Proceedings of the 17th International Conference on World Wide Web (WWW‘08), 2008. P. 111–120.
17. Ruder S., Ghaffari P., Breslin J. G. Deep Learning for Multilingual Aspect-based Sentiment Analysis // Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), 2016. P. 342–348.

18. Tamchyna A., Veselovska K. Recurrent Neural Networks for Sentence Classification // Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), 2016. P. 379–383.

19. Khalil T., El-Beltagy S. R. Deep Convolutional Neural Networks for Aspect Category and Sentiment Extraction // Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), 2016. P. 276–281.

20. Tarasov D. S. Deep Recurrent Neural Networks for Multiple Language Aspect-based Sentiment Analysis of User Reviews // Proceedings of the 21st International Conference on Computational Linguistics Dialog-2015, 2015. V. 2. P. 77–88.

21. Андрианов И. А., Майоров В. Д., Турдаков Д. Ю. Современные методы аспектно-ориентированного анализа эмоциональной окраски // Труды ИСП РАН, 2015. Т. 27, № 5. С. 5–22.

22. Qiu G. Opinion Word Expansion and Target Extraction Through Double Propagation // Computational Linguistics, 2011. Vol. 37, No. 1. P. 9–27.

23. Veselovska K. Using Hand-crafted Rules in Aspect Based Sentiment Analysis on Parsed Data // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014. P. 694–698.

24. Kiritchenko S., Zhu X., Cherry C., Mohammad S. Detecting Aspects and Sentiment in Customer Reviews // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014. P. 437–442.

25. Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, 2002. P. 417–424.

26. Tjong Kim Sang E. F., Meulder F.D. Introduction to the CoNLL-2003 Shared Task: Language independent Named Entity Recognition // Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, 2003. P. 142–147.

27. Wagner J. [et al.] Aspect-based Polarity Classification for SemEval Task 4 // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014. P. 223–229.

28. Zhang F., Lan M. A Combination Method and Multiple Features for Aspect Extraction and Sentiment Polarity Classification // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014. P. 252–258.

29. Mayorov V., Andrianov I. Syntactic and word2vec-based approach to aspect-based polarity detection in Russian // Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), 2016. P. 337–341.

30. Russian Semantic Similarity Evaluation. https://nlpub.ru/Russian_Semantic_Similarity_Evaluation

31. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality // Proceedings of NIPS, 2013. P. 3111-3119.

32. Manning C., Raghavan P., Schütze H. Introduction to information retrieval. Cambridge University Press. New York, 2008. P. 121.

33. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space // Proceedings of Workshop at ICLR, 2013.

34. Rong X. word2vec Parameter Learning Explained // arXiv preprint arXiv: 1411.2738, 2016.

35. McCulloch W. S., Pitts W. A logical calculus of the ideas immanent in nervous activity // Springer New York, 1943.

36. Короткий С. Нейронные сети: Алгоритм обратного распространения. http://www.gotai.net/documents-neural_networks.aspx

37. LeCun Y. [et al.] Gradient-based learning applied to document recognition // IEEE, 1998.

38. Russakovsky O., Deng J. [et al.] ImageNet Large Scale Visual Recognition Challenge // IJCV, 2015.

39. Srivastava N., Hinton G. [et al.] Dropout: A Simple Way to Prevent Neural Networks from Overfitting // Journal of Machine Learning Research, 2014. Vol. 15 Issue 6. P. 1929.

40. Союз // Русская грамматика. http://rusgram.ru/Союз

41. Friedman J. Greedy function approximation: a gradient boosting machine // The Annals of Statistics, 2001. V. 29. P. 1189–1232.

42. TensorFlow. https://www.tensorflow.org/

43. Attardi G., Sartiano D. Convolutional Neural Networks for Sentiment Classification // Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), 2016. P. 225–229.

44. RusVectōrēs: дистрибутивные семантические модели для русского языка. http://rusvectores.org/ru/

45. scikit-learn. Machine Learning in Python. http://scikit-learn.org/stable/index.html

46. Chawla N. [et al.] SMOTE: Synthetic Minority Over-sampling Technique // Journal of Artificial Intelligence Research 16, 2002. P. 321–357.

47. imbalanced-learn. https://github.com/scikit-learn-contrib/imbalanced-learn