УДК 004.056 doi: 10.15827/0236-235X.142.590-599 2023. Т. 36. № 4. С. 590–599

Идентификация фейк-новостей с помощью веб-ресурса на основе нейронных сетей

М.В. Тумбинская
Р.А. Галиев

Ссылка для цитирования
Тумбинская М.В., Галиев Р.А. Идентификация фейк-новостей с помощью веб-ресурса на основе нейронных сетей // Программные продукты и системы. 2023. Т. 36. № 4. С. 590–599. doi: 10.15827/0236-235X.142.590-599
Информация о статье
Поступила в редакцию: 01.06.2023 После доработки: 14.07.2023 Принята к публикации: 31.07.2023

Аннотация. Активное распространение информации средствами телекоммуникаций обусловило актуальность проблемы идентификация фейк-новостей на основе их заголовков. Целью данного исследования является повышение уровня достоверности информации пользователей в браузерной среде. Достижение ее возможно путем автоматизированного распознавания фейк-новостей на основе их заголовков в браузерной среде. В работе предложено использовать нейронные сети, которые позволяют проводить семантический анализ заголовков фейк-новостей и обрабатывать большие объемы данных. В статье рассмотрены популярные зарубежные и отечественные веб-ресурсы по идентификации фейк-новостей, их отличительные особенности. Проведены анализ и выбор нейросетевых моделей BERT-MLP, BERT-CNN, BERT-LSTM, которые были заложены в основу разработки веб-ресурса. Для обучения нейросетевых моделей использованы выборки русскоязычных заголовков новостей. Проведено экспериментальное исследование оценки качества и адекватности нейросетевых моделей на тестовых и валидационных выборках данных, в том числе на новостных заголовках различных тематик. Исходя из результатов оценки, разработанный веб-ресурс, в основу которого заложены модели BERT-CNN, BERT-GRU, BERT-LSTM, показал высокую результативность. Достоверность идентификации фейк-новостей составила 86,29 %. Практическая ценность работы заключается в том, что созданный веб-ресурс может быть использован в качестве инструмента для распознавания фейк-новостей специалистами по противодействию деструктивным политическим технологиям, а также специалистами по информационной безопасности. Для повышения точности классификации целесообразно расширить обучающую выборку данных, что позволит выявлять в данных более сложные зависимости.

Ключевые слова: фейк-новость, нейросетевая модель, BERT, CNN, GRU, LSTM, заголовок новости, веб-ресурс, нелегитимные новости

Введение. В современном мире информация является неотъемлемой частью жизни людей. Ее активное распространение происходит всеми возможными источниками: телевидение, радио, газеты, журналы, новостные сайты, социальные сети. Однако не вся информация достоверна. Распространяемые фейк-новости по своей структуре имеют большое сходство с официальными, однако их суть искажена, и они могут нанести серьезный ущерб как отдельным людям, вводя их в заблуждение, так и обществу в целом, создавая напряженность, влиять на общественное мнение, вызывать панику и хаос.

Так, весной 2020 года появились фейк-новости о распространении COVID-19 по сетям связи 5G. Это привело к последствиям в социальной сфере – часть вышек этих сетей в Великобритании были атакованы, вследствие чего у некоторых людей сложилось негативное мнение о данной технологии. Примером может служить и распространение фейк-новостей в процессе выборов в США в 2016 году. Под воздействием дезинформации в социальных сетях некоторые избиратели отдавали предпочтение одному из кандидатов. Таким образом, для общества и государства фейк-новости представляют потенциальную и реальную угрозу, поскольку в руках злоумышленника они являются высокоэффективным инструментом социального и психологического воздействия на человека, а также извлечения собственной выгоды.

Идентификация фейк-новостей достаточно изучена, но остается актуальной задачей.

Для борьбы с распространением фейк-новостей и снижения их воздействия создаются специальные веб-ресурсы, предназначенные для идентификации таких новостей. Они удобны и не требуют от пользователя каких-либо специальных знаний и навыков.

Предметом данного исследования являются информационные системы идентификации фейк-новостей. В качестве методов исследования используются нейросетевые технологии и машинное обучение. Ручной способ проверки новостей на основе их заголовков довольно трудоемкий, в связи с чем целесообразна разработка информационных систем идентификации фейк-новостей на основе нейросетевых технологий. В работе предложен веб-ресурс, в который заложены гибридные нейросетевые

модели BERT-MLP, BERT-CNN, BERT-LSTM, позволяющий в автоматизированном режиме идентифицировать фейк-новость по заголовку. Практическая значимость заключается в том, что разработанный веб-ресурс может быть использован в качестве инструмента для распознавания фейк-новостей специалистами по противодействию деструктивным политическим технологиям, а также специалистами по информационной безопасности. Для повышения точности классификации целесообразно расширить обучающую выборку данных, что позволит выявлять более сложные зависимости в данных.

Анализ веб-ресурсов по идентификации фейк-новостей

Для анализа существующих решений были выбраны десять популярных веб-ресурсов: ChatGPT, «Война с фейками», FactCheck, Politifact, Snopes, «Лапша Медиа», Oigetit Fake News Filter, Alt News, Full Fact, AFP Fact Check. Результаты сравнительного анализа позволяют сделать следующие выводы.

1) Проверку фейк-новостей на основании экспертной оценки осуществляют 80 % проанализированных веб-ресурсов. Это может привести к тому, что в зависимости от страны, в которой функционирует веб-ресурс, оценка может быть пристрастной, группа экспертов может не иметь доступа ко всей информации, необходимой для корректной оценки новости, оценка новости может приводить к временным задержкам в силу специфики ее выполнения экспертным методом.

2) Все проанализированные веб-ресурсы имеют возможность ввода требуемой для проверки заголовка новости, однако только 10 % рассмотренных веб-ресурсов позволяют пользователю получить достоверный результат оценки за короткое время.

3) Проверку русскоязычных новостей не поддерживают 70 % проанализированных веб-ресурсов. Это может вызвать затруднения у россиян при желании проверить новость на фейк.

4) Возможность определить уровень достоверности новости в процентах имеют только 20 % рассмотренных веб-ресурсов. Причем данная функция может дать более информативную оценку новости для пользователя по сравнению с классификационной оценкой.

5) Лишь 20 % проанализированных веб-ресурсов используют нейросетевые технологии для идентификации фейк-новостей. Применение данных технологий позволяет за короткий срок сформировать результат оценки введенного пользователем заголовка новости. Следует отметить, что использование комбинации работы нескольких нейронных сетей повышает точность и достоверность оценки фейк-новостей.

Анализ показал необходимость разработки нового веб-ресурса по идентификации заголовков фейк-новостей в браузерной среде, который обеспечивал бы возможность проверки русскоязычных фейк-новостей с отображением классификационной оценки отнесения заголовка новости к категории «истина» или «ложь» и с детализацией уровня достоверности новостного заголовка в процентах. Идентификация заголовков новостей должна осуществляться на основе интеграционной оценки работы нейронных сетей.

Анализ нейросетевых моделей для идентификации фейк-новостей

Поскольку идентификацию фейк-новостей можно интерпретировать как задачу бинарной классификации текстовой информации («истина» или «ложь»), для ее решения возможно использование алгоритмов машинного обучения с учителем средствами нейросетевых моделей [1]. Для формирования требований к выбору нейросетевых моделей данного типа рассмотрим исследования в предметной области. В работах [2–4] был проведен сравнительный анализ эффективности рекуррентных (RNN) и сверточных (CNN) нейронных сетей, а также нейронных сетей на основе архитектуры трансформеров (Transformer based) для решения задачи идентификации англоязычных фейк-новостей на тему COVID-19. Анализ работ [5–7] показал, что наилучшими нейросетевыми моделями являются модель BERT (точность оценки 97,71 %), а также гибридная модель BERT-CNN (точность оценки 97,52 %). В работе [6] предложена модель классификации текстовой информации на английском языке, комбинирующая результаты работы трех нейросетевых гибридных моделей. В первой модели использованы нейронная сеть с архитектурой трансформеров BERT и рекуррентная нейронная сеть с долгой краткосрочной памятью LSTM. Во второй модели также использован BERT, но уже с применением двух параллельных нейронных сетей LSTM (BiLSTM). Третья модель сформирована из модели BERT

и рекуррентной нейронной сети с управляемыми блоками GRU. Сделан вывод о том, что комбинированный подход имеет лучшую точность оценки (89,81 %) в сравнении с используемыми по отдельности моделями и другими методами машинного обучения. В [7] исследовано применение различных методов машинного обучения и нейронных сетей при обнаружении англоязычных заголовков фейк-новостей в социальных сетях. Показано, что наилучшими нейросетевыми моделями являются гибридная модель BERT-LSTM (точность 97,55 %) и модель BERT-CNN, именуемая в данной статье FakeBERT (98,90 %).

Таким образом, для эффективной идентификации фейк-новостей на основе их семантического анализа необходимо использовать следующие гибридные нейросетевые модели:

– BERT-CNN (основана на архитектуре трансформера BERT, а также сверточной нейронной сети CNN);

– BERT-GRU (основана на модели BERT, а также рекуррентной нейронной сети с управляемыми блоками GRU);

– BERT-LSTM (рекуррентная нейронная сеть с долгой краткосрочной памятью LSTM) [8, 9].

Кроме этого, для достижения лучших результатов необходимо использовать механизм комбинирования результатов работы представленных нейросетевых моделей.

Разработка веб-ресурса по идентификации фейк-новостей на основе нейронных сетей

Веб-ресурс был реализован в виде клиент-серверного веб-приложения. Разработка веб-ресурса была проведена в три этапа. На первом этапе разрабатывался веб-интерфейс. На втором – элементы серверной части, отвечающие за взаимодействие с клиентской частью, – маршруты и обработчики маршрутов Flask. На третьем этапе осуществлялись разработка и программная реализация модулей серверной части, отвечающих за обработку введенной пользователем информации и выдачу результатов классификации.

Взаимодействие клиентской и серверной частей построено таким образом, что после получения новостной ссылки клиентская часть отправляет запрос на серверную, в которой происходит ее дальнейшая обработка. В результате формируется классификационная оценка – отнесение обнаруженного заголовка новости к категории «истина» или «ложь» с детализацией уровня достоверности новостного заголовка в процентах. Результаты классификации отправляются на клиентскую часть в виде ответа и отображаются в веб-интерфейсе пользователя.

Серверная часть имеет последовательную архитектуру и включает в себя пять основных модулей.

1. *Модуль проверки корректности данных*. Используется для проверки введенной ссылки на соответствие общей структуре веб-ссылок. Кроме этого, модуль проверяет доступность источника ссылки путем отправки GET-запроса на получение содержимого страницы и удовлетворительного ответа от сервера источника. Это необходимо для общей проверки введенных пользователем данных, а также возможности дальнейшего извлечения заголовка новости. Если данные некорректны, модуль сообщает об этом и требует ввода повторно.

2. *Модуль проверки источника в реестре СМИ*. Необходим для проверки присутствия источника ссылки в списке зарегистрированных в РФ. Это позволяет пользователю получить дополнительную информацию о достоверности источника введенной ссылки. Проверка осуществляется путем извлечения доменного имени с введенной ссылки в виде текстовой строки и поиска данной строки в списке зарегистрированных СМИ. Список представляет собой файл с расширением .csv, который был предварительно сформирован путем сбора данных с официального реестра Роскомнадзора. В таблице присутствует информация об источниках, относящихся к одной из категорий и имеющих действующую лицензию на осуществление деятельности в области СМИ: сетевое издание, электронное периодическое издание, информационное агентство.

3. *Парсер заголовка новости, или модуль извлечения новостного заголовка*. Выполняет функцию извлечения заголовка новости с введенной ссылки на новость с веб-ресурса, социальной сети VK или мессенджера Telegram. Для извлечения новостного заголовка под каждый из указанных источников разработаны отдельные функции. Вследствие ограничения после извлечения заголовка новости данная строка проверяется на наличие слов, не являющихся русскими. Если такие слова обнаружены, модуль сообщает о их наличии и требует ввода данных повторно.

4. *Препроцессор, или модуль предварительной обработки новостного заголовка*. Выполняет операции преобразования текстового заголовка новости в формат, необходимый для

использования классификатора. К функциям препроцессора относятся:

– удаление пробелов, знаков препинания, а также спецсимволов;

– приведение всех слов заголовка новости к нижнему регистру;

– токенизация: разбиение заголовка новости на отдельные части (токены); добавление в начало и в конец последовательности специальных токенов: [CLS] для обозначения начала предложения (численное представление – 101) и [SEP] для обозначения конца последовательности (численное представление – 102); присвоение каждому токену специального числового идентификатора; формирование бинарной маски заголовка новости для уточнения длины заголовка и корректной работы классификатора.

Данный модуль возвращает последовательность числовых идентификаторов каждого токена и бинарную маску, которые далее передаются на вход классификатора.

5. *Классификатор, или модуль комбинирования результатов работы нейросетевых моделей*. Данный модуль предназначен для формирования классификационной оценки отнесения заголовка новости к категории «истина» или «ложь» с детализацией уровня достоверности новостного заголовка в процентах. В основе данного модуля – механизм комбинирования результатов работы гибридных нейросетевых моделей BERT-GRU, BERT-CNN и BERT-LSTM. Каждая из этих моделей обучена на специально подготовленной выборке данных, включающей как легитимные, так и нелегитимные заголовки новостей на русском языке.

В качестве входных данных классификатор получает предварительно обработанные данные из препроцессора, которые затем используются для решения задачи бинарной классификации (0 – «истина», 1 – «ложь»). Выходными выступают числовые вероятности бинарной классификации, лежащие в диапазоне от 0 до 1, над которыми затем выполняется операция усреднения. Если итоговая оценка находится в диапазоне от 0 до 0,5, заголовок относится к категории «истина», в противном случае – «ложь». Кроме этого, классификатор использует данную числовую оценку для детализации уровня достоверности новостного заголовка в процентах.

После того как серверная часть обработает запрос от клиентской части, она формирует ответ, который содержит данные о результатах классификации, и отправляет его на клиентскую часть.

Разработка клиентской части веб-ресурса. Пользователь имеет возможность проверить заголовок новости на фейк путем ввода в соответствующее текстовое поле ссылки (URL) на новость с веб-ресурса, социальной сети VK или мессенджера Telegram. Чтобы отправить введенную ссылку в виде HTTP-запроса POST на серверную часть и запустить алгоритм проверки заголовка новости на фейк, пользователю необходимо нажать на кнопку «Проверить».

Для отображения отрицательного ответа от сервера в процессе обработки введенного заголовка предусмотрено специальное поле, в котором в текстовом формате выводится одно из следующих сообщений: «Введена некорректная ссылка», «Веб-страница недоступна или не отвечает», «Обнаружены иностранные слова». Иначе в данном поле отображается: «Источник является официальным СМИ», то есть доменное имя источника ссылки присутствует в списке СМИ, зарегистрированных в Роскомнадзоре РФ, «Источник не является официальным СМИ» – в противном случае. После получения положительного ответа от серверной части предусмотрено текстовое поле для отображения результатов классификационной оценки отнесения обнаруженного заголовка новости к категории («истина» или «ложь»), с детализацей уровня достоверности заголовка новости в процентах. Кроме того, для детализации результатов работы каждой нейросетевой модели в отдельности рядом с кнопкой «Проверить» появляется кнопка «Подробнее». При нажатии на данную кнопку отображаются три текстовых поля, содержащих результаты работы нейросетевых моделей: BERT-CNN, BERT-GRU, BERT-LSTM (http://www.swsys.ru/uploaded/image/2023-4/1.jpg).

Разработка серверной части веб-ресурса. Для создания модуля взаимодействия клиентской и серверной частей разрабатываемого веб-ресурса был использован фреймворк Flask [10] на языке программирования Python. В процессе программной реализации серверной части были подключены и использованы следующие модули Python:

```
import re
import numpy as np
import pandas as pd
import requests
import vk_api
import tensorflow as tf
from urllib.parse import urlparse
from newspaper import Article
from bs4 import BeautifulSoup
```

```
   from flask import Flask, render_tem-
plate, request, jsonify
   from selenium import webdriver
   from selenium. webdriver.common.by
import By
   from transformers import TFBert-
Model, BertTokenizer
   from keras.models import load_model
   from keras.callbacks import Ear-
lyStopping
   from sklearn.model_selection import
train_test_split
```

Чтобы серверная часть смогла получать и обрабатывать запросы пользователя на получение содержимого веб-ресурса, а также введенную пользователем новостную ссылку, был инициализирован объект Flask с указанием маршрутов и обработчиков:

– '/': отображает главную страницу (веб-интерфейс), когда пользователь пытается получить доступ к веб-ресурсу (метод запроса GET); при использовании данного маршрута выполняется обработчик, который возвращает на клиентскую часть содержимое HTML-страницы;

– '/check_web_link': обрабатывает событие нажатия пользователем кнопки «Проверить» (метод запроса POST); связанный обработчик получает введенный пользователем текст, а затем последовательно выполняет запуск модулей проверки корректности данных и проверки источника в реестре СМИ; результаты выполнения функции отправляются на клиентскую часть в виде ответа в формате JSON;

– '/predict': обрабатывает событие нажатия пользователем кнопки «Проверить» (метод запроса POST), когда работа вышеизложенных модулей завершается успешно; обработчик получает введенный пользователем текст, а затем выполняет последовательный запуск модулей «Парсер» заголовка новости», «Препроцессор» и классификатора, результаты выполнения функции отправляются на клиентскую часть в виде ответа в формате JSON.

Сбор исходных данных для обучения нейросетевых моделей. Для обучения нейросетевых моделей, используемых в основе классификатора, были собраны как легитимные, так и нелегитимные заголовки новостей на русском языке. Сбор легитимных новостных заголовков осуществлялся с официальных новостных веб-ресурсов: сетевых изданий «Российская Газета», «РИА Новости», «Реальное время», IT-World.ru «РБК», информационного агентства «ТАСС», портала новостей «Новости Mail.ru», электронного периодического издания «БИЗНЕС Online».

Сбор нелегитимных новостных заголовков осуществлялся с различных веб-ресурсов и каналов мессенджера Telegram, в которых публикуются как сатирические, так и широко распространенные фейк-новости из Интернета. Перечень источников нелегитимных новостей: каналы мессенджера Telegram «Fake Control» и «Осторожно, ФЕЙК», веб-ресурсы «Война с фейками», «Лапша Медиа», «Панорама».

В основе функций сбора данных используются возможности подключаемых модулей: Selenium – для выполнения автоматизированного сбора данных, Pandas – для сохранения полученных данных в формат pandas.DataFrame. В случае сбора данных с общедоступных каналов мессенджера Telegram дополнительно используются возможности подключаемых модулей Requests для получения содержимого каналов в веб-версии мессенджера Telegram в формате .html и BeautifulSoup4 для обработки полученных данных и извлечения новостных заголовков.

Все собранные заголовки сохраняются в файле формата .csv. В данном файле содержится 40 000 новостных заголовков, в которых равное количество легитимных и нелегитимных заголовков новостей – по 20 000 каждого типа.

Общая структура нейросетевых моделей. В основе классификатора новостных заголовков лежат гибридные нейросетевые модели BERT-CNN, BERT-GRU, BERT-LSTM, которые выполняют классификационную оценку заголовков на основе двух категорий – «истина» и «ложь». Структура, а также значения гиперпараметров гибридных нейросетевых моделей были заимствованы в работах [1, 6].

Рассмотрим общую структуру используемых нейросетевых моделей.

Входными данными являются последовательность числовых идентификаторов новостного заголовка и бинарная маска новостного заголовка.

Каждая из гибридных нейросетевых моделей использует нейросетевой слой, основанный на архитектуре трансформеров BERT, для создания векторного представления каждого числового идентификатора в заголовке новости с учетом бинарной маски для уточнения длины заголовка. Для улучшения точности и сокращения времени обучения каждая из нейросетевых моделей использует заранее предобученную нейросетевую модель BERT, предобученную на русском корпусе слов – sberbank-ai/ruBert-base (https://huggingface.co/ai-forever/ruBert-base).

Полученные векторные представления затем передаются в нейросетевые слои в зависимости от конкретной гибридной модели (CNN, GRU или LSTM) для обнаружения характерных признаков каждой из категорий новостных заголовков. Таким образом, гибридные модели используют семантический анализ для выявления фейк-новостей.

Далее для обработки полученных из предыдущего слоя признаков используется полносвязный нейросетевой слой. Важно отметить, что первоначально в модели BERT-CNN этот слой отсутствовал, однако был добавлен вследствие улучшения точности распознавания обеих категорий новостных заголовков. В моделях BERT-GRU и BERT-LSTM [4] он присутствует изначально и остается неизменным.

Кроме этого, поскольку нейросетевые модели должны относить новостные заголовки к одной из двух категорий, ключевой задачей является выполнение бинарной классификации заголовков новостей (0 – «истина», 1 – «ложь»). Для этого в конце каждой модели используется полносвязный нейросетевой слой, содержащий один нейрон, выполняющий данную задачу.

В качестве выходных данных выступают числовые вероятности бинарной классификации, лежащие в диапазоне от 0 до 1.

Построение и обучение нейросетевых моделей. Для построения и обучения гибридных нейросетевых моделей BERT-CNN, BERT-GRU и BERT-LSTM используется следующие подключаемые модули: Transformers для использования нейросетевого слоя BERT, Tensorflow для построения общей архитектуры, обучения и сохранения нейросетевых моделей, Scikit-learn для разделения исходных данных на тренировочную и валидационную выборки.

Поскольку используемые нейросетевые модели могут обрабатывать данные только одинакового размера, в качестве максимальной длины новостного заголовка было выбрано значение 21 (рис. 1). Выбор осуществлялся с учетом необходимости минимизации длины заголовка для ускорения обучения нейросетевых моделей. Кроме этого, 97,35 % всех новостных заголовков имеют длину, не превышающую 21 слово.

Следует отметить, что в программной реализации нейросетевых моделей установлена максимальная длина последовательности, равная 23 токенам. Это связано с тем, что в процессе токенизации в начале и в конце обработанной последовательности добавляются

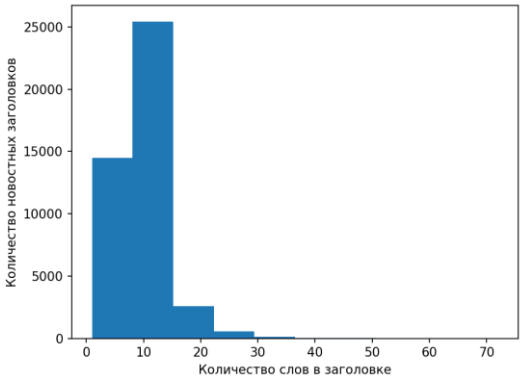


Рис. 1. Гистограмма распределения новостных заголовков в зависимости от количества слов

Fig. 1. News headline distribution histogram depending on word number

специальные токены [CLS] (численное представление – 101) и [SEP] (численное представление – 102).

В качестве обучающей выборки случайным образом были выбраны 80 % исходных новостных заголовков. Остальные 20 % использовались как валидационная выборка для оценки точности нейросетевых моделей по окончании каждой эпохи обучения и настройки значений гиперпараметров нейросетевых моделей. Стоит отметить, что при обучении выбранных нейросетевых моделей валидационная выборка использовалась для нахождения оптимального количества эпох обучения. Благодаря этому обучение нейросетевых моделей осуществлялось в течение двух эпох, поскольку именно на третьей эпохе в данных моделях возникает явление «переобучения» [11].

Программная реализация классификатора. В основе работы классификатора лежит принцип комбинирования результатов работы трех обученных гибридных нейросетевых моделей – BERT-CNN, BERT-GRU, BERT-LSTM. Классификатор получает на вход данные от препроцессора, а затем последовательно выполняет запуск функций предсказания каждой нейросетевой модели в отдельности для получения бинарной вероятности отнесения заголовка новости к одной из двух категорий («истина» или «ложь»). После этого классификатор выполняет функцию усреднения для получения итогового числового результата предсказания, лежащего в диапазоне от 0 до 1.

Если итоговая оценка находится в диапазоне от 0 до 0,5, заголовок относится к категории «истина», в противном случае – «ложь». Кроме этого, классификатор использует данную число-

вую оценку для детализации уровня достоверности новостного заголовка (в процентах).

Таким образом, классификатор возвращает значение классификационной оценки с детализацией уровня достоверности новостного заголовка в процентах от каждой нейросетевой модели, включая результат их комбинирования.

Экспериментальное исследование оценки качества разработанного веб-ресурса

Разработанный веб-ресурс предназначен для выполнения классификационной оценки новостных заголовков («истина» или «ложь»). Главным компонентом веб-ресурса, выполняющим данную классификацию, является модуль комбинирования результатов работы гибридных нейросетевых моделей BERT-CNN, BERT-GRU, BERT-LSTM.

Вследствие этого для проведения оценки качества необходимо проверить, насколько точно и адекватно нейросетевые модели, лежащие в основе веб-ресурса, справляются с задачей классификационной оценки легитимных и нелегитимных заголовков новостей. Для этого воспользуемся метриками оценки качества нейросетевых моделей (в процентах), предложенными в работах [1, 6, 12].

Метрика общей точности (Accuracy) отражает долю всех правильно классифицированных новостных заголовков от их общего числа:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

Метрика точности (Precision) отражает долю верно классифицированных нелегитимных новостных заголовков относительно общего числа заголовков новостей, отнесенных нейросетевой моделью к категории «ложь»:

$$Precision = \frac{TP}{TP + FP}.$$

Полнота (Recall) отражает долю верно классифицированных нелегитимных новостей относительно общего числа фактических нелегитимных новостей:

$$Recall = \frac{TP}{TP + FN}.$$

F1-мера (F1) является гармоническим средним между метриками точности и полноты и отражает компромисс между ними:

$$F = \frac{2(Precision \cdot Recall)}{Precision + Recall}.$$

В данных формулах используются следующие обозначения [12]: *TP* (true positive) – количество нелегитимных новостей, которые были правильно отнесены нейросетевой моделью к категории «ложь»; *TN* (true negative) – количество легитимных новостей, которые были правильно отнесены нейросетевой моделью к категории «истина»; *FP* (false positive) – количество легитимных новостей, которые были ошибочно отнесены нейросетевой моделью к категории «ложь»; *FN* (false negative) – количество нелегитимных новостей, которые были ошибочно отнесены нейросетевой моделью к категории «истина».

Результаты общей оценки нейросетевых моделей представлены в таблице. Оценка проводилась на валидационной выборке размером 8 000 новостных заголовков (по 4 000 на каждую категорию), а также на тестовой выборке, сформированной путем сбора 4 000 новостных заголовков (по 2 000 на каждую категорию). Сбор данных для тестовой выборки был сделан спустя две недели после сбора данных исходной (обучающей и валидационной) выборки.

Это необходимо для оценки возможности нейросетевых моделей осуществлять задачу идентификации заголовков, не использованных во время обучения и являющихся случайными для моделей. Стоит отметить, что в таблице также приводится оценка качества классификатора (RuFakeBERT), комбинирующего результаты работ приведенных гибридных моделей.

По итогам оценки разработанный модуль комбинирования результатов работы гибридных нейросетевых моделей BERT-CNN, BERT-GRU, BERT-LSTM имеет наилучшие показатели:

Результаты оценки нейросетевых моделей

Results of evaluating neural network models

Наименоменование модели	Валидационная выборка, %				Тестовая выборка, %			
	Acc	Pr	Rec	F1	Acc	Pr	Rec	F1
BERT-CNN	86,17	83,8	86,13	84,95	85,57	82,1	85,32	83,68
BERT-GRU	86,82	84,79	85,85	85,29	85,97	81,93	85,89	83,91
BERT-LSTM	86,48	83,39	86,51	84,92	86,54	83,81	85,44	84,62
RuFakeBERT	**87,21**	**84,83**	**86,87**	**85,67**	**86,7**	**83,87**	**85,93**	**84,89**

– на тестовой выборке результат общей точности классификатора составляет 87,21 %, F1-меры – 85,67 % (на основе метрик Precision – 84,83 %, Recall – 86,87 %);

– на валидационной выборке результат общей точности классификатора – 86,7 %, F1-меры – 84,89 % (на основе метрик Precision – 84,83 %, Recall – 86,87 %).

Для более детальной оценки качества классификатора был проведен анализ его способности точно идентифицировать каждую из категорий новостных заголовков («истина» или «ложь») отдельно друг от друга. Для этой оценки используются метрики [12]:

– отрицательного предсказанного значения (NPV), которое отражает долю верно классифицированных легитимных новостных заголовков относительно общего числа заголовков новостей, отнесенных моделью к категории «истина»:

$$NPV = \frac{TN}{TN + FN};$$

– положительного предсказанного значения (PPV), которое отражает долю верно классифицированных нелегитимных новостей относительно общего числа заголовков новостей, отнесенных моделью к категории «ложь»:

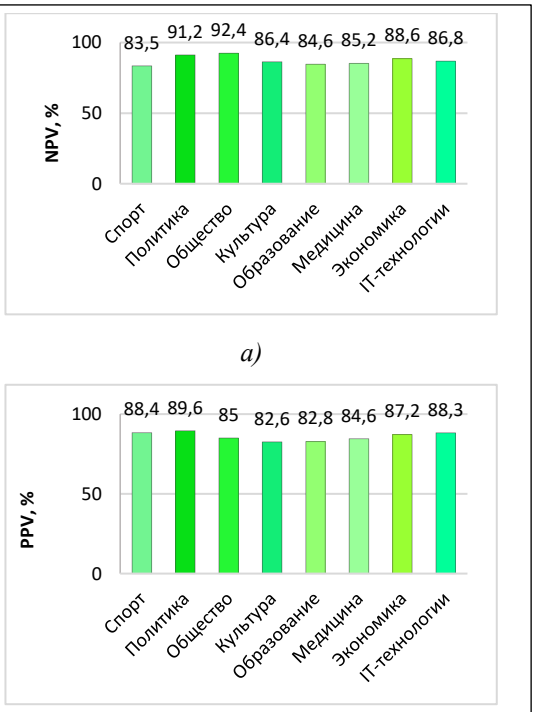
$$PPV = \frac{TP}{TP + FP}.$$

Для детализационной оценки качества классификатора как тестовая выборка легитимных и нелегитимных данных были выбраны новостные заголовки на различные темы: спорт, политика, общество, культура, медицина, образование, экономика, IT-технологии.

Детализационная оценка была проведена на тестовой выборке, состоящей из 4 000 новостных заголовков, по 500 заголовков каждой тематики. В каждой тематике было использовано равное количество легитимных и нелегитимных новостных заголовков. Результаты оценки представлены на диаграммах (рис. 2).

Таким образом, можно сделать вывод о том, что разработанный модуль более качественно и адекватно классифицирует легитимные новости на тему политики (89,6 %), а также нелегитимные новости на тему общества (92,4 %). Кроме этого, было выявлено, что самые низкие результаты оценки классификатор выдает при оценке легитимных новостных заголовков на тему культуры (82,6 %), а также нелегитимных спортивных новостей (83,5 %).

Для формирования итоговой оценки качества воспользуемся усредненными результатами нейросетевых моделей, а также классификатора по всем использованным метрикам: BERT-CNN – 85,09 %, BERT-GRU – 85,49 %, BERT-LSTM – 85,64 %, RuFakeBERT – 86,29 %.



а)

б)

Рис. 2. Гистограмма оценки качества классификатора на выборке легитимных новостных заголовков: а) по метрике NPV; б) по метрике PPV

Fig. 2. Histogram of classifier quality assessment on a sample of legitimate news headlines: a) NPV metric; b) PPV metric

Как видно, разработанный классификатор (RuFakeBERT) показал улучшение результатов оценки качества по сравнению с нейросетевыми моделями BERT-CNN, BERT-GRU и BERT-LSTM на 1,2 %, 0,8 % и 0,65 % соответственно.

Для повышения точности классификации целесообразно расширить обучающую выборку данных. Это позволит нейросетевым моделям, используемым в модуле классификатора, выявлять более сложные зависимости в данных. Такой подход улучшит способность моделей к обобщению и повысит точность предсказаний для свежих новостных заголовков.

Заключение

В работе представлен обзор решений в области идентификации фейк-новостей, предложен веб-ресурс по идентификации заголовков

фейк-новостей на русском языке на основе интеграционной работы нейронных сетей с отображением классификационной оценки отнесения заголовка новости к категории «истина» или «ложь», с детализацией уровня достоверности новостного заголовка в процентах, проведено экспериментальное исследование. Новизной является программная реализация инструмента идентификации фейк-новостей.

Результаты исследований показали, что на тестовой выборке общая точность классификатора составила 87,21 %, F1-меры – 85,67 %, на валидационной выборке – соответственно 86,7 % и 84,89 %. Была проведена дополнительная оценка качества классификатора на каждой из категорий новостных заголовков («истина» или «ложь») отдельно друг от друга с использованием метрик NPV и PPV. Так, данный модуль лучше всего справляется с классификацией легитимных политических новостных заголовков (89,6 %), а также нелегитимных общественных новостей (92,4 %). Худшие результаты модель показала при классификации легитимных новостей о культуре (82,6 %) и нелегитимных спортивных новостей (83,5 %). Разработанный классификатор RuFakeBERT показал улучшение результатов оценки качества по всем использованным метрикам по сравнению с нейросетевыми моделями BERT-CNN, BERT-GRU и BERT-LSTM на 1,2 %, 0,8 % и 0,65 % соответственно. Итоговая оценка качества классификатора составила 86,29 %.

Экспериментальные исследования показали необходимость увеличения объема выборки для обучения нейросетевых моделей для более качественного анализа данных и достоверного процесса идентификации. В дальнейшем планируются разработка системы поддержки принятия решений для автоматического распознавания фейк-новостей и соответствующая доработка веб-ресурса.

Список литературы

1. Alghamdi J., Lin Y., Luo S. A comparative study of machine learning and deep learning techniques for fake news detection. Information, 2022, vol. 13, no. 12, art. 576. doi: 10.3390/info13120576.

2. Мухаматханов Р.М., Михайлов А.А., Баянов Б.И., Тумбинская М.В. Классификация DDos-атак на основе нейросетевой модели // Прикладная информатика. 2019. Т. 14. № 1. С. 96–103.

3. Wang H., Tian K., Zhengjiang Wu, Wang L. A short text classification method based on convolutional neural network and semantic extension. IJCIS, 2021, vol. 14, no. 1, pp. 367–375. doi: 10.2991/ijcis.d.201207.001.

4. Тумбинская М.В., Баянов Б.И., Рахимов Р.Ж., Кормильцев Н.В., Уваров А.Д. Анализ и прогнозирование вредоносного сетевого трафика в облачных сервисах // Бизнес-информатика. 2019. Т. 13. № 1. С. 71–81. doi: 10.17323/1998-0663.2019.1.71.81.

5. Тумбинская М.В. Анализ и прогнозирование сетевых атак типа SYN-FLOOD на web-ресурсы // Информатизация образования и науки. 2018. № 4. С. 61–68.

6. Tan K.L., Lee C.P., Lim K.M., Anbananthen K.S.M. Sentiment analysis with ensemble hybrid deep learning model. IEEE Access, 2022, vol. 10, pp. 103694–103704. doi: 10.1109/ACCESS.2022.3210182.

7. Kaliyar R.K., Goswami A., Narang P. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 2021, vol. 80, no. 8, pp. 11765–11788. doi: 10.1007/s11042-020-10183-2.

8. Xing Y., Xiao C. A GRU model for aspect level sentiment analysis. J. Phys.: Conf. Ser., 2019, vol. 1302, art. 032042. doi: 10.1088/1742-6596/1302/3/032042.

9. Triyadi I., Prasetiyo B., Nikmah T.L. News text classification using Long-Term Short Memory LSTM algorithm. J. Soft Comput. Explor., 2023, vol. 4, no. 2, pp. 79–86. doi: 10.52465/joscex.v4i2.136.

10 Grinberg M. Flask Web Development: Developing Web Applications with Python. CA, Sebastopol, O'Reilly Media Publ., 2018, 292 p.

11. Ying X. An overview of overfitting and its solutions. J. Phys.: Conf. Ser., 2022, vol. 1168, art. 022022. doi: 10.1088/1742-6596/1168/2/022022.

12. De Diego I.M., Redondo A.R., Fernandez R.R. General performance score for classification problems. Applied Intelligence, 2022, vol. 52, no. 10, pp. 12049–12063. doi: 10.1007/s10489-021-03041-7.

Software & Systems doi: 10.15827/0236-235X.142.590-599 2023, vol. 36, no. 4, pp. 590–599

Neural network-based approach for identifying fake news

Marina V. Tumbinskaya
Renat A. Galiev

For citation
Tumbinskaya, M.V., Galiev, R.A. (2023) ‘Neural network-based approach for identifying fake news’, *Software & Systems*, 36(4), pp. 590–599 (in Russ.). doi: 10.15827/0236-235X.142.590-599

Article info
Received: 01.06.2023 After revision: 14.07.2023 Accepted: 31.07.2023

Abstract. Currently, telecommunication facilities actively disseminate news. Identification of fake news based on their headlines has become a relevant problem. The purpose of this study is to increase a user information reliability level in browser environment. Automated recognition of fake news based on their headlines in the browser environment can achieve this goal. The paper proposes using neural networks that allow analyzing fake news headlines semantically and processing large amounts of data. The article examines popular foreign and domestic web resources for identifying fake news, their distinctive features. The authors carried out an analysis and selected neural network models BERT-MLP, BERT-CNN, BERT-LSTM to be a basis for developing a web resource. They used samples of Russian-language news headlines to train neural network models. An experimental study was carried out to assess the quality and adequacy of neural network models on test and validation data samples including news headlines on various topics. The developed web resource based on the BERT-CNN, BERT-GRU, BERT-LSTM models showed good results. The reliability of identifying fake news was 86,29%. The practical value of the work is in the fact that the developed web resource can be used as a tool for recognizing fake news by specialists in counteracting destructive political technologies, as well as information security specialists. To improve classification accuracy, it is advisable to expand a training data set, which allows identifying more complex dependencies in data.
Keywords: fake news, neural network model, BERT, CNN, GRU, LSTM, news headline, illegitimate news

References

1. Alghamdi, J., Lin, Y., Luo, S. (2022) 'A comparative study of machine learning and deep learning techniques for fake news detection', *Information*, 13(12), art. 576. doi: 10.3390/info13120576.
2. Mukhamatkhanov, R., Mikhailov, A., Bayanov, B., Tumbinskaya, M. (2019) 'Classification of DDoS-attacks based on the neural network model', *J. of Applied Informatics*, 14(1), pp. 96–103 (in Russ.).
3. Wang, H., Tian, K., Zhengjiang, Wu, Wang, L. (2021) 'A short text classification method based on convolutional neural network and semantic extension', *IJCIS*, 14(1), pp. 367–375. doi: 10.2991/ijcis.d.201207.001.
4. Tumbinskaya, M.V., Bayanov, B.I., Rakhimov, R.Zh., Kormiltcev, N.V., Uvarov, A.D. (2019) 'Analysis and forecast of undesirable cloud services traffic', *Business Informatics*, 13(1), pp. 71–81 (in Russ.). doi: 10.17323/1998-0663. 2019.1.71.81.
5. Tumbinskaya, M.V. (2018) 'Analysis and forecasting of SYN-FLOOD network attacks to web-resources', *Informatization of Education and Sci.*, (4), pp. 61–68 (in Russ.).
6. Tan, K.L., Lee, C.P., Lim, K.M., Anbananthen, K. (2022) 'Sentiment analysis with ensemble hybrid deep learning model', *IEEE Access*, 10, pp. 103694–103704. doi: 10.1109/ACCESS.2022.3210182.
7. Kaliyar, R.K., Goswami, A., Narang, P. (2021) 'FakeBERT: Fake news detection in social media with a BERT-based deep learning approach', *Multimedia Tools and Applications*, 80(8), pp. 11765–11788. doi: 10.1007/s11042-020-10183-2.
8. Xing, Y., Xiao, C. (2019) 'A GRU model for aspect level sentiment analysis', *J. Phys.: Conf. Ser*., 1302, art. 032042. doi: 10.1088/1742-6596/1302/3/032042.
9. Triyadi, I., Prasetiyo, B., Nikmah, T.L. (2023) 'News text classification using Long-Term Short Memory (LSTM) algorithm', *J. Soft Comput. Explor.*, 4(2), pp. 79–86. doi: 10.52465/joscex.v4i2.136.
10. Grinberg, M. (2018) *Flask Web Development: Developing Web Applications with Python*. CA, Sebastopol: O'Reilly Media Publ., 292 p.
11. Ying, X. (2022) 'An overview of overfitting and its solutions', *J. Phys.: Conf. Ser*., 1168, art. 022022. doi: 10.1088/1742-6596/1168/2/022022.
12. De Diego, I.M., Redondo, A.R., Fernandez, R.R. (2022) 'General performance score for classification problems', *Applied Intelligence*, 52(10), pp. 12049–12063. doi: 10.1007/s10489-021-03041-7.

Авторы
Тумбинская Марина Владимировна [1], к.т.н., доцент, tumbinskaya@inbox.ru
Галиев Ренат Алмазович [1], студент, renat5390@yandex.ru

[1] Казанский национальный исследовательский технический университет им. А.Н. Туполева, г. Казань, 420111, Россия

Authors
Marina V. Tumbinskaya [1], Cand. of Sci. (Engineering), Associate Professor, tumbinskaya@inbox.ru
Renat A. Galiev [1], Student, renat5390@yandex.ru

[1] Kazan National Research Technical University named after A.N. Tupolev, Kazan, 420111, Russian Federation