

заседание Президиума Российской академии наук
«Искусственный интеллект в контексте информационной безопасности»

Технологии искусственного интеллекта и безопасность информационного пространства

Воронцов Константин Вячеславович

[k.v.vorontsov @ phystech.edu](mailto:k.v.vorontsov@phystech.edu)

(д.ф.-м.н., профессор РАН, МГУ — МФТИ — ФИЦ ИУ РАН)

23 ноября 2021

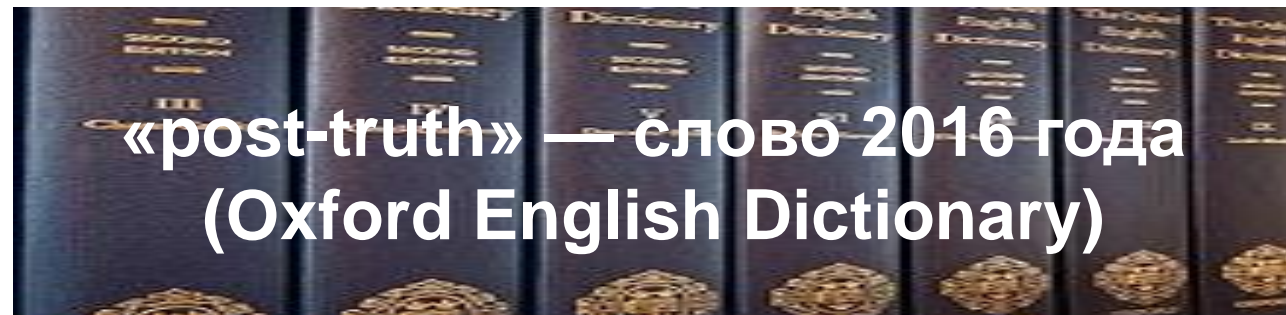
Безопасность и Искусственный интеллект

Различные аспекты ИИ в контексте информационной безопасности:

- Обеспечение безопасности (неатакуемости) систем *искусственного интеллекта* и *машинного обучения*
- Обеспечение безопасности и этичности применения систем *искусственного интеллекта* людьми и организациями
- Применение технологий *искусственного интеллекта* для обеспечения безопасности компьютерных систем
- Применение технологий *искусственного интеллекта* для обеспечения информационной безопасности государства и общества

Явление и политика постправды (post-truth)

- Факты становятся менее значимы, чем эмоции и личные убеждения
- Ложь, повторяясь, находит сторонников даже после её разоблачения
- «Ложь летит, а Истина хромает вслед за ней» (*Джонатан Свифт*)
- Фейковые новости способны формировать общественное мнение



- Предпосылки постправды: IT-технологии вырвались из под контроля
- Как использовать технологии ML/NLP, чтобы нейтрализовать угрозы?

В чем опасность постправды

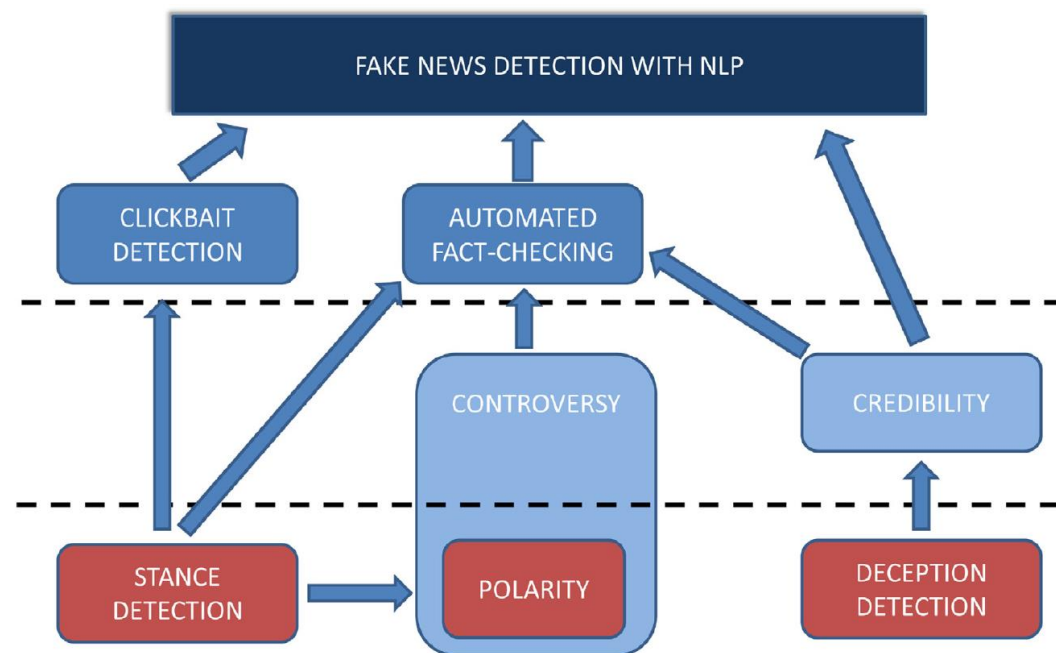
- Постправда часто маскируется под «другие грани истины»
- Распространение и легитимизация ложной (мифологизированной, идеологизированной) картины мира
- Обесценивание истины путём постепенной подмены информирования «инфотейментом»
- Разрушение социокультурного кода



- **Постправда — это инструмент «мягкой силы» в гибридных войнах**

Область исследований «Fake News Detection»

1. Deception Detection
выявление обмана в тексте новости
2. Automated Fact-Checking
автоматическая проверка фактов
3. Stance Detection
выявление позиции за/против запроса (claim)
4. Controversy Detection
выявление и кластеризация разногласий
5. Polarization Detection
классификация позиций по многим темам
6. Clickbait Detection
выявление противоречий заголовка и текста
7. Credibility Scores
оценка достоверности источника или новости



E.Saquete, D.Tomás, P.Moreda, P.Martínez-Barco, M.Palomar. Fighting post-truth using natural language processing: A review and open challenges. Expert Systems With Applications, Elsevier, 2020.

1. Deception Detection (выявление обмана)

- **История:** более 50 лет исследований в психологии и криминологии
- **Задача** классификации текста на два класса: *обман / не обман*
- **Обучающие выборки:**
 - Контролируемый эксперимент: люди *врут / не врут* на заданную тему
 - Материалы судебных заседаний (датасет DECOUR)
 - Отзывы на товары/услуги, проверяемые с помощью краудсорсинга
- **Признаки** – лингвистические маркеры (Linguistic-Based Cues, LBC)
- **Критерии:** Accuracy или F-мера 70–92% в зависимости от задачи
- На небольших датасетах классический ML лучше и проще DL
- Проблема переноса моделей на другие датасеты

Типы лингвистических маркеров

Манипулятивные и суггестивные приёмы

- многословие: плеоназмы, лишние слова, тавтологии, расщепления сказуемого
- избыточные повторы слов и фраз
- повышенная когнитивная сложность текста, перегруженные синтаксические конструкции
- повышенная экспрессивность, преобладание негативной тональности
- категоричность, психологическое давление

Уход от личной ответственности

- безличные глаголы, глаголы абстрактной семантики, модальные глаголы, объективация
- неконкретность, уклончивость, безличность, неопределённость высказываний

Подача информации

- оторванность от контекста: пониженная детализация места, времени, событий
- упрощение, пониженное лексическое разнообразие, лексическая недостаточность
- замалчивание фактов, сообщение ложных сведений (fact-checking, см. далее)

2. Automated Fact-Checking (проверка фактов)

- **История:** ручной fact-checking давно используется в журналистике
- **Задача** классификации текста целиком, по порядковой шкале:
True, Mostly True, Half True, Mostly False, False
- **Обучающие выборки:**
 - Платформы для проверки фактов: Politifact, FullFact, FactCheck и др.
 - Соревнования: CLEF-2018,19,20,21, FEVER, SemEval (Rumour-Eval)
 - Датасеты: NELA-GT-2018,19, FakeNewsNet, Snopes и др.
- **Вспомогательная задача:** стоит ли отправлять текст на проверку?
Три класса: *Non-Factual Sentence, Unimportant, Check-Worthy*
(пример: ClaimBuster, <https://idir.uta.edu/claimbuster>, 2015)

3. Stance Detection (выявление позиции)

- **История:** задача textual entailment (текстового следования) – классификация пар текстов «текст $t \Rightarrow$ гипотеза h » на три класса: « h следует из t », « h противоречит t », « h не относится к t »
- **Задача:** классификация текста h относительно запроса (claim) t : *agree, disagree, discusses (позиция не высказана), unrelated*
- **Обучающие выборки:**
 - SNLI: 570K пар предложений: entail, contradict, independent
 - Датасеты: Emergent, SemEval-2016 6A(stance), FakeNewsChallenge FNC-1
- **Критерии:** F1-мера до 97% на новостях; Accuracy до 68% на Twitter

4. Controversy / 5. Polarization Detection

Две специальные разновидности задачи Stance Detection

- **Controversy Detection** (выявление полемики, разногласий):
 - кластеризация мнений без учителя
 - выделение сообществ сторонников каждого мнения в социальной сети
 - количественное оценивание объёма и динамики сообществ
- **Polarization Detection** (выявление поляризованности общества):
 - выявление разногласий по совокупности запросов или тем
- **Обучающие выборки:**
 - Датасеты социальных сетей, обычно Twitter
 - Википедия
- **Критерии:** Accuracy 73–83% (на Википедии, методом kNN)

6. Clickbait Detection (обнаружение кликбейта)

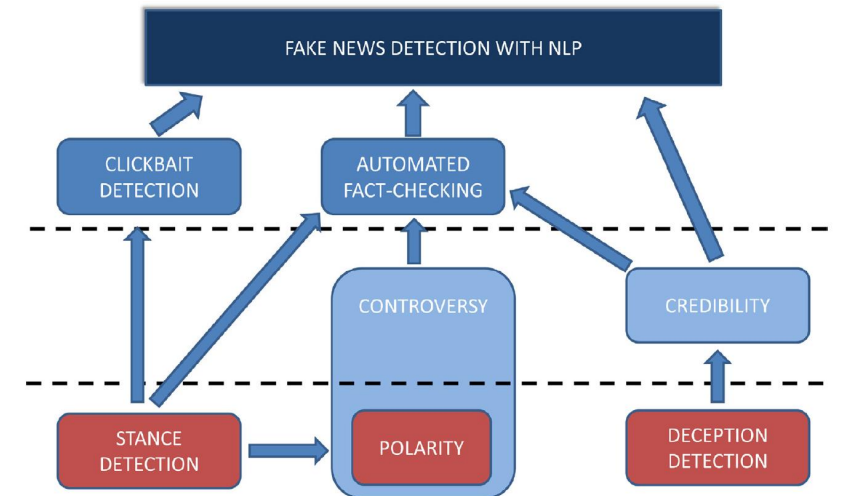
- **История:** задача появилась в 2016 году. Обнаружение заголовков или ссылок-приманок, не соответствующих сути контента
- **Задача:** классификация пары «заголовок, текст» на два класса
Задача аналогична Textual Entailment и Stance Detection
- **Признаки:** гиперболизация, противоречия, web-трафик
- **Обучающие выборки:**
 - Датасеты: Webis-Clickbait 2017 (32К заголовков) и др.
 - Соревнование: Clickbait challenge 2017
- **Критерии:** F1-мера до 68%; Accuracy до 86%

7. Credibility Scores (Оценивание надёжности)

- **История:** старая задача в социологии, психологии, маркетинге
- **Задача:** оценить уровень доверия (credibility, trustworthiness) для источника (СМИ, блогера, пользователя) или отдельной новости
- **Признаки:**
 - распространение ненадёжного контента (spam, deception, fake и др.)
 - вероятность быть ботом (по диспропорции рассылок и качеству контента)
 - стиль контента, геолокация и образовательный уровень читателей
- **Обучающие выборки:**
 - много несопоставимых датасетов, отсутствует «золотой стандарт»
- **Критерии:** AUC до 89%; accuracy до 81%; MSE до 0.33
 - много критериев, не хватает методологического единства

Чего-то не хватает...

1. **Fake News** – не единственный и не самый сильный инструмент политики постправды.
2. **Пропаганда** использует не только фейки, но и полуправду, замалчивание, манипулятивные воздействия и т.д.
3. **Информационные войны** нацелены на разрушение социокультурного кода и сложившейся общественной идеологии.
 - Как распознавать манипулятивные воздействия и идеологические атаки?
 - Как находить разногласия и замалчивание?
 - Насколько расширится типология задач?

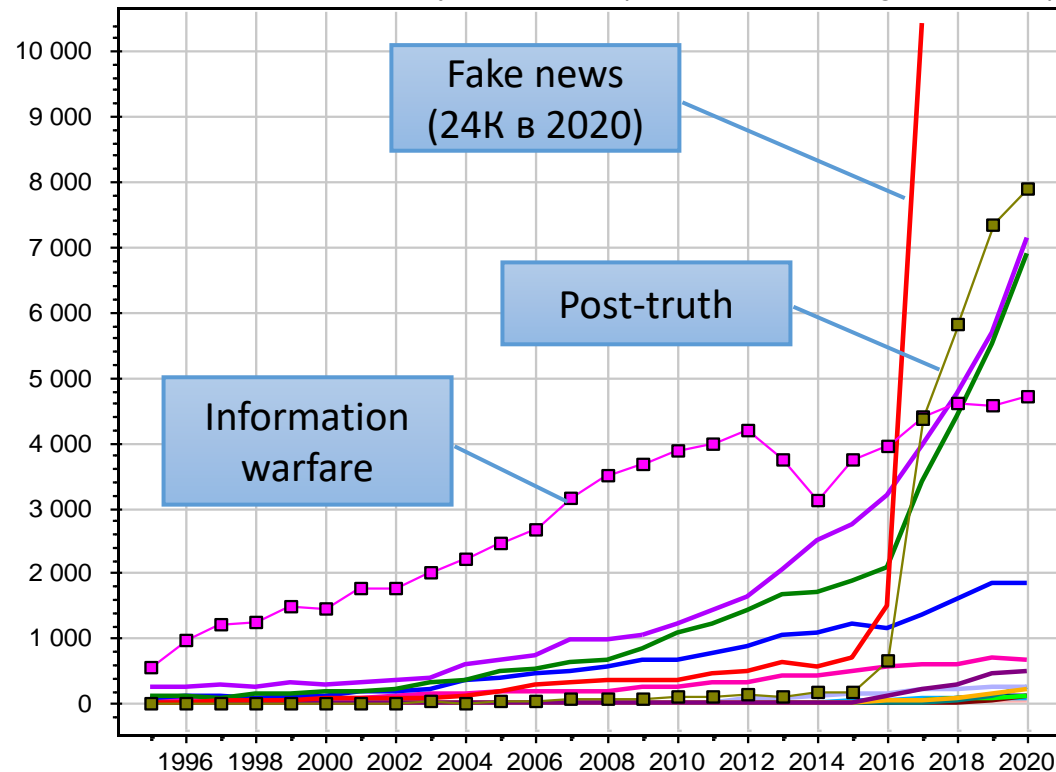


E.Saquete, D.Tomás, P.Moreda, P.Martínez-Barco, M.Palomar.
Fighting post-truth using natural language processing: A review and open challenges. *Expert Systems With Applications*, Elsevier, 2020.

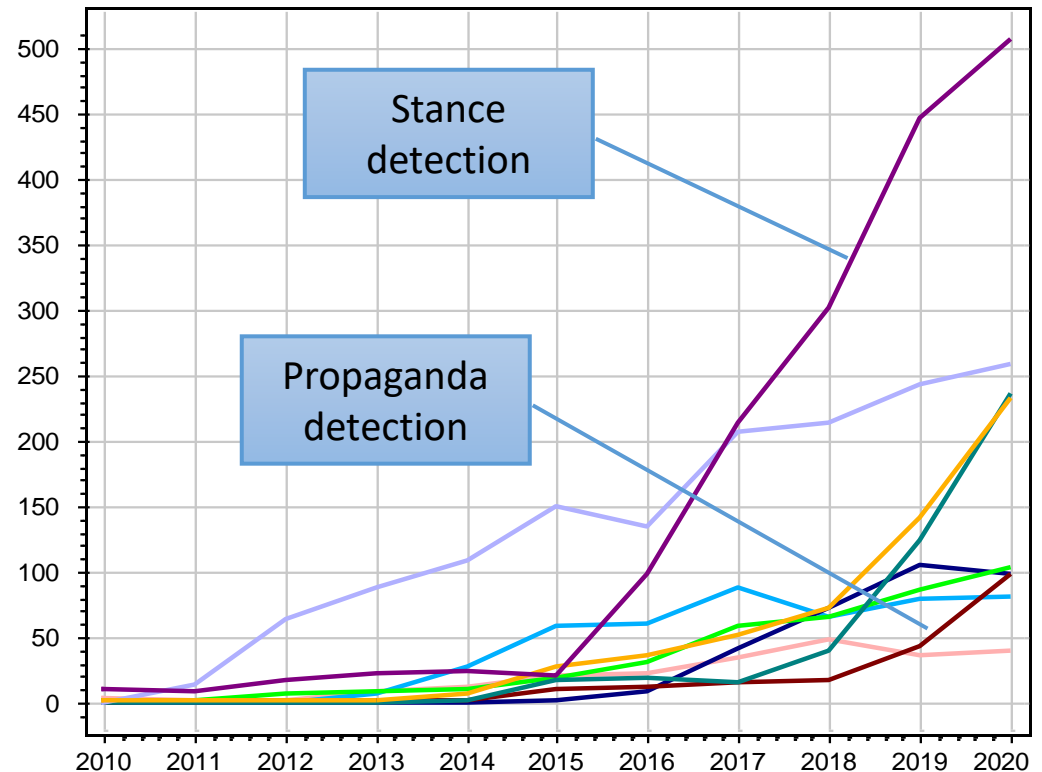
Fake News и близкие тренды исследований

(библиометрический анализ по данным Google Scholar)

Число публикаций (по данным Google Scholar)















Новые тренды последних 10 лет



- post-truth
- information warfare
- fake news
- political polarization
- fact checking
- language manipulation
- deception detection
- stance detection
- rumor detection
- misinformation detection
- hoax detection
- propaganda detection
- clickbait detection
- controversy detection
- deceptive opinion spam
- virality prediction

Типология потенциально опасного дискурса и система подзадач ML/NLP для его детекции

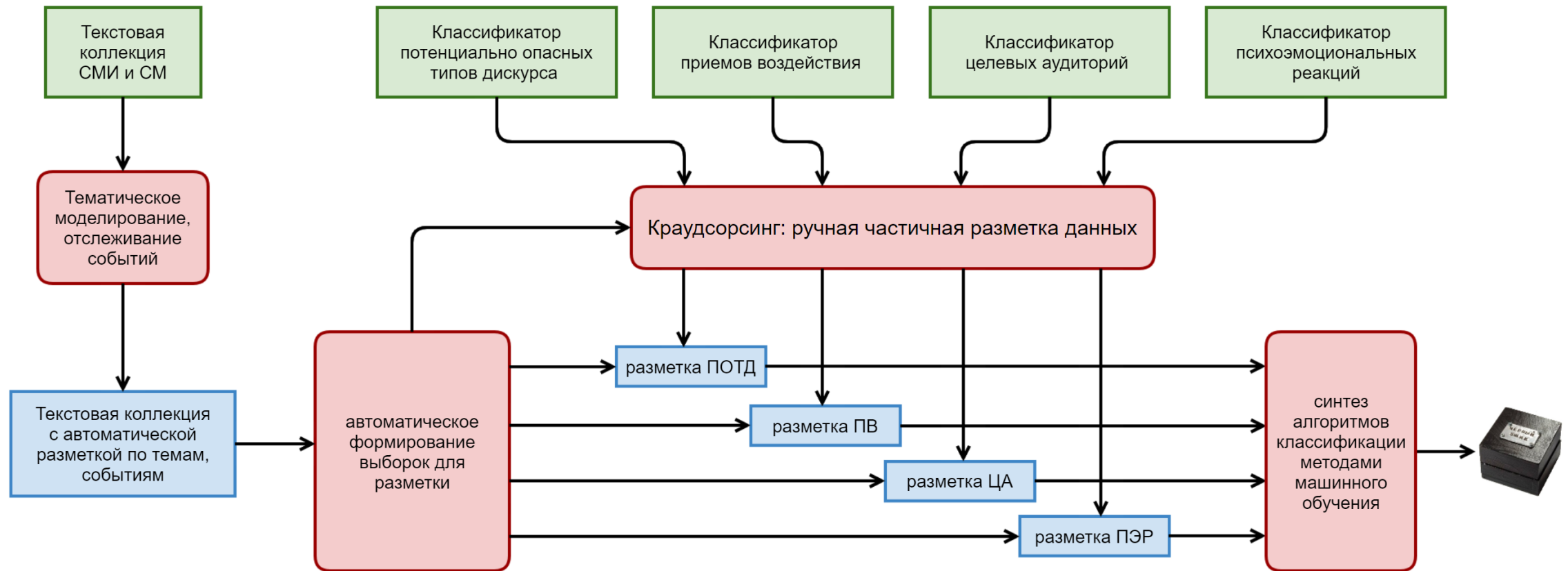
воздействия → **фейки** → **пропаганда** → **инф.война**

1.  детекция приёмов манипулирования
2.  детекция замалчивания
3.  детекция обмана (deception detection), слухов (rumors d.), мистификаций (hoaxes d.)
4.  детекция кликбэйта (clickbait detection)
5.  автоматическая проверка фактов (auto fact-checking)
6.  детекция позиции (stance d.), противоречий (controversy d.), поляризации (polarization d.)
7.  выявление конструктов картины мира: идеологем, мифологем
8.  оценивание возможных психо-эмоциональных реакций
9.  выявление целевых аудиторий воздействия
10.  оценивание и предсказание скорости распространения (virality prediction)
11.  оценивание достоверности источников (credibility scores)
12.  детекция прямой агрессии (угрозы, призывы, провокации, вербовка, экстремизм)

Четыре основных типа подзадач ML/NLP

- 1. Классификация текста (сообщения/предложения) целиком**
 - deception detection, fact-checking, text credibility
- 2. Классификация пары текстов**
 - stance, controversy, polarization, clickbait detection
 - выявление противоречий, разногласий, замалчивания
- 3. Разметка текста (выделение и классификация фрагментов)**
 - поиск лингвистических маркеров (linguistic-based cues) в тексте
 - детекция приёмов манипулирования
 - выявление конструкторов картины мира: мифологем, идеологем
 - выявление психо-эмоциональных реакций и целевых аудиторий
- 4. Кластеризация или тематическое моделирование**
 - кластеризация мнений по заданной теме (controversy detection)
 - выявление поляризованных мнений (polarization detection)
 - выявление мнений как сочетаний слов, их семантических ролей и тональностей
 - выявление «картин мира» – устойчивых сочетаний суждений и идеологем

Разметка текстовых данных как способ формализации гуманитарных знаний



Пример. Поляризация мнений о событии

... Президент Петр Порошенко заявил, что Россия де-факто конфисковала украинские предприятия, которые находятся на неподконтрольной Киеву территории. Сегодня ДНР и ЛНР "национализировали" украинские предприятия ... При этом Кремль защитил конфискацию предприятий в ЛДНР ... Украина потребует расширить санкции ... За все эти действия обязательно наступит наказание. Украина потребует расширения санкций на тех, кто украл украинские предприятия ... *(Kiev opinion)*

... По словам Захарченко, Киев встретит свой "ужасный конец"... Киев возьмется за ум, и в целях спасения собственной промышленности снимет блокаду ... Обстановка, которую искусственно создала Украина с блокадой Донбасса, вынудила ... кошмарит свой народ ... если в Киеве были приняты какое-либо постановление ... положительные результаты, как в республиках, так и в России... Если им удастся сместить Порошенко и при этом не развалить Украину, то все вернется на свои места ... *(Moscow opinion)*

Subject

Object

Agent

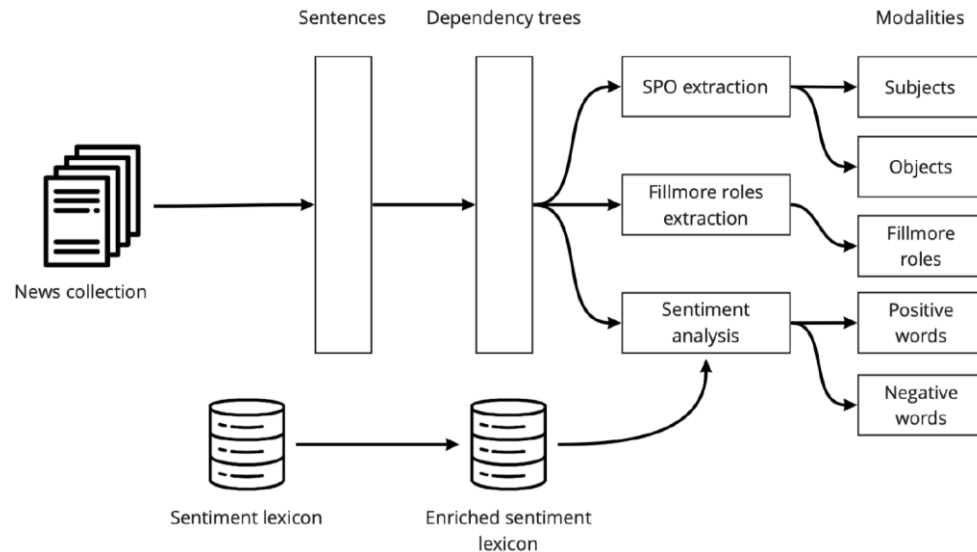
Locative

Negative lexicon

Dependent word

- Слова «Порошенко», «Россия», «Украина» встречаются одинаково часто
- «Порошенко» — субъект в первом тексте и объект во втором
- «Россия» — агент в первом тексте и локация во втором
- Негативная тональность: «Россия», «Кремль» в 1-ом, «Киев», «Украина» во 2-ом

Пример. Поляризация мнений о событии



Modalities	<i>Pr</i>	<i>Rec</i>	<i>F1</i>
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
All	0.77	0.97	0.86

LPR Business

Modalities	<i>Pr</i>	<i>Rec</i>	<i>F1</i>
TF-IDF	0.57	0.97	0.72
SPO	0.56	0.99	0.72
FR	0.67	0.97	0.79
Sent	0.56	0.55	0.55
SPO+FR	0.72	0.99	0.83
SPO+Sent	0.57	0.99	0.72
FR+Sent	0.73	0.97	0.83
All	0.77	0.94	0.85

Paris Trump

- Мнение формализуется как устойчивое сочетание слов, терминов, объектов и субъектов, их семантических ролей по Филлмору и их тональных окрасок
- Все они используются в тематической модели как отдельные модальности

Feldman D. G., Sadekova T. R., Vorontsov K. V. [Combining Facts, Semantic Roles and Sentiment Lexicon in A Generative Model for Opinion Mining](#). Computational Linguistics and Intellectual Technologies. Dialogue 2020.

Резюме

- Противостояние угрозам политики постправды – социально значимая задача, миссия и вызов для научно-технологического сообщества ML/NLP
- Задача *Fake News Detection* расширяется до выявления всех видов потенциально опасного дискурса (манипуляций, пропаганды, информационной войны)
- Эти задачи вполне решаемы современными средствами ML/NLP
- Решение требует междисциплинарного подхода, объединения усилий политологов, журналистов, лингвистов, психологов, AI-инженеров

Воронцов Константин Вячеславович

[k.v.vorontsov @ phystech.edu](mailto:k.v.vorontsov@phystech.edu)