УДК 004.855.5

Д. В. Желябин
ПАО «Ростелеком», Москва, e-mail: zhe_dv@mail.ru

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧИ NLP КЛАССИФИКАЦИИ ТЕКСТА НА ОСНОВЕ АНАЛИЗА СЕМАНТИКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Ключевые слова: машинное обучение, искусственный интеллект, семантика естественного языка, классификация текста, векторное представление, прямое кодирование, облако слов, логистическая регрессия.

В статье рассматриваются основные методы машинного обучения для решения бизнес-задачи NLP классификации текста на основе анализа семантики естественного языка. В условиях конкуренции оперативный контроль за потоками поступающих данных является жизненно необходимым. Значительные объемы данных становятся причиной для поиска ответа на сложные аналитические задачи, результат решения которых способен оказать влияние на руководителей и определить векторы развития и направления роста бизнеса на ближайшую и долгосрочную перспективы. Актуальной темой последнего времени является практическая возможность применения методов машинного обучения для решения поставленных бизнес-задач. Одной из наиболее востребованных является задача, связанная с пониманием текста и его дальнейшей обработки. Автор, в рамках работы в ПАО «Ростелеком», разработал модели машинного обучения, основанные на анализе семантики естественного языка для классификации наименований доходных закупок в целях подготовки аналитической отчетности и ее оперативного предоставления высшему руководству для принятия управленческих решений. В заключении проведен расчет экономической эффективности проекта разработки модели машинного обучения в рамках анализа доходной части, выражающейся в экономии денежных средств.

D. V. Zhelyabin
OJSC Rostelecom, Moscow, e-mail: zhe_dv@mail.ru

APPLICATION OF MACHINE LEARNING METHODS TO SOLVE THE NLP TEXT CLASSIFICATION PROBLEM BASED ON ANALYSIS OF SEMANTICS OF NATURAL LANGUAGE

Keywords: machine learning, artificial intelligence, natural language semantics, text classification, words embedding, one hot encoding, word cloud, logistic regression.

The article deals with the main methods of machine learning for solving the business problem of NLP text classification based on the analysis of natural language semantics. Operational control over the flow of incoming data is vital in a competitive environment. Large amounts of data cause you to search for answers to complex analytical tasks. The result of solving problems affects managers, as well as determines the vectors of business development in the future. A recent hot topic is the possibility of using machine learning methods to solve business problems. One of the most important tasks is to understand and process the text. The author developed machine learning models for classifying the names of profitable purchases in order to prepare analytical reports for top management. In conclusion, the calculation of economic efficiency of the project development of a machine learning model to analyze revenues, expressed in monetary saving as well.

Введение

В ходе работы в Департаменте бизнес-анализа Корпоративного центра ПАО «Ростелеком» со стороны высшего руководства были выдвинуты требования по подготовке периодических аналитических отчетов, связанных с оценкой результатов работы сотрудников департаментов по работе с корпоративным и государственным сегментами с доходными закупками в целях выполнения которых появилась необходимость в разработке модели машинного обучения, служащей основой для классификации текстовых наименований доходных закупок.

Цель исследования

Целью исследования является разработка и оценка моделей машинного обучения, основанных на анализе семантики естественного языка для классификации наименований доходных закупок в целях подготовки аналитической от-

четности и ее оперативного предоставления высшему руководству для принятия управленческих решений.

Материал и методы исследования

В ходе исследования использовались данные наименований закупок с официального сайта единой информационной системы в сфере закупок. В процессе исследования применялись методы машинного обучения, такие как логистическая регрессия, градиентный бустинг, ближайшие соседи.

Результаты исследования и их обсуждение

Бизнес-задача состоит в анализе доходных закупок, а именно в выявлении среди них профильных закупок, чьи наименования принадлежат тематике «Связь» – предоставление доступа в интернет, оказание услуг мобильной связи, предоставление цифровых каналов связи и т.д. На основе выявленных профильных закупок руководство получает информацию о том, каким заказчикам не удалось оказать услуги связи по причине отсутствия сетевых ресурсов, о потенциальных потерях среди профильных закупок, а также о том, стоит ли и далее принимать участие в тех или иных закупках. Существуют большие временные затраты на формирование отчетов, что объясняется объемами закупок, поступающих для анализа. Проведен расчет затрат для проведения ручной классификации закупок (табл. 1).

Следует отметить, что при подготовке отчетов об участии в закупках более 80% времени затрачивается только на проведение классификации закупок, что создает проблемные ситуации, связанные с оперативным предоставлением отчета руководству. В целях сокращения временных затрат на 70% и уменьшения трудозатрат в 15 раз было положено начало разработке модели машинного обучения для классификации наименований закупок, в том числе на основе анализа семантики естественного языка. Языком программирования выбран высокоуровневый язык программирования Python, а средой разработки – служба Microsoft Azure Notebooks. Далее будет рассмотрен процесс построения модели машинного обучения согласно методологии Knowledge Discovery in Databases (KDD) и будет включать следующие этапы [1, с. 4-5]:

1. Сбор данных – извлечение исходных данных в виде текстового корпуса;
2. Предобработка данных – применение методов очистки данных;
3. Data Mining – выбор метода машинного обучения и построение модели для извлечения знаний из данных;
4. Оценка – измерение эффективности модели машинного обучения.

Постановку задачи в формальных терминах машинного обучения можно выразить следующим образом [2, с. 13]: Задано множество документов $D = \{d_1,\ldots,d_{|D|}\}$ и множество классов $C = \{c_1,\ldots,c_{|C|}\}$. Неизвестная целевая функция $\Phi: D \cdot C \rightarrow \{0, 1\}$ задается формулой $\Phi(d_j, c_i) = \begin{cases} 0, \text{ если } d_j \notin c_i \\ 1, \text{ если } d_j \in c_i \end{cases}$. Необходимо построить классификатор $\Phi': D \cdot C \rightarrow \{0, 1\}$, максимально близкий к функции Φ. Другими словами, необходимо произвести бинарную классификацию текстовой информации с помощью способа машинного обучения с учителем.

Таблица 1

Стоимость выполнения операций по подготовке аналитических отчетов

Операция ручной классификации закупок	Частота выполнения, раз/мес.	Трудоемкость, минут	Денежные затраты, руб (ставка 1 000 руб./час)
Отчет об участии в закупках по профилю «Мобильная связь»	1	1 440	24 000
Отчет о выигранных закупках по профилю «Связь»	4	440	29 333
Отчет о закупках по профилю «Связь» с отказами от участия	1	1 500	25 000
Итого		4 700	78 333

В данном случае необходима обработка текстовых данных на естественном языке, реализующаяся с помощью технологии Natural Language Processing (NLP), характеризующаяся использованием последовательных операций по обработке исходного текста для его дальнейшего преобразования в целях извлечения полезной информации [3, с. 53]. Для решения задач NLP требуется наличие подготовленных текстовых коллекций, называемых текстовыми корпусами. В целях выполнения первого этапа методологии KDD извлечена выборка наименований закупок объемом 16 809 закупок, а также проведена ручная разметка наименований на два класса: закупка профильная и закупка не профильная. На этапе предобработки данных необходимо произвести очистку текстового корпуса на основе приведения слов к нижнему регистру, токенизации по словам для разделения предложения наименования закупки на слова-компоненты, удаления стоп-слов для избежания шума в данных, а также проведения лемматизации слов, который заключается в приведении слова к лемме, к его канонической форме. Для совершения данного процесса необходимо воспользоваться морфологическими анализаторами, среди которых присутствует pymorphy2, способный для входящего слова в ходе совершения морфологического разбора произвести в том числе его нормализацию [4, с. 1]. В ходе выполнения второго этапа методологии KDD произведена предобработка исходного текстового корпуса (рис. 1).

Необходимо разбить предобработанный текстовый корпус на обучающую и тестовую выборки. В данном случае произведено разделение на две выборки в соотношении 7:3. Далее следует решить, каким способом будет происходить кодирование нормализованных наборов слов, так как модели машинного обучения способны работать с числовыми представлениями. Существует два основных способа для преобразования слов в векторы – прямое кодирование и векторное представление. Прямое кодирование основывается на понятии «мешок слов», который заключается в проведении векторизации элементов текстового корпуса (т.н. документов текстового корпуса), в результате которой размерность векторов определяется через мощность словарного запаса текстового корпуса, а элементы векторов равны количеству вхождений того или иного слова из словаря для элемента текстового корпуса. Составлен уникальный словарь для обучающей выборки мощностью 7 739 слов. Для уменьшения размерности мешка слов осуществлено исключение слов из уникального словаря, имеющих частотность десять слов и менее. Рассмотрение облака слов для низкочастотных слов, сформированного с помощью библиотеки WordCloud, реализованной на Python, дает основание для понимания факта их низкой ценности с точки зрения наличия в словаре (рис. 2).

В результате оптимизации словаря его мощность составляет 1 105 слова, что положительно скажется на размерности мешка слов и эффективности работы модели в дальнейшем, и представляет из себя ценные слова для модели.

	Наименование закупки	Класс
0	"оказание услуг безлимитного доступа к информа...	1
1	"-оказание услуг по корпоративной подвижной ра...	1
2	"оказание услуг связи - присоединение и пропус...	1
3	поставка аппаратных криптошлюзов для организац...	0
4	"электронный аукцион на оказание услуг по прод...	0
5	""оказание услуг местной телефонной связи	1
6	оказание услуг по предоставлению права использ...	0
7	"""оказание услуг по разработке и внедрению си...	0
8	оказание услуг сотовой связи	1
9	услуги технической поддержки программного обес...	0

	Наименование закупки: леммы	Класс
0	[оказание, услуга, безлимитный, доступ, информ...	1
1	[оказание, услуга, корпоративный, подвижный, р...	1
2	[оказание, услуга, связь, присоединение, пропу...	1
3	[поставка, аппаратный, криптошлюз, организация...	0
4	[электронный, аукцион, оказание, услуга, продл...	0
5	[оказание, услуга, местный, телефонный, связь]	1
6	[оказание, услуга, предоставление, право, испо...	0
7	[оказание, услуга, разработка, внедрение, сист...	0
8	[оказание, услуга, сотовый, связь]	1
9	[услуга, технический, поддержка, программный, ...	0

Рис. 1. Фрагмент размеченного текстового корпуса до и после предобработки

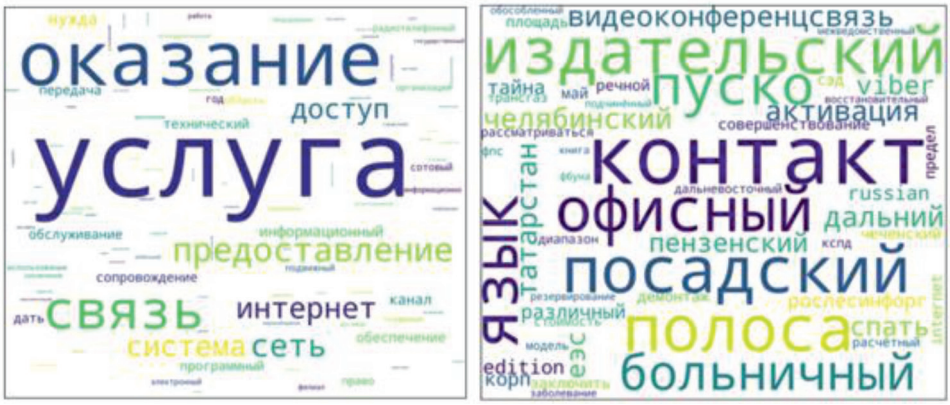

Рис. 2. Облака высокочастотных и низкочастотных слов

На основе оптимизированного словаря сформирован мешок слов для обучающей выборки текстового корпуса. Количество измерений мешка слов определяется через мощность оптимизированного словаря, а количество строк определяется через количество документов обучающей выборки текстового корпуса. Сформированный мешок слов имеет размерность мощности словаря. Вторым способом преобразования слов является векторное представление. В случае использования векторного представления слов необходимо применять специальные дистрибутивно-семантические модели для появления возможности формирования набора плотных векторов. В основе одной из таких моделей, Word2Vec, содержится гипотеза о необходимости анализа ближайших контекстных слов, количество которых определяется задаваемым параметром скользящего окна, предназначенного для выявления закономерностей в рамках контекстного окружения слова [5, с. 82]. На основе заданного скользящего окна происходит продвижение по всему текстовому корпусу и вычисляется попарная встречаемость целевого слова с окружающими его словами. В результате совершенного продвижения формируются значения частоты встречи слов во всем текстовом корпусе, а также значение количества раз, которое слово скользящего окна находилось в контекстном окружении целевого слова. Далее происходит вычисление весов слов контекстного окружения по отношению к целевому слову. В данном случае векторы будут построены таким образом, что близкие по контексту слова, будут находиться на наиболее близком расстоянии. Расстоянием между векторами двух слов будет являться косинусное сходство, вычисляемое по следующей формуле [6, с. 104]:

$$cosine\,similarity = \frac{(A,B)}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}.$$

Одним из параметров модели будет размерность векторного представления, минимальная частота слов для отбора в словарь, а также размер скользящего окна. Минимальная частота слов для отбора в словарь будет равняться также десяти, а размер скользящего окна трем, а размерность векторов около двухсот. Результатом является способность модели на основе полученного слова, представленного в виде входного вектора, предсказать результат распределения вероятностей данного слова быть в контексте с другими на всем их множестве. Для иллюстрации вышесказанного осуществим нахождение близких по контексту слов через вычисление косинусного сходства (табл. 2).

Таблица 2

Вероятность появления в контекстном окружении слов

Входное слово	Контекстное окружение	Вероятность появления слов, %
сотовый	подвижный	96
	радиотелефонный	90
	мобильный	84
интернет	высокоскоростной	70
	широкополосный	69
	представление	68

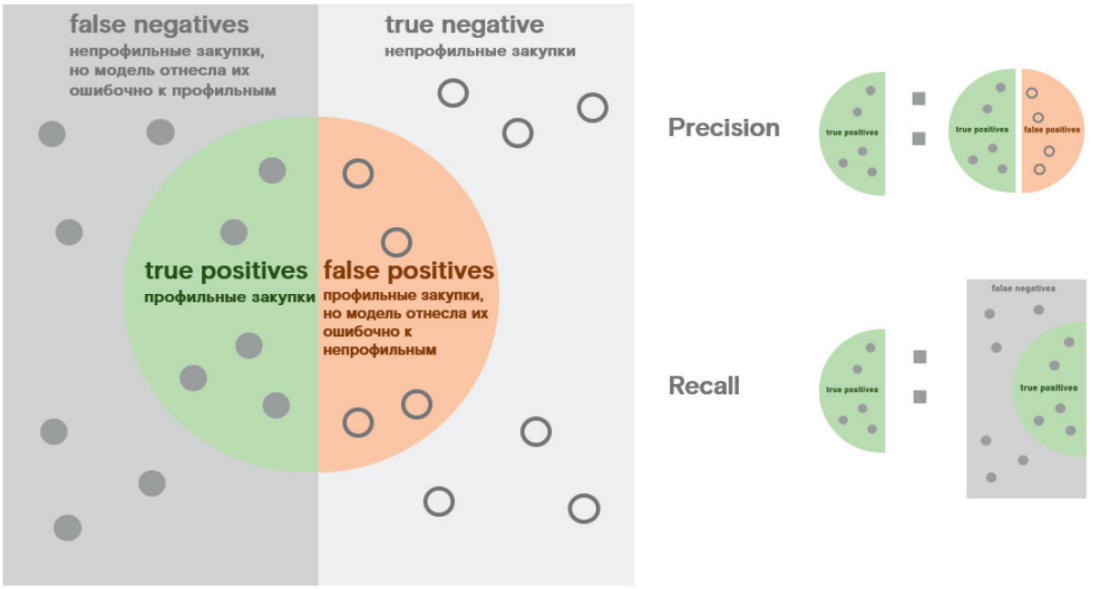

Рис. 3. Метрики для оценки модели машинного обучения

После совершения полной предобработки данных следует третий этап методологии KDD – Data Mining. Выбор методов машинного обучения объясняется их практической применимостью для решения задач классификации [7, с. 11-12]:

1. Логистическая регрессия (LR) – при применении данного метода происходит определение вероятности, с которой входное значение относится к определенному классу, в частности, в случае бинарной логистической регрессии осуществляется разделение исходного пространства границей на две части.

2. Градиентный бустинг (XGB) – при применении данного метода происходит построение предсказаний на основе слабых моделей, которые объединяются в ансамбль и на каждой итерации их последовательного применения предсказательная способность всей модели повышается.

3. Ближайшие соседи (KNN) – при применении данного метода входные значения классифицируются в зависимости от принадлежности к одному из классов его ближайших, соседних, значений на основе определения расстояния между входным значением и уже классифицированными значениями.

После разработки модели машинного обучения предстоит этап оценки моделей на основе метрик [8, с. 12] (рис. 3).

Получены результаты оценки модели машинного обучения на тестовой выборке, а также рассчитаны метрики (табл. 3).

Таблица 3

Оценка моделей машинного обучения на тестовой выборке, %

Способ кодирования	Прямое кодирование			Векторное представление		
Метод машинного обучения / Наименование метрики	LR	XGB	KNN	LR	XGB	KNN
Precision	97,2	96,6	97,5	94,6	95,9	96
Recall	95,4	95,7	92,4	93,3	95,1	93,6
F-Measure	96,3	96,2	94,9	94	95,5	94,8

Таблица 4

Расчет затрат на классификацию закупок с использованием модели машинного обучения

Операция ручной классификации закупок	Частота выполнения, раз/мес.	Трудоемкость, минут	Денежные затраты, руб (ставка 1 000 руб./час)
Отчет об участии в закупках по профилю «Мобильная связь»	1	57	950
Отчет о выигранных закупках по профилю «Связь»	4	20	1 333
Отчет о закупках по профилю «Связь» с отказами от участия	1	60	1 000
Итого		197	3 283

Таблица 5

Расчет экономии средств

Показатели	Затраты		Абсолютные показатели затрат, мес.	Относительное изменение затрат, %	Индекс изменения затрат
	До	После			
Трудоемкость	T_0, минут	T_1, минут	$\Delta T = T_0 - T_1$	$K_T = \frac{\Delta T}{T_0} \times 100\%$	$Y_T = \frac{T_0}{T_1}$
	4 700	197	4 503	96	24
Стоимость	C_0, руб./мес.	C_1, руб./мес.	$\Delta C = C_0 - C_1$	$K_C = \frac{\Delta C}{C_0} \times 100\%$	$Y_C = \frac{C_0}{C_1}$
	78 333	3 283	75 050	96	24

Необходимо отметить тот факт, что способ прямого кодирования показывает более высокие значения метрик как на обучающей, так и на тестовых выборках, чем способ векторного представления, что можно объяснить заданными правилами описания объекта закупки, регламентирующиеся Федеральными законами N 44-ФЗ и N 223-ФЗ. Другими словами, закупки, размещаемые на электронных площадках, именуются согласно общероссийскому классификатору продукции по видам экономической деятельности. В данном случае, модели машинного обучения на основе способа прямого кодирования демонстрируют лучшие результаты, так как содержащийся в его структуре мешок слов располагает возможностью кодирования структуры классификатора, насколько это возможно исходя из объема выборки и состава уникального словаря. Так как процесс обучения модели машинного обучения характеризуется итеративностью, в частности, в плане работы с исходным текстовым корпусом, для дальнейшего возможного улучшения показателей метрик появляется задача в дополнении исходного текстового корпуса новыми наименованиями непрофильных закупок в целях увеличения словарного запаса.

Выводы

В ходе разработки моделей машинного обучения и их применения на основе новых, ранее не известных для моделей, данных в рамках подготовки аналитических отчетов достигнуты следующие результаты:

1. Сокращение временных затрат на проведение классификации до 96%, что составляет индекс их изменения в 24 раза;

2. Повышение исполнительской дисциплины – полное исключение фактов нарушения контрольных сроков в ходе подготовки аналитического отчета;

3. Сокращение стоимости операций до 96%, что составляет индекс их изменения в 24 раза.

Произведен расчет затрат на классификацию закупок с использованием модели машинного обучения и расчет экономии средств (табл. 4-5).

Библиографический список

1. Borja Molina-Coronado, Usue Mori, Alexander Mendiburu, José Miguel-Alonso, «Survey of Network Intrusion Detection Methods from the Perspective of the Knowledge Discovery in Databases Process» [Electronic resource] // arXiv, Cornell University, 2020. URL: https://arxiv.org/abs/2001.09697v1 (accessed: 01.06.2020).

2. Бровкин К.Е., Раскатова М.В. Исследование методов машинного обучения для классификации неструктурированных текстовых документов // Международный журнал информационных технологий и энергоэффективности. 2019. Т. 4. № 2 (12). С. 12-17.

3. Искендеров Р. И., Волкова О. Р. Технология NLP при машинном обучении информационной системы службы технической поддержки. Наука сегодня: вызовы, перспективы и возможности: материалы научно-практической конференции, Вологда, 2019. Вологда, научный центр «Диспут», 2019. С. 52-55.

4. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages [Electronic resource]: arXiv, Cornell University, 2015. URL: https://arxiv.org/abs/1503.07283 (accessed: 07.06.2020).

5. Башков А.С., Соломенцев Я.К. Использование векторных методов представления слов в задачах выявления трендов // Вестник Российского нового университета. 2019. № 2. С. 80-88.

6. Filipyev A. Item-Based Recommender System with Statistical Learning for Unauthorized Customers // Системный анализ в науке и образовании. 2019. № 1. С. 102-111.

7. Калытюк И.С., Французова Г.А., Гунько А.В. К вопросу выбора методов предикативного анализа данных социальных медиа // Автоматика и программная инженерия. 2019. № 4. С. 9-17.

8. Архипов В.А. Сравнительный анализ метрик качества для моделей бинарной классификации на примере кредитного скоринга // Вестник Алтайской академии экономики и права. 2019. № 9. С. 12-15.