Автоматизация бинарной классификации текстов английского языка по варианту языка и жанру с применением технологии искусственных нейронных сетей

К. А. Сидоров, email: kirs53@bk.ru
А. Д. Коротких, О.В. Донина

Воронежский государственный университет

Аннотация. *В рамках данного исследования был построен бинарный классификатор текстов английского языка по варианту и жанру с применением технологии искусственных нейронных сетей. Данные для классификации выгружены из корпусов английского языка Марка Дэвиса: Global Web-based English (GloWbE) и News on the Web (NOW). Моделирование ИНС реализовывалось с помощью открытых библиотек TensorFlow Hub и Keras на языке программирования Python.*

Ключевые слова: *искусственная нейронная сеть, вариант языка, жанр текста, автоматическая классификация.*

Многие исследования в сфере компьютерной лингвистики сегодня дают ученым возможность автоматизации сложных процессов, ранее выполнявшихся вручную и требующих большого количества различных ресурсов.

Актуально в последний годы стало обращение к технологиям, разрабатываемым в сфере искусственного интеллекта. В исследованиях активно применяются искусственные нейронные сети (ИНС), представляющие собой раздел искусственного интеллекта, в котором для обработки сигналов используются явления, аналогичные происходящим в нейронах живых существ. Важнейшие особенности сети состоят в способности параллельной обработки информации всеми звеньями, способности к обучению и обобщению накопленных знаний, а также в возможности ее реализации с применением технологии сверхбольшой степени интеграции [1].

Чаще всего ИНС применяются в рамках компьютерной лингвистики, например, для классификации больших массивов лингвистических данных.

В данном исследовании нами проводилась автоматизация классификации текстов английского языка [2]. Для этой цели были реализованы две модели бинарной классификации для определения: а) варианта английского языка, б) жанра текста на английском языка.

В качестве основы была взята нейронная сеть, классифицирующая отзывы к кинофильмам на положительные и отрицательные (представлена в учебном руководстве от Tensorflow [3]). Это одна из наиболее простых в реализации ИНС, эффективно справляющихся с задачей бинарной классификации.

Обучение (с учителем) и тестирование ИНС происходило: в первом случае, на фрагментах текстов из корпуса NOW, относящиеся к двум вариантам английского языка: британскому и американскому, во втором, к новостному (news) и интернет-дискурсу (web) из корпуса GloWbE. В данной работе для удобства проведения классификации понятия дискурс и жанр были приравнены друг к другу. Новостной дискурс соотносится с обозначенным нами жанром «новости», а интернет-дискурс с жанром «веб».

Все исходные данные получены в рамках проекта кафедры ТиПЛ «Криптоклассы английского языка». Разметка проводилась вручную. Итоговый результат был согласован между 3-мя разметчиками. Общее количество данных составило 528626 примеров. Для повышения качества бинарной классификации в документах была удалена вся метаинформация, не относящаяся к теме исследования.

Разбивка фрагментов текстов по вариантам языка и жанрам представлена в таблицах ниже (табл. 1, 2).

Таблица 1

Разбивка примеров по вариантам языка

Метка	Вариант языка	Количество примеров
1	Английский британский	68 223
0	Английский американский	70 547
ВСЕГО		138 770

Таблица 2

Разбивка примеров по жанрам

Метка	Жанр	Количество примеров
1	Новости	252 463
0	Веб	137 393
ВСЕГО		389 856

Использованная модель классификатора состоит из трех слоев. Для упрощения процесса обработки, во входном слое (TensorFlow Hub) происходит преобразование предложений в их векторное представление посредством Saved Model. Полученные векторы проходят через

скрытый полносвязный слой, состоящий из 16 нейронов. В выходном слое активационной функцией выступает сигмоида, значения которой находятся в пределах от нуля до единицы. В качестве функции потерь используется перекрестная энтропия, в качестве оптимизатора – Adam (Adaptive Moment Estimation).

Распределение исходных данных на выборки для классификации по варианту языка представлено в таблице ниже (табл. 3).

Таблица 3

Распределение данных на выборки. Вариант языка

	Выборка			ВСЕГО
	Обучающая	Валидационная	Тестовая	
Количество примеров, ед.	83 262	27 754	27 754	138 770
Количество примеров, %	60%	20%	20%	100%

Экспериментальным методом были выбраны следующие показатели нейронной сети: 30 эпох, 128 образцов в одном батче. Лучшие результаты тестирования составили: 0.892 (точность) и 0.259 (коэффициент потерь) (рис. 1, 2).

Изменение показателей приводило к переобучению, снижению точности и увеличению коэффициента потерь.

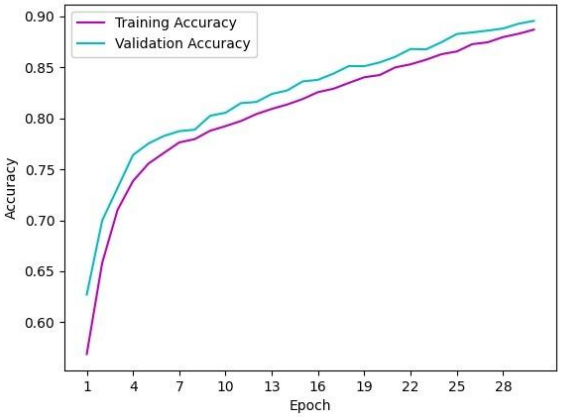

Рис. 1. Варианты языка. График зависимости точности от номера эпохи.

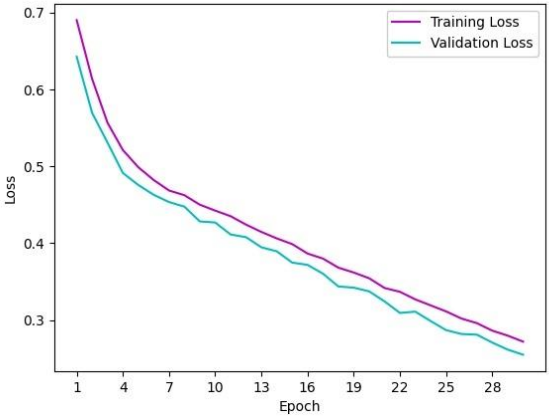

Рис. 2. Варианты языка. График зависимости коэффициента потерь от номера эпохи.

Для проверки корректности работы построенной модели был проведен эксперимент. На вход нейронной сети поступали 30% неверно и 70% верно размеченных примеров обучающей выборки. Прочие показатели не менялись (табл. 3). В результате коэффициент потерь увеличился и превысил единицу, а точность составила 0.609 (рис. 3, 4). Подобное ухудшение значений свидетельствует о правильной работе модели в условиях эксперимента.

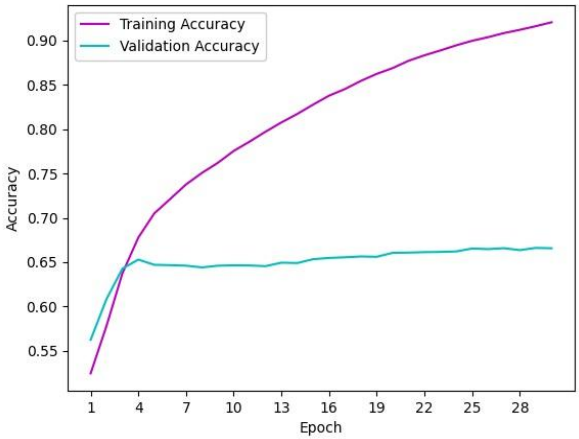

Рис. 3. Варианты языка. Эксперимент. График зависимости точности от номера эпохи.

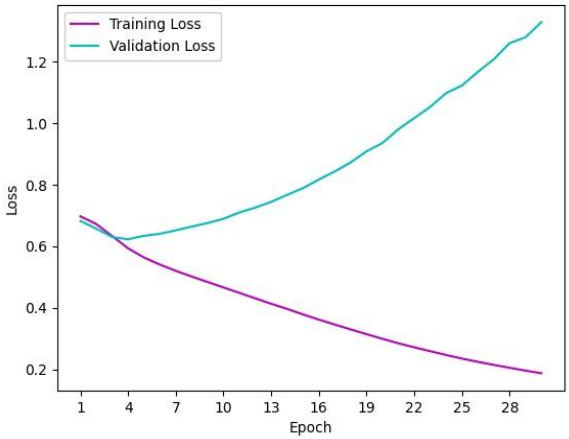

Рис. 4. Варианты языка. Эксперимент. График зависимости коэффициента потерь от номера эпохи.

Распределение исходных данных на выборки для классификации по жанру представлено в таблице ниже (табл. 4).

Таблица 4

Распределение данных на выборки. Жанр

	Выборка			ВСЕГО
	Обучающая	Валидационная	Тестовая	
Количество примеров, ед.	233 913	77 971	77 971	389 856
Количество примеров, %	60%	20%	20%	100%

В данном случае было установлено количество эпох равнялось 20, объем батча составил 256 образцов. Показатели были выбраны экспериментальным путём. Во время тестирования наилучшие значения показателей разнялись 0.892 (точность) и 0.259 (коэффициент потерь) (рис. 5, 6).

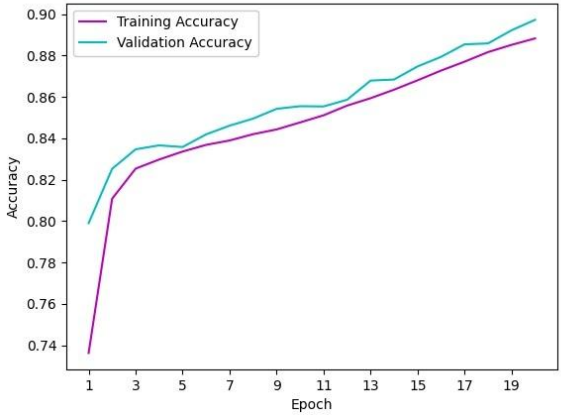

Рис. 5. Жанр. График зависимости точности от номера эпохи

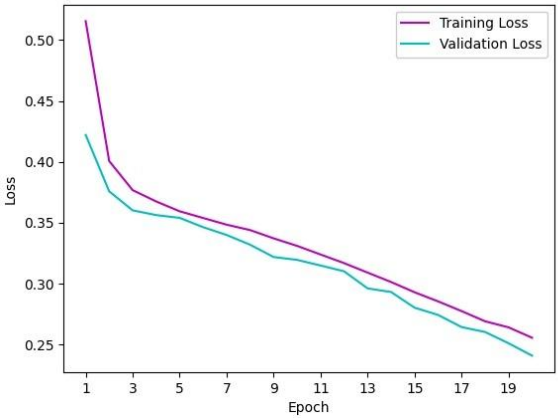

Рис. 6. Жанр. График зависимости коэффициента потерь от номера эпохи

Обе реализации модели достигли высоких показателей точности (более 85%) и низкого значения коэффициента потерь (менее 0.3). При сравнении наших результатов с результатами аналогичных исследований, посвященным идентификации варианта языка с использованием различных функций (мешка слов, N-грамм, векторной и др. моделей представления текста) [4] и жанровой классификации текстов с помощью свёрточных и рекуррентных нейронных сетей [5], где наилучший результат достигал 71,5% и 72,12% соответственно, можно сделать вывод о хорошей производительности представленной искусственной нейронной сети.

Список литературы

1. Нейронные сети для обработки информации / Пер. с польского И.Д. Рудинского. – М.: Финансы и статистика, 2002. С. 13.

2. Саймон Х. Нейронные сети: полный курс. – 2-е – М.: «Вильямс», 2006. – с.1104.

3. Учебное руководство от Tensorflow [Электронный ресурс]. – Режим доступа: https://www.tensorflow.org/tutorials/keras/text_classification_with_hub

4. Rangel F., Rosso P., Pothhast M., Stein B. Gender and Language Variety Identification in Twitter. M.: PAN at CLEF, 2017.

5. Батраева И. А. Использование анализа семантической близости слов при решении задачи определения жанровой

принадлежности текстов методами глубокого обучения // Вестник Томского государственного университета. 2020. №50. С. 14-22.