УДК 004.912

ВЕРИФИКАЦИЯ РУССКОЯЗЫЧНЫХ ИНТЕРНЕТ-ТЕКСТОВ РАЗЛИЧНЫХ ЖАНРОВ

Лагутина Ксения Владимировна, к.т.н., старший преподаватель,
Лагутина Надежда Станиславовна, к. ф.-м. н., доцент
Ярославский государственный университет им. П.Г. Демидова
Lagutina Ksenia, PhD, lagutinakv@mail.ru
Lagutina Nadezhda, PhD, lagutinans@rambler.ru
P.G. Demidov Yaroslavl State University

В работе рассматривается задача классификации русскоязычных текстов по жанрам с использованием лингвистических характеристик. Текст моделируется вектором числовых характеристик уровня символов, слов и ритма. В качестве классификатора используется нейронная сеть Bi-LSTM. Верификация жанров показывает высокое качество решения задачи с F-мерой от 90 % до 98 %.

Ключевые слова: *автоматическая обработка текста, стилометрия, кластеризация, итар*

Одним из разделов компьютерной лингвистики является стилометрия, занимающаяся количественной оценкой лингвистических характеристик текстов естественного языка. К этому разделу относится задача классификации текстов по жанрам, таким как новости, реклама, отзывы и рецензии, научные статьи и т. п. Жанр можно рассматривать как набор документов, имеющих одинаковые стилистические свойства [1, с. 28]. Он позволяет определить функции документа и его коммуникативный контекст. Классификация текстов по жанрам является важной задачей в информационном поиске, синтаксическом и семантическом анализе текста, автоматической аннотации документа, машинном переводе [2, с. 1584]. Жанровая классификация текста полезна при обнаружении спама и более быстром поиске нужной информации, связана с особенностями грамматики, синтаксиса языка и смысла слов.

Основой решения описанной проблемы является построение модели текста как многомерного вектора числовых характеристик. Далее, с помощью алгоритмов машинного обучения, выполняется классификация текстовых жанров на основе обучающего набора документов с соответствующими метками. В качестве классификаторов часто используются нейронные сети различной архитектуры или алгоритмы обучения с учителем, такие как машины опорных векторов, наивный байесовский классификатор, деревья решений. Этот подход хорошо изучен для английского языка и значительно меньше для других естественных языков [3, с. 674].

Авторы работы поставили задачу исследовать возможность жанровой классификации русскоязычных текстов на основе лингвистических характеристик. Для этого были выделены следующие подзадачи: собрать корпус Интернет-текстов различных жанров, кластеризовать их на основе построенных характеристических векторов, сопоставить кластеры с жанрами, верифицировать жанры, т. е. отделить друг от друга с помощью бинарной классификации.

Для кластеризации и верификации авторами самостоятельно был собран корпус из 16000 текстов восьми жанров, по 2000 текстов на жанр. Жанры включают в себя рекламные тексты, комментарии из социальных сетей, новости, описания торговых компаний, блоги с сайта Хабр, научные статьи, рецензии на фильмы с сайта Кинопоиск, посты ВКонтакте.

Тексты моделировались при помощи лингвистических характеристик трёх типов:

- характеристики уровня символов: частоты букв и знаков препинания, средняя длина предложения в символах и словах, средняя длина слова в символах;
- характеристики уровня структуры: n-граммы частей речи, n = 1,2,3,4;
- характеристики уровня ритма:

--плотность ритма — количество появлений в тексте ритмического средства, разделённое на количество предложений. Ритмические средства включают в себя анафору, эпифору, симплоку, анадиплозис, эпаналепсис, многосоюзие, диакопу, эпизевксис, хиазм, апозиопеза, повторяющиеся вопросительные и восклицательные предложения, аллитерацию и ассонанс. Алгоритмы их поиска и подсчёта описаны в предыдущей работе авторов [5, с. 245];

--доли различных частей речи среди слов, составляющих ритмические средства.

С помощью указанных лингвистических данных текст представлялся как вектор чисел, который мог включать в себя как набор характеристик одного уровня, так и конкатенацию векторов характеристик двух или трёх уровней.

Вектора всего корпуса текстов были собраны в общую двумерную матрицу, которая служила исходными данными для алгоритма кластеризации UMAP. Это один из популярных подходов к кластеризации, который уменьшает размерность данных до двумерного пространства и позволяет их визуализировать на координатной плоскости в виде точек. Реализация данного метода в Python-библиотеке umap позволяет разметить точки по заданным категориям, в данном случае по жанрам.

Значения гиперпараметров UMAP были подобраны вручную: 50 соседей с минимальным расстоянием 0,5. В качестве меры близости рассматривались три метрики: расстояние Евклида, расстояние Чебышёва и корреляция. Последняя позволила получить самые наглядные кластеры, представленные на графиках на рисунке.

Наиболее хорошо тексты были кластеризованы при помощи характеристик уровня символов и комбинации всех трёх уровней, наименее хорошо — уровня структуры. Рекламные тексты (ad) формируют много отдельных маленьких кластеров, кластеры остальных жанров накладываются друг на друга, но тем не менее, их можно выделить. Следует отметить, что вектора содержат по несколько десятков характеристик, так что упрощённое двумерное представление текстов показало, что жанры могут отделяться друг от друга на основе выбранных характеристик.

Для того чтобы подтвердить результаты кластеризации, было решено валидировать жанры, то есть провести восемь экспериментов по классификации. В каждом эксперименте выбирался жанр, и тексты классифицировались как принадлежащие данному жанру или нет.

Исходными данными выступали те же вектора характеристик всех трёх уровней, так как для них кластеризация показала лучшие результаты. В качестве алгоритма классификации была выбрана рекуррентная нейронная сеть со слоем двунаправленной LSTM с 64 блоками и выходным слоем из 8 нейронов, использующим сигмоиду как функцию активации Softmax для бинарной классификации и Adam как оптимизатор. Количество эпох обучения составляло 25, размер батча был равен 20. Такая нейронная сеть является одним из самых популярных классификаторов в компьютерной лингвистике.

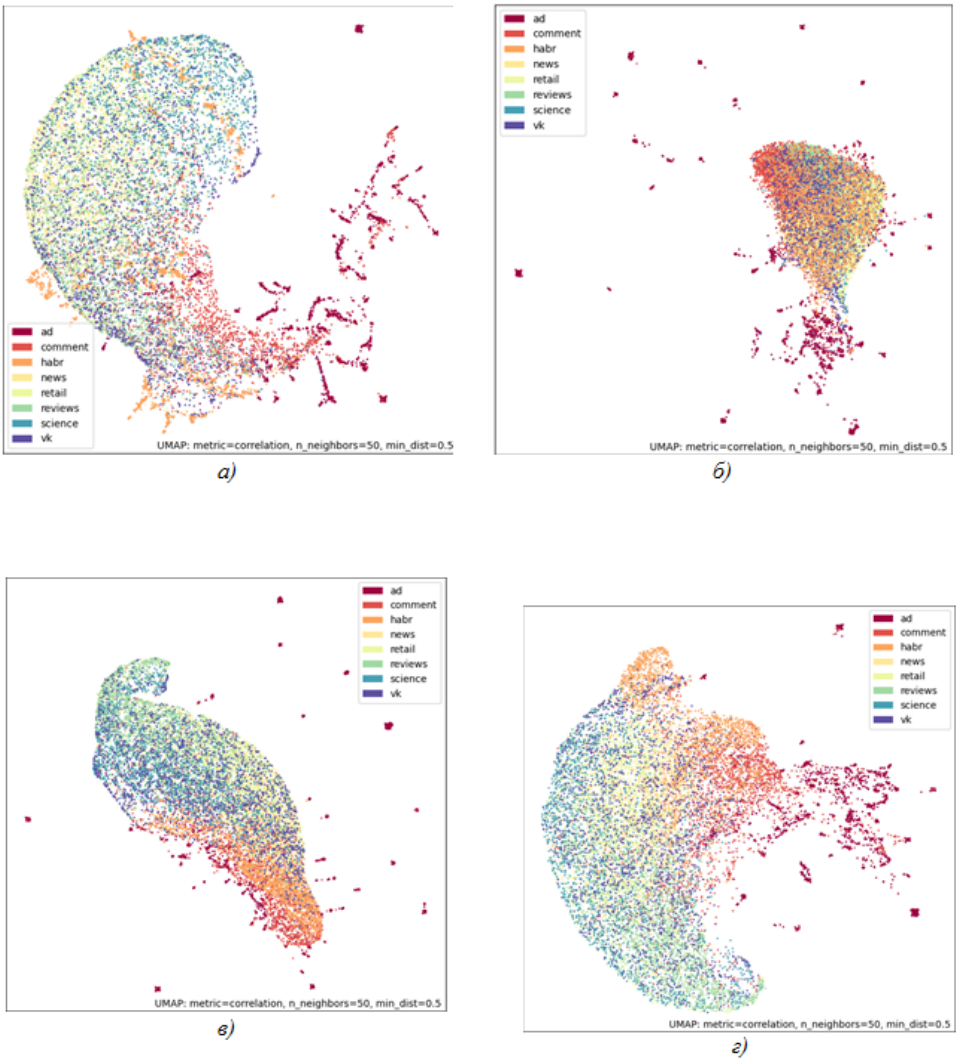

Рисунок – Кластеризация текстов различных жанров с помощью UMAP, характеристики уровня: а) символов, б) структуры, в) ритма, г) всех трёх

Для классификации корпус текстов был разделён случайным образом на обучающую, валидационную и тестовую выборки в отношении 60%:20%:20%. На основе качества классификации валидационной выборки были подобраны описанные выше гиперпараметры, приводящие к лучшим результатам на данном корпусе. На тестовой выборке был проведён итоговый эксперимент, результаты которого описаны в таблице.

Таблица – Верификация текстов различных жанров с помощью лингвистических характеристик трёх уровней

Жанр	Точность	Полнота	F-мера
Хабр	96.7	95.2	96.0
Научные статьи	94.4	95.3	94.8
Описания компаний	93.8	94.0	93.9
Реклама	98.3	99.1	98.7
Новости	95.7	96.2	96.0
ВК	91.6	89.9	90.8
Отзывы	96.9	97.3	97.1
Комментарии	91.9	88.5	90.2
Среднее по жанрам	94.9	94.4	94.7

Оценка качества выполнялась с помощью стандартных метрик: точность, полнота и F-мера. Программные средства для постановки данных экспериментов написаны на языке программирования Python и используют библиотеки Scikit-Learn 1.2.2 и Keras 2.12.0.

Результаты экспериментов показывают, что наиболее хорошо от других жанров отделяются рекламные тексты, что подтверждает результаты кластеризации. В среднем тексты верифицируются с F-мерой 94.7 %, выше среднего верифицируются Хабр-блоги, научные статьи, рекламные и новостные тексты, а также отзывы. Наиболее низко, но тем не менее с достаточно хорошим значением F-меры 90 % верифицируются посты ВК и комментарии в соцсетях. Скорее всего такое качество связано с тем, что эти тексты наиболее разнообразны по авторскому стилю. Для остальных жанров, кроме отзывов, имеются профессиональные правила касательно стилистики: научные, научно-популярные, журналистские или маркетинговые. Отзывы собраны с общего сайта и посвящены одной тематике — фильмам — и для них в русском языке также имеются общепринятые правила написания, что и обуславливает высокое качество верификации — 97.1 %.

Таким образом, исследование показывает, что лингвистические характеристики позволяют верифицировать жанры русскоязычных Интернет-текстов с высоким качеством. Направлением для будущих исследований может быть верификация и классификация жанровых текстов при помощи других моделей, например, нейросетевых.

Работа поддержана стипендией Президента Российской Федерации для молодых ученых и аспирантов, осуществляющих перспективные научные исследования и разработки по приоритетным направлениям модернизации российской экономики: № СП-2109.2021.5.

Список использованных источников

1. Onan A. An ensemble scheme based on language function analysis and feature engineering for text genre classification //Journal of Information Science. – 2018. – Vol. 44. – №. 1. – P. 28-47.

2. Kuzman T., Rupnik P., Ljubešić N. The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild //Proceedings of the Thirteenth Language Resources and Evaluation Conference. – 2022. – P. 1584-1594.

3. Toshevska M., Gievska S. A Review of Text Style Transfer Using Deep Learning //IEEE Transactions on Artificial Intelligence. – 2022. – Vol. 3. – №. 05. – P. 669-684.

4. Lagutina K. Authorship verification of literary texts with rhythm features /K. Lagutina [et al.]. – Proceedings of the 28th Conference of Open Innovations Association FRUCT. – IEEE, 2021. – P. 240-251.