

Заседание Экспертного совета по развитию цифровой экономики, технологий и инноваций Молодёжного парламента при Государственной думе ФС РФ
«Противодействие деструктивной идеологии (нацизм, терроризм, криминал)»

Технологии искусственного интеллекта против фейков, постправды и информационных войн

Воронцов Константин Вячеславович

k.v.vorontsov@mlsa-iai.ru

д.ф.-м.н., профессор РАН,
зав. лаб. Машинного обучения и семантического анализа
Института искусственного интеллекта МГУ,
зав. каф. Машинного обучения и цифровой гуманитаристики МФТИ,

17 февраля 2023

Ментальная (когнитивная) информационная война

«Если в классических войнах целью является уничтожение живой силы противника, в современных кибервойнах — уничтожение инфраструктуры противника, то **целью новой войны является уничтожение самосознания, изменение ментальной — цивилизационной основы — общества противника**. Я бы назвал этот тип войны — ментальным»

*Андрей Ильницкий, советник Министра обороны РФ,
«Арсенал Отечества» № 1 (51) за 2021*

Тексты, медиаобъекты — новый вид вооружения — культурологического

Задачи: мониторинг, своевременное обнаружение, противодействие деструктивным идеологиям, пропаганде, фейкам, постправде

Технологии вырвались из-под контроля?

- Эффект «неопровержимой лжи» в информационном пространстве
- Эффект «информационных пузырей» в системах поиска и рекомендаций
- Генераторы фейковых ответов (ChatGPT), голоса, видео (DeepFake)
- Достижения гуманитарных наук в управлении массовым сознанием

Какие технологии вернут контроль?

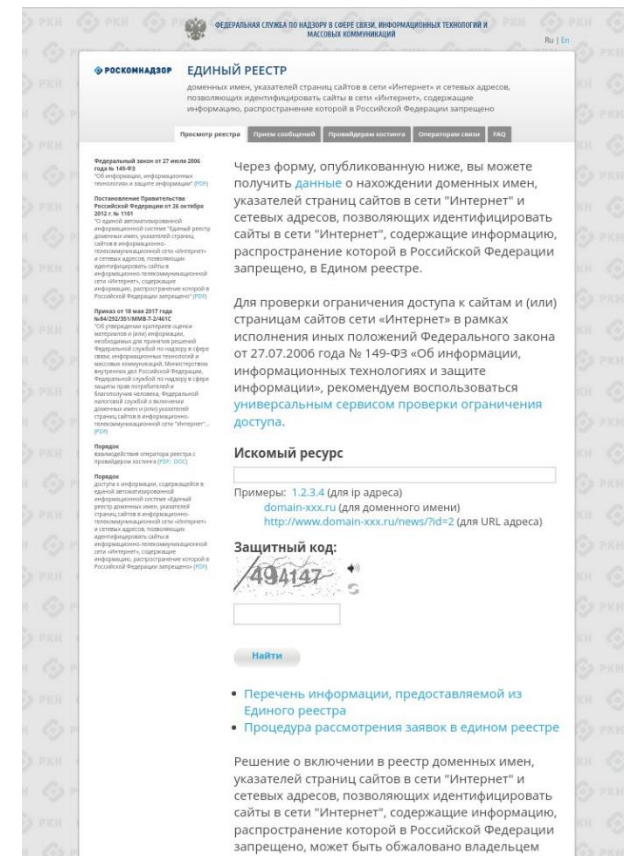
Три уровня защиты: государство, профсообщества, граждане

- Мониторинг напряжённости и угроз в информационном пространстве
- Формализация гуманитарных знаний (базы фактов, разметки и т.д.)
- Сервисы поиска пропаганды, антипропаганды и их источников

Типы запрещённой информации (149-ФЗ)

Примеры типов информации для мониторинга и внесения в **Единый реестр запрещённых сайтов** (всего более 50 типов)

- Детская порнография
- Незаконный оборот наркотиков
- Призывы к самоубийству
- Вовлечение несовершеннолетних
- Экстремистские материалы
- Пропаганда проституции
- Экономические преступления
- Взрывчатые вещества
- Поддельные документы
- Уклонение от армии
- Порнографические материалы
- Государственная тайна
- Контрафактная продукция
- Оружие
- Шоплифтинг (кража в магазинах)
- Пересечение государственной границы
- Специальные технические средства
- Пропаганда бродяжничества
- Сильнодействующие препараты
- Браконьерство
- Продажа человеческих органов
- Дискриминация, национализм
- Соккрытие трупа
- Зацепинг
- Срыв выборов
- Продажа курсовых, дипломных работ
- Руферы
- АУЕ, криминальный образ жизни
- ...



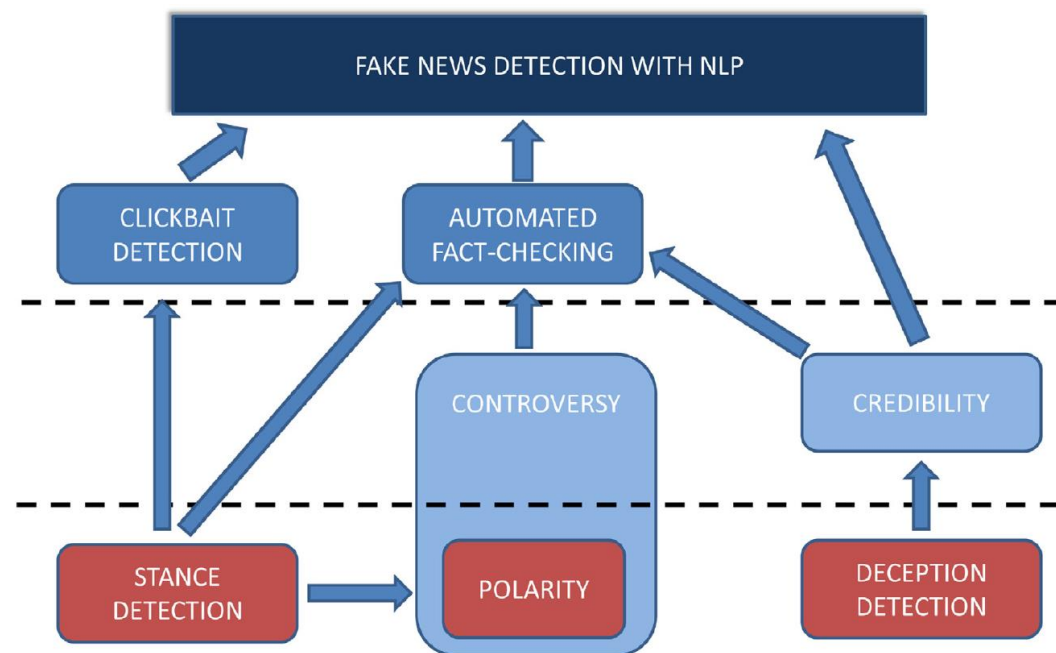
Типы не запрещённой, но потенциально опасной информации — «серая зона»

Примеры типов информации, представляющих угрозу для государства и общества. В ходе НИР было описано 114 угроз (не включая методички Шарпа), список открытый...

- Обесценивание гражданской идентичности
- Фейки об истории России
- Оскорбительные арт-акции о России
- Обесценивание достижений науки, культуры, искусства и спорта России
- Культурологические угрозы: перегибы политкорректности, мигрантофобия
- Тунеядство, обесценивание труда
- Когнитивные искажения, обесценивание, ксенофобия в учебной литературе
- Модифицированные методы Джина Шарпа
- Обесценивание нацпроектов
- Пропаганда эмиграции из России
- Игромания
- Стигматизация денег, пениафобия
- Трансформация семейных ценностей: полиамория, промискуитет, чайлдфри, суррогатное материнство
- Стигматизация больных и инвалидов
- Газлайтинг, буллинг, моббинг, хейтинг в детской и молодежной среде
- Романтизация, героизация разрушительного асоциального поведения
- Неэстетичное поведение, пропаганда девиаций стиля поведения
- Псевдолечение, лжелечение

Область исследований «Fake News Detection»

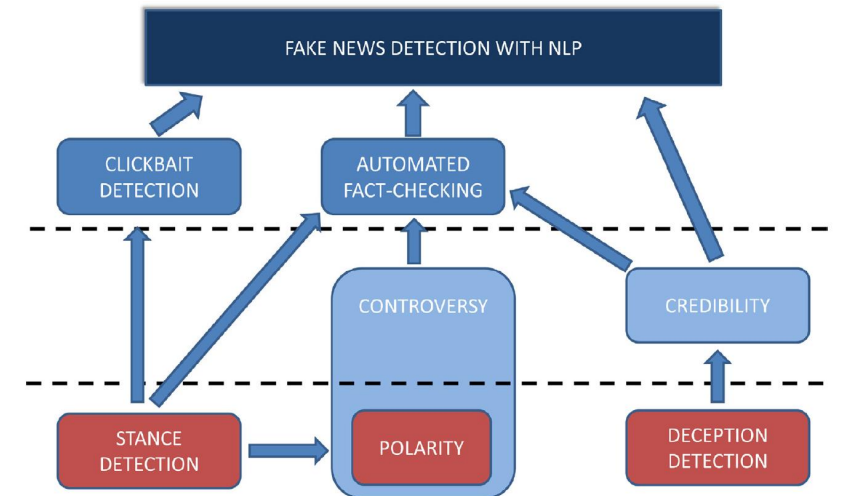
1. Deception Detection
выявление обмана в тексте новости
2. Automated Fact-Checking
автоматическая проверка фактов
3. Stance Detection
выявление позиции за/против запроса (claim)
4. Controversy Detection
выявление и кластеризация разногласий
5. Polarization Detection
классификация позиций по многим темам
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.

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. **Ментальные войны** нацелены на разрушение социокультурного кода, общественной идеологии и морали.
 - 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.

Типология деструктивного дискурса и система подзадач ML/NLP для его детекции

воздействия → фейки → пропаганда → инфо-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 конструкторов картины мира: ценностей, идеологем, мифологем
8.  оценивание возможных психо-эмоциональных реакций реципиента
9.  выявление целевых аудиторий воздействия
10.  оценивание и предсказание скорости распространения (virality prediction)
11.  оценивание достоверности источников (credibility scores)
12.  детекция деструктивных воздействий (угроз, провокаций, вербовки, экстремизма)

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 идеологем, ценностей, элементов социокультурного кода
 - выявление психо-эмоциональных реакций и целевых аудиторий
 - выявление мнений, тональных оценочных суждений
- 4. Кластеризация или тематическое моделирование**
 - кластеризация мнений по заданной теме (controversy detection)
 - выявление поляризации общественного мнения (polarization detection)

Задача детекции приёмов манипулирования

Структура манипуляции:

- фрагмент-мишень
- фрагмент-воздействие
- тип манипуляции

Пример из СМИ:

«**Зеленский** просто **играет роль президента, а не является президентом**^[обесценивание], – считает экс-депутат Верховной рады Борислав Береза»

Типы манипуляций (всего 18 типов):

- негативизация (обесценивание, дисфемизмы, ярлыки, депрессивы и т.п.)
- позитивизация (героизация, эвфемизация, лозунги и т.п.)
- деавторизация (замалчивание источника, маскировка под ссылку и т.п.)
- паралогизация (алогизм, ложное следование, подмена тезиса и т.п.)

Классификация приёмов манипулирования

1. Негативизация

- 1.1 Навешивания ярлыков
- 1.2 Дисфемизмы
- 1.3 Аналогия с негативным объектом
- 1.4 Антифразис
- 1.5 Прием обесценивания
- 1.6 Негативирующая гиперболизация
- 1.7 Моделирование негативного сценария
- 1.8 Вкрапление депрессивов

2. Позитивизация

- 2.1 Эвфемизация
- 2.2 Лозунговые слова и словосочетания
- 2.3 Позитивирующая гиперболизация

3. Деавторизация

- 3.1 Маскировка под ссылку на авторитет
- 3.2 Ссылки на неопределенный источник
- 3.3 Ссылки на неназванных свидетелей

4. Паралогизация

- 4.1 Ложная причинно-следственная связь
- 4.2 Прием «после этого не значит поэтому»
- 4.3 Подмена тезиса
- 4.4 Высказывание о состоянии другого

Детекция пропаганды (propaganda detection)

Чтобы выявлять пропаганду, нужно иметь модель пропаганды:

1. *Подмена и/или дополнение фактов мнениями*
2. *Фрагментирование: часть фактов замалчивается*
3. *Деконтекстуализация: изымается контекст, без которого корректное понимание смысла фактов невозможно*
4. *Реконтекстуализация: конструируется новый контекст, выгодный манипулятору*

Подзадачи ML/NLP:

- Выделение и различение фактов и мнений
- Выявление замалчиваний путём сравнения с другими источниками
- Выявление идеологем, используемых для реконтекстуализации

Обучающие выборки:

- Тексты новостей с размеченными фрагментами (факты, мнения, идеологемы)

Методология применения ИИ для детекции

- Отбор и подготовка выборки медиаобъектов для разметки (на регулярной основе, в потоковом режиме)
- Разработка классификаторов для разметки
- Разметка медиаданных (фрагмент + метка + связь + комментарий)
— **магистральный путь формализации гуманитарных знаний**
- Построение моделей разметки на основе ИИ (автоматическая разметка — это форма объяснимого ИИ)
- Оценивание моделей разметки в сравнении с экспертами
- Внедрение моделей в системы мониторинга, поиска, рекомендаций

Методология разметки обучающих выборок

Пик научной фантастики (и советской, и западной) пришелся на 1960–1970-е годы. Однако в 1970-х годах этот жанр начал постепенно затухать и сходить на нет, уже в 1980-х на Западе начинает набирать силу жанр фэнтези. Конечно же, это неслучайно. Именно 1960-е годы стали пиком научно-технического прогресса в XX веке. К тому времени закончилась первая половина XX столетия, за эти полсотни лет было изобретено столько, что все казалось возможным, верилось, что прогресс будет нарастать по экспоненте. **1960-е — это мир безудержного социального и культурно-технического оптимизма.** Человек полетел в космос, запустил искусственные спутники и задумался об освоении других планет.

Но этот порыв человечества в будущее создавал определенную угрозу для власти имущих как на Западе, так и в Советском Союзе. И уже в 1960-е годы перед сотрудниками Тавистокского института изучения человека в Великобритании (причем по иронии судьбы он располагается в графстве Девоншир, рядом с дартмурскими болотами, где разыгрывалась мрачная драма «Собаки Баскервилей» Конан Дойля) **была поставлена задача притормозить научно-технический прогресс путем внедрения определенных информационно-психологических и организационных моделей.** В частности, стартовала работа по созданию молодежных и женских субкультур и движений (именно в это время как по заказу появились The Beatles, The Rolling Stones, стал развиваться экологизм).

Одна из главных задач, поставленных перед Тавистоком, звучала так: to stamp out the cultural optimism of the 1960s (искоренить, вырубить, вытравить культурный оптимизм 1960-х годов). А **научная фантастика, особенно советская, безусловно, была оптимистической по своему настрою.** Некоторые менее оптимистические ноты (не могу их назвать пессимистическими, но они выглядели более сложными, чем просто оптимизм) прослеживались у ряда писателей в соцлагере, в частности в книгах Станислава Лема (достаточно почитать его «Астронавтов» и «Магелланово облако»). Однако общий настрой советской фантастики до середины 1960-х годов был преимущественно оптимистичным — это видно и по творчеству братьев Стругацких, и по романам Ивана Ефремова.

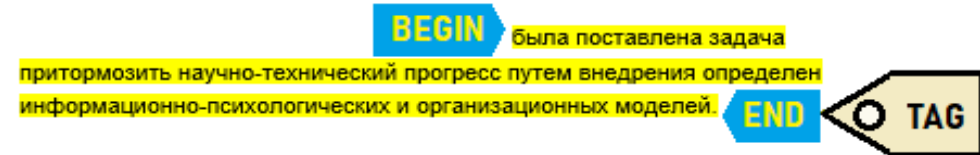
Первый доклад Римскому клубу (он создан в 1968 году) назывался «Пределы роста». В нем утверждалось, что человечество в своем индустриальном развитии достигло пределов, избыточно давит на природную среду, надо тормозить промышленно-экономическое развитие, перейдя к «нулевому росту». То есть 50 процентов всех средств должно идти на нейтрализацию негативных последствий, который несет индустриальное развитие.

Разметка состоит из элементов

Элемент разметки может содержать любое число фрагментов, затекстов и тегов

Метки выбираются из словаря меток

Фрагмент задаётся началом и концом, может иметь одну или более меток:



Коммент может выбираться из словаря фраз или генерироваться по контексту, может иметь одну или более меток

Методология оценивания (ПРО//ЧТЕНИЕ)

- В основе методики — сравнение пар разметок текста: «модель \leftrightarrow эксперт», «эксперт-1 \leftrightarrow эксперт-2», на основе оптимального сопоставления их элементов
- Согласованность разметок (A,B) измеряется многими критериями, вычисляется их средневзвешенная согласованность $Con(A,B)$
- СТАР (Средняя Точность Алгоритмической Разметки) — средняя по выборке $Con(A,E)$ разметки модели A и разметки эксперта E
- СТЭР (Средняя Точность Экспертной Разметки) — средняя по выборке $Con(E1,E2)$ разметок двух экспертов, E1 и E2
- ОТАР = СТАР / СТЭР, если больше 100%, то модель лучше экспертов

Подытожим

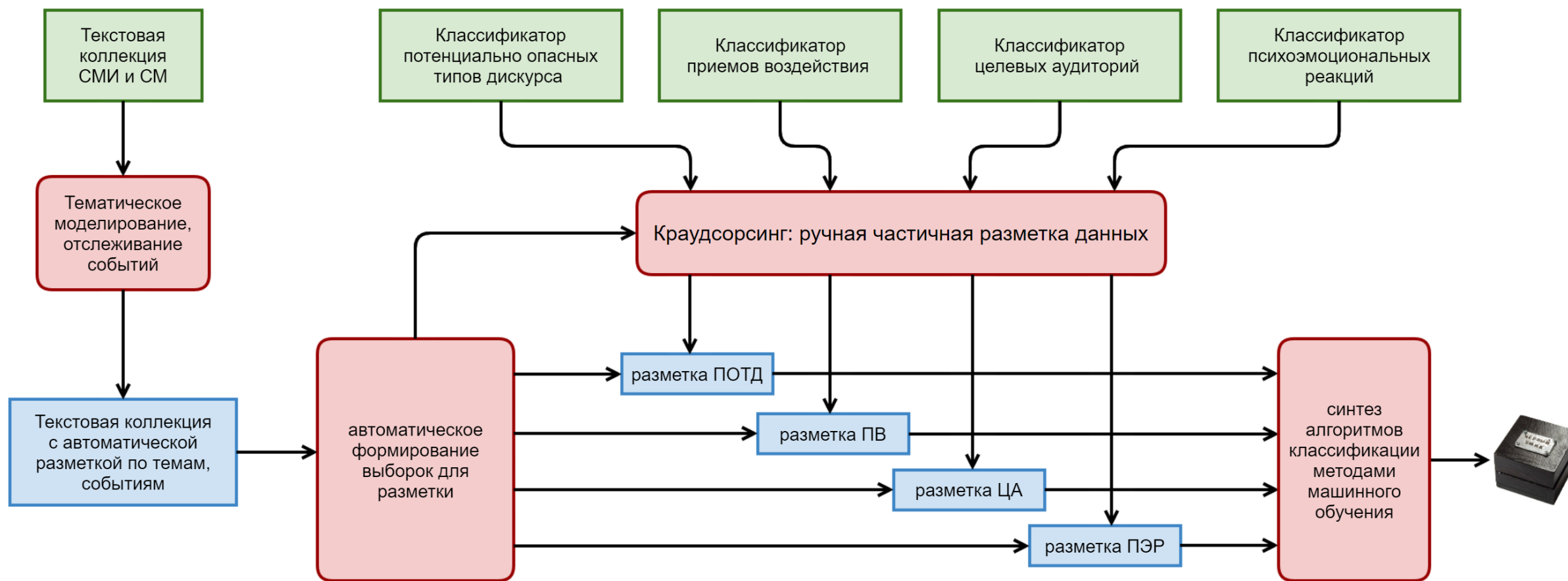
- Противостояние деструктивным идеологиям в эпоху ментальных войн – миссия и вызов для междисциплинарного научного сообщества AI & DH
- Задачи детекции деструктивного дискурса (манипуляций, поляризации и др.) вполне решаемы современными средствами ML/NLP
- Разметка текстовых данных — магистральный путь формализации гуманитарных знаний в технологиях ИИ

Воронцов Константин Вячеславович

[k.v.vorontsov @ mlsa-iai.ru](mailto:k.v.vorontsov@mlsa-iai.ru)

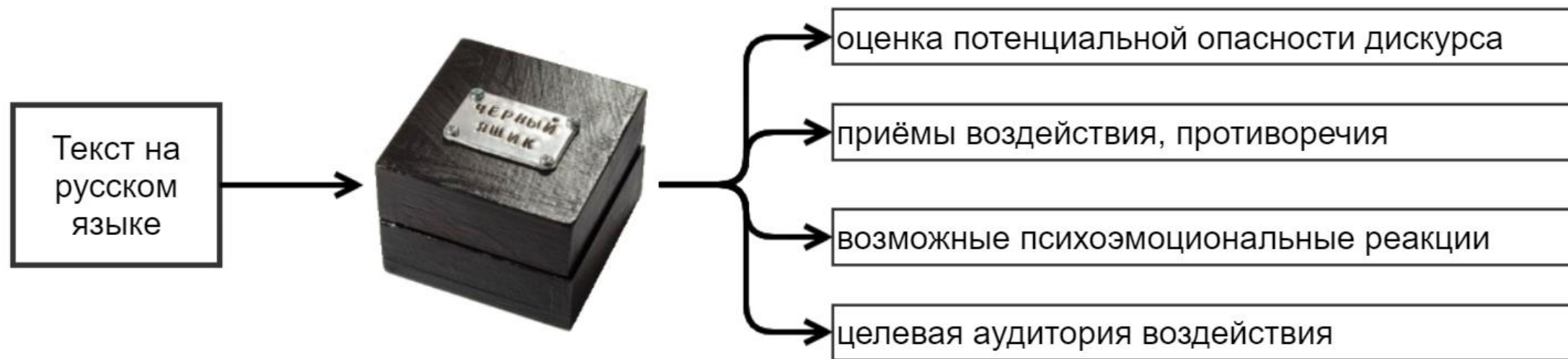
Дзен: «Цивилизационная идеология»

Разметка текстовых данных — магистральный путь формализации гуманитарных знаний



На выходе — модель классификации угроз в медийном информационном пространстве

Модель, обученная по размеченным обучающим выборкам, может быть использована в автоматическом режиме для мониторинга и фильтрации деструктивного дискурса в информационном пространстве



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

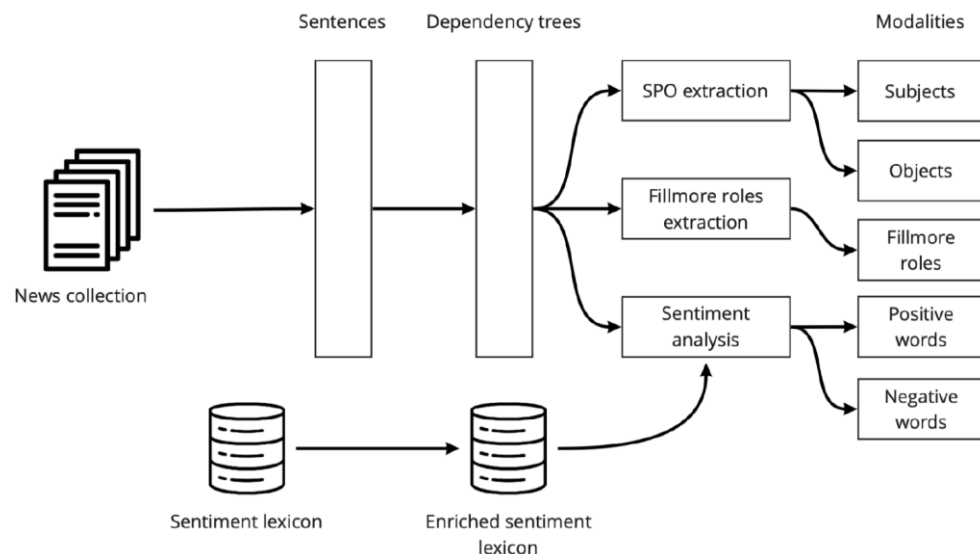
Locative

Negative lexicon

Dependent word

- Слова «Порошенко», «Россия», «Украина» встречаются одинаково часто
- «Порошенко» — субъект в первом тексте и объект во втором
- «Россия» — агенс 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

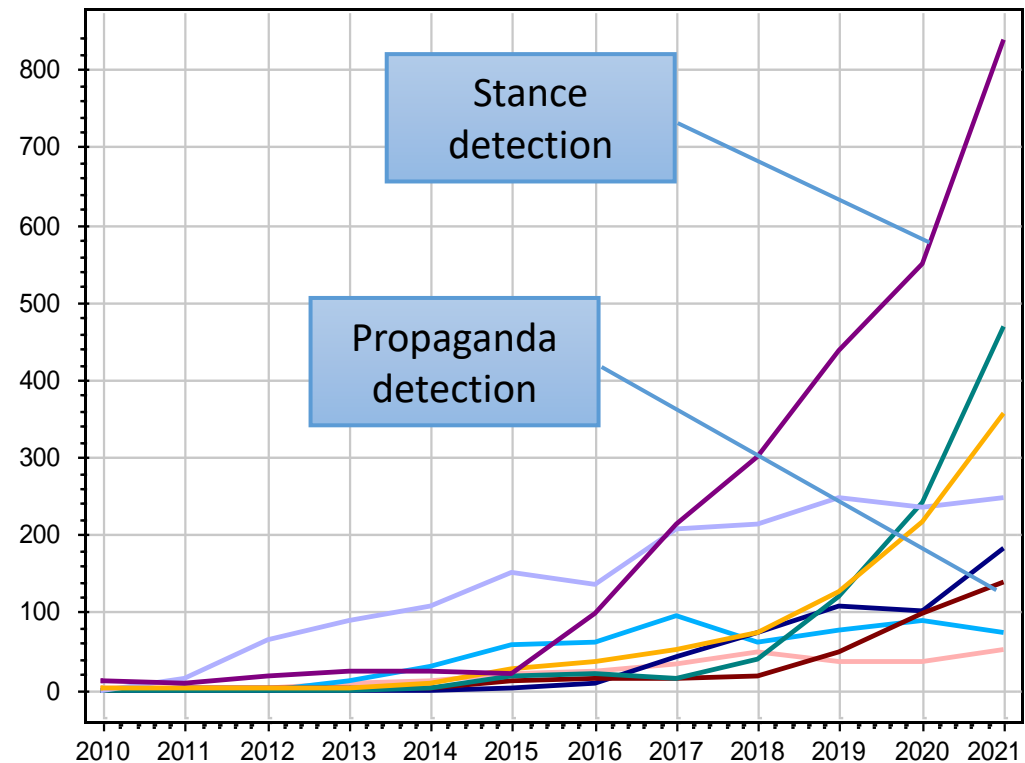
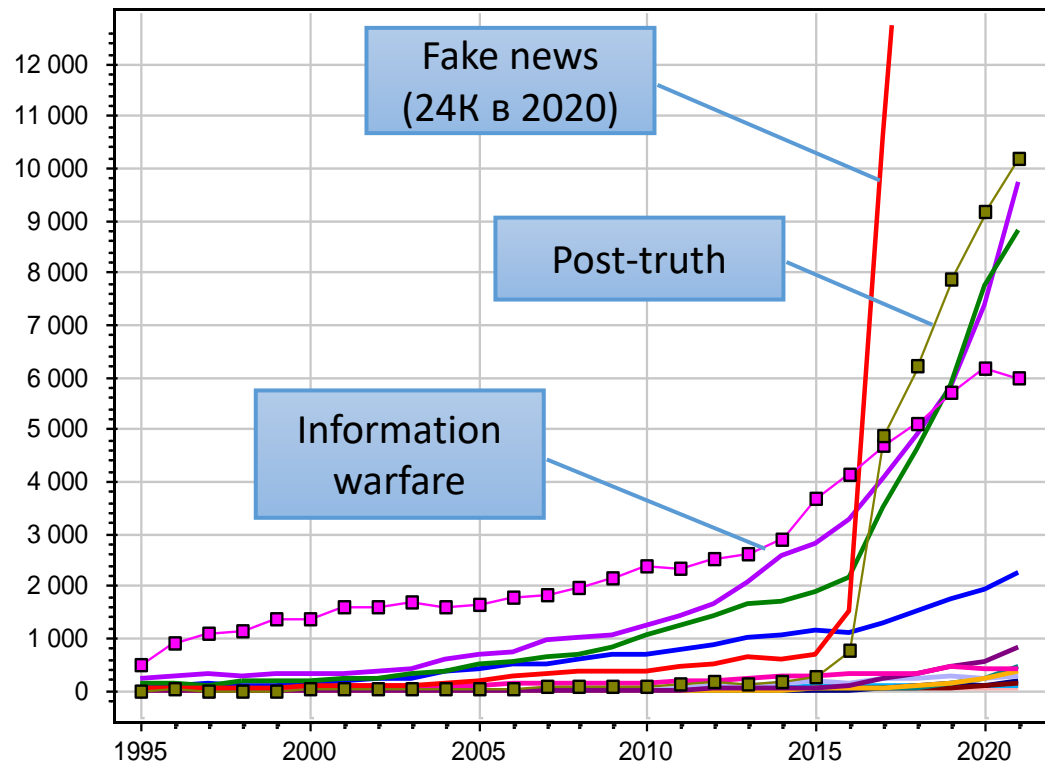
Feldman D. G., Sadekova T. R., Vorontsov K. V. [Combining Facts, Semantic Roles and Sentiment Lexicon in A Generative Model for Opinion Mining](#). Computational Linguistics and Intellectual Technologies. Dialogue 2020.

Fake News и смежные области 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
 ■ propaganda detection
- clickbait detection
 ■ controversy detection
 ■ deceptive opinion spam
 ■ virality prediction

Постановка задач в машинном обучении (ML)

Этап №1 – обучение (train)

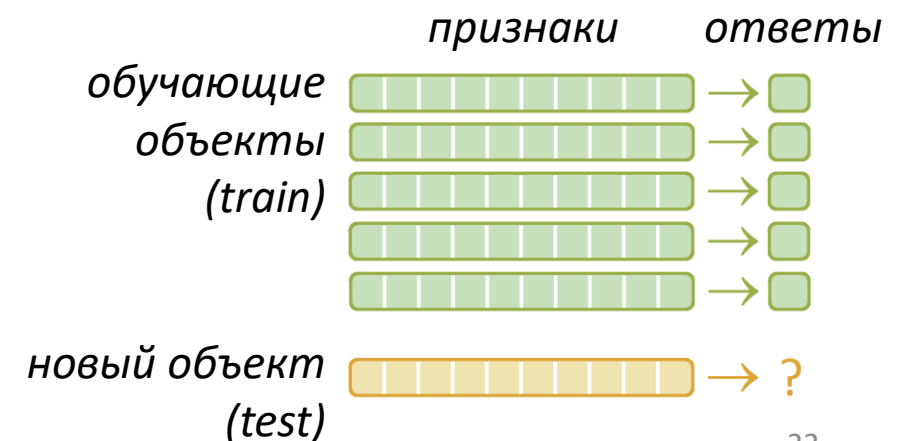
- **На входе:**
обучающая выборка пар «объект → ответ»
- **На выходе:**
модель, предсказывающая ответ по объекту

Задача поставлена, если у неё есть «ДНК»:

- **Дано**
- **Найти**
- **Критерий**

Этап №2 – применение (test)

- **На входе:**
данные – новый объект
- **На выходе:**
предсказание ответа на новом объекте



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