семинар Центра прикладного анализа больших данных
Томского государственного университета

Задачи выявления речевых манипуляций и поляризации общественного мнения в новостных текстах

Воронцов Константин Вячеславович

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

k.v.vorontsov@phystech.edu

Постправда (post-truth, «слово года» 2016)

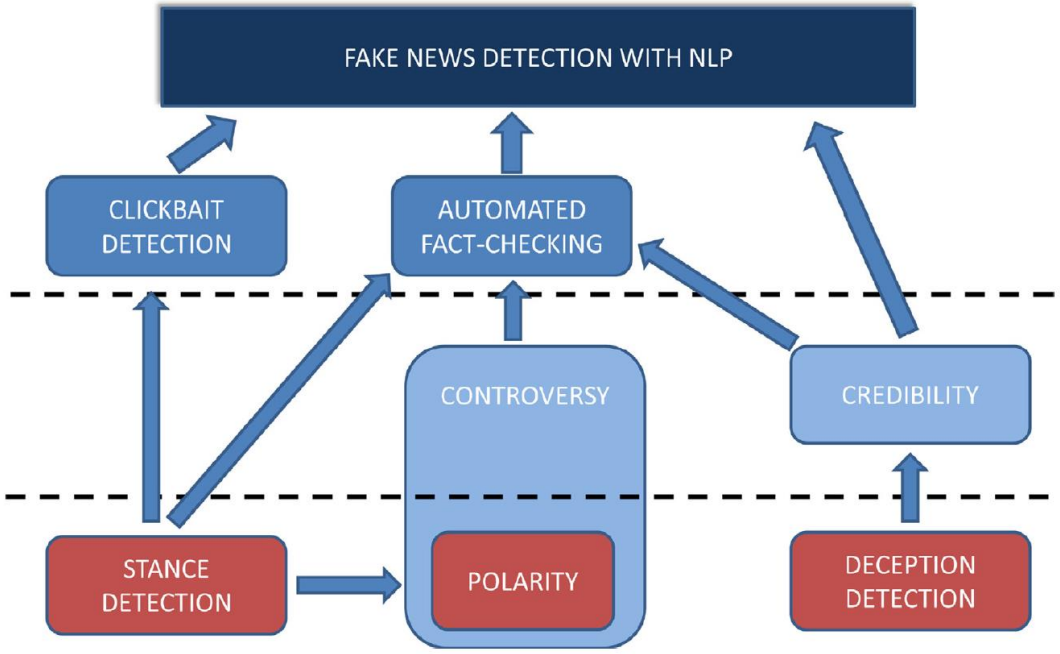
- маскируется под «другие грани истины»
- порождает явления «неопровержимой лжи» и «информационных пузырей»
- обесценивает истину подменой информирования «инфотейментом»
- создаёт ложные идеологизированные и мифологизированные картины мира
- разрушает социокультурный код
- становится инструментом «мягкой силы» в гибридных войнах

Предпосылки явления постправды

- **Психологические:**
  - для людей факты менее значимы, чем эмоции и личные убеждения
  - люди охотнее распространяют ложь и негатив, чем правду и позитив
  - люди подвержены даже таким грубым приёмам пропаганды, как повторение
- **Политические:**
  - теория пропаганды Уолтера Липпмана вытеснила теорию Джона Дьюи в 60-х
  - постправда — удобный инструмент «мягкой силы» в гибридных войнах
- **Технологические:**
  - интернет увеличил скорость распространения информации и охват аудитории
  - появились технологии генерации фейковых новостей, изображений, видео
  - СМИ лишились рекламных бюджетов и функции «четвёртой власти»

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.

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 (LBC)

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

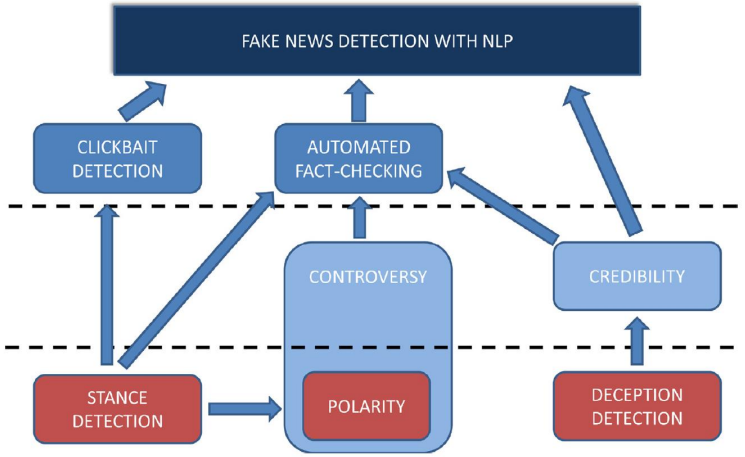
Подача информации

- оторванность от контекста: пониженная детализация места, времени, событий
- упрощение, пониженное лексическое разнообразие, лексическая недостаточность

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.

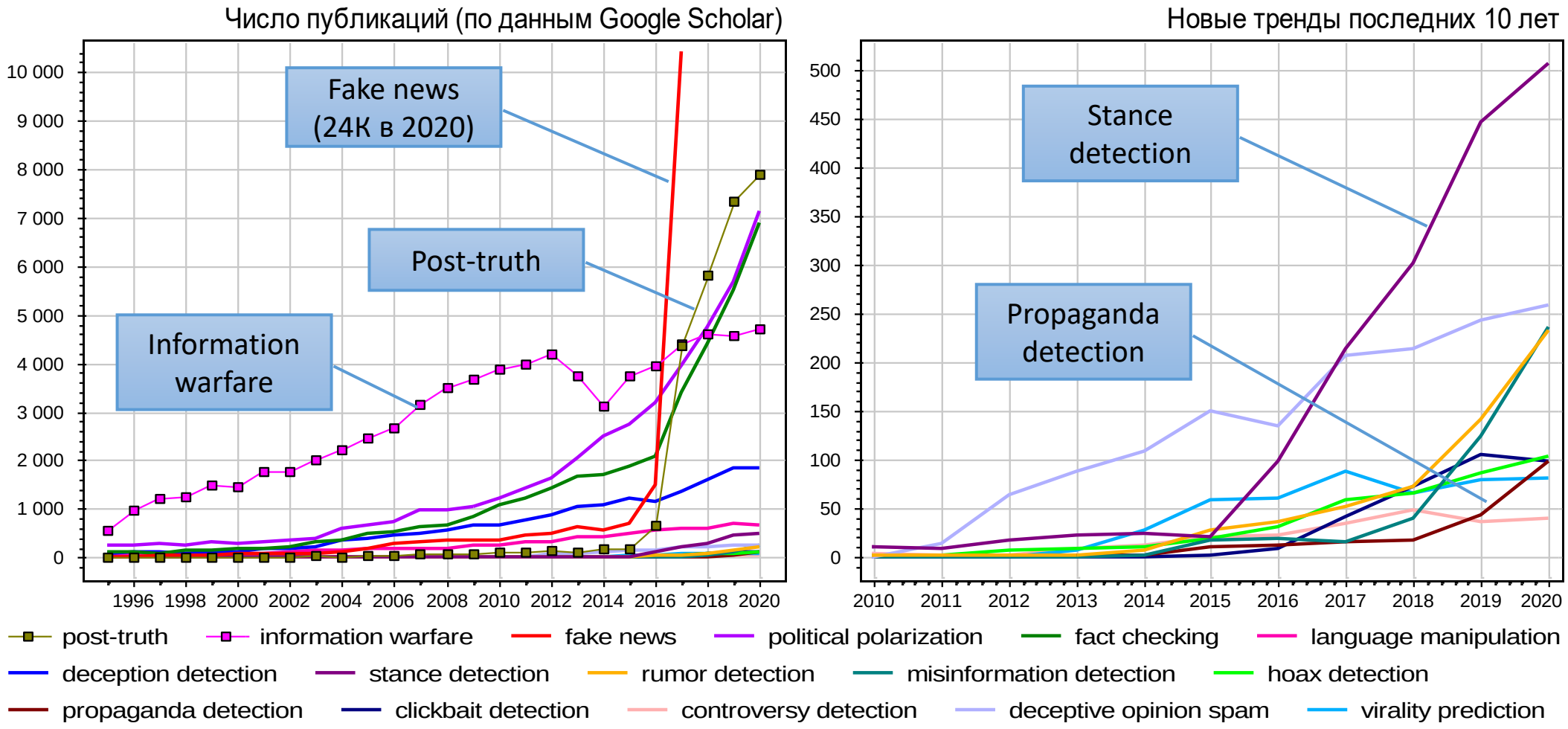
- Как распознавать манипулятивные воздействия и идеологические атаки?
- Как находить противоречия и замалчивание?
- Насколько расширится типология задач?



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)



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
   - выявление приёмов манипулирования
   - выявление конструктов 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 «картин мира» – устойчивых сочетаний суждений и идеологем

Выявление 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:

- 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.)

Детализация приёмов манипулирования

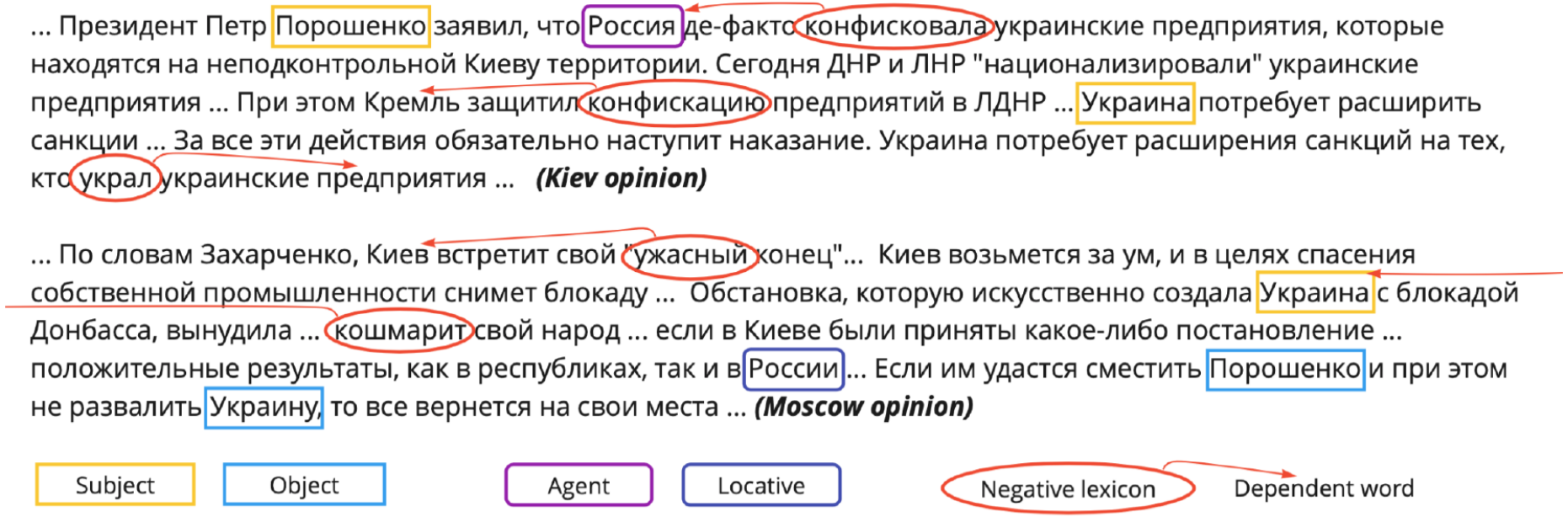
- обесценивание, троллинг, газлайтинг, буллинг, остракизм
- гиперболизация
- эвфемизм, нейтрализация, смягчение, замена языковых табу
- дисфемизм, придание негативной смысловой нагрузки
- метафоризация
- отвлечение внимания
- замалчивание
- отсутствие ссылок на источники
- отмывание пропаганды (обращение к менее надёжному источнику)
- создание образа врага
- дискредитация ценностей
- запугивание, речь ненависти
- …

Детализация приёмов манипулирования

(демагогические приёмы, логические уловки, эксплуатация когнитивных искажений)

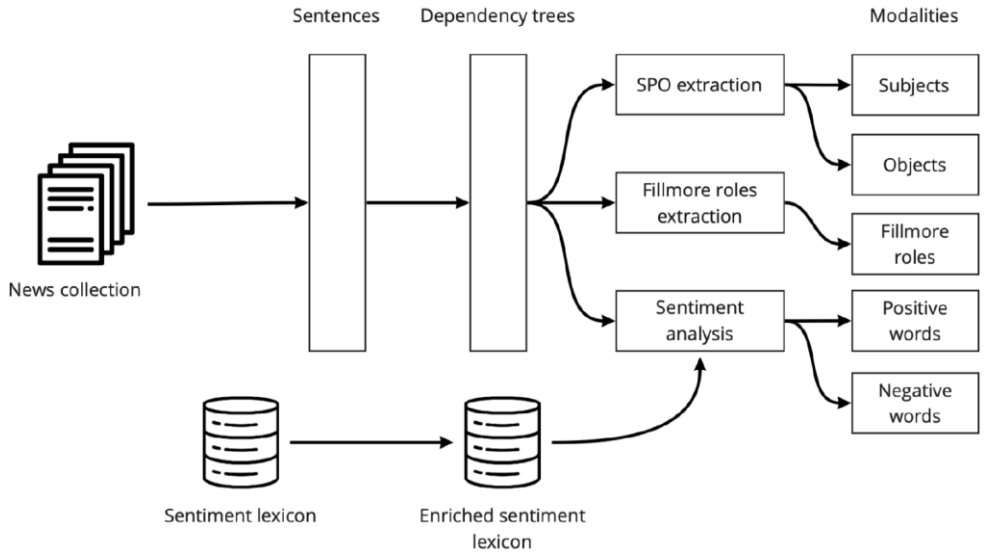
- переход на личности (ad hominem)
- безосновательные оскорбления
- перенос критики, «сведение к Гитлеру»
- аргументация к мнению большинства (argumentum ad populum)
- подмена тезиса (ignoratio elenchi, «соломенное чучело», straw man)
- предвзятая интерпретация
- концентрация на частностях
- апелляция к очевидности, ложная авторитетность
- ложная гордость слушателя («всем известно», «давно доказано»)
- аргумент к незнанию, неосведомлённости (argumentum ad ignorantiam)
- ложная пресуппозиция
- ложная альтернатива, ложная дилемма
- …

Выявления поляризации мнений по теме



- 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

Задача выделения мнений в теме или событии



Modalities	*Pr*	*Rec*	*F*1
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	*Pr*	*Rec*	*F*1
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, именованных сущностей, 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.

Выявление 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, образующих реконтекстуализацию

Обучающая выборка:

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

Сухой остаток

- Прикладные области детекции поляризации и манипулирования:
  - мониторинг «точек информационной напряжённости» в обществе
  - отношение общества к климатической повестке
  - отношение общества к пандемии COVID-19
  - распространение лженаучных и конспирологических теорий
- Технологии ML/NLP требуют формирования размеченных выборок
- Это магистральный путь формализации гуманитарных знаний
- Междисциплинарный подход: объединения усилий AI-инженеров, лингвистов, психологов, политологов, журналистов
- Более подробный доклад — на конференции ММРО-2021:
  https://www.youtube.com/watch?v=pPIsC38i8JQ&t=2700s