MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 31, NO. 3, 2024
journal homepage: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Methods of Implicit Aspect Detection in Russian Publicism Sentences

A. Y. Poletaev[1], I. V. Paramonov[1], E. M. Kolupaev[1]

DOI: 10.18255/1818-1015-2024-3-226-239

[1]P.G. Demidov Yaroslavl State University, Yaroslavl, Russia

MSC2020: 68T50
Research article
Full text in Russian

Received July 1, 2024
Revised July 25, 2024
Accepted July 31, 2024

The paper compares performance of various methods of automatic implicit aspect detection in publicism sentences in Russian. The task of implicit aspect detection is an auxiliary task in the aspect-oriented sentiment analysis. The experiments were conducted on a corpus of sentences extracted from political campaign materials. The best results, with F1-measure reaching 0.84, were obtained using the Navec embeddings and classifiers based on the support vector machine method. Fairly high results, with F1-measure reaching 0.77, were obtained using the bag-of-words model and the naive Bayesian classifier. Other methods showed lower performance. It was also revealed during the experiments that the detection quality can differ significantly between the aspects. The detection quality is the highest for the aspects associated with characteristic marker words, for example, “health car” and “holding elections”. More general aspects, such as “quality of governance”, are detected with the worst quality.

Keywords: aspect detection; implicit aspects; sentiment analysis; publicism

INFORMATION ABOUT THE AUTHORS

Poletaev, Anatoliy Y. (corresponding author)	ORCID iD: 0000-0003-0116-4739. E-mail: anatoliy-poletaev@mail.ru Post-graduate student
Paramonov, Ilya V.	ORCID iD: 0000-0003-3984-8423. E-mail: ilya.paramonov@fruct.org PhD, Associate professor
Kolupaev, Egor M.	ORCID iD: 0009-0006-4312-2413. E-mail: kolupaew.eg@yandex.ru Student

Funding: Russian Science Foundation (Project no. 23-21-00495).

For citation: A. Y. Poletaev, I. V. Paramonov, and E. M. Kolupaev, “Methods of implicit aspect detection in Russian publicism sentences”, *Modeling and Analysis of Information Systems*, vol. 31, no. 3, pp. 226–239, 2024. DOI: 10.18255/1818-1015-2024-3-226-239.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 31, № 3, 2024
сайт журнала: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Методы определения неявно упоминаемых аспектов в публицистических предложениях на русском языке

А. Ю. Полетаев[1], И. В. Парамонов[1], Е. М. Колупаев[1]

DOI: 10.18255/1818-1015-2024-3-226-239

[1]Ярославский государственный университет им. П.Г. Демидова, Ярославль, Россия

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 1 июля 2024 г.
После доработки 25 июля 2024 г.
Принята к публикации 31 июля 2024 г.

В работе сравнивается качество работы различных методов определения неявно упоминаемых аспектов социально-экономической жизни в публицистических предложениях на русском языке. Задача определения неявно упоминаемых аспектов является вспомогательной для задач аспектно-ориентированного анализа тональности. Эксперименты проводились на корпусе предложений, извлечённых из политической агитации. Лучшие результаты, с F1-мерой, достигающей 0.84, были получены с использованием эмбеддингов Navec и классификаторов, основанных на методе опорных векторов. Достаточно высокие результаты, с F1-мерой до 0.77, были получены при использовании модели «мешок слов» и наивного байесовского классификатора. Остальные методы показали более низкие результаты. Также в ходе экспериментов было выявлено, что качество определения различных аспектов может достаточно сильно отличаться. Лучше всего определяются аспекты, с которыми в речи связаны характерные слова-маркеры, например, «здравоохранение» и «проведение выборов» Хуже всего определяются упоминания достаточно общих аспектов, таких как «качество управления».

Ключевые слова: определение аспектов; неявные аспекты; анализ тональности; публицистический стиль

ИНФОРМАЦИЯ ОБ АВТОРАХ

Полетаев, Анатолий Юрьевич (автор для корреспонденции)	ORCID iD: 0000-0003-0116-4739. E-mail: anatoliy-poletaev@mail.ru Аспирант
Парамонов, Илья Вячеславович	ORCID iD: 0000-0003-3984-8423. E-mail: ilya.paramonov@fruct.org Канд. физ.-мат. наук, доцент
Колупаев, Егор Михайлович	ORCID iD: 0009-0006-4312-2413. E-mail: kolupaew.eg@yandex.ru Студент

Финансирование: Российский научный фонд (проект № 23-21-00495).

Для цитирования: A. Y. Poletaev, I. V. Paramonov, and E. M. Kolupaev, “Methods of implicit aspect detection in Russian publicism sentences”, *Modeling and Analysis of Information Systems*, vol. 31, no. 3, pp. 226–239, 2024. DOI: 10.18255/1818-1015-2024-3-226-239.

Введение

Анализ тональности — направление компьютерной лингвистики, изучающее автоматическое определение выраженного в тексте авторского отношения [1]. В зависимости от поставленной задачи анализ может производиться на разных уровнях, например, на уровне текста, абзаца или отдельного предложения. Кроме того, может выполняться как анализ общей тональности (определение отношения к теме текста в целом), так и аспектно-ориентированный анализ (определение отношения к конкретным аспектам темы текста) [2].

Поскольку в отдельно взятом предложении обычно затрагиваются не все аспекты обсуждаемой темы, а только часть из них, при аспектно-ориентированном анализе тональности возникает вспомогательная задача: по данному предложению определить, какие из аспектов в нём упоминаются [3]. Эта задача может решаться как для явных упоминаний аспектов (explicit aspects), так и для всех упоминаний, включая неявные (implicit aspects). Если в предложении можно выделить конкретную группу слов, называющих аспект, то аспект упоминается явно. Иначе речь идёт о неявном упоминании. Например, в предложении «Кандидат Михаил Петров пообещал жителям Приволжска, что к их городу будет построено современное шоссе» явно упоминаются аспекты социально-экономической жизни «Михаил Петров» и «Приволжск» (следуя [4], авторы данной статьи рассматривают все именованные сущности как аспекты) и неявно упоминаются аспекты «проведение выборов» и «состояние дорожной сети».

Данная работа посвящена сравнению методов определения неявно упоминаемых аспектов в публицистических предложениях на русском языке и выявлению наилучшего качества, которого можно достичь с их помощью. Более формально: исследовались методы решения следующей задачи: дано предложение S и набор аспектов $A_1, \ldots, A_n$; требуется для каждого аспекта A_i определить, упоминается ли он в предложении S. Эта задача может быть переформулирована как n задач бинарной классификации: дано предложение S и аспект A_i; требуется определить, упоминается ли A_i в S или нет. В рамках исследования рассматривались публицистические тексты, посвящённые социально-экономической жизни некоторого региона, при этом список рассматриваемых аспектов был составлен вручную на основе собранного корпуса. Необходимо отметить, что задача выделения аспектов (aspect extraction), т. е. автоматического составления списка всех аспектов, упоминаемых в тексте, выходит за рамки данной работы.

Важная особенность данной работы — проведение экспериментов на корпусе предложений, извлечённых из материалов предвыборной агитации кандидатов на различные должности государственного и муниципального управления. Поскольку в подобных материалах упоминаются многие аспекты социально-экономической жизни [5], это даёт возможность сравнить сложность определения упоминаний различных аспектов.

1. Обзор существующих методов

Задача определения неявно упоминаемых аспектов в тексте (иногда называемая задачей определения категорий аспектов — aspect category extraction, в противовес задаче определения терминов аспектов — aspect term extraction, которые упоминаются в тексте явно [2]) является не самой распространённой решаемой задачей в области анализа тональности, однако для английского языка существует множество устоявшихся методов её решения. Основными из них являются: методы, основанные на правилах; методы тематического моделирования; методы обучения без учителя или с частичным обучением, в т. ч. основанные на кластеризации и совместной частоте встречаемости; методы машинного обучения и нейросетевые классификаторы. Рассмотрим их кратко, а более полную информацию можно найти в обзорах [3, 6, 7].

Идея методов, основанных на правилах, заключается в определении некоторого набора шаблонов для фрагментов текста, позволяющих фиксировать появление аспектов. Лучше всего такой под-

ход работает для явно упоминаемых аспектов либо в предметных областях со стандартизованной структурой предложений. В целом для неявно упоминаемых аспектов данный подход не является популярным ввиду сложности реализации.

Также для выделения неявно упоминаемых аспектов могут использоваться методы кластеризации, однако их применение может быть затруднено ввиду сложности интерпретации аспектов, соответствующих построенным кластерам. Когда выделяемые аспекты известны заранее, может быть полезным подход с частичным обучением, когда используется некоторое количество слов, описывающих аспекты, в качестве «затравочных», а затем используются техники, основанные на использовании совместной частоты встречаемости слов в корпусе.

Методы тематического моделирования — это вероятностные методы, нацеленные на идентификацию слов, характерных предложениям с конкретными неявно упоминаемыми аспектами. Их эффективность для рассматриваемой задачи обусловлена близостью последней к задаче тематического моделирования.

Наиболее обширный класс методов включает в себя методы машинного обучения (в т. ч. глубокого), а также нейросетевые классификаторы. Как правило, данные методы дают неплохие результаты для типичных предметных областей аспектно-ориентированного анализа тональности, таких как твиты и отзывы на различные продукты. Типичные значения F-меры в таких задачах обычно варьируются в диапазоне 70–85 %.

Некоторым исследователям удаётся добиться очень высоких результатов за счёт построения гибридных классификаторов, использующих различные методы. Например, в работе [8] авторам удалось достичь F-меры около 96 % на открытом наборе данных соревнования SemEval 2015 ABSA Dataset.

К сожалению, для русского языка существуют лишь единичные работы, посвящённые задаче определения неявно упоминаемых аспектов для анализа тональности.

В частности, в работе [9] предложен алгоритм определения категорий аспектов (по существу неявно упоминаемых аспектов) на основе построения семантического графа для заданной предметной области с последующим использованием этого графа при кластеризации анализируемых текстов. На корпусе отзывов о ресторанах соревнования SemEval-2016 (Task 5) точность определения категорий «ресторан», «обслуживание», «интерьер», «еда» с применением разработанного алгоритма составила 65–72 %.

В работе [10] рассматривается нейросетевая модель, предназначенная для одновременного определения аспектов в предложениях, фрагментов, к ним относящимся, и информации о тональности. Используются абстрактные конечные автоматы для извлечения сущностей и рекуррентные нейронные сети для их классификации. Для русского языка модель тестировалась на отзывах на товары AliExpress и позволила получить F-меру, равную 72 %.

Ввиду крайней недостаточности исследований, посвящённых русскому языку, представляется актуальным проведение исследований практически с любыми моделями и методами для решения задачи выделения неявно упоминаемых аспектов. Авторы настоящей статьи остановились на применении наиболее популярных для англоязычных текстов традиционных классификаторов в совокупности с доступными для русского языка эмбеддингами.

2. Корпуса текстов

Перед началом экспериментов был собран корпус предложений из материалов предвыборной агитации на русском языке, и для каждого предложения этого корпуса были указаны упоминаемые аспекты социально-экономической жизни. Список аспектов был составлен исходя из двух принципов. Во-первых, аспект должен упоминаться в достаточно большом числе предложений корпуса, иначе будет сложно оценить качество определения упоминаний этого аспекта. Во-вторых, включаемые в список аспекты не должны быть слишком общими, т. е. должны характеризовать какую-то

Table 1. The number of sentences containing a particular aspect

Таблица 1. Количество предложений, в которых упоминаются аспекты

Аспект	Первоначальный корпус	Расширенный корпус
развитие промышленности	317	618
качество и стоимость услуг ЖКХ	401	611
проведение выборов	279	581
качество управления	308	459
качество здравоохранения	215	365
состояние жилья	198	297
развитие сельского хозяйства	204	294
состояние дорожной сети	214	286
газификация	167	275
строительство атомной электростанции	141	222

конкретную сторону социально-экономической жизни. Например, в предвыборной агитации часто говорится об обществе в целом («работаю на пользу общества», «общество меня понимает»), но если выделить «общество» как аспект, то придётся считать, что он неявно упоминается практически во всех предложениях политической агитации.

В результате в список попали следующие аспекты: развитие промышленности, качество и стоимость услуг ЖКХ, проведение выборов, качество управления, качество здравоохранения, состояние жилья, развитие сельского хозяйства, состояние дорожной сети, газификация, строительство атомной электростанции.

Для первого этапа экспериментов был собран и размечен первоначальный корпус из 2050 случайно выбранных из предвыборной агитации предложений. Среди них оказались как предложения, в которых упоминается более чем один аспект, так и предложения, в которых не упоминается ни одного аспекта из списка. Подробная информация о том, в каком числе предложений упоминался тот или иной аспект, приведена в таблице 1. Как можно видеть, в первоначальном корпусе каждый аспект упоминается как минимум в 141 предложении, а большая часть аспектов — как минимум в 200 предложениях.

Для второго этапа экспериментов и изучения того, как размер корпуса влияет на качество определения упоминаний аспектов, первоначальный корпус был расширен до 3400 предложений. Новые предложения подбирались таким образом, чтобы каждый аспект упоминался хотя бы в 200 предложениях. Подробная информация о том, в каком количестве предложений расширенного корпуса упоминается тот или иной аспект, также приведена в таблице 1.

3. Постановка экспериментов

В этом разделе описаны методы, применённые авторами на различных этапах решения задачи определения неявно упоминаемых в предложении аспектов социально-экономической жизни.

3.1. Предобработка предложений

На этапе предобработки предложения преобразуются так, чтобы метод классификации мог более эффективно их обработать. В данной работе использовались следующие техники предобработки:

- токенизация — преобразование строки предложения к списку токенов — его слов;
- лемматизация — замена каждого токена на начальную форму соответствующего слова;
- стемминг — замена каждого токена на грамматическую основу соответствующего слова;
- удаление из предложения всех токенов короче трёх символов.

Токенизация необходима для работы практически всех методов автоматической обработки текста, поэтому она проводилась всегда.

Цель лемматизации и стемминга — абстрагироваться от конкретных словоформ и облегчить для классификатора выделение связей терминов с аспектами [11]. Лемматизация позволяет сохранять больше информации о связях между словами, чем стемминг, что может повлиять на качество классификации. При проведении экспериментов могла применяться либо лемматизация, либо стемминг, либо ни одна из этих техник. Для лемматизации использовался морфологический анализатор из состава библиотеки Natasha, а для стемминга — Snowball Stemmer из состава библиотеки NLTK [12].

Цель удаления из предложений слов короче трёх символов — избавиться от служебных слов и коротких местоимений, а также чисел, которые могут создавать «шум» во входных данных и затруднять классификацию [13]. Эта техника применялась или не применялась вне зависимости от того, использовались ли лемматизация и стемминг.

3.2. Векторизация предложений

На этапе векторизации для каждого предложения строилось его векторное представление. Для этого использовались два метода с использованием эмбеддингов.

Первый метод — получение векторов-эмбеддингов для каждого слова предложения с последующим усреднением этих векторов [14]. Использовались эмбеддинги: Word2Vec [15, 16] длины 100, предобученный на русскоязычной Википедии и дообученный на предложениях из политической агитации; FastText [17] длины 512, обученный на предложениях из политической агитации; Navec (основанный на GloVe) [18, 19] длины 300 из состава проекта Natasha, причём использовались два варианта: обученный на текстах художественной литературы и обученный на новостных текстах.

В рамках второго метода для получения векторов предложений использовались эмбеддинги Doc2Vec [20], обученные на предложениях из политической агитации и 2750 случайно выбранных предложениях из СМИ.

3.3. Классификация предложений

На этом этапе обучались различные классификаторы для решения задачи бинарной классификации: дано предложение в виде списка токенов; требуется определить, упоминается ли в предложении определённый аспект из списка аспектов социально-экономической жизни. При этом использовалась пятикратная кросс-валидация.

Эксперименты проводились со следующими классификаторами:

- наивный байесовский классификатор (Naive Bayes Classifier);
- классификатор гребневой регрессии (Ridge Classifier);
- классификатор на основе логистической регрессии (Logistic Regression Classifier);
- классификатор на основе стохастического градиентного спуска (SGD Classifier);
- классификатор на основе метода опорных векторов с линейным ядром (Linear SVC);
- классификатор на основе метода опорных векторов с ядром на основе радиально-базисной функции (RBF SVC);
- дерево решений (Decision Tree Classifier);
- случайный лес (Random Forest Classifier).

Наивный байесовский классификатор использовал модель «мешка слов», а все остальные классификаторы — векторизацию предложений с помощью эмбеддингов.

Для случайного леса эксперименты проводились с максимальной глубиной от 3 до 7, а также без ограничений на максимальную глубину. Реализации классификаторов были взяты из библиотеки scikit-learn [21].

Table 2. Top 5 results with Naive Bayes with extended corpus

Таблица 2. 5 лучших результатов с Naive Bayes с расширенным корпусом

Классификатор	Число слов в «мешке слов»	F_1
Bernoulli	200	0.77
Multinomial	200	0.76
Bernoulli	300	0.75
Multinomial	300	0.74
Multinomial	400	0.73

Table 3. Best result with Naive Bayes with extended corpus

Таблица 3. Лучший результат с байесовским классификатором с расширенным корпусом

Аспект	Точность	Полнота	F_1
проведение выборов	0.77	0.83	0.80
качество управления	0.74	0.71	0.73
качество здравоохранения	0.83	0.71	0.77
строительство атомной электростанции	0.88	0.89	0.88
развитие промышленности	0.85	0.91	0.88
развитие сельского хозяйства	0.71	0.76	0.73
состояние жилья	0.64	0.75	0.69
качество и стоимость услуг ЖКХ	0.86	0.79	0.82
состояние дорожной сети	0.76	0.68	0.72
газификация	0.71	0.66	0.68

Для сравнения качества классификаторов использовалась усреднённая F_1-мера классификации по всем подвыборкам кросс-валидации и по всем аспектам. Также для каждого аспекта вычислялись усреднённые точность, полнота и F_1-мера. Это делалось для того, чтобы можно было оценить, упоминания каких аспектов выделяются лучше, а каких — хуже.

4. Результаты

4.1. Модель «мешка слов»

При экспериментах с наивным байесовским классификатором были опробованы несколько вариаций классификатора: классический байесовский классификатор (Gaussian Naive Bayes), классификатор для полиномиальных моделей (Multinomial Naive Bayes), классификатор с алгоритмом дополнения (Complement Naive Bayes) и классификатор для многомерных моделей (Bernoulli Naive Bayes). Также варьировалось число слов, формирующих «мешок слов» — от 200 до 2000 с шагом 100. Пять лучших результатов приведены в таблице 2. Все они были получены с использованием стемминга и без удаления служебных слов.

Метрики обнаружения упоминаний конкретных аспектов для наилучшего результата приведены в таблице 3. Как можно видеть, упоминания всех аспектов, кроме «состояния жилья» и «газификации», обнаруживаются с F_1-мерой не ниже 0.7, а для многих она даже превышает 0.8.

4.2. Эмбеддинги Word2Vec

При всех экспериментах с Word2Vec использовались эмбеддинги, предобученные на русскоязычной Википедии и дообученные на предложениях из политической агитации в течение 5, 10, 20 или 30 эпох. Пять лучших результатов приведены в таблице 4. Все они были получены при дообучении эмбеддингов в течение 10 эпох и без использования стемминга. Наилучший результат был получен без использования лемматизации, однако 4 других среди 5 лучших — с её использованием. Лучшие результаты показала модель с использованием лемматизации и 10 эпохами обучения.

Table 4. Top 5 results with Word2Vec embedding with extended corpus

Таблица 4. 5 лучших результатов с эмбеддингами Word2Vec с расширенным корпусом

Лемматизация	Удаление коротких токенов	Классификатор	F_1
Нет	Нет	Logistic Regression	0.66
Да	Да	RBF SVC	0.65
Да	Да	Linear SVC	0.61
Да	Да	Logistic Regression	0.61
Да	Да	Ridge Regression	0.60

Table 5. Best result with Word2Vec embedding with extended corpus

Таблица 5. Лучший результат с эмбеддингами Word2Vec с расширенным корпусом

Аспект	Точность	Полнота	F_1
проведение выборов	0.78	0.93	0.84
качество управления	0.47	0.8	0.59
качество здравоохранения	0.70	0.85	0.77
строительство атомной электростанции	0.41	0.78	0.54
развитие промышленности	0.66	0.88	0.76
развитие сельского хозяйства	0.56	0.80	0.66
состояние жилья	0.44	0.81	0.57
качество и стоимость услуг ЖКХ	0.51	0.80	0.62
состояние дорожной сети	0.59	0.83	0.69
газификация	0.47	0.83	0.60

С данной моделью наилучший результат дал классификатор, основанный на логистической регрессии. Подробные данные о метриках выделения отдельных аспектов представлены в таблице 5. В среднем результат оказался хуже, чем при использовании наивного байесовского классификатора. Тем не менее, стали гораздо лучше определяться упоминания аспекта «проведение выборов»: F_1-мера возросла на 0.20 и достигла 0.84.

4.3. Эмбеддинги Doc2Vec

В ходе экспериментов эмбеддинги Doc2Vec обучались на предложениях политической агитации и случайно выбранных 2750 предложениях из СМИ. В ходе экспериментов варьировались следующие параметры:

- длина эмбеддинга: 64, 128, 256 или 512;
- алгоритм обучения: распределённая память (distributed memory, DM) или распределённый мешок слов (distributed bag of words, DBOW);
- минимальное число вхождений слова в обучающий набор: 0, 5 или 15;
- число эпох обучения: 10, 30 или 50.

Пять лучших результатов, полученных с использованием Doc2Vec, приведены в таблице 6. Во всех случаях проводилось обучение по алгоритму распределённого мешка слов, с использованием лемматизации и без использования стемминга, только на словах, которые встречались в обучающем наборе хотя бы 5 раз, в течение 50 эпох. Средние результаты получились немного хуже, чем при использовании Word2Vec, при этом с теми же классификаторами использовались вектора эмбеддингов гораздо большей длины.

Подробные метрики для выделения различных аспектов у лучшего результата приведены в таблице 7. Как можно видеть, практически все они также немного хуже, чем при использовании Word2Vec, однако, Doc2Vec дал гораздо лучшее качество для определения упоминаний аспекта «строительство атомной электростанции».

Table 6. Top 5 results with Doc2Vec embedding with extended corpus

Таблица 6. 5 лучших результатов с эмбеддингами Doc2Vec с расширенным корпусом

Размерность вектора	Удаление коротких токенов	Классификатор	F_1
512	Нет	Linear SVC	0.65
256	Да	Linear SVC	0.63
256	Да	RBF SVC	0.61
512	Нет	RBF SVC	0.59
256	Нет	Logistic Regression	0.59

Table 7. Best result with Doc2Vec embedding with extended corpus

Таблица 7. Лучший результат с эмбеддингами Doc2Vec с расширенным корпусом

Аспект	Точность	Полнота	F_1
проведение выборов	0.76	0.86	0.81
качество управления	0.53	0.73	0.61
качество здравоохранения	0.51	0.77	0.61
строительство атомной электростанции	0.61	0.80	0.69
развитие промышленности	0.83	0.83	0.83
развитие сельского хозяйства	0.49	0.71	0.58
состояние жилья	0.46	0.73	0.56
качество и стоимость услуг ЖКХ	0.78	0.79	0.78
состояние дорожной сети	0.43	0.73	0.54
газификация	0.43	0.77	0.55

4.4. Эмбеддинги FastText

При проведении экспериментов была использована реализация алгоритма обучения эмбеддингов FastText из библиотеки gensim. Обучение производилось на корпусе предложений из политической агитации. Эксперименты проводились как на предложениях из первоначального корпуса, так и из расширенного. По пять лучших результатов для каждого из корпусов приведены в таблице 8. Все они были получены с использованием лемматизации, без использования стемминга и без удаления коротких токенов. Лучший результат на расширенном корпусе был получен с использованием классификатора на основе метода опорных векторов. Метрики качества определения упоминаний каждого из аспектов представлены в таблице 9. Как можно видеть, с увеличением объёма корпуса качество определения хорошо выделявшихся аспектов (например, «проведение выборов», «развитие промышленности») увеличилось достаточно слабо, в пределах 5 %, а качество определения упоминаний аспекта «строительство АЭС» даже снизилось. Для плохо определявшихся аспектов, например, «газификация» или «состояние жилья» рост был гораздо более существенным, вплоть до 16 %.

4.5. Эмбеддинги Navec

При проведении экспериментов были использованы две модели Navec. Первая модель hudlit-12B-500K-300d-100q, обученная на предложениях художественной литературы, содержит векторы для 500 тысяч слов. Вторая модель news-1B-250K-300d-100q обучена на текстах новостей и содержит векторы для 250 тысяч слов. Эксперименты проводились как на предложениях из первоначального корпуса, так и из расширенного. По пять лучших результатов на каждом из корпусов приведены в таблице 10. Все они были получены с использованием лемматизации, без использования стемминга и при удалении коротких токенов. Как можно видеть, увеличение количества предложений приводит к существенному росту качества определения упоминаний аспектов. При этом необхо-

Table 8. Top 5 results with FastText embedding with initial and extended corpora

Таблица 8. 5 лучших результатов с эмбеддингами FastText с изначальным и расширенным корпусами

Первоначальный корпус			Расширенный корпус		
Модель	Классификатор	F_1	Модель	Классификатор	F_1
FastText	RBF SVC	0.75	FastText	RBF SVC	0.79
	Linear SVC	0.74		Linear SVC	0.78
	Logistic Regression	0.72		Logistic Regression	0.76
	Ridge Regression	0.72		Ridge Regression	0.74
	SGD	0.65		SGD	0.71

Table 9. Best results with FastText embedding with initial and extended corpus

Таблица 9. Лучшие результаты с эмбеддингами FastText с изначальным и расширенным корпусами

Аспект	Первоначальный корпус			Расширенный корпус		
	Точность	Полнота	F_1	Точность	Полнота	F_1
проведение выборов	0.84	0.87	0.86	0.91	0.92	0.91
качество управления	0.66	0.80	0.73	0.69	0.83	0.75
качество здравоохранения	0.72	0.85	0.78	0.78	0.85	0.81
строительство атомной электростанции	0.89	0.84	0.87	0.85	0.82	0.83
развитие промышленности	0.77	0.86	0.82	0.83	0.89	0.86
развитие сельского хозяйства	0.63	0.80	0.71	0.62	0.81	0.71
состояние жилья	0.52	0.74	0.61	0.61	0.77	0.68
качество и стоимость услуг ЖКХ	0.72	0.84	0.78	0.77	0.86	0.81
состояние дорожной сети	0.76	0.76	0.76	0.70	0.80	0.75
газификация	0.60	0.75	0.66	0.79	0.85	0.82

димо отметить, что в обоих случаях лучший результат был получен с помощью использования метода опорных векторов с ядром RBF и эмбеддингов, обученных на новостях. Результаты обнаружения упоминаний отдельных аспектов представлены в таблице 11. Лучше всего определяются упоминания аспектов «проведение выборов», «качество здравоохранения», «строительство атомной электростанции» и «развитие промышленности», а хуже всего — аспектов «качество управления» и «состояние жилья».

Нужно особо отметить, что при использовании эмбеддингов, обученных на текстах художественной литературы, существенно снижается качество определений аспекта «газификация», так как в них отсутствует вектор для слова «газификация» и его производных.

5. Обсуждение результатов

В таблице 12 представлены 5 лучших результатов, полученных после экспериментов со всеми 4 описанными моделями.

Полученный с помощью наивного байесовского классификатора уровень качества со средней F_1-мерой, равной 0.77 оказался достаточно высоким. Результат существенно выше, с F_1-мерой, достигающей 0.84, был получен только при использовании эмбеддингов Navec и классификатора на основе метода опорных векторов. Из этого следует, что в задаче обнаружения упоминаний аспектов достаточно хорошие результаты могут достигаться даже с помощью достаточно примитивных методов при условии достаточно большого корпуса.

На этапе предобработки к повышению качества стабильно приводила лемматизация слов; стемминг и удаление служебных слов иногда приводили к повышению качества, а иногда — наоборот.

Table 10. Top 5 results with Navec embedding with initial and extended corpora

Таблица 10. 5 лучших результатов с эмбеддингами Navec с изначальным и расширенным корпусами

Первоначальный корпус			Расширенный корпус		
Модель	Классификатор	F_1	Модель	Классификатор	F_1
news	RBF SVC	0.69	news	RBF SVC	0.84
news	Ridge Regression	0.69	hudlit	RBF SVC	0.81
news	Linear SVC	0.69	news	Logistic Regression	0.78
news	Logistic Regression	0.68	news	Linear SVC	0.77
hudlit	RBF SVC	0.66	hudlit	Logistic Regression	0.75

Table 11. Best Navec results with initial and extended corpora for every aspect

Таблица 11. Лучшие результаты Navec с первоначальным и расширенным корпусами для каждого аспекта

Аспект	Первоначальный корпус			Расширенный корпус		
	Точность	Полнота	F_1	Точность	Полнота	F_1
проведение выборов	0.76	0.81	0.78	0.92	0.95	0.94
качество управления	0.56	0.65	0.60	0.73	0.84	0.78
качество здравоохранения	0.82	0.72	0.76	0.87	0.90	0.88
строительство атомной электростанции	0.88	0.67	0.76	0.95	0.91	0.93
развитие промышленности	0.73	0.73	0.73	0.87	0.93	0.90
развитие сельского хозяйства	0.68	0.62	0.65	0.72	0.84	0.78
состояние жилья	0.62	0.58	0.60	0.66	0.84	0.74
качество и стоимость услуг ЖКХ	0.73	0.71	0.72	0.79	0.88	0.83
состояние дорожной сети	0.86	0.69	0.77	0.79	0.81	0.80
газификация	0.70	0.52	0.60	0.78	0.82	0.80

Table 12. Best results

Таблица 12. Лучшие результаты

Модель	Классификатор	F_1
Navec news	RBF SVC	0.84
Navec hudlit	RBF SVC	0.81
FastText	RBF SVC	0.79
FastText	Linear SVC	0.78
Navec news	Logistic Regression	0.78

При использовании эмбеддингов во-первых, важно, чтобы они были обучены на достаточно большом корпусе — эмбеддинги нужно обучать на корпусе существенно большего объёма, чем тот, на котором обучается классификатор. Во-вторых, качество классификации повышается, если эмбеддинги обучены на корпусе меньшего объёма, но более близкой предметной области: новостные тексты ближе к политической агитации, чем художественная литература, и поэтому при использовании Navec, обученного на новостных текстах, качество было существенно выше, несмотря на меньший объём обучающего корпуса.

На качество классификации влияют ещё несколько факторов. Во-первых, это выбор самого классификатора. Метод опорных векторов и логистическая регрессия дают стабильно более высокие результаты, чем гребневая регрессия и классификаторы на основе решающих деревьев. Другой фактор — число предложений с упоминанием аспекта, на которых обучается классификатор. Аспекты «развитие промышленности» и «развитие сельского хозяйства» схожи, они оба обладают достаточно устойчивым набором слов-маркеров, но из-за того, что предложений с упоминанием

развития промышленности в два раза больше, чем с упоминанием развития сельского хозяйства, его упоминания обнаруживаются как минимум на 10 % лучше. Также очень серьёзный прирост качества был получен при расширении корпуса, причём этот прирост в большинстве случаев произошёл за счёт роста полноты, а не точности, т. е. классификаторы смогли лучше обучиться тому, в каких именно предложениях может упоминаться тот или иной аспект.

Сложность обнаружения упоминаний различных аспектов может очень сильно отличаться. Проще всего обнаруживаются упоминания аспектов, у которых есть свои характерные слова-маркеры, например, «строительство АЭС», «здравоохранение», «проведение выборов». Сложнее всего обнаруживаются упоминания аспектов «качество управления» и «состояние жилья», потому что они достаточно общие, у них нет своей характерной лексики, и их легко спутать с другими аспектами.

Нужно отметить, что удачная комбинация классификатора и типа эмбеддингов показывает примерно одинаково высокое качество обнаружения упоминаний всех аспектов, а неудачная — высокое качество для отдельных легко обнаруживаемых аспектов, и значительно более низкое — для всех остальных.

Нужно обратить внимание, что всех экспериментах, кроме экспериментов с FastText, полнота превышает точность, то есть классификаторы обнаруживают упоминания аспекта чаще, чем эксперт-разметчик. Это может быть вызвано тем, что почти все аспекты социально-экономической жизни тесно связаны между собой, в отличие, например, от аспектов товаров или услуг.

Заключение

В работе было проведено сравнение методов определения неявно упоминаемых аспектов в публицистических предложениях на русском языке. Эксперименты проводились на корпусе предложений из текстов предвыборной агитации. Они показали, что наилучшие результаты с F_1-мерой, равной 0.84, получаются при использовании эмбеддингов Navec, обученных на новостных текстах, и классификатора на основе опорных векторов с ядром на основе радиально-базисной функции. При этом достаточно хорошие результаты, с F_1-мерой, равной 0.77, могут быть получены при использовании наивного байесовского классификатора и представления предложений в виде «мешка слов».

Также в ходе экспериментов было выявлено, что сложность определения различных аспектов достаточно сильно отличается. Лучше всего определяются упоминания аспектов, для которых характерна определённая лексика. Такими аспектами являются, например, «строительство АЭС» или «проведение выборов». Хуже всего определяются упоминания общих аспектов, таких, как «качество управления». При этом лучшие методы позволяют достаточно хорошо определять упоминания как легкообнаружимых, так и труднообнаружимых аспектов, тогда как остальные показывают хорошие результаты для легкообнаружимых и плохие — для труднообнаружимых.

References

[1] B. Liu, *Sentiment Analysis and Opinion Mining*. Springer, 2022, 167 pp.

[2] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, “A survey on aspect-based sentiment analysis: Tasks, methods, and challenges”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 11, pp. 11 019–11 038, 2022. DOI: 10.1109/TKDE.2022.3230975.

[3] M. M. Truşcă and F. Frasincar, “Survey on aspect detection for aspect-based sentiment analysis”, *Artificial Intelligence Review*, vol. 56, no. 5, pp. 3797–3846, 2023.

[4] A. Naumov, R. Rybka, A. Sboev, A. Selivanov, and A. Gryaznov, “Neural-network method for determining text author’s sentiment to an aspect specified by the named entity”, in *CEUR Workshop Proceedings*, vol. 2648, 2020, pp. 134–143.

[5] E. V. Sergeeva, "Features of speech exposure in the preelection media discourse", in *Aktual'nye problemy gumanitarnogo znaniya v tekhnicheskom vuze*, in Russian, 2021, pp. 237–239.

[6] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and challenges of aspect-based sentiment analysis: A comprehensive survey", *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 845–863, 2020. DOI: 10.1109/TAFFC.2020.2970399.

[7] P. K. Soni and R. Rambola, "A survey on implicit aspect detection for sentiment analysis: Terminology, issues, and scope", *IEEE Access*, vol. 10, pp. 63 932–63 957, 2022. DOI: 10.1109/ACCESS.2022.3183205.

[8] B. Mohammed *et al.*, "Hybrid approach to extract adjectives for implicit aspect identification in opinion mining", in *11th International Conference on Intelligent Systems: Theories and Applications (SITA)*, IEEE, 2016, pp. 1–5. DOI: 10.1109/SITA.2016.7772284.

[9] A. O. Kornej and E. N. Kryuchkova, "Semantiko-statisticheskij algoritm opredeleniya kategorij aspektov v zadachah sentiment-analiza", *Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskie nauki*, no. 6 (216), pp. 66–74, 2020, in Russian. DOI: 10.18522/2311-3103-2020-6-66-74.

[10] E. I. Gribkov and Y. P. Ekhlakov, "Nejrosetevaya model' na osnove sistemy perekhodov dlya izvlecheniya sostavnyh ob'ektov i ih atributov iz tekstov na estestvennom yazyke", *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki*, vol. 23, no. 1, pp. 47–52, 2020, in Russian. DOI: 10.21293/1818-0442-2020-23-1-47-52.

[11] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text preprocessing for text mining in organizational research: Review and recommendations", *Organizational Research Methods*, vol. 25, no. 1, pp. 114–146, 2022. DOI: 10.1177/1094428120971683.

[12] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[13] U. Naseem, I. Razzak, and P. W. Eklund, "A survey of pre-processing techniques to improve short-text quality: A case study on hate speech detection on Twitter", *Multimedia Tools and Applications*, vol. 80, pp. 35 239–35 266, 2021. DOI: 10.1007/s11042-020-10082-6.

[14] J. Coates and D. Bollegala, "Frustratingly easy meta-embedding – computing meta-embeddings by averaging source word embeddings", in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Association for Computational Linguistics, 2018, pp. 194–198. DOI: 10.18653/v1/N18-2031.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, 2013. arXiv: 1301.3781v3 [cs.CL].

[16] I. Yamada *et al.*, "Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, 2020, pp. 23–30. DOI: 10.18653/v1/2020.emnlp-demos.4.

[17] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification", in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Association for Computational Linguistics, 2017, pp. 427–431. DOI: 10.48550/arXiv.1607.01759.

[18] A. Kukushkin. "Navec – kompaktnye embeddingi dlya russkogo yazyka". in Russian. (2020), [Online]. Available: https://natasha.github.io/navec/ (visited on 08/11/2024).

[19] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation", in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

[20] Q. Le and T. Mikolov, "Distributed representations of sentences and documents", in *International conference on machine learning*, PMLR, 2014, pp. 1188–1196.

[21] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.