ИНТЕЛЛЕКТУАЛЬНЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

УДК 007:681.512.2

ВЕКТОРИЗАЦИЯ ТЕКСТА НА ОСНОВЕ ICF+ ОНТОЛОГИИ В АНСАМБЛЯХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ ЭЛЕКТРОННЫХ РЕСУРСОВ

И. Ю. Каширин, д.т.н., профессор кафедры ВПМ РГРТУ, Рязань, Россия;
orcid.org/0000-0003-1694-7410, e-mail: igor-kashirin@mail.ru

Рассматривается оригинальная технология проектирования и применения моделей машинного обучения, а также их ансамблей для классификации и сложного анализа англоязычных политических текстов отечественных и прозападных электронных средств массовой информации. Рассматривается сквозной пример программной реализации в среде Python v.3.10, Anaconda v.2.1.

При программной реализации технологии используются: поисковый ретривер, Python паттерны, интеллектуальная вставка спецтокенов. Эффективность представленной технологии подтверждается серией практических экспериментов на примере решения задачи бинарной классификации новостных статей по идеологической направленности на прозападные и пророссийские. Результаты исследования будут полезны в задачах прогнозирования кризисных политических ситуаций.

***Целью работы** как научной статьи является представление специалистам в области искусственного интеллекта новой, разработанной автором работы, технологии онтологической векторизации политических новостей, позволяющей анализировать и прогнозировать социальные ситуации различных уровней подробностей.*

***Ключевые слова**: Bert-модели, онтологические модели, векторизация текста, токенайзер, ретривер, политические новости, ансамбли ML-моделей, прогнозирование, семантическое сходство.*

DOI: 10.21667/1995-4565-2024-90-41-53

Введение

Фейк-новости западных средств массовой информации (СМИ) стали обыденным инструментом информационной войны, социальное значение которой невозможно переоценить. Сегодня США контролируют 86 % СМИ, 8 % контролирует Япония. Результатом этих факторов стала потеря политического суверенитета многими странами мира, включая Западную Европу.

Современный уровень IT-средств интеллектуального анализа дает возможность применять ML-модели (machine learning models) для идентификации электронных новостных материалов по следующим основным классам:

- фейк-новости [1];
- токсичные тексты [2];
- материалы гневного характера [3];
- тексты, провоцирующие беспорядки [4];
- статьи с положительным или отрицательным настроением [5];
- комбинированные информационные классы [6].

Главную роль в моделях перечисленных аспектов интеллектуального анализа играет корпус материалов [7], на которых ML-модель прошла предобучение (pre-training) и дообучение (fine tuning). Более 90 % корпусов, находящихся в общедоступных репозиториях [8, 9]

датасетов и нейросетевых моделей, представляют собой подборку прозападных источников информации. Если же речь идет о больших данных, то они предполагают наличие в корпусах противоречивых и в большей части недостоверных сведений. Еще бо́льшая зависимость от таких источников у самых мощных моделей, какими являются генеративная языковая модель Яндекса YaLM и языковая модель ruGPT3XL от Сбера, а также западная модель OpenAI с триллионом параметров [10]. Эти модели могут рассматриваться как инструментарий для высоких технологий. Вследствие этого нейросети большой мощности вполне заслуженно мотивируют привлекательность не только для молодежи, но и для опытных специалистов. Однако обучение и обслуживание таких моделей требуют затрат в миллионы долларов, в то время как этими моделями не может быть решен вопрос о правдивости политических публикаций в СМИ.

Изложенные обстоятельства показывают актуальность исследований в сфере интеллектуального анализа (Data Mining) существенно идеологически разнонаправленных материалов электронных СМИ из разных стран мира. При этом можно опираться на ML-модели малой и средней мощности, поскольку их можно ограничивать локальными тематическими областями новостных материалов.

Такие исследования предлагается проводить на основе:

– анализа и использования существующих ML-моделей;
– модификации существующих моделей;
– разработки новых эффективных принципов организации моделей.

Комплексирование перечисленных основ проектирования должно привести к синергетическому эффекту получения многополярных, сбалансированных и более объективных *средств мониторинга информационных электронных ресурсов*. В качестве одного из новых принципов проектирования таких средств в настоящей статье предлагается использовать *онтологическую векторизацию политических текстов в ансамблях ML-моделей*. Получение нового способа и технологии мониторинга электронных ресурсов является целью исследования, рассматриваемого в настоящей статье. В качестве задачи, решение которой покажет эффективность новой технологии, ставится получение ML-модели для двоичной классификации англоязычных новостных статей на «статьи западных СМИ» и «статьи российских СМИ».

Теоретическая часть

Архитектура системы мониторинга электронных ресурсов

Идея использования онтологий для векторизации текстов исходит из гипотезы востребованности учителей. Обучение нейросетевых моделей на больших данных дает возможность смоделировать естественно-языковый диалог с пользователем. Так, например, дети раннего возраста иногда удивляют «взрослостью» своих суждений, весьма отдаленно понимая, о чем идет речь в реальности. Умение индуктивного и дедуктивного вывода, выстраивания продуктивных родовидовых и причинно-следственных таксономий в единую картину мира приходит лишь с возрастом и при наличии талантливого учителя. Чем лучше учитель, тем лучше результат.

Для моделей машинного обучения роль хорошего учителя могут играть существующие модели знаний, которых наработано в области искусственного интеллекта большое множество [11]. Одним из весьма распространенных подходов являются онтологические модели, основанные на дескриптивной логике [12, 13] и выраженные в схемах разметки RDF и OWL. Эти схемы были получены в рамках технологии Semantic Web, продвигаемой W3C-консорциумом.

Достоинством этой концепции является появление множества продуманных пространств имен, разработанных международными школами инженерии знаний. В рамках этих пространств реализованы структуризации и определения огромного числа понятий для различных предметных областей. К недостаткам можно отнести нередкую взаимную противоречивость существующих пространств имен и ограниченность OWL-конструкций предикативной

семантикой, сужающей полные теоретико-множественные возможности. В то же время онтологии весьма пригодны для описания главных схем знаний – таксономий. Применению OWL немало способствует редактор онтологий Protege, весьма развитый в его последних версиях.

Упрощение OWL-описаний для их автоматической обработки моделями глубокого обучения целесообразно производить средствами мощной библиотеки spaCy с инструментарием Word2vec, а также WordNet.

Общая архитектура автоматизированной системы мониторинга электронных ресурсов приведена на рисунке 1.

В архитектуре выделено шесть явно выраженных этапов создания и применения технологии мониторинга информационных ресурсов, в частности электронных публикаций СМИ на английском языке. Англоязычность не является ограничивающим фактором, поскольку результаты мониторинга можно автоматически переводить на русский язык. В то же время публикация мониторингового online-ресурса в международной сети будет способствовать устранению существующего перекоса аналогичных инструментальных средств в сторону прозападной идеологии.

Web-скрапинг электронных средств массовой информации

Web-скрапинг можно определить как этап добычи данных для обучающего корпуса англоязычных текстовых материалов СМИ. Для этого автором статьи был разработан модуль сбора тематических корпусов CorpusMining v.2.0, использующий инструментарий Googlesearch и BeautifulSoup4 в среде Python v.3.10, Anaconda v.2.1. Модуль имеет следующие параметры, состав которых стандартизован с множеством существующих NLP-программ:

```
    #  Входные параметры:
    #  WebSites           -  список web-ресурсов, с которых начинается поиск
    #  MediaNames      -  список изданий, в которых преимущественно предлагается
вести поиск
    #  Query                  -  шаблон запроса
    #  Xlsx_filenames -  выходной файл в формате xlsx
    # Поля выходного датасета:
    #
thors,   date,          text,          website
    #  Индекс, Издание, Заголовок статьи, Авторы, Дата публикации, Текст статьи,
Адрес web-ресурса
```

Рисунок 1 – Архитектура системы мониторинга электронных ресурсов
Figure 1 – Architecture of electronic resource monitoring system

С помощью CorpusMining были сформированы тематические корпуса для последующего обучения ML-моделей, общим числом 5000 политических статей. Базовыми электронными ресурсами чаще всего были западные СМИ:

```
    West Web Sites =          ['cnn.com',     'politico.com', 'bloomberg.com',
'aljazeera.com',
                      'nytimes.com', 'msnbc.com', 'theguardian.com', 'indianex-
press.com']
```

и отечественные англоязычные СМИ:

```
              East    Web     Sites    =    ['sputnikglobe.com',         'RT.com',
'en.kremlin.ru', 'rbth.com', 'meduza.io/en',
               'tass.com',          'interfax.com',    'government.ru/en/news/',
'en.globalaffairs.ru', 'rossiyasegodnya.com'] .
```

Для последующей работы в рамках системы мониторинга информационных ресурсов можно использовать готовые ретриверы, которые выделяют наиболее информативные тексты из существующих корпусов. В экспериментах с CorpusMining был применен Haystack 2.0, имеющий следующие пять этапов обработки текстов:

– Installing Haystack (подключение библиотек Heystack);
– Initializing the DocumentStore (подготовка хранилища корпуса документов);
– Preparing Documents (предварительная обработка документов хранилища);
– Initializing the Reader (инициализация селектора докуменов);
– Creating the Retriever-Reader Pipeline (создание и применение ретривера).

В рассматриваемом исследовании документами являлись информационные материалы из корпуса статей, разделенные не только по тематике, но и во времени публикации. Такое разделение предназначено для решения задач исследования динамики политических процессов.

Создание тематических онтологий

Способ разработки тематических онтологий в системе мониторинга информационных ресурсов политической тематики, как и архитектура проектирования всей системы, является инновацией. Основной задачей использования онтологий является специальная разметка текста в процессе его токенизации на последующих этапах работы системы. Токенизация является основным процессом векторизации текстов, поэтому здесь речь можно вести об онтологической векторизации новостных статей.

При проектировании тематической онтологии выделяются родовидовые и причинно-следственные таксономии, отдельные объекты, отношения и свойства предметной области, которые составляют существо фактов, содержащихся в тексте новости СМИ. В качестве примера приведем рисунок 2 с графическим представлением онтологии для темы «акт агрессии».

При построении этой онтологической схемы был использован редактор онтологий Protege 4.02.

Центральным отношением для темы «акт агрессии» выбрана ICF-триада «conflict (attack,revenge)» («конфликт(нападение, месть)»).

Согласно концепции ICF-отнологий [14, 15], базовым отношением для построения основной таксономии является ICF – отношение, определяющееся с помощью следующей композиционной формулы:

$$ICF(X, y, z) = IsA(X, y) \cap IsA(X, z) \cap Cont(y, z) \cap Form(X, y) \cap Form(X, z) .$$

Здесь все отношения семантически определены гегелевской триадой: IsA – родовидовое отношение, Cont – отношение содержательной противоположности концептов (семантически отличающееся от disjoint в концепции OWL), Form – отношение «проявляться в форме». Отношение Form позволяет интерпретировать триаду как взаимное наследование концептами y и z свойства друг друга посредством концепта-предка X. Это чем-то напоминает ромбовидное наследование в современной версии языка C++. В нашем прикладном примере отношение ICF проявляется в том, что концепт «конфликт» выражается в форме «нападения» или «мести», хотя акт мести есть нападение, а нападение может быть местью.

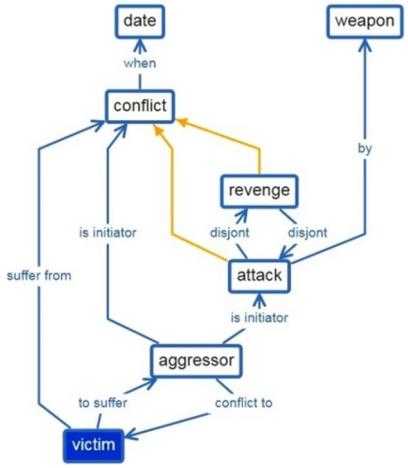

Рисунок 2 – Фрагмент онтологии для темы «акт агрессии»
Figure 2 – Ontology fragment for the «act of aggression» topic»

Например, прозападные СМИ трактуют практически любые атаки США как месть за какую-либо провинность. Остальные отношения примера определяют существование таких фактов, как «конфликт предполагает жертвы», «в конфликте используется оружие» и «конфликт имеет привязку к дате».

Теперь можно использовать полученный фрагмент онтологии для интеллектуальной векторизации, например, следующего текста, процитированного из электронного ресурса «cnn.com»:

«The United States on Friday conducted major airstrikes on dozens of targets across Iraq and Syria in retaliation for a drone attack in Jordan last month that killed three US troops.»

(«Соединенные Штаты в пятницу нанесли крупные авиаудары по десяткам целей по всей территории Ирака и Сирии в отместку за атаку беспилотника в Иордании в прошлом месяце, в результате которой погибли трое американских военнослужащих.»)

В соответствии с рассматриваемой технологией мониторинга информационных ресурсов для анализа приведенного текста используется библиотека spaCy с импортом nlp-парсера и построителя теоретико-графовых отношений networkx:

```
import spacy
import networkx as nx
nlp = spacy.load("en_core_web_sm")
```

Результатом их применения будет множество графов, которые в сокращении представлены фрагментом, изображенным на рисунке 3.

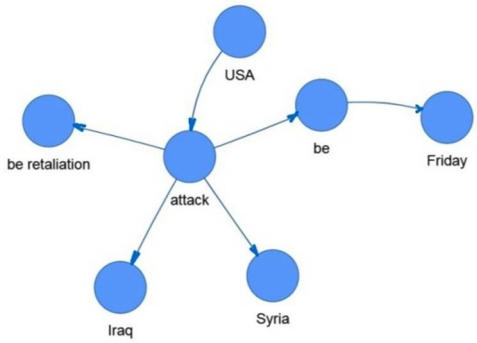

Рисунок 3 – Фрагмент прецедента онтологии «акт агрессии»
для новости из СМИ «CNN.com»
Figure 3 – A fragment of «act of aggression» ontology precedent for news
from «CNN.com»

Фактически этот граф является прецедентом OWL-онтологии, приведенной на рисунке 2.

При необходимости применение механизма паттернов (pattern из библиотеки spacy.matcher) дает возможность вывести дополнительные отношения. Например, из того,

что сработал паттерн `[{"OP":"*"},[{"ENT_TYPE": "GPE"}, {"LEMMA": "conduct"}, {"OP":"*"}, {"LEMMA": "strike"}]`,следует, что вхождение леммы GPE – это географический объект, являющийся агрессором, а присутствие ключевых слов из списка ["drone", "plane", "tank"] позволяет ассоциировать их с концептом «weapon» (оружие). При включении этих гиперонимов в онтологию сформируется новый фрагмент, представленный рисунком 4. Такая онтологическая модель существенно проще OWL-описания общей онтологии, изображенной на рисунке 2. Кроме того, она дает возможность не просто анализировать тексты новостных статей в целом, но выделять в них отдельные факты, которые затем можно исследовать на достоверность.

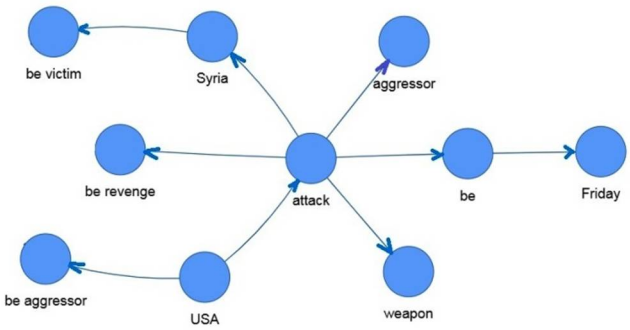

Рисунок 4 – Фрагмент онтологии с включением гиперонимов
Figure 4 – A fragment of the ontology with the inclusion of hyperonyms

Однако работа с фактами остается за рамками настоящей статьи, поскольку нашей главной целью остается онтологическая векторизация всего текста.

Получение паттернов для векторизации текста

Векторизация естественно-языковых конструкций используется в ML-моделях для преобразования текста в числовые кортежи. В рассматриваемом подходе влияние на алгоритм векторизации достигается за счет вставки в текст спецтокенов с помощью паттернов библиотеки spacy.matcher до начала применения токенайзера нейронной сети.

На этом этапе используется токенайзер модели bert-base-cased, на последующих этапах используется и сама эта модель (pytorch_model.bin) с подэтапами тренинга и тестирования качества классификации текстов.

Кроме механизма паттернов в технологии проектирования и эксплуатации системы мониторинга используются еще два эффективных средства:

Word2vec − способ построения сжатого пространства векторов слов, использующий нейронные сети и представленный программным инструментарием технологий семантической векторизации;

WordNet − семантический тезаурус английского языка, разработанный в Принстонском университете и представленный программным инструментарием с лексической базой данных.

Оба этих инструментария имеют русскоязычные подмножества, а WordNet имеет еще и конструктивно-расширенный российский аналог RuWordNet [16]. Имея в словарном запасе около 200 тысяч синсетов, эти средства дают возможность быстрого поиска синонимов, антонимов, гипонимов, гиперонимов, меронимов и других концептуальных отношений на уровне слов и других текстовых конструкций. Наиболее частой актуальной задачей, решаемой с помощью таких механизмов, является *установление семантической близости* слов, словосочетаний (именных групп) и предложений.

В рассматриваемой здесь технологии перечисленный инструментарий используется для выявления семантических ассоциаций, позволяющих рассматривать текстовые конструкции как более частные или более общие по отношению друг к другу. Это дает возможность выделять словарные категории семантически схожих классов из прикладной области политиче-

ских новостных материалов. Каждый такой класс получает собственную уникальную метку, которая используется в разметке текста в качестве спецтокена.

Приведем небольшие примеры Python-кода для обнаружения семантического сходства слов:

```
is_synonyms("attack", "aggress")                    >       True
 is_hypernym("attacker", "aggressor")             >       True
tokens = nlp(u'aggress airstrike target')
for token1 in tokens:
 for token2 in tokens:
  print(token1.text, token2.text, token1.similarity(token2))
> aggress aggress 1.0
> aggress airstrike 0.5176423788070679
> aggress target 0.16260215640068054
> airstrike aggress 0.5176423788070679
> airstrike airstrike 1.0
> airstrike target 0.3199193775653839
> target aggress 0.16260215640068054
> target airstrike 0.3199193775653839
> target target 1.0
```

Следует заметить, что при большой мощности лексических множеств Word2vec и WordNet они в настоящее время содержат далеко не все пары из отношений семантической близости. В нашем случае это заставляет вводить дополнения к имеющимся тезаурусам из элементов отношений сферы политической терминологии.

Пример дополнения специфических семантических категорий с переводом на русский язык представлен таблицей 1.

Таблица 1 – Дополнительные семантические категории

Table 1 – Additional semantic categories

№	Категории	Тип категории	Слова, именные группы категории
1	Verb-know	синсет глаголов	знал, был предупрежден, оповещен
2	Verb-forced	синсет глаголов	были вынуждены, вынуждены
3	Verb-begin	синсет глаголов	начали, открыли, создали, высадились
4	Verb-end	синсет глаголов	закончили, покинули, вывели, убрали
5	Started_first	паттерн	начали первыми
6	Provoked	паттерн	спровоцировали, вызвали ответный
7	Beneficiary	паттерн	за этим стоят, этим управляют, являются выгодоприобретателями
8	WereVictims	паттерн	появились (есть, были) жертвы
9	Attacked	паттерн	атаковали, напали, совершили налет, высадились, обстреляли, бомбили, взорвали, нарушили границы, нанесли удары, нанесли удары
10	Destroyed-killed	синсет глаголов	уничтожили, убили, ликвидировали
11	Lied	паттерн	солгали, слукавили, было неправдой,обманули, схитрили, не выполнили, оказалосьложью, смакронили
12	Struck-back	паттерн	ответили на, отомстили, нанесли ответный удар
13	Noun-weapon	синсет сущест-вительных	беспилотник, танк, самолет, артиллерия,миномет, гаубица, гранатомет
14	GeneralMilitary	общий синсет	военный, цель, оружие, война,военнослужащий
15	Subjugate	паттерн	навязывать идеологию, подчинять, подминать, лишать самостоятельности, нарушать
16	Domination	паттерн	дискриминация, мировое господство, фашизм
17	IndependenceForFight	паттерн	бороться за независимость, биться за освобождение, добиваться самостоятельности
18	Menace	паттерн	угрожали, обещали ответить, обещали отомстить, отомстим, ответ будет, в отместку
19	Dangerous	паттерн	опасная ситуация, опасность

Такие дополнения могут формироваться с помощью обучения нейросетей, но в онтологической таксономии много смысловых оттенков, являющихся частными случаями или фрагментами ситуативных семантико-синтаксических конструкций, парадигм.

Кроме того, паттерны позволяют сформировать обучающий корпус для автоматического построения таксономий.

Для приведенного ранее текста из СМИ cnn.com про атаку США способы сопоставления слов текста с концептами онтологии даны в таблице 2. Для краткости приводится пример с двумя спецтокенами: «[MLTR]» и «[EVNT]».

Таблица 2 – Способы сопоставления (matching) концептов и слов текста

Table 2 – Ways of matching text concepts and words

Спецтокен	Слова	Способ сопоставления	Результат
[MLTR]	conducted (нанесли)	pattern	True
[MLTR]	airstrikes (авиаудары)	similarity	aggress airstrike 0.5176796913146973
[EVNT]	targets (цели)	pattern	True
[EVNT]	retaliation (месть)	similarity	event retaliation 0.4425693154335022
[MLTR]	drone (беспилотник)	pattern	True
[MLTR]	killed (убито)	pattern	True
[EVNT]	troops (погибшие)	similarity	event troop 0.600471556186676

Далее можно использовать полученные в результате всех сопоставлений следующие паттерны:

```
    patterns_military = [
             [{"ENT_TYPE": "GPE"},   {"OP":"*"},{"LEMMA": "conduct"}, {"OP":"*"},
{"LEMMA": "strike"}],
            [{"LEMMA": "conduct"}, {"OP":"*"}, {"LEMMA": "airstrike"}],
            [{"LEMMA": "conduct"}, {}, {"LEMMA": "airstrike"}],
            [{"LEMMA": "conduct"}],
            [{"LEMMA": "airstrike"}],
            [{"LEMMA": "kill"}],
            [{"LEMMA": "drone", "OP":"?"}, {"LEMMA": "attack"}]
    ]
    matcher.add("MLTR", patterns_military)
    patterns_event = [
       [{"LOWER": "in", "OP":"*"}, {"LEMMA": "retaliation", "OP":"{1}"}, {"LOW-
ER": "for", "OP":"*"}],
      [{"LEMMA": "target"}],
      [{"LEMMA": "troop"}]
    ]
    matcher.add("EVNT", patterns_event)
```

Результирующим предложением со вставленными спецтокенами будет такое:

```
> USA Friday [VERB] [MLTR]conducted major [MLTR]airstrikes dozens of
[EVNT]targets across Iraq and Syria in [EVNT]retaliation for a drone
[MLTR]attack in Jordan last month that [VERB] [MLTR]killed three US
[EVNT]troops.
```

После полученной разметки предложение будет отправлено на токенизацию, где токенизатор `bert-base-cased` будет предварительно модернизирован:

```
    from transformers import BertTokenizer
    # Загрузка существующего токенизатора
    tokenizer = BertTokenizer.from_pretrained('bert-base-cased')
    # Копирование токенизатора для последующего обновления
    new_tokenizer = tokenizer.save_pretrained("tokenizer_directory")
    # Добавление новых токенов в словарь
    new_special_tokens_dict    =    {'additional_special_tokens':    ['[MLTR]',
'[EVNT]']}
    new_tokenizer.add_special_tokens(new_special_tokens_dict)
    # Сохранение обновленного токенизатора
    new_tokenizer.save_pretrained("tokenizer_directory")
```

Дообучение Bert-моделей. Сравнение с ранее полученными результатами

Для применения в задаче мониторинга электронных информационных ресурсов в рассматриваемой технологии используются Bert-модели различных модификаций. Перечень моделей, давших наилучшие результаты, сведен в таблицу 3.

Таблица 3 – Список моделей, задействованных в технологии
Table 3 – List of models involved in the technology

Наименование модели в репозитории Hugging Face	Функция	Время обучения	Занимаемая память	Точность F1 «восток»	Точность F1 «запад»
hamzab-roberta-fake-news-classificatione	Определение фейк-новости	48 мин	394 Мб	0.5453	0.8150
GPA-roberta-base-openai-detector	Применение GPT	2 мин	475 Мб	0.7027	0.7192
SkolkovoInstitute/russian_toxicity_ classifier	Токсичность текста	2,5 мин	681 Мб	0.7059	0.6714
martin-ha-toxic-comment-model	Токсичность текста	4,5 мин	1,423 Гб	0.7894	0.7547
IYuIdeology-bert-base-cased-correct	Идеология	2 мин	1,240 Гб	0.9190	0.8115
IYuIdeology-bert-base-cased-ontlg	Идеология	2 мин	1,241 Гб	0.9302	0.8520
twitter-roberta-sentiment	Настроение текста	3,1 мин	586 Мб	0.6944	0.6780
facebook-roberta-hate-speech-dynabench-r4-target	Гневный текст	2 мин	597 Мб	0.6912	0.6800

Приведенные модели предобучены их авторами на больших наборах данных и предназначены для классификации или регрессии при решении разных задач, имеющих, тем не менее, общие технологические основы обработки текстов на естественном языке. Как было замечено ранее, автор использовал собственный корпус из 5000 политических статей, извлеченных из электронных сервисов с помощью инструментария CorpusMining.

Задача бинарной классификации ставилась как определение принадлежности естественно-языкового текста к одному из классов: «статья западного СМИ» или «статья российского СМИ». В таблице 3 эти классы обозначены соответственно как «запад» или «восток».

При проведении практических экспериментов были выделены материалы следующих тем: «Putin vs Biden», «HAMAS vs IDF», «Tucker Carlson» и другие. На этих материалах было произведено дообучение («fine tuning») всех перечисленных моделей.

Одна из моделей, `IYuIdeology-bert-base-cased-correct`, была изначально взята автором статьи как базовая модель `bert-base-cased` и предобучена исключительно на корпусе из выделенных 5000 статей. Усредненные характеристики точности моделей также приведены в таблице 3. Там же приведены характеристики модели `IYuIdeology-bert-base-cased-ontlg`, полученной на основе `IYuIdeology-bert-base-cased-correct` применением рассмотренной технологии онтологической векторизации.

Сложный анализ политических текстов. Вывод результатов

Использование ансамблей ML-моделей, изначально предобученных для анализа текстов по различным формам человеческого восприятия, дает возможность решать задачи политического прогнозирования. С учетом, что СМИ различных стран, как правило, отражают идеологию политической верхушки, к этим задачам можно отнести:

- выделение кластеров идеологически близких государств;
- прогнозирование затухания старых и возникновения новых конфликтов;
- выявление новых политических тенденций;
- определение близости позиций государств в сферах локальных тематик;
- выявление информационной зависимости одних СМИ от других.

Для решения перечисленных задач проектируется онтология метазнаний, представленная рисунком 5 и в форме теоретико-графовых отношений networkx Python (рисунок 6).

Тексты новостных статей, собранных ретривером, при помещении их в общий корпус для последующего обучения ML-моделей дополняются признаками «дата публикации», «политическая тема», «наименование СМИ», «государственная принадлежность». Кроме того, по отдельным темам в корпусе можно вычислить функционально-зависимый признак «возрастание/уменьшение количества публикаций». Эти признаки достаточны для решения перечисленных в настоящем параграфе задач.

Задача *выделения кластеров идеологически близких государств* решается с помощью вычисления усредненной характеристики F1 для ансамбля моделей, спрогнозировавших отнесение множества публикаций к СМИ какого-то из двух государств. Для большей точности решения задачи можно воспользоваться регрессионными формами Bert-моделей.

Для *прогнозирования затухания старых и возникновения новых конфликтов* достаточно вычислить уже упомянутую характеристику «возрастание/уменьшение количества публикаций» и дополнить вычисления повышением уровня гневности и токсичности материалов СМИ во времени.

Выявление новых политических тенденций является наиболее сложной задачей и определяется косвенно на основе выявления выраженного концептуального дрейфа ML-моделей с ухудшением во времени их точности.

Решение задачи *определения близости позиций государств в сферах локальных тематик* аналогично решению выделения кластеров идеологически близких государств при рассмотрении публикаций исключительно одной локальной темы.

Выявление информационной зависимости одних СМИ от других основано на анализе публикаций на краткосрочных отрезках времени и требует высокой трудоемкости работ датасайнтистов. Здесь потребуется анализ метаданных при разделении корпуса текстов не только тематически, но и в краткосрочном времени (1-2 суток). Доминирующим может считаться СМИ, предоставившее новую информацию с четко очерченной идеологической направленностью (высокой характеристикой точности F1) раньше других СМИ той же идеологической направленности.

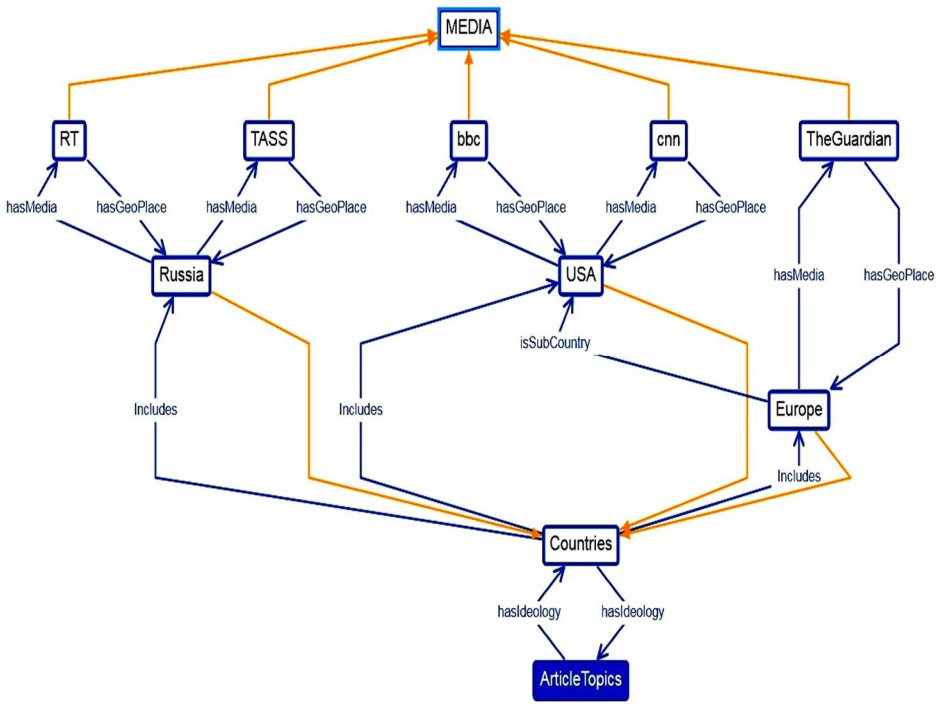

Рисунок 5 – Онтология метазнаний об идеологической ориентации СМИ
Figure 5 – Ontology of meta-knowledge about media ideological orientation

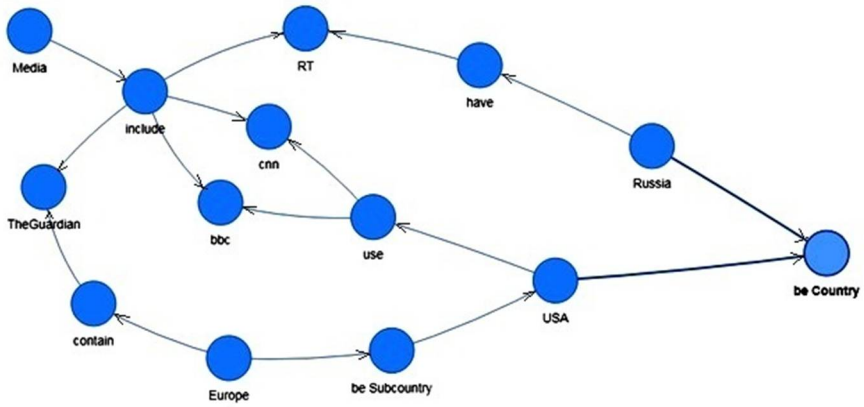

Рисунок 6 – Онтология метазнаний в networkx
Figure 6 – Ontology of meta-knowledge in networkx

Заключение

1. Результаты проведенного исследования показали эффективность применения онтологической векторизации, которая позволила улучшить характеристику точности базовой Bert-модели F1 на 1-4 %. Сравнительные характеристики с существующими предобученными международно-известными моделями были также улучшены, о чем свидетельствует таблица 3.

2. Использование ансамбля моделей определяло не только принадлежность англоязычных новостей к прозападным или пророссийским СМИ, но и психологическую окраску текста с дополнительным учетом присутствия в тексте GPT-фрагментов или признаков фейка. Это дает возможность сделать вывод о большей токсичности текстов класса «запад», но большей гневности и правдивости класса «восток».

3. Разделение материалов по тематике («Путин vs Байден», «Хамас vs Цахал», «Tucker Carlson») дает возможность оценить близость идеологических взглядов востока и запада на различные темы. Такое разделение при достаточности обучающего корпуса для СМИ разных стран мира позволяет производить кластеризацию стран на идеологические коалиции как в целом, так и по отдельным политическим направлениям.

4. Временно*е* разделение материалов по датам («2023», «2024») позволяет прослеживать положительные и отрицательные тенденции в дрейфе идеологий разных стран, прогнозируя сближение и разделение политических позиций государств, вплоть до назревания конфликтных ситуаций.

5. Временной фактор является главным признаком идентификации информационной зависимости одних СМИ от других и, как следствие, одних стран от других.

6. Выявление ретривером (интеллектуальным инструментарием фильтрации корпусов) изменения мощности тематических множеств во времени определяет как разжигание, так и потерю интереса государств к различным политическим событиям.

Библиографический список

1. **Анастасьев А.А., Асташкин М.С., Агафонов П.А., Каширин И.Ю.** Определение достоверности новостей с использованием Ml-моделей, основанных на знаниях. IIASU'23 – Artificial intelligence in management, control, and data processing systems. Proceedings of the II All-Russian scientific conference (Moscow, April 27-28, 2023): In 5 volumes. Moscow, Publishing House «KDU», 2023, vol. 2, pp. 21-27.

2. **Deshpande A., Murahari V., Rajpurohit T., Kalyan A., Narasimhan K**. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. arXiv preprint arXiv:2304.05335.

3. **Ocampo N., Sviridova E., Cabrio E., Villata S.** 2023. Anin-depth analysis of implicit and subtle hate speech messages // In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 1989-2005.

4. **Hartvigsen T., Gabriel S., Palangi H., Sap M., Ray D., Kamar E.** 2022. Toxigen: Alargescale-machine-generated dataset forad versarial and implicit hatespeech detection // In Proceeding sof the 60th Annual Meetingof the Association for Computational Linguistics (vol. 1: Long Papers), pp. 3309-3326.

5. **Bizzoni Y., Moreira P., Thomsen M.R., Nielbo K.** 2023. Sentiment almatters-predicting literary quality by sentimentanalysis and stylometric features // In Proceedings of the13th Workshop on Computational Approaches to Subjectivity, Sentiment, &Social Media Analysis, pp. 11-18, Toronto, Canada. Association for Computational Linguistics.

6. **Chiarcos C., Apostol E.-S., Kabashi B., Truică C.-O.** 2022. Modelling frequency, attestation, and corpus-398 based information with OntoLex-FrAC // In Proceedings of the 29th International Conference on 400 Computational Linguistics, pp. 4018-4027.

7. **Trevor J.,** 2008. From archive to corpus: transcription and annotation in the creation of signed language corpora // Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation, pp. 16-29.

8. Международный репозиторий для анализа данных и оригинальных технологических решений. [Электронный ресурс]. 2024. Дата обновления: 10.02.2024. URL: https://www.kaggle.com/ (дата обращения: 17.02.2022).

9. The platform where the machine learning community collaborates on models, datasets, and applications. [Электронный ресурс]. 2024. Дата обновления: 10.01.2024. URL: https://huggingface.co (дата обращения: 18.01.2024).

10**. Fedus W., Zoph B., Shazeer N.** Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv preprint arXiv: 2101.03961 (2021).

11. **Каширин Д.И., Каширин И.Ю.** Модели представления знаний в системах искусственного интеллекта // Вестник Рязанского государственного радиотехнического университета. 2010. № 31, с. 36-43.

12. **Bader Ed.ByF.**, **Calvanese D., MacGuinness D., Nardi D., Schneider P.P**. The DescriptionLogics Handbook. Theory, Implementation and Applications. Cambridge University Press, 2003.

13. **Kashirin I.Yu., Filatov I.Yu.** Formalized Description Of Intuitive Perception Of Spatial Situations // 2019 8th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 2019, pp. 1-4.

14. **Каширин И.Ю**. Иерархические числа для проектирования ICF-таксономий искусственного интеллекта // Вестник Рязанского государственного радиотехнического университета. 2020. № 71. С. 71-82.

15. **Каширин И.Ю.** Идентификация достоверности новостей с помощью моделей машинного обучения // Вестник Рязанского государственного радиотехнического университета. 2023. № 83. С. 36-47.

16. **Bochkarev V.V, Solovyev V.D.** Properties of the network of semantic relations in the Russian language based on the RuWordNet data. Journal of Physics: Conference Series 1391. 2019. pp. 1-5. DOI:10.1088/1742-6596/1391/1/012052.

UDC 007:681.512.2

VECTORIZATION OF TEXT BASED ON ICF+ ONTOLOGY IN ENSEMBLES OF ML MODELS TO CLASSIFY ELECTRONIC RESOURCES

I. Yu. Kashirin, Dr. Sc. (Tech.), Professor of the Department of VPM, Ryazan State Technical University, Ryazan, Russia;
orcid.org/0000-0003-1694-7410, e-mail: igor-kashirin@mail.ru

The original technology of designing and applying machine learning models, as well as their ensembles, for classification and complex analysis of English-language political texts of domestic and pro-Western electronic media is considered. An end-to-end example of software implementation in Python v.3.10, Anaconda v.2.1 is considered. In technology software implementation, the following components are used: search retriever, Python patterns, intelligent insertion of special tokens. The effectiveness of the technology presented is confirmed by a series of practical experiments using the example of solving the problem of binary classifi-

cation of news articles by ideological orientation into pro-Western and pro-Russian. The results of the study will be useful in forecasting crisis political situations.

***The aim** of the work is to present a new way of ontological vectorization of political news, which allows analyzing and predicting social situations of various levels of detail. of political news, which allows analyzing and predicting social situations of various levels of detail.*

***Keywords:** Bert models, ontological models, text vectorization, tokenizer, retriever, political news, ensembles of ML models, forecasting, semantic similarity.*

DOI: 10.21667/1995-4565-2024-90-41-53

References

1. **Anastasiev A.A., Astashkin M.S., Agafonov P.A., Kashirin I.Y.** Determining the reliability of news using knowledge-based Ml models. *IIASU'23 – Artificial intelligence in management, control, and data processing systems. Proceedings of the II All-Russian scientific conference* (Moscow, April 27-28, 2023: In 5 volumes. Moscow, Publishing House «KDU»), 2023, vol. 2, pp .21-27.

2. **Deshpande A., Murahari V., Rajpurohit T., Kalyan A., Narasimhan K**. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. arXiv preprint arXiv:2304.05335.

3. **Ocampo N., Sviridova E., Cabrio E., Villata S.** 2023. Anin-depth analysis of implicit and subtle hate speech messages. *In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1989-2005.

4. **Hartvigsen T., Gabriel S., Palangi H., Sap M., Ray D., Kamar E.** 2022. Toxigen: Alargescalemachine-generated dataset forad versarial and implicit hatespeech detection. *In Proceeding sof the 60th Annual Meetingof the Association for Computational Linguistics* (vol. 1: Long Papers), pp. 3309-3326.

5. **Bizzoni Y., Moreira P., Thomsen M.R., Nielbo K.** Sentiment almatters-predicting literary quality by sentimentanalysis and stylometric features. *In Proceedings of the13th Workshop on Computational Approaches to Subjectivity, Sentiment, &Social Media Analysis*, 2023. pp. 11-18, Toronto, Canada. Association for Computational Linguistics.

6. **Chiarcos C., Apostol E.-S., Kabashi B., Truică C.-O.** Modelling frequency, attestation, and corpus-398 based information with OntoLex-FrAC. *In Proceedings of the 29th International Conference on 400 Computational Linguistics*, 2022, pp. 4018-4027.

7. **Trevor J.** From archive to corpus: transcription and annotation in the creation of signed language corpora. *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, 2008, pp. 16-29.

8. An international repository for data analysis and original technological solutions. [electronic resource]. 2024. Date of update: 02/10/2024. URL: https://www.kaggle.com / (date of access: 02/17/2022).

9. The platform where the machine learning community collaborates on models, datasets, and applications. [Elektronnyj resurs]. 2024. Data obnovleniya: 10.01.2024. URL: https://huggingface.co (data obrashcheniya: 18.01.2024).

10**. Fedus W., Zoph B., Shazeer N.** Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv preprint arXiv*:2101.03961 (2021).

11. **Kashirin D.I., Kashirin I.Yu.** Models of knowledge representation in artificial intelligence systems. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2010, no. 31, pp. 36-43. (in Russian).

12. **Bader Ed.ByF.**, **Calvanese D., MacGuinness D., Nardi D., Schneider P.P**. The DescriptionLogics Handbook. Theory, Implementation and Applications. *Cambridge University Press*, 20031З.

13. **Kashirin I.Yu., Filatov I.Yu.** Formalized Description Of Intuitive Perception Of Spatial Situations. *2019 8th Mediterranean Conference on Embedded Computing (MECO*), Budva, Montenegro, 2019, pp. 1-4.

14. **Kashirin I.Yu.** Hierarchical numbers for designing the ICF taxonomy of artificial intelligence. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2020, no. 71, pp. 71-82. (in Russian).

15. **Kashirin I.Yu.** Identification of the reliability of news using machine learning models. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta.* 2023, no. 83, pp. 36-47. (in Russian).

16. **Bochkarev V.V, Solovyev V.D.** Properties of the network of semantic relations in the Russian language based on the RuWordNet data. *Journal of Physics*: *Conference Series 1391*. 2019, pp. 1-5. DOI:10.1088/1742-6596/1391/1/012052.