Т. В. Трифонов, И. Е. Андреев, А. В. Глазкова
Тюменский государственный университет, г. Тюмень
УДК 004.912

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ТОНОВОЙ КЛАССИФИКАЦИИ ТЕКСТОВ

Аннотация. В данной статье происходит сравнение методов машинного обучения для тоновой классификации текстов с помощью методов опорных векторов (SVM), ИНС с ячейками LSTM, ИНС с ячейками BiLSTM. Сравнение выбранных методов оценивается процентом верно предсказанных классов на одной и той же тестовой выборке.
Ключевые слова: emotional detection from text, emotion classification, обработка естественного языка, тоновая классификация текстов.

Введение

Обработка естественного языка (Natural Language Processing, NLP) – это обширная область IT связанная с использованием компьютеров для анализа естественных языков, к которой относятся такие дисциплины, как распознавание и обработка речи, выделение смысловых отношений, категоризация документов, а также реферирование и аннотирование текста. Но все эти виды анализа основаны на относительно небольшой группе базовых методик: разделение потока текста на фрагменты – токенизация (tokenization), определение границ предложений, классификация и выделение отношений (между элементами текста). [2]

На вход в систему обработки естественного языка приходит последовательность символов (см. рис. 1). Далее, происходит определение основных характеристик входной последовательности слов (определение частей речи, приведение к нормальным формам и т.д). После морфологической обработки строится синтаксическое дерево. Далее происходит семантический анализ, который направлен на работу с фактических восприятием текстовой информации. В завершении,

производится анализ текста целиком, а именно: выделяется основная тема текста, подтемы текста, скрытый смысл и т.д.

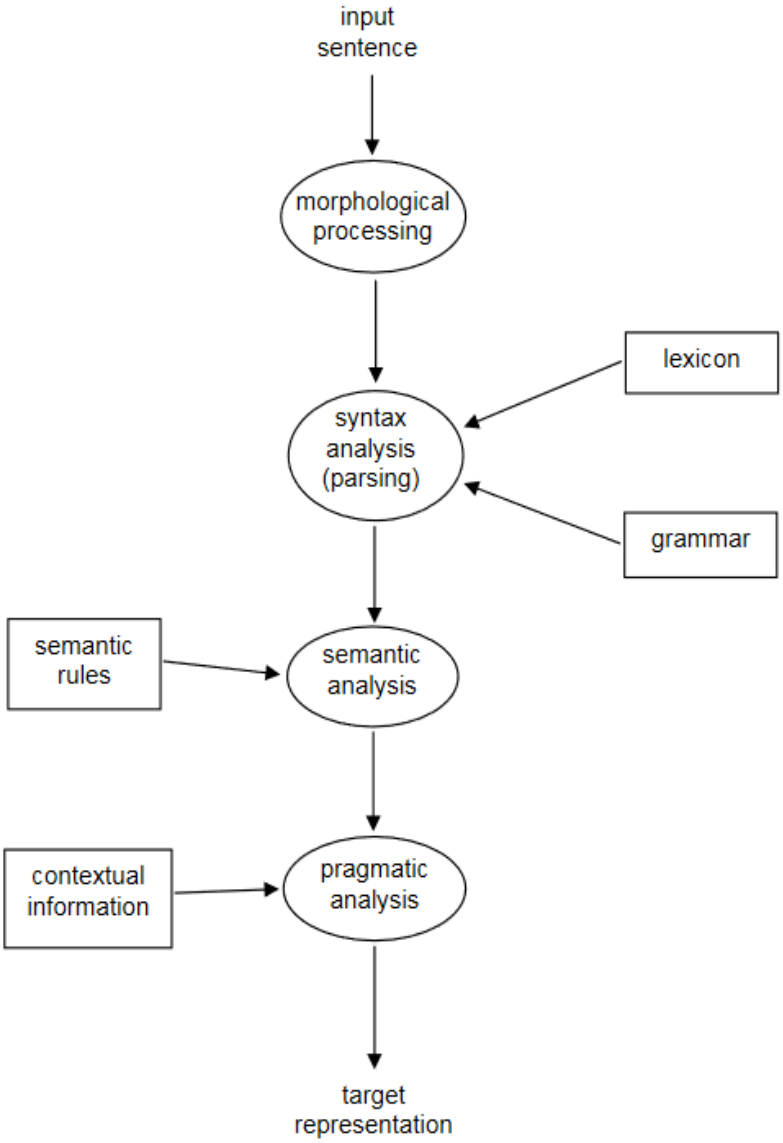

Рис. 1. Общая схема обработки естественного языка.

В данной работе будет подробно рассмотрен такой раздел NLP, как семантический анализ текстов.

Семантический анализ является этапом в последовательности действий алгоритма автоматического понимания текстов, заключающийся в выделении семантических отношений, формировании семантического представления текстов. Один из возможных вариантов представления семантического представления — структура, состоящая из «текстовых фактов».

Подраздел семантического анализа, основанный на выделении эмоциональной составляющей текста называется тональным анализом.

Тональный анализ является необходимым инструментом, для корректного анализа текста.

Классификация данных – общая задача машинного обучения (machine learning), в этом направлении применяются методы оптимизации и аналитической геометрии.

В настоящее время высокую эффективность демонстрируют подходы к проведению автоматической классификации, связанные с использованием рекуррентных нейросетевых моделей. Эксперименты, посвященные сравнению данных подходов с существовавшими ранее методами машинного обучения, показывают, как правило, превосходство рекуррентных нейронных сетей [4-5]. Если говорить о детерминированных методах классификации, в задачах обработки текстов достаточно широко применяется метод опорных векторов (Support Vector Machine, SVM), который успешно используется во многих практических приложениях [6-7]. В нашей работе мы проводим сравнение эффективности метода опорных векторов и двух архитектур рекуррентных нейронных сетей на примере задачи тоновой классификации текстов.

Методология

a) Support Vector Machine

Метод опорных векторов заключается в переводе исходных векторов в пространство более высокой размерности и поиске разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы.

Подход, описанный выше и показанный на рисунке (см. рис. 2), обобщается на многомерный случай.

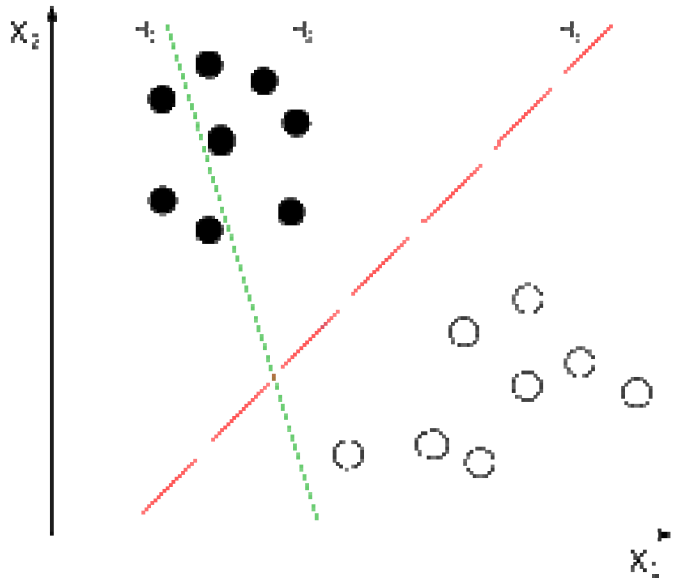

Рис. 2. H1, H2, H3 – гиперплоскости.
H3 – гиперплоскость максимальной разности.

b) Рекуррентные нейронные сети

Нейронная сеть — это математическая модель, а также ее программные или аппаратные реализации, построенная в некотором смысле по образу и подобию сетей нервных клеток живого организма.

Как и линейные методы классификации и регрессии, по сути нейронные сети выдают ответ вида: $y(x,w) = f(\sum_{j=1}^{M} w_j \phi_j(x))$, где f — нелинейная функция активации, w — вектор весов, Φ — нелинейные базисные функции. Обучение нейронных сетей состоит в настройке весов а также базисных функций. Нейронные сети отличаются по типу ячеек и архитектуре. Нейронная сеть требует набор подготовленных данных для обучения (в дальнейшем датасет с англ. «набор данных»). [1]

Так как не во всем наборе входных данных содержится полезная информация эти данные нуждаются в обработке и выделении признаков. Мы используем такое представление текста, как Word2Vec. Данное представление является обученной моделью, которая на вход принимает слово, а на выходе мы имеем вектор, фиксированной длины. Модель обучена на корпусе текстов Araneum [8], со следующими параметрами:

кол-во слов – 10 миллиардов;

объем словаря – 196.465 слов;

алгоритм – Continuous Skipgram;

размерность вектора – 600.

Рекуррентные нейронные сети (*Recurrent neural network*; *RNN*) — вид нейронных сетей, где связи между элементами образуют направленную последовательность (см. рис. 3). Благодаря этому появляется возможность обрабатывать серии событий во времени или последовательные пространственные цепочки.

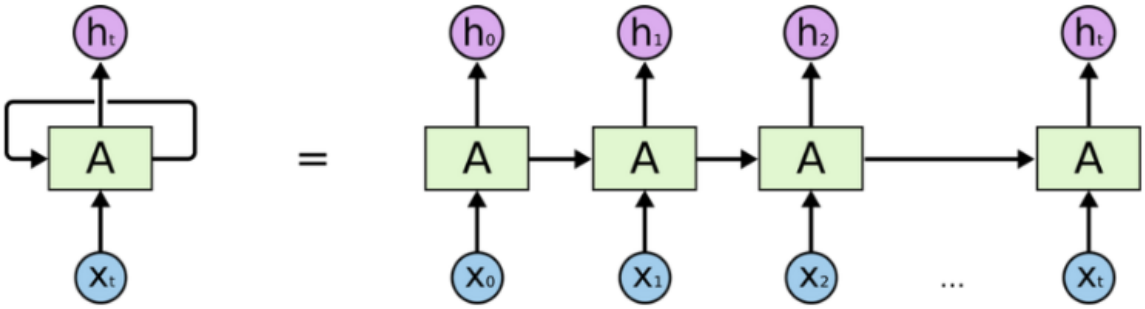

Рис. 3. Развертка рекуррентной нейронной сети

Рекуррентные нейронные сети отличаются друг от друга реализацией самой ячейки.

Сети с долгой краткосрочной памятью *(long short term memory, LSTM)* стараются решить проблему потери информации, используя фильтры и явно заданную клетку памяти. У каждого нейрона есть клетка памяти и три фильтра: входной, выходной и забывающий (см. рис. 4).

LSTM разработаны специально, чтобы избежать проблемы долговременной зависимости.

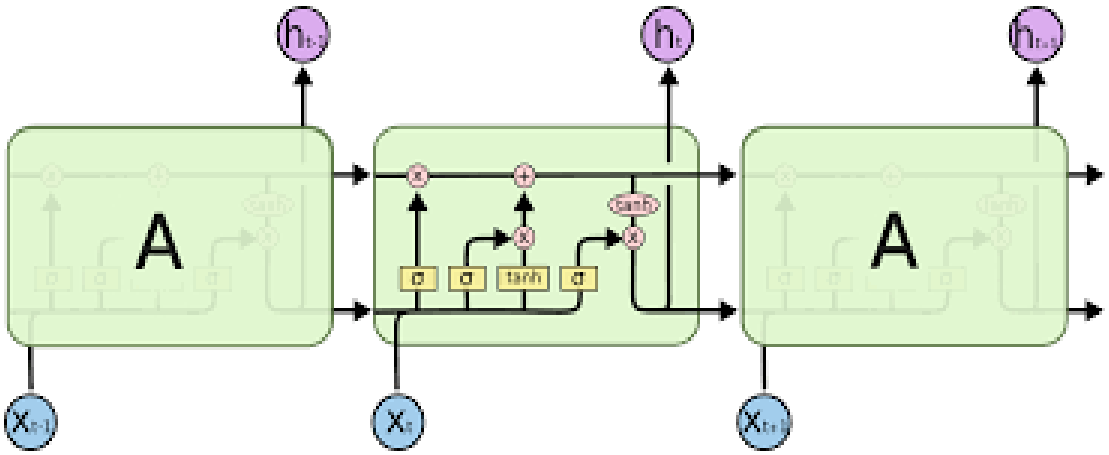

Рис. 4. Схема ячейки LSTM

Двунаправленная сеть долгой краткосрочной памяти (BiLSTM) комбинирует классическую LSTM-сеть, которая обрабатывает последовательность данных от её начала до конца, с другой LSTM-сетью, которая рассматривает последовательность в обратном порядке, то есть получает данные с предыдущих и последующих итераций.

Эксперимент

а) Описание данных.

Для сравнения методов машинного обучения была выбрана задача бинарной классификации. В качестве датасета [3] была выбрана выборка, состоящая из постов и двух меток:

“0” – негативное сообщение

“1” – позитивное сообщение

пользователей социальной сети Twitter.com.

Входными данными для моделей являются вектора размерностью [600, 1], полученные с помощью представлений Word2Vec из сообщений.

Задача моделей – верно предсказать негативным или позитивным является полученный текст на входе.

Датасет имеет объем 48.34 мб, в котором содержится 226914 записей (114.991 позитивных и 111.923 негативных), из которых 200000 записей было использовано для обучения, а 26914 для проведения тестирования.

b) Параметры обучения моделей.

Выбор параметров обучения моделей и структуры нейронных сетей был проведен экспериментально. Значения параметров, использованные при проведении сравнения моделей, указаны в таблице 1.

Таблица 1. Параметры обучения

Модель	**Параметры обучения ***	**Архитектура**
LSTM	batch_size = 10 num_hidden = 32 activation = “softmax” loss = “categorical crossentropy” optimizer = “adam”	1. LSTM (32 нейрона на скрытом слое) 2. Dropout (0.5) 3. LSTM (32 нейрона на скрытом слое) 4. Dropout (0.5) 5. LSTM (32 нейрона на скрытом слое) 6. Dropout (0.5) 7. Dense (2 выходных нейрона)
Bi_LSTM	batch_size = 10 num_hidden = [32, 64] activation = “softmax” loss = “categorical crossentropy” optimizer = “adam”	1. BiLSTM (32 нейрона на скрытом слое) 2. Dropout (0.5) 3. BiLSTM (64 нейрона на скрытом слое) 4. Dropout (0.5)
SVM	kernel=”rbf”	-

***** *Примечание*:

batch_size – параметр, отвечающий за количество примеров, которые необходимо обработать сети прежде тем, чем обновить весовые коэффициенты;

num_hidden – количество нейронов на скрытом слое;

activation – функция активации на выходном слое;

loss – функция ошибки;

optimizer – метод оптимизации весовых коэффициентов сети;

kernel – функция ядра.

Анализ результата

Качество модели оценивалось по метрике Accuracy на тестовой выборке. Данная метрика была выбрана, так как количество примеров разных классов в тестовой выборке сбалансировано.

Результат обучения метода SVM составил 0,74.

Сравниваются результаты между моделями рекуррентных сетей, в зависимости от числа эпох.

Таблица 2. Сравнение моделей

Эпохи	LSTM	BiLSTM
100	0,79	0,88
200	0,84	0,91
300	0,86	0,92

Исходя из данных, представленных в таблице и результата работы метода опорных векторов, можно сделать следующие выводы:

1. Алгоритмы глубокого обучения работают лучше, чем стандартные методы машинного обучения, в том случае, когда у нас имеется достаточное количество данных

2. Bidirectional ячейки в задаче тональной классификации текстов работают лучше, чем ячейки, анализирующие входную последовательность только в одну сторону

Заключение

Глубокие модели машинного обучения в последние годы ушли далеко вперед по сравнению со стандартными алгоритмами машинного обучения. Это связано с тем, что глубокие сети способны выделять самостоятельно очень большое число скрытых параметров, внутри входных данных. А количество и качество скрытых параметров напрямую связано с качеством распознавания, т.к корректный анализ зависит именно от их числа. Число параметров, в свою очередь, зависит непосредственно от глубины сети и корректности конструирования архитектуры сети, а также очень важным фактором является объем и качество датасета.

Современные архитектуры глубоких нейронных сетей, зачастую, уступают стандартным методам машинного обучения, при решении менее-комплексных задач с плохой выборкой данных, т.к стандартные методы требуют меньшего объема датасета для корректной работы.

СПИСОК ЛИТЕРАТУРЫ

1. Goodfellow I., Bengio Y., Courville Deep Learning. 2016. 802 с.
2. Daniel Jurafsky and James H. Martin “Speech and language processing” // Stanford, 2018.
3. URL: https://web.Stanford.edu/~jurafsky/slp3/ed3book.pdf (дата обращения: 14.04.19).
4. Рубцова Ю.В. Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. 2015. №1(109). С. 72-78.
5. Arkhipenko K., Kozlov I., Trofimovich J., Skorniakov K., Gomzin A., Turdakov D. Comparison of Neural Architectures for Sentiment Analysis of Russian Tweets // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”. 2016. С. 50-58.
6. Глазкова А.В. Автоматический поиск фрагментов, содержащих биографическую информацию, в тексте на естественном языке // Труды Института системного программирования РАН. 2018. Т. 30. № 6.. С. 221-236.
7. Pratama B.Y., Sarno R. Personality classification based on Twitter text using Naive Bayes, KNN and SVM // 2015 International Conference on Data and Software Engineering (ICoDSE). IEEE, 2015. С. 170-174.
8. Дмитриев А.С., Соловьев И.С., Заболеева-Зотова А.В. Извлечение взаимосвязей между объектами и терминами в текстах на экономическую тематику // Известия Волгоградского государственного технического университета. 2015. № 13 (177). С. 55-60.
9. Корпус русских текстов Araneum // 2017.
10. URL: http://ucts.uniba.sk/aranea_about/_russicum.html (дата обращения: 14.04.19).