УДК 004.912 **DOI** 10.15622/sp.52.3

А.В. Глазкова

ПОДХОД К ПРОВЕДЕНИЮ КЛАССИФИКАЦИИ ТЕКСТОВ НА ОСНОВАНИИ ВОЗРАСТНЫХ ГРУПП ИХ АДРЕСАТОВ

Глазкова А.В. **Подход к проведению классификации текстов на основании возрастных групп их адресатов.**

Аннотация. В статье рассматривается новый подход к выполнению классификации текстов, учитывающий наличие различных типов классификационных признаков (бинарных, номинальных, порядковых и интервальных). Особенность представленного подхода состоит в поэтапном проведении классификации, которое дает возможность не приводить разнотипные признаки, характеризующие текст, к единому диапазону. Также в статье предлагается набор классификационных признаков для проведения классификации русскоязычных текстов на основании их предполагаемой возрастной аудитории.

В работе описывается вычислительный эксперимент с использованием текстов, включенных в Национальный корпус русского языка. Выборка включает в себя заведомо качественные и максимально разнообразные тексты, написанные на русском языке. Документы, входящие в состав рассматриваемой выборки, разделены в соответствии с мнениями экспертов-лингвистов на две категории — взрослые и детские. Таким образом, для обучения и тестирования использовались тексты, размеченные экспертами.

В статье приведены значения точности классификации текстов, полученные в результате проведения серии экспериментов по автоматическому определению возрастных категорий адресатов текста (для кого написан текст — для детей или для взрослых).

Ключевые слова: извлечение информации, классификация текстов, обработка естественного языка, признаки текста.

1. Введение. Вопросы классификации текстовой информации особенно актуальны в связи с постоянным увеличением количества электронных ресурсов. Механизмы обработки и классификации текстов (в частности, представленные в работах [1-6]) позволяют уменьшить объем необходимой к просмотру информации и ограничить содержимое результата поискового запроса в соответствии с требуемыми характеристиками. Так, в статье С. Н. Карповича [5] рассмотрен алгоритм многозначной классификации текстов на основании вероятностных тематических моделей. В работе А. О. Шумской [6] предложен метод определения автоматически сгенерированных текстов, который может применяться для отсечения поискового спама.

В настоящее время наблюдается повышенный интерес исследователей и разработчиков к возможностям классификации текстов на основании характеристик, общих для текстов, адресованных одной возрастной аудитории. Актуальность решения задач, связанных с определением предполагаемой группы читателей, для которых написан текст, обоснована введением возрастных ограничений на содержимое Интернет-ресурсов, развитием систем электронного обучения, а также малой освещенностью обозначенной проблемы в работах российских ученых.

В работах зарубежных ученых тема установления характеристик адресата текста раскрыта более широко. При этом исследования выполнялись преимущественно для англоязычных корпусов (в частности, E. Shriberg и соавторы [7], S.V. Ravuri и A. Stolcke [8]). Следовательно, в настоящее время не существует общепринятого мнения о том, какой набор классификационных признаков может быть положен в основу данной классификации для русскоязычных текстов. Работы N. Jovanovic и H. op den Akker [9], H. Lee и соавторов [10], K. Santosh и соавторов [11] посвящены определению адресатов реплик мультидиалога или группового чата. В качестве признаков, служащих идентификатором адресата, в данных работах выступают личные, неопределенные и притяжательные местоимения, личные имена участников диалога, последовательность высказываний и т.д. Стоит отметить, что наборы признаков, посвященных выявлению адресата текста, представленные указанными авторами, характеризуются узкой предметностью. Очевидно, что при проведении классификации текстов общей тематики перечисленные признаки не будут являться информативными.

К работам близкой тематики следует также отнести статьи A. Pentel [12-13], посвященные идентификации возраста автора текста. В указанных работах в качестве классификационных признаков используются среднее количество символов в словах, среднее количество слов в предложении, среднее количество грамматических основ в предложении, доля «сложных» слов (состоящих из n и более символов), доля односложных слов, показатели удобочитаемости текстов [14].

В данной работе предлагается подход к проведению классификации текстов на основании возрастных групп адресатов (то есть выявлению возрастной группы читателей, для которой написан текст). Актуальность решения рассматриваемой задачи продиктована развитием систем электронного обучения, необходимостью совершенствования механизмов поиска в онлайн-библиотеках и электронных каталогах, а также введением на законодательном уровне возрастных ограничений на информационную продукцию и необходимости соответствующей нормировки информационных ресурсов [15].

Особенностью предлагаемого в работе подхода выступает поэтапное выполнение классификации, позволяющее избежать приведения значений признаков, выраженных в разных шкалах, к единому диапазону. Также в работе описываются формирование набора классификационных признаков для русскоязычных текстов и вычислительный эксперимент, использующий предложенный подход.

2. Типы признаков. Формирование набора признаков текстов, подлежащих классификации на основании их возрастной аудитории,

подразумевает наличие характеристик разных типов. С метрической точки зрения в работе рассматривались следующие типы признаков:

– бинарные: $\{0,1\}$ (например, наличие/отсутствие специальной лексики в тексте).

– номинальные: конечное множество значений (литературная форма: рассказ, повесть, роман; жанр).

– порядковые: конечное упорядоченное множество значений (период создания; уровень образования аудитории).

– интервальные: интервальное значение (число сложных синтаксических конструкций; число предложений).

Некоторые из бинарных, номинальных и порядковых признаков, участвующих в процессе классификации, могут не влиять на принадлежность текста категории (например, структурный тип текста: проза или поэзия). В то время как влияющие признаки данных типов могут выполнять различные функции:

– либо представлять собой маркеры, ограничивающие круг категорий, с которыми сопоставляется текст;

– либо являться специфическим маркером наличия дополнительных уточняющих признаков.

Так, присутствие ненормативной лексики в тексте однозначно говорит о том, что данный текст не предназначен читателям младших возрастных групп. С другой стороны, бинарный признак, характеризующий наличие иллюстраций в документе, нуждается в уточнениях, касающихся типа изображений. С большой долей вероятности текст, содержащий графики, не может быть адресован младшей возрастной аудитории.

3. Алгоритм классификации. Присутствие маркеров в признаковом описании текста позволяет проводить классификацию поэтапно.

Шаг 1. В первую очередь проводится определение степени влияния бинарных, номинальных и порядковых признаков.

Шаг 2. Далее оценивается наличие значений, однозначно указывающих на категорию адресатов или ограничивающих круг категорий, которым может принадлежать текст.

Шаг 3. Если на втором шаге категория текста не была определена однозначно, то на основании значений оставшихся признаков, измеряемых в интервальной шкале (средняя длина предложений, количество многосложных слов и др.), проводится сопоставление текстов с рассматриваемыми категориями.

Рассмотрим следующий пример. Пусть классификация проводится по категориям $\{K_1, K_2, K_3, K_4, K_5\}$, которые характеризуются критическими значениями признаков $\{S_1, S_2, S_3, S_4, S_5, S_6\}$. Критиче-

ское значение признака для категории представляет собой возможный диапазон значений этого признака для текстов данной категории.

Описания категорий представлены в таблице 1.

Таблица 1. Наборы критических значений признаков для категорий $\{K_1, K_2, K_3, K_4, K_5\}$

Категория	S_1	S_2	S_3	S_4	S_5	S_6
K_1	1	проза	тип_1	1	[0, 0,5)	[0, 0,5)
K_2	1	проза	тип_1	0	[0,5, 1]	[0, 0,25)
K_3	1	поэзия	тип_1	0	[0, 0,5)	[0,5, 1]
K_4	0	проза	тип_1	0	[0, 0,5)	[0, 0,5)
K_5	1	поэзия	тип_2	0	[0, 0,5)	[0, 0,5)

В данном примере признаки имеют следующие характеристики:

1) S_1 и S_4 — бинарные признаки, $D_{S_1} = D_{S_4} = \{0,1\}$;
2) S_2 — номинальный, $D_{S_2} = \{"поэзия", "проза"\}$;
3) S_3 — порядковый, $D_{S_3} = \{"тип\_1", "тип\_2"\}$;
4) S_5 и S_6 — интервальные, $D_{S_5} = D_{S_6} = [0,1]$;

где D_{S_k} — множество допустимых значений признака, $k = \overline{1,6}$.

Требуется определить категорию, к которой относится текст T, определяемый набором значений признаков следующим образом: $F_T = (1, "проза", "тип\_1", 0, 0.25, 0.75)$. Совпадения значений классификационных признаков текста и категорий выделены курсивом в таблице 2.

Таблица 2. Совпадения значений признаков текста T с критическими значениями категорий $\{K_1, K_2, K_3, K_4, K_5\}$

Категория	S_1	S_2	S_3	S_4	S_5	S_6
K_1	*1*	*проза*	*тип_1*	1	*[0, 0,5)*	[0, 0,5)
K_2	*1*	*проза*	*тип_1*	*0*	[0,5, 1]	[0, 0,25)
K_3	*1*	поэзия	*тип_1*	*0*	*[0, 0,5)*	*[0,5, 1]*
K_4	0	*проза*	*тип_1*	*0*	*[0, 0,5)*	[0, 0,5)
K_5	*1*	поэзия	тип_2	*0*	*[0, 0,5)*	[0, 0,5)

В рамках предложенного алгоритма выполняются следующие действия.

Шаг 1. Будем считать, что для решаемой задачи выполняются условия:

1) если значение признака S_1 равно 1, то текст однозначно не может быть отнесен к категориям, для которых значение критического значения признака S_1 равно 0;

2) значение S_2 не оказывает влияния на принадлежность текста категории (признак S_2 может быть исключен из дальнейшего рассмотрения для сокращения размерности признакового пространства; в данном примере он приводится в качестве иллюстрации того, что признаки могут оказывать различную степень влияния на принадлежность текста категориям);

3) значение S_3 само по себе не оказывает влияния на принадлежность текста категории. Однако тексты, для которых $S_3 = "mun\_1"$ и $S_4 = 0$, не могут относиться к категориям, для которых $S_4 = 1$.

Шаг 2. Если учесть принятые условия, круг категорий, к которым можно отнести рассматриваемый текст, может быть сужен (таблица 3). Согласно условию 1, текст T не может быть отнесен к категории K_4. Исходя из условия 3, из рассмотрения исключена категория K_1.

Таблица 3. Список рассматриваемых категорий (шаг 2)

$F_T = (1, "проза", "mun\_1", 0, 0.25, 0.75)$						
Категория	S_1	S_2	S_3	S_4	S_5	S_6
K_2	*1*	*проза*	*mun_1*	*0*	[0,5, 1]	[0, 0,25)
K_3	*1*	поэзия	*mun_1*	*0*	*[0, 0,5)*	*[0,5 ,1]*
K_5	*1*	поэзия	тип_2	*0*	*[0, 0,5)*	[0, 0,5)

Шаг 3. В ходе сопоставления значений признаков текста критическим значениям интервальных признаков категорий определяется соответствие текста оставшимся категориям.

4. Корпус текстов. С целью проверки эффективности предложенного алгоритма был проведен вычислительный эксперимент на множестве текстов, написанных на естественном языке. В ходе вычислительного эксперимента использовались база данных «Морфологический стандарт Национального корпуса русского языка» и «База данных метатекстовой разметки Национального корпуса русского языка» (коллекция детской литературы)» [16].

Тексты, составляющие Национальный корпус русского языка [17], размечены по различным лингвистическим параметрам. Базы содержат заведомо качественные и максимально разнообразные тек-

сты на русском языке, возрастная категория потенциальных читателей которых (взрослая или детская) определена на основании мнений экспертов. Объем выборки составляет 532 текста художественной литературы и 510 текстов детской литературы. В базах данных представлены тексты 372 авторов. Во время обучения и тестирования выборка n раз разбивалась на обучающую и контрольную. Далее были посчитаны средние значения точности по всем разбиениям.

Распределение текстов по длине (по количеству слов) представлено на рисунке 1, по году создания — на рисунке 2. Минимальная длина текстов, входящих в базы данных, составляет 30 слов. По данным, представленным на графике (рисунок 1), видно, что более 60% текстов (623 текста) имеют длину не более 500 слов. Средняя длина текста в корпусе составляет 471 слово.

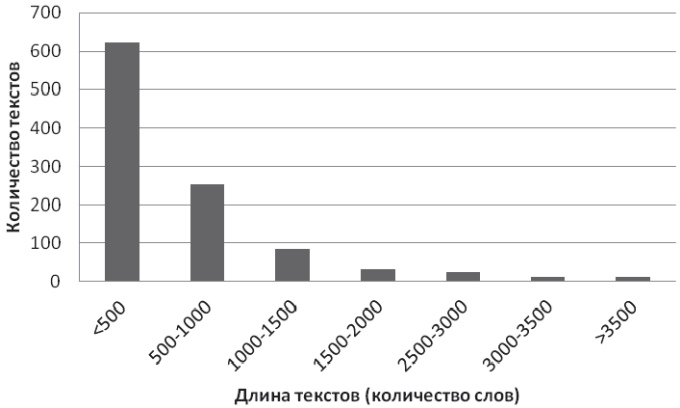

Рис. 1. Распределение длины текстов в корпусе

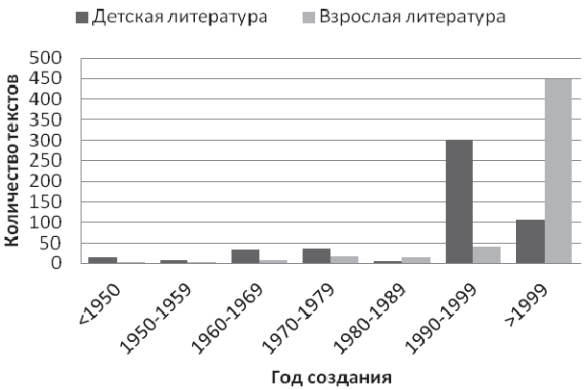

Рис. 2. Распределение длины текстов в корпусе

Для проведения отбора классификационных признаков также использовались две выборки текстов, находящихся в открытом доступе на сайте Национального корпуса русского языка [17]. Первая выборка представлена художественными текстами различных жанров (историческая проза, приключения, документальная проза и т.д., кроме детской литературы; всего 5 902 документа, 9 332 659 предложений, 94 538 056 слов), вторая — детской литературой (всего 632 документа, 547 735 предложений, 4 742 627 слов).

В исследовании, в соответствии с предоставленной для эксперимента выборкой, используется деление текстов на детские и взрослые.

5. Поиск классификационных признаков. В первую очередь для формирования набора бинарных, номинальных и порядковых признаков были использованы данные корпусной разметки текстов. Перечень исходных характеристик текстов представлен в таблице 4. В столбце «Возможные значения» приведены возможные значения характеристики для текстов обучающей выборки.

Таблица 4. Характеристики исходных данных

Характеристика	Шкала измерения	Возможные значения
Возрастная аудитория	Номинальная	взрослая; детская
Год создания	Порядковая	1829-2005 (с рядом исключений, отсутствующих в базе)
Жанр	Номинальная	заметка; интервью; мемуары; очерк; рассказ; сказка; статья; учебная литература; фельетон
Источник	Номинальная	газета; журнал; книга
Наличие иллюстраций	Бинарная	присутствуют; отсутствуют
Пол автора	Бинарная	женский; мужской
Тип иллюстрации	Номинальная	график; рисунок; фотография
Функциональный стиль	Номинальная	научный; публицистический; художественный
Хронотоп	Бинарная	ирреальный мир; реальный мир

Таким образом, тексты, входящие в обучающую выборку, изначально разделены на взрослые и детские.

Характеристика «Возрастная аудитория» является меткой классов, ее значения использовались при обучении и тестировании классификатора. Остальные характеристики, представленные в таблице 4, послужили основой для формирования признакового пространства.

Далее был выполнен поиск признаков, значения которых не встречаются в явном виде в разметке (в основном, количественных). Для этого была предварительно проведена фильтрация слов, входящих в рассматриваемые тексты, которая включает в себя следующие этапы:

1) отсечение стоп-слов (союзов, междометий, местоимений, чисел, отдельно стоящих букв, общеупотребительных предлогов и вводных слов);

2) отсечение слов, входящих в более чем m (%) документов. В вычислительном эксперименте значение варьировалось от 70 до 80 процентов документов.

Затем слова текстов каждой из рассматриваемых выборок были представлены в виде множества лексем, объединяющих в себе словоформы каждого встречающегося в тексте слова и соответствующих им частотностей. При формировании множества лексем диминутивы (слова, имеющие уменьшительный аффиксы, — например, «домик», «ключик» и т.п.) рассматривались как отдельные лексемы [18]. Таким образом, были организованы модели bag-of-words [19] для каждой категории текстов. Сами же исходные тексты были представлены в виде набора предложений (рисунок 3).

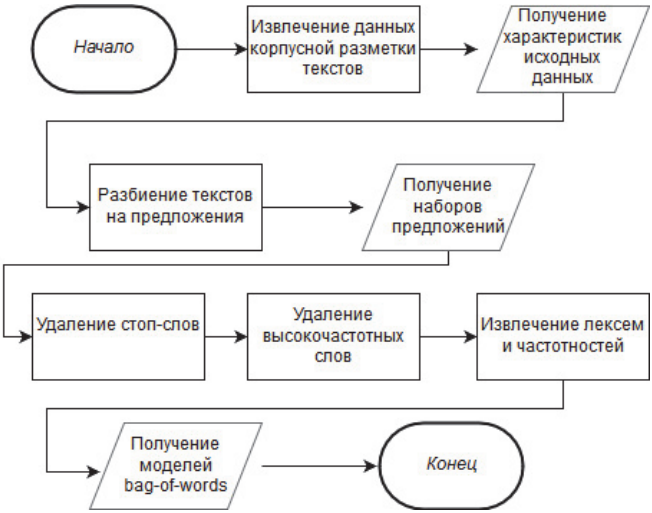

Рис. 3. Представление текстов

В целях проведения первого шага классификации в качестве признаков, предположительно являющихся маркерами возрастной аудитории, представляется возможным выделить характеристики, описание которых приведено в таблице 5.

Таблица 5. Перечень признаков, являющихся маркерами возрастной аудитории

Признак	Тип признака	Значение	Взрослые тексты	Детские тексты
Наличие ненормативной лексики	Бинарный	Присутствует	1	0
Наличие специальной лексики	Бинарный	Присутствует	0,95	0,05
Наличие специальных символов	Бинарный	Присутствует	0,96	0,04
Жанр	Номинальный	Мемуары	1	0
		Сказка	0	1
Тип иллюстраций	Номинальный	График	0,98	0,02

В столбцах «Взрослые тексты» и «Детские тексты» приводится доля текстов обучающей выборки, относящихся к данной категории, от общего количества текстов, обладающих данным значением признака.

Отличие в количестве значений признаков, приведенных в таблицах 4 и 5, связано с тем, что в таблице 5 присутствуют только те значения, которые являются маркерами возрастной аудитории, то есть такие, присутствие которых с определенной долей вероятности (>0,9) говорит о принадлежности текста категории.

Величины, представленные в столбцах «Взрослые тексты» и «Детские тексты», были получены на основании анализа обучающей выборки. Соответственно, они обусловлены составом корпуса, предоставленного для анализа. Так, значение «Сказка» признака «Жанр», согласно данным анализа обучающей выборки, говорит об однозначной принадлежности текста детской возрастной аудитории. В то время как на практике значение «Сказка» может соответствовать произведениям для взрослых (например, произведениям братьев Стругацких, Дж. К. Роулинг, Дж. Р. Р. Толкина).

Необходимо заметить, что некоторые значения признаков не являются однозначными маркерами возрастной аудитории. В частности, для текстов, имеющих значение «График» признака «Тип иллюстра-

ций», принадлежность к взрослой возрастной аудитории определяется с вероятностью 0,98. Однако, поскольку в данном эксперименте производится деление текстов на взрослые и детские, этой неточностью можно пренебречь. В то же время при проведении классификации детских текстов по более конкретным возрастным категориям выраженность значения данного признака может служить маркером более взрослой аудитории или требовать введения уточняющих признаков.

В рамках данного исследования в качестве маркеров использовались, в частности, признаки «Наличие ненормативной лексики» и «Наличие специальной лексики». В перспективе возможно выделение слов, словоформ или других специфических признаков, характерных лишь для одной возрастной группы, и их использование в качестве маркера.

При поиске ненормативной и специальной лексики использовались словари, составленные вручную на основании нескольких открытых онлайн-источников.

После проведения анализа текстов оценивается наличие значений, однозначно указывающих на категорию адресатов. На основании полученных представлений текстов (моделей bag-of-words и наборов предложений) были оценены значения количественных признаков двух выборок. Для приведения значений к единому диапазону было дополнительно выполнено нормирование. Нормирование проводилось по формуле [20]:

$$x_{ij} = \frac{y_{ij} - \overline{y}_j}{\sigma_j}, \sigma_j \neq 0, \quad (1)$$

где x_{ij} — нормированное значение признака, y_{ij}, y_{ij} — значение признака до нормирования, $\overline{y}_j$ — среднее значение признака, измеренного для всех объектов выборки, σ_j — среднеквадратическое отклонение значений признака; i — количество текстов выборки, j — количество признаков.

Исходные значения количественных признаков до проведения нормирования и их среднеквадратические отклонения (σ) приведены в таблице 6. Веса признаков определяются значениями их информативности, полученными при помощи метода накопленных частот.

Таблица 6. Значения количественных признаков

Признак	Возраст читателей		Среднеквадратическое отклонение	Весовой коэффициент
	взрослый	детский		
Средняя длина слов текста (кроме стоп-слов)	8,35	6,11	0,8	0,231
Среднее количество слов в предложении	11,41	6,2	3,5	0,189
Количество многосложных слов в тексте (состоящих из более чем трех слогов) (%)	22,95	13,91	7,53	0,181
Количество особых глагольных форм в тексте (%)	3	2,09	0,59	0,084
Среднее количество грамматических основ в предложении	2,47	1,81	0,7	0,079
Количество числительных в тексте (%)	3,1	2,59	0,61	0,076
Доля простых предложений с двумя главными членами (относительно простых предложений. %)	64,5	67,3	4,22	0,048
Доля служебных слов (%)	27,85	23,03	5,69	0,047
Количество глаголов в тексте (%)	20,18	21,79	1,71	0,044
Количество прилагательных в тексте (%)	11,24	10,83	1,31	0,031

6. Эксперимент. В результате проведения вычислительного эксперимента получены значения, характеризующие точность классификации, выполненной с помощью предложенного подхода.

Обучение классификаторов проводилось на 70% текстов имеющихся выборок, тестирование проводилось на оставшихся 30%.

После проведения n разбиений исходной выборки на обучающую и контрольную ($n=5$) были вычислены средние значения по всем разбиениям. Сравнение показало, что допустимой границей отсечения часто встречающихся слов является $m=70$.

На первом шаге классификации было оценено наличие значений признаков, однозначно определяющих, к какой из представленных категорий относится текст (признаков-маркеров). Далее, если такие значения были найдены, тексту назначалась категория и он считался классифицированным. В том случае, когда категория не была определена при помощи признаков-маркеров, проводилась классификация текста на основании интервальных признаков, для осуществления которой вычислялись расстояния между вектором значений интервальных признаков и центрами масс категорий.

Расчет данных расстояний может быть произведен с помощью любой метрики, предназначенной для вычисления меры близости объектов, представленных в виде наборов классификационных признаков. Анализ данных показал, что в рамках данного эксперимента признаки объектов, между которыми устанавливается мера сходства, являются статистически зависимыми. При этом числовая оценка их значимости определяется весовыми коэффициентами. В этом случае в качестве меры близости текстов может быть принято расстояние Махаланобиса.

Тогда расстояние между текстом и центром масс категории R, представленным в виде вектора средневзвешенных значений признаков, определяется следующим образом:

$$\rho\left(F_{T_i}, R\right) = \sqrt{(F_{T_i} - R)^T \Lambda^T C^{-1} (F_{T_i} - R)}, \qquad (2)$$

где Λ — матрица весовых коэффициентов; C — матрица ковариации; R — вектор, характеризующий расположение центра масс категорий.

Вектор, характеризующий расположение центра масс категорий, может быть:

1) либо представлен в виде вектора средневзвешенных значений признаков:

$$R = \frac{\sum_{j=1}^{M} k_j F_{T_j}}{M}, \qquad (3)$$

где M — число текстов данной категории, входящих в обучающую выборку, $1 \le M \le L$; k_j — весовой коэффициент доверия тексту обучающей выборки, $k_j > 0$, $\sum_{j=1}^{L} k_j = 1$;

2) либо рассчитан с учетом порогового значения весового коэффициента, которое обозначено k_{min} :

$$R_{(L)} = \begin{cases} \dfrac{(L-1)R_{(L-1)} + k_L F_{T_L}}{L}, k_L \geq k_{min}; \\ R_{(L-1)}, k_L < k_{min} \end{cases} \qquad (4)$$
$$R_{(1)} = k_1 F_{T_1}.$$

Пороговое значение k_{min} может быть определено на основании мнений экспертов или экспериментально. В случае, когда коэффициент доверия тексту обучающей выборки не превышает порогового значения ($k_L < k_{min}$), корректировка значений, составляющих вектор R , не будет производиться. Следовательно, тексты контрольной выборки имеют коэффициент доверия $k_L < k_{min}$. В свою очередь коэффициенты k_L текстов обучающей выборки превышают пороговое значение k_{min} . Таким образом, на основании множества обучающих текстов может вычисляться значения, характеризующие центр масс категорий.

7. Результаты эксперимента. Результат классификации характеризуются количественной характеристикой точности классификации [21], то есть отношением числа верно классифицированных документов к общему количеству документов, отнесенных к данной категории:

$$Percision = \frac{TP}{TP + FP} * 100\%, \qquad (5)$$

где TP — количество верно классифицированных документов, FP – количество документов, ошибочно отнесенных к категории.

Объем контрольной выборки составил 313 текстов, объем обучающей — 729 текстов (то есть 30% и 70% всех имеющихся текстов соответственно). Данное соотношение выбрано на основании рекомендаций, представленных в работах [22-23]. Разбиение множества всех текстов на обучающую и контрольную выборки проводилось n раз. В качестве точности классификации указаны средние значения по всем разбиениям.

Точность классификации для подхода, представленного в данной работе, составила 74,16% (среднеквадратическое отклонение — 5,88%). При этом точность классификации составила 73,33% на вы-

борке детских текстов и 75,19% — на выборке взрослых текстов (таблица 7).

Таблица 7. Результаты классификации

Выборка текстов	Точность классификации (%)	Среднеквадратическое отклонение (%)
Взрослые тексты	73,33	5,98
Детские тексты	75,19	5,62
Обе выборки (взрослые и детские тексты)	74,16	5,88

В используемых для эксперимента выборках присутствовали тексты трех функциональных стилей: научного, публицистического и художественного. Значения точности классификации для подвыборок соответствующих стилей приведены в таблице 8.

Таблица 8. Точность классификации для текстов, распределенных по функциональным стилям

Функциональный стиль	Количество детских текстов в подвыборке (%)	Точность классификации (%)	Среднеквадратическое отклонение (%)
Научный	48,2	65,11	4,67
Публицистический	17,8	80,21	5,85
Художественный	64,6	78,45	5,34

Как видно из таблицы 8, наименьшую точность классификации — 65,11% — метод демонстрирует на подмножестве научных текстов. Это можно объяснить тем, что как взрослые, так и детские научные тексты (в частности учебно-научная и научно-популярная литература) имеют ряд черт, характерных для данного функционального стиля [24]:

1) использование специальной лексики и специфических языковых средств;

2) соответствие нормам русского литературного языка;

3) высокая частотность причастных и деепричастных оборотов, сложных синтаксических конструкций.

Значимым результатом эксперимента является величина ошибок, полученных для использованных в работе методов. Если считать целью классификации фильтрацию текстов, не предназначенных детской возрастной аудитории (отсечение текстов, адресованных взрослым читателям), то в данном эксперименте можно рассмотреть ошибки двух типов:

1) ошибка первого рода (доля случаев, когда текст, адресованный детской возрастной группе, не был отнесен к категории детских текстов);

2) ошибка второго рода (доля случаев, когда текст, адресованный взрослой возрастной группе, был отнесен к категории детских текстов).

Величина ошибки первого рода при проведении классификации всей имеющейся тестовой выборки составила 26,67%, ошибки второго рода – 24,81%.

Значения ошибок первого и второго рода для различных выборок текстов приведены в таблице 9.

Таблица 9. Значения ошибок первого и второго рода

Цель классификации	Ошибка первого рода (%)	Ошибка второго рода (%)
Отсечение взрослых текстов (на всей выборке)	26,67	24,81
Отсечение взрослых текстов (художественные тексты)	21,73	21,15
Отсечение взрослых текстов (публицистические тексты)	19,33	18,88
Отсечение взрослых текстов (подвыборка новостных текстов, жанр — «заметка»)	19,14	18,85

Для решения практической задачи проведения возрастной классификации информационной продукции (отсечения текстов, адресованных взрослым читателям) особую важность имеет минимизация ошибок второго рода.

8. Заключение. В работе представлен подход к проведению классификации текстов на основании возрастных категорий их адресатов. Автором предложен набор классификационных признаков, которые могут послужить основой такой классификации. Приведены результаты эксперимента, иллюстрирующие точность классификации, проведенной с помощью рассмотренного подхода. В сравнении с зарубежными работами, посвященными установлению характеристик адресата текста и выполненными с использованием англоязычных корпусов, продемонстрированы близкие значения точности классификации.

Эксперимент проводился на текстах двух возрастных категорий читателей — взрослой и детской, что обусловлено разметкой корпуса, предоставленного для исследования. В дальнейшем возможна апробация подхода на большем количестве классов текстов. Для этого необходимо собрать достаточное количество текстов, адресованных различным категориям читателей. Например, тексты могут быть разделе-

ны на категории «детские», «подростковые» «взрослые» или «0+», «6+», «12+», «16+», «18+».

Важным преимуществом предложенного в данной работе подхода является отсутствие необходимости приводить к единому диапазону значения разнотипных признаков (требуется лишь проведение нормирования значений интервальных признаков в случае, если категория не определяется признаками-маркерами). Кроме того, данный подход позволяет на шаге 2 исключить из рассмотрения ряд категорий, что дает возможность сократить временные затраты на проведение классификации текста на основании интервальных признаков.

К сложностям рассматриваемого подхода относится необходимость проведения предварительного экспертного анализа признаков с целью выявления признаков-маркеров. Разработка путей преодоления указанной трудности является одним из перспективных направлений данной работы. Также в дальнейшем планируется провести сравнение временных затрат предложенного подхода с подходами, использующие разнотипные признаки и требующие их приведения к единому диапазону.

Представленный подход реализован в рамках создания программного комплекса автоматической классификации текстов [25], о чем имеется соответствующее свидетельство о регистрации программы для ЭВМ [26].

Литература

1. *Усталов Д.А., Гольдштейн М.Л.* Распределенная инструментальная среда словарного морфологического анализа для обработки русского языка // Вестник ЮФУ. Математическое моделирование и программирование. 2012. №27. С. 119–127.
2. *Рубцова Ю.В.* Разработка и исследование предметно независимого классификатора текстов по тональности // Труды СПИИРАН. 2014. №5. С. 59–77.
3. *Тутубалина Е.В.* Совместная вероятностная тематическая модель для идентификации проблемных высказываний, связанных нарушением функциональности продуктов // Труды ИСП РАН. 2015. №4. С. 111–128.
4. *Астраханцев Н.А., Федоренко Д.Г., Турдаков Д.Ю.* Методы автоматического извлечения терминов из коллекции текстов предметной области // Программирование. 2015. №6. С. 33–52.
5. *Карпович С.Н.* Многозначная классификация текстовых документов с использованием вероятностного тематического моделирования ml-PLSI // Труды СПИИРАН. 2016. №4. С. 92–104.
6. *Шумская А.О.* Метод определения искусственных текстов на основе расчета меры принадлежности к инвариантам // Труды СПИИРАН. 2016. №6. С. 104–121.
7. *Shriberg E., Stolcke A., Ravuri S.V.* Addressee detection for dialog systems using temporal and spectral dimensions of speaking style // Proceedings of Interspeech. 2013. pp. 2559–2563.
8. *Ravuri S.V., Stolcke A.* Neural Network Models for Lexical Addressee Detection // Proceedings of Interspeech. 2014. pp. 298–302.

9. *Jovanovic N., op den Akker H.* Towards automatic addressee identification in multi-party dialogues // Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue. 2004. pp. 89–92.
10. *Lee H., Stolcke A., Shriberg E.* Using out-of-domain data for lexical addressee detection in human-human-computer dialog // Proceedings of North American ACL Human Language Technology Conference. 2013. pp. 215–219.
11. *Santosh K., Shekhar M., Varma V.* Author Profiling: Predicting Age and Gender from Blogs // CLEF 2013 evaluation labs and workshop. 2013. pp. 23–26.
12. *Pentel A.* Effect of different feature types on age based classification of short texts // 6th International Conference on Information, Intelligence, Systems and Applications (IISA). 2015. pp. 1–7.
13. *Pentel A.* Automatic Age Detection Using Text Readability Features // CEUR Workshop Proceedings. 2015. pp. 40–45.
14. *Lorge I.* Predicting readability // Teachers college record. 1944. №45. pp. 404–419.
15. Федеральный закон Российской Федерации № 436-ФЗ «О защите детей от информации, причиняющей вред их здоровью и развитию». 2010. URL: http://www.consultant.ru/document/cons_doc_LAW_108808 (дата обращения: 11.02.2016).
16. «База данных метатекстовой разметки Национального корпуса русского языка» (коллекция детской литературы)». 2014.
17. Национальный корпус русского языка. URL: http:// ruscorpora.ru (дата обращения 26.01.2016).
18. *Ахапкина Я.Э. и др.* Проблемы функциональной грамматики. Принцип естественной классификации // М.: Языки славянской культуры. 2013. 507 с.
19. *Jurafsky D., Martin J.H.* Speech and Language Processing (2nd Edition) // Upper Saddle River. New Jersey: Prentice Hall. 2009. 975 p.
20. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: классификация и снижение размерности // М.: Финансы и статистика. 1989. 607 с.
21. *Шокин Ю.И., Федотов А.М., Барахнин В.Б.* Проблемы поиска информации // Новосибирск: Наука. 2010. 220 с.
22. *Кафтанников И.Л., Парасич А.В.* Проблемы формирования обучающей выборки в задачах машинного обучения // Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. 2016. Т. 16. №3. С. 15–24.
23. *Canavet O., Fleuret F.* Efficient sample mining for object detection // Proceedings of the Asian Conference on Machine Learning (ACML). 2014. pp. 48–63.
24. *Введенская Л.А., Кашаева Е.Ю., Павлова Л.Г.* Русский язык и культура речи. Учебное пособие для вузов для бакалавров и магистрантов / Под ред. С.А. Осташова // М.: Феникс. 2016. 539 с.
25. *Глазкова А.В.* Интеллектуальная система автоматического определения категории потенциальных адресатов текста // Программные продукты и системы. 2016. №3. С. 85–89.
26. *Глазкова А.В.* Расчёт оценки степени близости категорий текстов при решении задач классификации электронных документов. Свидетельство о регистрации ПрЭВМ №6164622015. 2015.

Глазкова Анна Валерьевна — кандидат технических наук, ассистент кафедры программного обеспечения института математики и компьютерных наук, Тюменский государственный университет (ТюмГУ). Область научных интересов: обработка естественного языка, машинное обучение, нейронные сети, классификация текстов, прикладная лингвистика. Число научных публикаций — 18. anna_glazkova@yahoo.com; ул. Перекопская, 15а, каб. 210, Тюмень, 625003; р.т.: +79091826371, Факс: +7(345)2640135.

DOI 10.15622/sp.52.3

A.V. GLAZKOVA

AN APPROACH TO TEXT CLASSIFICATION BASED ON AGE GROUPS OF ADDRESSEES

Glazkova A.V. **An Approach to Text Classification based on Age Groups of Addressees.**

Abstract. The article deals with a new approach to text classification considering the existence of different types of classification features (binary, nominal, ordinal and interval).

The specialty of the approach is a phased classification process, which makes it possible to not cause different types of classification features to a single range. The author describes a computational experiment using texts included in Russian National Corpus and suggests the set of classification features for Russian text classification based on the age of theirs supposed readers. Text documents included in the sample are divided into two categories – for adults and for children, — according to the views of experts.

Keywords: information extraction, text classification, natural language processing, text attributes.

Glazkova Anna Valer'evna — Ph.D., assistant of software department of Institute of mathematics and computer science, Tyumen State University (TSU). Research interests: natural language processing, machine learning, neural networks, text classification, applied linguistics. The number of publications — 18. anna_glazkova@yahoo.com; 15a, Perekopskaya street, Tyumen, 625003, Russia; office phone: +79091826371, Fax: +7(345)2640135.

References

1. Ustalov D.A., Gol'dshteyn M.L. [A Distributed Dictionary - Based Morphological Analysis Framework for Russian Language Processing]. *Vestnik YuFU. Matematicheskoe modelirovanie i programmirovanie – Bulletin of the South Ural State University, Series "Mathematical Modelling, Programming & Computer Software"*. 2012. vol. 27. pp. 119–127. (In Russ.).
2. Rubtsova Yu.V. [Research and Development of Domain Independent Sentiment Classifier]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2014. vol. 5. pp. 59–77. (In Russ.).
3. Tutubalina E.V. [Sentiment-based Topic Model for Mining Usability Issues and Failures with User Products]. *Trudy ISP RAN – The Proceedings of the ISP RAS*. 2015. vol. 4. pp. 111–128. (In Russ.).
4. Astrakhantsev N.A., Fedorenko D.G., Turdakov D.Yu. [Methods for automatically extracting terms from a collection of domain texts]. *Programmirovanie – Programming.* 2015. vol. 6. pp. 33–52. (In Russ.).
5. Karpovich S.N. [Multi-Label Classification of Text Documents using Probabilistic Topic Model ml-PLSI]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2016. vol. 4. pp. 92–104. (In Russ.).
6. Shumskaya A.O. [Method of the artificial texts identification basen on the calculation of the belonging measure to the invariants]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2016. vol. 6. pp. 104–121. (In Russ.).
7. Shriberg E., Stolcke A., Ravuri S.V. Addressee detection for dialog systems using temporal and spectral dimensions of speaking style. Proceedings of Interspeech. 2013. pp. 2559–2563.
8. Ravuri S.V., Stolcke A. Neural Network Models for Lexical Addressee Detection. Proceedings of Interspeech. 2014. pp. 298–302.

9. Jovanovic N., op den Akker H. Towards automatic addressee identification in multi-party dialogues. Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue. 2004. pp. 89–92.
10. Lee H., Stolcke A., Shriberg E. Using out-of-domain data for lexical addressee detection in human-human-computer dialog. Proceedings of North American ACL Human Language Technology Conference. 2013. pp. 215–219.
11. Santosh K., Shekhar M., Varma V. Author Profiling: Predicting Age and Gender from Blogs. CLEF 2013 evaluation labs and workshop. 2013. pp. 23–26.
12. Pentel A. Effect of different feature types on age based classification of short texts. 6th International Conference on Information, Intelligence, Systems and Applications (IISA). 2015. pp. 1–7.
13. Pentel A. Automatic Age Detection Using Text Readability Features. CEUR Workshop Proceedings. 2015. pp. 40–45.
14. Lorge I. Predicting readability. Teachers college record. 1944. vol. 45. pp. 404–419.
15. Federal'nyj zakon Rossijskoj Federacii № 436-FZ «O zashhite detej ot informacii, prichinjajushhej vred ih zdorov'ju i razvitiju» [The Federal Law of the Russian Federation № 436-FZ «On protection of children from information harmful to their health and development»]. 2010. Available at: http://www.consultant.ru/document/cons_doc_LAW_108808 (accessed: 11.02.2016). (In Russ.).
16. «Baza dannyh metatekstovoj razmetki Nacional'nogo korpusa russkogo jazyka» (kollekcija detskoj literatury)» [«The Russian National Corpus database of metatext markup" (a collection of children's literature)»]. 2014. (In Russ.).
17. Nacional'nyj korpus russkogo jazyka [The Russian National Corpus]. Available at: http:// ruscorpora.ru (accessed 26.01.2016). (In Russ.).
18. Akhapkina Ya.E. et al. *Problemy funkcional'noj grammatiki. Princip estestvennoj klassifikacii* [Problems of functional grammar. The principle of natural classification]. M.: Jazyki slavjanskoj kul'tury. 2013. 507 p. (In Russ.).
19. Jurafsky D., Martin J.H. Speech and Language Processing (2nd Edition). Upper Saddle River. New Jersey: Prentice Hall. 2009. 975 p.
20. Aivazyan S.A., Buchstaber V.M., Enukov I.S., Meshalkin L.D. *Prikladnaja statistika: klassifikacija i snizhenie razmernosti* [Applied Statistics: Classification and Dimension Reduction]. M.: Finansy i statistika. 1989. 607 p. (In Russ.).
21. Shokin Yu.I., Fedotov A.M., Barakhnin V.B. *Problemy informacionnogo poiska* [Problems of information retrieval]. Novosibirsk: Nauka. 2010. 220 p. (In Russ.).
22. Kaftannikov I.L., Parasich A.V. Problems of training set's formation in machine learning tasks. *Vestnik juzhno-ural''skogo gosudarstvennogo universiteta. Serija: komp'juternye tehnologii, upravlenie, radiojelektronika – Bulletin of the South Ural State University. Series: Computer technologies, automatic control & radioelectronics*. 2016. vol. 16. no. 3. pp. 15–24. (In Russ.).
23. Canavet O., Fleuret F. Efficient sample mining for object detection. Proceedings of the Asian Conference on Machine Learning (ACML). 2014. pp. 48–63.
24. Vvedenskaja L.A., Kashaeva E.Ju., Pavlova L.G. *Russkij jazyk i kul'tura rechi. Uchebnoe posobie dlja vuzov dlja bakalavrov i magistrantov* [Russian language and culture of speech. Textbook for high schools for undergraduate and graduate students]. M.: Feniks. 2016. 539 p. (In Russ.).
25. Glazkova A.V. [Intelligent system for automatic identification of text addressee category]. *Programmnye produkty i sistemy – Software & Systems*. 2016. vol. 3. pp. 85–89. (In Russ.).
26. Glazkova A.V. [Calculation of assessing proximity word categories in solving problems of classification of electronic documents]. Certificate of state registration of computer programs no. 6164622015. 2015. (In Russ.).