УДК 81’33

Донина Ольга Валерьевна
кандидат филологических наук,
доцент кафедры теоретической
и прикладной лингвистики
Воронежский государственный
университет
г. Воронеж, Россия

Olga Donina
PhD in Philology, Associate Professor,
Department of Theoretical
and Applied Linguistics
Voronezh State University
Voronezh, Russia
olga-donina@mail.ru

ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ МЕТАФОР

Данное исследование направлено на оценку эффективности нейронных сетей для задач бинарной классификации текста, в частности, для выявления наличия метафоры в предложениях. Для проведения сравнительного анализа были разработаны и протестированы шесть классификаторов, половина из которых основывалась на классических

моделях машинного обучения (наивный байесовский классификатор, логистическая регрессия, метод опорных векторов), а другая половина – на нейросетевых архитектурах (рекуррентная нейронная сеть, сверточная нейронная сеть, глубокая нейронная сеть). В качестве датасета для обучения и тестирования классификаторов был использован корпус размеченных вручную примеров метафорической сочетаемости, содержащий 166 765 предложений. Для оценки качества классификации текста использовались такие метрики, как precision, recall, F1-score и accuracy, при этом приоритетной оценкой являлся F1-score. Результаты показали, что нейронные сети демонстрируют значительное преимущество в выявлении метафор в предложениях по сравнению с методами машинного обучения. Особенно выделяется глубокая нейронная сеть (DNN), которая достигает высоких значений precision, recall и F1-score для обоих классов (наличие/отсутствие метафоры). Это может быть связано с ее способностью извлекать сложные зависимости из данных и строить более глубокие иерархические представления. Также стоит отметить, что метод опорных векторов (SVM) показывает неплохие результаты, хотя его эффективность немного ниже нейронных сетей. Проведение исследований в области автоматической классификации текстов с использованием нейронных сетей открывает дорогу к усовершенствованию способов автоматического извлечения метафор в текстах, что имеет важное значение для задач обработки естественного языка и лингвистического анализа.

К л ю ч е в ы е с л о в а: *классификация текста; нейронные сети; машинное обучение; Natural Language Processing; автоматическое выявление метафор.*

APPLYING MACHINE LEARNING FOR AUTOMATIC METAPHOR EXTRACTION

The objective of this study was to assess the efficacy of neural networks in the context of binary text classification tasks, with a particular focus on the detection of metaphors in sentences. A comparative analysis was conducted using six classifiers, with half based on classical machine learning models (naive Bayesian classifier, logistic regression, support vector machine) and the other half on neural network architectures (recurrent neural network, convolutional neural network, deep neural network). A corpus of manually labelled examples of metaphorical combinability containing 166,765 sentences was used as a dataset for training and testing the classifiers. Metrics such as precision, recall, F1-score and accuracy were employed to assess the quality of text classification, with F1-score being the primary metric. The results demonstrated that neural networks exhibited a pronounced advantage in detecting metaphors in sentences relative to machine learning methods. The deep neural network (DNN) in particular exhibited notable performance, achieving high precision, recall and F1-score values for both classes (presence/absence of metaphor). This may be attributed to its capacity to extract intricate interdependencies from data and construct more profound hierarchical representations. It is also noteworthy that the support vector machine (SVM) demonstrates commendable outcomes, although its performance is slightly inferior to that of neural networks. Conducting research in the domain of automatic text classification using neural networks paves the way for the advancement of automated metaphor extraction techniques, which has significant implications for natural language processing and linguistic analysis tasks.

K e y w o r d s: *text classification; neural networks; machine learning; natural language processing; automatic metaphor detection.*

Цель данного исследования заключалась в оценке эффективности нейронных сетей для задач бинарной классификации текста. Для проведения сравнительного анализа были разработаны и протестированы шесть класси-

фикаторов. Половина из них основывалась на классических моделях машинного обучения: наивном байесовском классификаторе (NBC), логистической регрессии (LR) и методе опорных векторов (SVM). Другая же половина опиралась на нейросетевые архитектуры, которые включали в себя модели рекуррентной (RNN), сверточной (CNN) и глубокой нейронной сети (DNN).

Задачей бинарной классификации было выявление наличия метафоры в предложениях. В качестве датасета для обучения и тестирования классификаторов был использован корпус размеченных вручную примеров метафорической сочетаемости, содержащий 166 765 предложений. Ручная разметка проводилась в основном студентами факультета романо-германской филологии Воронежского государственного университета, обучающихся на направлении «Фундаментальная и прикладная лингвистика», в рамках учебной практики в период 2016–2020 гг. [1; 2]. В данном корпусе наличие метафоры размечалось как «1», а отсутствие метафорического употребления – как «0» [3]. Эта метка, присвоенная каждому примеру, представляла собой ключевую информацию, которую мы использовали для обучения и оценки эффективности методов классификации.

Далее было необходимо утвердить методы оценки качества классификации текста. Существует несколько метрик, которые используются для оценки качества классификации текста [4]:

Precision (точность) – показатель того, какая доля предсказанных положительных (метафорических) классов действительно является положительными.

Recall (полнота) – показатель того, какая доля всех положительных (метафорических) текстов была правильно предсказана.

F1–score – среднее гармоническое между точностью и полнотой. Данная метрика учитывает и точность, и полноту, что позволяет оценить баланс между ними.

Accuracy (точность классификации) – показатель того, какая доля всех предсказанных классов (как метафорических, так и нет) была предсказана правильно.

В качестве приоритетной оценки использовалась метрика F1–Score, которая лучше, чем Accuracy, отражает результаты при сильном перевесе в классах [5; 6].

Датасет был разделен на обучающую и тестовую выборки в пропорции 70 % на 30 %, что оказалось наиболее оптимальным процентным соотношением для имеющегося объема данных [7].

Предобработка включала в себя токенизацию, лемматизацию, приведение слов к нижнему регистру и удалению стоп-слов с использованием Python-библиотек pandas, numpy и NLTK [8; 9].

Результаты работы классификаторов отображены в таблице ниже.

Сравнение методов машинного обучения

		precision	recall	f1–score
Наивный байесовский классификатор	0	0.83	0.89	0.86
Наивный байесовский классификатор	1	0.80	0.72	0.76
Логистическая регрессия	0	0.86	0.88	0.87
Логистическая регрессия	1	0.80	0.77	0.79
Метод опорных векторов	0	0.86	0.90	0.88
Метод опорных векторов	1	0.83	0.78	0.80
Рекуррентная нейронная сеть	0	0.88	0.91	0.89
Рекуррентная нейронная сеть	1	0.84	0.81	0.82
Сверточная нейронная сеть	0	0.86	0.91	0.88
Сверточная нейронная сеть	1	0.84	0.76	0.80
Глубокая нейронная сеть	0	0.95	0.95	**0.95**
Глубокая нейронная сеть	1	0.92	0.92	**0.92**

В целом модель имеет довольно высокую точность в определении текстов без метафор (метка 0), однако она менее точно распознает тексты с метафорами (метка 1). Это может быть связано с тем, что класс 1 может содержать более разнообразные выражения и структуры, которые сложнее обнаружить модели [10].

Проведя сравнительный анализ результатов классификации текстов, мы пришли к выводу, что нейронные сети демонстрируют значительное преимущество в выявлении метафор в предложениях по сравнению с классическими методами машинного обучения, что подтверждает их способность извлекать количественный смысл из сложных или неточных данных [11]. Судя по значениям F1–score, который является ключевым показателем в задачах бинарной классификации, нейронные сети, включая RNN, CNN и DNN, достигают более высоких значений как для класса 0 (отсутствие метафоры), так и для класса 1 (наличие метафоры).

Особенно выделяется глубокая нейронная сеть (DNN), которая демонстрирует высокие значения precision, recall и f1–score для обоих классов. Это может быть связано с ее способностью извлекать сложные зависимости из данных и строить более глубокие иерархические представления [12]. Также стоит отметить, что и метод SVM показывает неплохие результаты, хотя его эффективность немного ниже нейронных сетей. Это может быть связано с тем, что SVM хорошо подходит для выявления нелинейных зависимостей в данных и рассматривается как один из лучших контролируемых алгоритмов машинного обучения для построения бинарного классификатора [13]

Проведение исследований в области автоматической классификации текстов с использованием нейронных сетей открывает дорогу к усовершенствованию способов автоматического извлечения метафор в текстах.

ЛИТЕРАТУРА

1. Донина О. В. Реализация концепции корпусного исследования лексики в ходе учебной практики бакалавров лингвистики // Территория науки. 2017. № 4. С. 173–177.

2. Борискина О. О., Донина О. В. Корпусные исследования в контексте современных технологий обучения языку // Лингвориторическая парадигма : теоретические и прикладные аспекты. 2017. № 22–2. С. 154–158.

3. Донина О. В., Дмитриев Д. С. Возможность использования методов машинного обучения для автоматического выявления стертых метафор // Лингвистический форум 2020: Язык и искусственный интеллект. Институт языкознания РАН. 2020. С. 83–84.

4. Донина О. В. Автоматизация лингвистических исследований. Издательский дом Воронежского государственного университета. 2022. 125 с.

5. Donina O. V. How to use machine learning to automatically detect dead metaphors // RaAM14. Conference Book of Abstracts. 2021. С. 247–248.

6. Донина О. В. Выявление метафорической сочетаемости методами машинного обучения // Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация. 2022. № 4. С. 128–143.

7. Становов В. В. Многоагентный алгоритм проектирования баз нечетких правил для задач классификации // Сибирский аэрокосмический журнал. 2015. № 4. С. 842–848.

8. Сидоров К. А., Коротких А. Д., Донина О. В. Автоматизация бинарной классификации текстов английского языка по варианту языка и жанру с применением технологии искусственных нейронных сетей // Информатика: проблемы, методы, технологии. Воронеж, 2021. С. 1508–1514.

9. Возможности использования искусственных нейронных сетей для классификации текстов по варианту языка и жанру / К. А. Сидоров, А. Д. Коротких, О. В. Донина, А. А. Пендюрина // Математика и междисциплинарные исследования. Пермь, 2020. С. 189–193.

10. Донина О. В. Применение методов Data Mining для решения лингвистических задач // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2017. № 1. С. 154–160.

11. Нейроны в нейронных сетях / В. М. Панарин, К. В. Гришаков, А. А. Маслова, О. В. Гришакова, А. В. Архипов // Известия ТулГУ. Технические науки. 2023. № 2. С. 438–443.

12. Березин С. А., Бондаренко И. Ю. Выделение именованных сущностей из текстов распорядительных документов с помощью глубоких нейронных сетей // Системная информатика. 2020. № 16. С. 137–148.

13. Шибайкин С. Д., Никулин В. В., Аббакумов А. А. Анализ применения методов машинного обучения компьютерных систем для повышения защищенности от мошеннических текстов // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. 2020. № 1. С. 29–40.