# Выявление пропагандистских текстов в корпусе новостных публикаций

Р.И. Мухамедиев, О.Г. Филатова, К.О. Якунин

***Аннотация* — В статье продемонстрированы возможности использования тематического моделирования (topic modeling) для идентификации пропаганды в СМИ. В современных условиях усиливающегося информационного противостояния между странами пропаганда и контрпропаганда выходят на первый план, так как государствам необходимо оградить своих граждан от различных информационных угроз, обеспечить их безопасность, что является обязательным условием для дальнейшего развития государства. А для этого, прежде всего, необходимы исследовательские проекты, тестирующие методы выявления пропаганды. Один из таких проектов, ориентированный на применение систем искусственного интеллекта в различных прикладных областях исследований на стыке машинного обучения, обработки естественного языка и изучения социума, представлен в статье. Описанный подход для выявления столь семантически нечеткого явления, как пропаганда, предлагается впервые.**

**Предлагаемый метод включает четыре основных этапа: формирование разделов корпуса, расчет тематической модели единого корпуса, расчет оценок дисбаланса корпусов по каждой теме; экстраполяция результатов оценки дисбаланса на все документы. Метод прошел перекрестную проверку на помеченной экспертом подвыборке из 1 тыс. новостей и показал достаточно высокий результат классификации. Оценка гармонической меры (F1-Score) от 0.72 до 0.94 в зависимости от выбранного порога.**

***Ключевые слова* — Тематическое моделирование, автоматическая классификация текстов, автоматическая обработка текстов, СМИ, пропаганда.**

## I. Введение

Известно, что пропаганда существовала всегда, а сам термин используется с 1622 года (когда Папа римский основал «Священную конгрегацию пропаганды веры»). Однако исследования пропаганды приобретают особую актуальность именно сейчас, в эпоху постправды, фейковых новостей и обострения информационного противоборства между государствами. Стремительные преобразования в сфере коммуникаций открывают новую исследовательскую лакуну и требуют изучения [1].

Представленный ниже инициативный исследовательский проект осуществлялся в апреле-июне 2020 г. Исследование выполнено международным коллективом, в который входили медиа-эксперты (О.Г. Филатова, Россия; Дж. М. Ионеску, Румыния) и группа исследователей из Казахстана под руководством профессора Р. И. Мухамедиева. В проекте использовался опыт участия в течение последних лет в серии проектов [2-3], ориентированных на применение систем искусственного интеллекта в различных прикладных областях исследований на стыке машинного обучения, обработки естественного языка и изучения социума. Под пропагандой понимается систематическое информационное влияние субъекта пропаганды на целевые аудитории для достижения заранее определенных целей.

Основная цель данной статьи – продемонстрировать возможности использования подхода, основанного на тематическом моделировании (topic modeling) для идентификации пропаганды в СМИ. Отметим, что, по нашему мнению, описанный подход для выявления столь семантически нечеткого явления, как пропаганда, предлагается впервые.

Статья получена 15 марта 2021 г.
Мухамедиев Равиль Ильгизович, Satbayev University, (email: r.mukhamediev@satbayev.university)
Филатова Ольга Георгиевна, Санкт-Петербургский государственный университет (email: o.filatova@spbu.ru)
Якунин Кирилл Олегович, Satbayev University (email: yakunin.k@mail.ru)

## II. Обзор подходов к анализу больших объемов текстовой информации и выявления текстов с пропагандистским содержанием

Анализ больших объемов текстовой информации в настоящее время обеспечивается методами автоматической обработки естественно языковых текстов (Natural Language Processing – NLP). Эти технологии позволяют пользователям получать информацию из больших объемов текстовых данных [4], обеспечивают анализ контента [5], персонализированный доступ к новостям [6], и даже поддерживают их производство и распространение [7-8]. Ключевыми аспектами, позволившими получить впечатляющие результаты в области автоматической обработки текстов на естественном языке, являются, согласно современным исследованиям [9], достижения в развитии методов машинного обучения, многократное увеличение вычислительной мощности, наличие большого объема лингвистических данных и развитие понимания структуры естественного языка в приложении к социальному контексту.

Проблема автоматической классификации текстов с пропагандистским содержанием исследуется в ряде работ [10-11], однако объём исследований на порядок меньше, чем в области анализа тональности (sentiment analysis), как и объём размеченных по пропаганде корпусов.

Следует отметить, задача выявления пропаганды вначале ставилась как часть задачи выявления фейковых

сообщений, позднее она выделилась в отдельную задачу.

При этом, поскольку новостная публикация может содержать как объективную информацию, так и пропаганду, текущие исследования сосредоточены на попытках выявления пропагандиского контента в отдельных предложениях (sentence level).

Так, в одной из недавних работ [12] представлен подход, основанный на «мешке слов» (bag-of-words), позволяющий классифицировать пропаганду на уровне отдельных предложений, с показателем качества F1 Score ~0.6. В работе [13] представлен результат по выявлению пропагандистских техник в текстах на уровне F1 Score =0.553. Для достижения подобного в общем то посредственного для бинарной задачи показателя качества используются весьма тяжеловесные модели трансформеров таких как BERT [14].

Сравнительный анализ использования последних моделей трансформеров ELMo, BERT, RoBERTa для выявления предложений с пропагандистским содержанием приведен в работе [15]. Модифицированные модели BERT-BiLSTM-Capsule [16].

Детальное описание решения классификационной задачи по выявлению текстов с пропагандистским содержанием приведено в [17]. В указанной работе задача решается как на уровне всего текста, так и на уровне отдельных предложений. При этом применяется относительно простой классификатор.

Появление сразу нескольких работ, начиная с конца 2019 года показывает взрывной рост интереса исследователей к задаче выявления пропаганды с помощью наиболее совершенных и сложных моделей машинного обучения.

Вместе с тем, одним из методов, продуктивно применяемых в области NLP, является тематический анализ или тематическое моделирование. Тематическое моделирование – метод, основанный на статистических характеристиках коллекций документов, который применяется в задачах автоматического реферирования, извлечения информации, информационного поиска и классификации [18]. Смысл данного подхода заключается в интуитивном понимании того, что документы в коллекции образуют группы, в которых частота встречаемости слов или сочетаний слов различается.

Расцвет данного направления исследований пришелся на 2012-2013 годы, после чего в 2019 году количество публикаций с термином «тематическое моделирование» (topic modeling) уменьшилось более чем вдвое (181 тыс. в 2013 и 78 тыс. в 2019, по данным scholar.google.com). В тоже время, тематическое моделирование, как метод решения прикладных задач в области обработки текстов остается в поле зрения исследователей. На рисунке 1 показаны результаты поиска научных публикаций за 10-ти летний период. Кривая сверху показывает изменение ежегодного количества публикаций с термином «topic modeling», кривая в середине для публикаций, в которых встречаются два термина одновременно: «topic modeling» и «sentiment analysis», нижняя кривая для сочетания «topic modeling» и «propaganda».

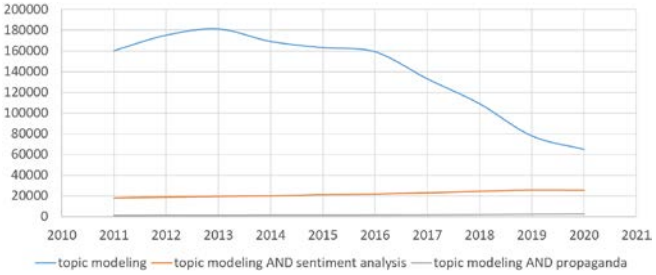


Рисунок 1. Изменение ежегодной публикационной активности в области тематического моделирования

Анализируя представленный график публикационной активности можно сделать вывод о том, что тематическое моделирование отличается хорошей проработанностью алгоритмов и методов, которые активно применяются для решения задач автоматической обработки текстов, в том числе, для решения задач классификации текстов.

Основой современных тематических моделей является статистическая модель естественного языка. Вероятностные тематические модели описывают документы (M) дискретным распределением на множестве тем (T), а темы – дискретным распределением на множестве терминов [19]. Другими словами, тематическая модель определяет, к каким темам относится каждый документ и какие слова образуют каждую тему. Кластеры документов, относящихся к совокупности тем, формируемых в процессе тематического моделирования, в частности, позволяют решать задачи синонимии и полисемии терминов [20].

Нужно отметить, что тематическое моделирование не дает хороших результатов в решении семантически противоречивых задач и, как многие методы, основанные на статистической модели естественного языка, не способно эффективно решать задачи выявления иронии, сарказма и т.п. Тем не менее, для решения задачи оценки групп текстов или публикаций отдельных СМИ тематическое моделирование, как показывают полученные нами результаты, может быть достаточно точным инструментом оценки общего уровня пропагандистского контента.

## III. Дизайн и результаты исследования

Предложенный нами подход отличается от существующих тем, что анализ происходит на более высоком уровне абстракции – на уровне так называемых топиков (тем), к которым тексты могут иметь большее или меньшее отношение, в отличие от других исследований, где анализ проводится на уровне отдельных слов и фраз. Отметим сразу ограничение метода. - Предложенный подход требует большого объёма корпуса (как минимум сотни тысяч документов)

Метод был предложен и апробирован [21] в первую очередь в связи с тем, что классический подход к классификации текстов предполагает наличие значительного объёма размеченных вручную текстов из заданного корпуса (в зависимости от подхода и корпуса может потребоваться от тысяч до сотен тысяч и даже миллионов размеченных вручную текстов). В то время, как для анализа тональности и решения других «традиционных» задач существует большое количество размеченных корпусов достаточного объёма из самых разных областей (посты в соц. сетях, отзывы и обзоры,

комментарии т. п.), имеется много множество задач с явной нехваткой размеченных данных – к таким задачам как раз и относится идентификация пропаганды, а также, например, социальной значимости, резонансности (популярности) публикаций и др. Поскольку во многих гуманитарных областях, включая социологию, политологию, психологию, существует потребность в автоматической классификации текстов, не ограничиваясь задачами определения тональности, нами была предложена модель, требующая минимальный объём ручной разметки (в случае данного исследования проводится разметка новостных источников), либо вообще без ручной разметки. Последний вариант возможен в случае, когда есть некое явное свойства публикаций, которое сильно связано (коррелирует) с целевым (неявным) свойством. Например, в случае если целевое свойство – это потенциальная популярность/резонансность публикации, то его можно связать с объективными показателями вовлечённости пользователей (просмотры, комментарии, лайки, репосты).

Предложенная модель также может быть рассмотрена как альтернативный подход к использованию принципа Transfer Learning, в том смысле, что модель использует эффективное векторное представление (embedding), основываясь на большом объёме неразмеченных данных. Следовательно, нужно отметить, что даже те документы, которые не могут быть отнесены к определённому подкорпусу (или для которых явное свойство, например вовлечённость пользователей, неизвестно), всё ещё могут быть использованы на этапе тематического моделирования для получения более эффективных векторных представлений.

Предложенный метод состоит из четырёх этапов:

1. Формирование корпуса текстов и его разделение на подкорпусы, используя некое явное (объективное) свойство публикаций (в случае данной работы – новостной источник).
2. Расчёт тематической модели полного корпуса.
3. Оценка меры межкорпусного тематического дисбаланса.
4. Экстраполяция полученных оценок дисбаланса на все документы корпуса, включая те, для которых значение явного свойства (см. этап 1) неизвестно (например, когда уровень пропагандистского содержания новостного источника оценить затруднительно или он неизвестен).

Далее остановимся на каждом этапе подробнее.

***Этап 1. Формирование корпуса текстов***

Нами был сформирован корпус новостных публикаций из открытых русскоязычных новостных источников. Предложенный метод предполагает, что корпус должен быть разделён на два или более отдельных корпусов, основываясь на некое явное свойство (в данном случае – новостной источник публикации) с целью выявить особенности, позволяющие определить некое неявное целевое свойство (в данном случае идентификации пропаганды в тексте публикации).

Русскоязычные СМИ в Российской Федерации представлены государственными холдингами, контролирующими и курирующими деятельность данных СМИ, частными или «независимыми СМИ» (так называемые в России «оппозиционные» или «либеральные» медиа) и иностранными СМИ публичной дипломатии, вещающими на русском языке, принадлежащими правительствам США, Великобритании, Германии и Франции, которые в России зарегистрированы в качестве иностранного агента в соответствии с Федеральным Законом 426-ФЗ от 02.12.2019 «О внесении изменений в Закон Российской Федерации "О средствах массовой информации"» и Федеральный закон «Об информации, информационных технологиях и о защите информации». Существуют также два государственных российских СМИ публичной дипломатии, Sputnik и RT, из-за внутреннего регулирования их деятельности, их целевая аудитория – это не граждане России, а граждане других государств. Для более точного исследования мы решили проанализировать пропагандистские СМИ с явно субъективной риторикой и сравнить их с теми СМИ, чья риторика является наиболее объективной.

Международные вещатели являются частью системы государственной публичной дипломатии, поэтому их риторика явно пропагандистская. Их цель – это продвижение имиджа и политических интересов государства в иностранном гражданском обществе. Поэтому мы решили проанализировать такие СМИ, как RT, Sputnik, Радио Свобода, Настоящее время и Deutsche Welle.

С другой стороны, не во всех частных русскоязычных СМИ отсутствует пропаганда. Если у государственных средств массовой информации риторика по отношению к правящему политическому классу является положительной, то так называемые оппозиционные СМИ направлены явно против властей. Поэтому, мы решили не анализировать такие медиа, как «Дождь», «Медуза» или «Новая газета» из-за выраженной антиправительственной пропаганды.

Вместо них мы решили проанализировать такие СМИ, которые менее вовлечены непосредственно в политическую жизнь страны, и которые больше ставят акцент на бизнес-среду или экономику, поскольку именно такие СМИ имеют более объективную риторику. Поэтому, мы проанализировали одно довольно нейтральное информагентство (Interfax), три бизнес-ориентированных СМИ (РБК, Ведомости, Бизнес ФМ) и одну интернет-газету (Lenta.ru).

Таким образом, исходя из соображений, изложенных выше, корпус, состоящий из 428180 публикаций за 2018 – 2020 гг., был разделён на два корпуса, которые составили публикации из упомянутых источников:

1) Пропагандистские публикации (346440 публикаций):
   a. RT
   b. Настоящее время
   c. Радио свобода
   d. Deutsche Welle
   e. Sputnik
2) Условно-объективные публикации (81740 публикаций):
   a. Ведомости
   b. Interfax
   c. Lenta.ru
   d. Бизнес FM
   e. РБК

***Этап 2. Тематическое моделирование***

Для построения тематической модели корпуса документов применяют: вероятностный латентно-семантический анализ (PLSA), ARTM (Additive regularization of topic models) [22] и, весьма популярное, латентное размещение Дирихле (LDA) [23-24]. LDA может быть выражен следующим равенством:

$$p(w,m) = \sum_{t \in T} p(w \mid t,m)\, p(t \mid m)$$
$$= \sum_{t \in T} p(w \mid t)\, p\,(t \mid m) = \sum_{t \in T} \varphi_{wt}\theta_{tm}$$

представляющим сумму смешанных условных распределений по всем темам множества T, где $p(w \mid t)$ условное распределение слов в темах, $p(t \mid m)$ условное распределение тем по новостям. Переход с условного распределения $p(w \mid t,m)$ на $p(w \mid t)$ осуществляется за счет гипотезы условной независимости, согласно которой появление слов в новостях m по теме t зависит от темы, но не зависит от новости m, и есть общее для всех новостей. Данное соотношение справедливо, исходя из допущений об отсутствие необходимости сохранения порядка документов (новостей) в корпусе и порядка слов в новости, помимо этого, метод LDA предполагает, что компоненты $\varphi_{wt}$ и $\theta_{tm}$ порождены непрерывным многомерным вероятностным распределением Дирихле. Целью алгоритма, является поиск параметров $\varphi_{wt}$ и $\theta_{tm}$, путем максимизации функции правдоподобия с соответствующей регуляризацией

$$\sum_{m \in M} \sum_{w \in m} n_{mw} ln \sum_{t \in T} \varphi_{wt}\theta_{tm} + R(\varphi,\theta) \to max$$

$n_{mw}$ – число вхождений слова $w$, в новость $m$, $R(\varphi,\theta)$ – логарифмический регуляризатор.

Для определения оптимального количества тематических кластеров T, часто применяется метод максимизации значения когерентности, рассчитанной с применением UMass метрики [25]

$$U\ \left(w_i, w_j, \varepsilon\right) = log\,\frac{M\left(w_i, w_j\right) + \varepsilon}{M(w_j)}$$

Где $M\left(\mathrm{w_i}, \mathrm{w_j}\right)$ –количество документов, содержащих слова $\mathrm{w_i}$ и $\mathrm{w_j}$, а $\mathrm{M(w_j)}$ - количество документов содержащих только слово $\mathrm{w_j}$. На основе указанной меры вычисляется значение когерентности отдельного тематического кластера

$$Coh\ (W_k)\ = \sum_{(\mathrm{w_i w_j}) \in \mathrm{W}} U\ (\mathrm{w_i w_j}, \varepsilon)$$

Где $W_{\mathrm{k}}$- множество слов кластера, ε коэффициент сглаживания обычно равный 1.

Чем больше документов с двумя словами относительно документов содержащих только одно слово, тем выше значение когерентности отдельного топика. В итоге выбирают такое количество тематических кластеров при котором достигается максимум усредненного значения когерентности

$$Coh\ (W_k)\ = \underset{\mathrm{T}}{\mathrm{argmax}} \frac{1}{T} \sum_{k \in \mathrm{T}} Coh\ (W_k)$$

Расширением LDA является ARTM, который реализован виде библиотеки с открытым исходным кодом BigARTM. Эта библиотека и была применена авторами для тематической классификации. Описание ARTM приведено, в том числе, в работе [28].

***Этап 3. Оценка межкорпусного дисбаланса***

Следующий этап – определение межкорпусного дисбаланса в распределении новостных публикаций разных корпусов в рамках каждого отдельного топика. Эта мера дисбаланса рассматривается как оценка влияния принадлежности к данному топику на целевой показатель (пропаганда), поскольку изначальное разделение на корпусы было проведено на основании явного объективного свойства (новостной источник), принимая во внимание соображения, позволяющие утверждать, что существует присущий этому разделению дисбаланс между пропагандой и условно-объективной информацией между этими двумя корпусами.

Формула меры дисбаланса:

$$D_{t_i c_j} = \frac{\sum_k w_{d_k t_i c_j}}{\sum_k \sum_l w_{d_k t_l c_j}} \Big/ \sum_m \sum_k \sum_l w_{d_k t_l c_m}$$

В данной формуле $D_{t_i c_j}$ – мера дисбаланса представленности документов из корпуса $c_j$ в топике $t_i$, а $w_{d_k t_l c_m}$ –вес принадлежности документа $d_k$ из корпуса $c_m$ к топику $t_l$.

***Этап 4. Экстраполяция оценок межкорпусного дисбаланса на все документы корпуса и валидация результатов классификации***

Последним этапом предложенного метода является применение полученной тематической модели и оценок дисбаланса для получения классификации каждого отдельного документа. Для этого существует две основные причины:

Несмотря на то, что были выбраны пропагандистские и условно-объективные источники, распределение пропагандистского содержания в новостях, безусловно, неравномерно, то есть пропагандистские публикации могут быть опубликованы в условно-объективных источниках и наоборот.

Как упоминалось выше, не все источники можно отнести к одному из двух корпусов, поскольку пропагандистское содержание определённых источников может быть сложно оценить.

Для агрегации оценок межкорпусного дисбаланса с весами отношения документов к каждому топику может быть применено несколько подходов:

1. Просто взвешенное среднее – именно этот подход использовался для получения описываемых в этой главе результатов

2. Байесовский подход к агрегации – этот подход рассматривает субъективные вероятности отношения

документа к заданному критерию. Преимущества подхода описаны в ряде публикаций [26-28].

3. Semi-supervised подход (полуобучаемый) [29] дает возможность предобучить модель на результатах, полученных путём применения описываемого подхода, а затем провести дообучение (fine tuning) модели на вручную размеченном наборе текстовых данных для увеличения качества работы модели.

Для проведения валидации предложенной модели была сформирована репрезентативная случайная выборка из 1000 публикаций из оригинального корпуса, которые были исключены из процесса тематического моделирования и расчёта мер межкорпусного дисбаланса. Эти публикации были вручную размечены экспертами по шкале Лайкерта от –2 до +2, где –2 – это условно объективная публикация, а +2 – пропаган-дистская.

Затем модель была применена к этой выборке для расчёта метрик качества работы модели. Шкала Лайкерта была линейно нормализована в интервале от 0 до 1, экстраполированные оценки пропаганды также были нормализованы от 0 до 1. На основании полученных значений была рассчитана корреляция Пирсона. Коэффициент корреляции Пирсона показывает меру взаимосвязи между экспертной разметкой и результатами модели. Коэффициент может варьироваться от 0 до 1, где 0 – полное отсутствие взаимосвязи, а 1 – полная чёткая связь между двумя показателями. При этом в гуманитарных исследованиях корреляция, выше, чем 0.2–0.3 на выборке достаточного объёма, считается доказательством наличия слабой, но достоверной связи между показателями.

Затем объектам были назначены классы – публикации с оценкой выше 0.5 были отнесены к классу «пропаганда», а ниже – к классу «объективные». Эти данные были использованы для расчёта метрик качества классификации – точность (precision) и ROC AUC [30].

## IV. Метрики оценки качества классификации

Оценка результатов классификации требует адекватных метрик, позволяющих численно оценить качество метода. Часто упоминаемый перечень метрик оценки классификаторов следующий [3]:

Accuracy
Precision
Recall
F1-score
F-score
Precision-Recall curve
ROC curve

Простейшая метрика (accuracy) – процент (доля) правильно классифицированных примеров. Однако, в случае классов, существенно неравных по количеству объектов, используют следующие показатели: «точность» (precision), «полнота» (recall), и обобщающий показатель гармоническое среднее – F1 Score.

При расчете показателей подсчитывают случаи правильной работы классификатора - True positive (TP) и True negative (TN) и неправильной работы - False negative (FN) и False positive (FP). FN или ошибка первого рода, возникает тогда когда объект классификации ошибочно отнесен к негативному классу, являясь на самом деле позитивным. FP или ошибка второго рода, наоборот, признак излишне оптимистического, или неосторожного, классификатора, то есть модель предсказала положительный результат для отрицательного объекта. Тогда

**Precision:** $P = \frac{TP}{(TP+FP)}$.

**Recall:** $R = \frac{TP}{(TP+FN)}$.

**F1 Score:** $F1Score = \frac{2*P*R}{(P+R)}$

.

Кривая Receiver Operating Characteristic (ROC), иногда называемая кривой ошибок, связывает показатели TP и FP. На рисунке 2 показана ROC кривая для почти идеального классификатора.

Для обобщения ROC кривой используется оценка площади под кривой - Area Under the Curve (AUC), которая изменяется от 1 (наилучший вариант классификатора) до 0.5 (наихудший случай).

ROC AUC является обобщённой метрикой оценки моделей машинного обучения, устойчивой к дисбалансу классов.

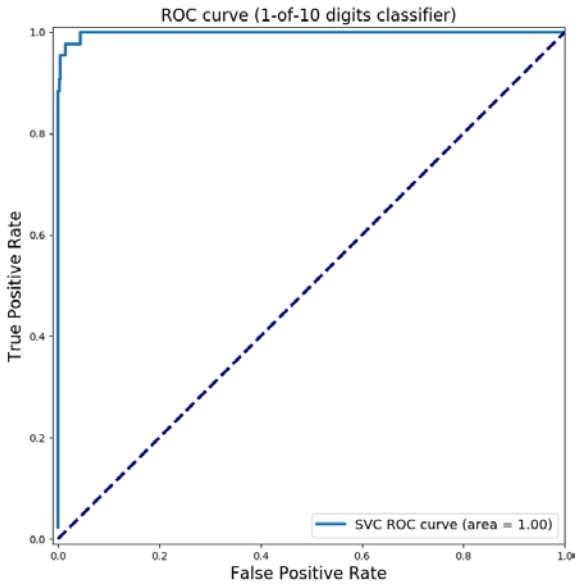


Рисунок 2. ROC кривая оптимального классификатора

ROC AUC 0.5 является индикатором того, что модель не обладает предсказательной способностью (работает полностью случайно). ROC AUC больше 0.6 считается показателем наличия слабой предсказатель-ной способности, а ROC AUC выше 0.9 – показатель очень высокого качества распознавания. Например, в исследованиях, связанных с медицинской диагностикой, обычно считают модель приемлемой для использования на практике при ROC AUC выше 0.9-0.95 [31].

Результаты, полученные на этапе 4 оценены с использованием

## V. Результаты и обсуждение

Результаты, полученные в ходе классификации приведены в таблице 1. В таблице порог классификации показывает, какие нормированные значения результатов работы модели рассматривались как результаты с достаточной уверенностью. Например «<0.4 или >0.6» означает, что значения между 0.4 и 0.6 относились к классу «неизвестно/спорно», результаты меньше 0.4 к

классу «объективные публикации», выше 0.6 – к классу «пропаганда». Это связано с тем, что при экспертной разметке выяснилось, что многие документы затруднительно отнести точно к тому или иному классу, соответственно необходимо было проверить, что предложенная модель позволяет отличить не только пропаганду от не-пропаганды, но и пропаганду от спорных/трудных для классификации статей.

Таблица 1. Результаты верификации модели

| Порог классификации | Корреляция Пирсона | Гармоническая мера (F1-Score (micro)) | ROC AUC |
|---|---|---|---|
| Без порога | 0.47 | 0.72 | 0.73 |
| <0.4 или >0.6 | 0.61 | 0.81 | 0.8 |
| <0.3 или >0.7 | 0.81 | 0.94 | 0.95 |

Даже без такого порога модель демонстрирует достаточную предсказательную способность (ROC AUC 0.73, F1-Score= 72%), тогда как с порогом «<0.3 или >0.7» предсказательная способность очень высокая (ROC AUC 0.95, F1-Score= 94%). Это означает, что модель работает значительно более точно на предельных значениях, но она в любом случае позволяет отличить и те новости, которые затруднительно отнести к классу «пропаганда».

Следует добавить также, что объективные новости распознаются чуть лучше, чем пропаганда, хотя и незначительно. Можно заметить, что recall (охват) гораздо больше у объективного класса (0.83 против 0.58). Это значит, что модель из всех объективных новостей нашла 83% таких новостей, а из всех пропагандистских – только 58% таких новостей.

Точность (precision) при этом примерно одинаковая (0.71 и 0.74). То есть из тех новостей, которые модель назвала объективными, 71% процент действительно таковыми являются, а из тех новостей, которые модель назвала пропагандистскими, таковых 74 процента.

Таким образом, предложенная модель позволяет получить высокую предсказательную способность, при минимально возможном объёме ручной экспертной разметки. Предлагается использовать высокоуровневую разметку корпуса по некоему явному свойству, которое должно иметь достаточную связь (корреляцию) с целевым неявным свойством. В данном случае разделение происходило по новостным источникам, однако могут быть использованы и другие варианты разделения на корпусы, включая автоматические.

**Ограничения метода.** Нужно отметить, что тематическое моделирование не дает хороших результатов в решении семантически противоречивых задач и, как многие методы, основанные статистической модели естественного языка, не способно эффективно решать задачи выявления иронии, сарказма и т.п.

Тематическое моделирование не способно решить задачу выявления пропаганды на уровне отдельных предложений.

Тем не менее, для решения задачи оценки групп текстов или публикаций отдельных СМИ тематическое моделирование на наш взгляд способно представить обобщенную картину состояния с довольно высокой точностью. Кроме того, полученный результат можно рассматривать как отправную точку при сравнении с более сложными моделями.

Описанный метод применялся с некоторыми дополнениями для оценки социально-значимого контента новостных публикаций [28] и оценки дисбаланса новостных публикаций о возобновляемых источниках энергии [32].

## VI. Заключение

В работе описан метод классификации новостных текстов с пропагандистским содержанием на основе тематической модели корпуса новостных документов. Метод включает четырехэтапный процесс:

1. Формирование корпуса текстов и его разделение на подкорпусы, используя некое объективное свойство публикаций (в случае данной работы – новостной источник).
2. Расчёт тематической модели полного корпуса.
3. Оценка меры межкорпусного тематического дисбаланса.
4. Экстраполяция полученных оценок дисбаланса на все документы корпуса.

Валидация работы модели на случайной выборке из 1000 новостей, показала высокую предсказательную способность – точность от 64% до 88% в зависимости от порога классификации. Проведенная работа ожидаемо показала, что оценивать уровень пропаганды довольно трудно – как для экспертов, так и для машины.

Конечно, встречаются очень яркие образцы пропагандистского контента, но, тем не менее, часто заведомо «пропагандистские» СМИ пишут не только пропагандистские статьи. И, наоборот, не все объективные средства массовой информации всегда объективны. К тому же уровень журналистского мастерства разный в разных СМИ и часто встречаются публикации крайне низкого качества, затрудняющие понимание смысла. Поэтому в дальнейшем следует более тщательно подходить к отбору конкретных массмедиа, а также – постепенно расширять перечень СМИ и уточнять, какие СМИ относятся к какому корпусу. Можно пробовать брать для эксперимента такие СМИ, например, где будет больше крайних вариантов или расширять базу, чтобы туда попадало больше таких крайних, ярких вариантов. И в идеальном случае каждый текст должны распознавать хотя бы 3 эксперта, что могло бы улучшить точность идентификации пропаганды (по сути дела, методом голосования уточнять экспертную разметку). Экспертные оценки в первую очередь нужны для того, чтобы провалидировать работу модели, но в перспективе такая экспертная разметка отдельных документов может использоваться для дообучения модели.

Машина не может пока на 100% заменить человека, но она может значительно сократить трудозатраты экспертов по выявлению пропагандистского контента. Использованный нами метод ограничен – корпусом документов, текущим набором тематик и т.п. Тем не менее, при всех ограничениях, получен достаточно высокий результат классификации. Учитывая то, что для получения такого результата не требуется большой размеченный корпус, метод может быть сравнительно легко использован на практике.

Возможности для продолжения исследований включают применение предложенного метода к другим задачам классификации текстов, таким как тональность,

социальная значимость, резонансность, оценка и моделирование изменения тематических групп во времени, в том числе, в сочетании с другими способами агрегации оценок статей.

Корпус размеченных новостных текстов, использованный для валидации результатов работы метода, можно получить по ссылке https://www.dropbox.com/s/gku10klwgztxd1w/propaganda%20-%20sample%20blind.xlsx?dl=0

## Библиография


[1] Филатова О.Г. Пропаганда в эпоху ботов, троллей и fake-news: теоретические подходы и прикладные исследования // Стратегические коммуникации в бизнесе и политике. – 2018. – Т. 1 (4). – С. 86-94.

[2] Barakhnin V.B., Muhamedyev R.I. , Mussabaev R.R., Kozhemyakina O.Yu., Issayeva A., Kuchin Ya.I., Murzakhmetov S.B., Yakunin K.O. Methods to identify the destructive information // Journal of Physics: Conf. Series. – 2019. – V. 1117. – 10 p. URL: http://dx.doi.org/10.1088/1742-6596/1117/1/012001.

[3] Muhamedyev R. Machine learning methods: An overview // Computer Modelling & New Technologies. – 2015. – Vol. 19 (6). – C. 14-29.

[4] Korencÿic D., Ristov, S., Sÿnajder, J. Document-based topic coherence measures for news media text // Expert Systems with Applications. – 2018. – Vol. 114. – P. 357-373.

[5] Neuendorf K. A. The content analysis guidebook. Sage. – 2016.

[6] Steinberger J., Ebrahim M., Ehrmann M., Hurriyetoglu A., Kabadjov M., Lenkova P., Steinberger R., Tanev H., VGŸzquez S., Zavarella V. Creating sentiment dictionaries via triangulation // Decision Support Systems. – 2012. – Vol. 53 (4). – P. 689-694.

[7] Clerwall C. Enter the robot journalist: Users' perceptions of automated content // Journalism Practice. –2014. – Vol. 8. – P. 519–531.

[8] Popescu O., Strapparava C. Natural Language Processing meets Journalism // Proceedings of the 2017 EMNLP Workshop. Copenhagen, Denmark: Association for Computational Linguistics. – 2017.

[9] Hirschberg J., Manning C. D. Advances in natural language processing // Science. – 2015. – Vol. 349 (6245). – P. 261–266.

[10] Barrón-Cedeno A. et al. Proppy: A system to unmask propaganda in online news //Proceedings of the AAAI Conference on Artificial Intelligence. – 2019. – Т. 33. – №. 01. – С. 9847-9848.

[11] Barrón-Cedeno A., Jaradat I., Da San Martino G., Nakov P. Proppy: Organizing the news based on their propagandistic content // Information Processing & Management. 2019. Vol. 56 (5). – P. 1849-1864.

[12] Da San Martino G., Yu S., Barrón-Cedeno A., Petrov R., Nakov P. Fine-grained analysis of propaganda in news article . In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). – 2019. – P. 5640–5650.

[13] Altiti O., Abdullah M., Obiedat R. JUST at SemEval-2020 Task 11: Detecting Propaganda Techniques Using BERT Pre-trained Model // Proceedings of the Fourteenth Workshop on Semantic Evaluation. – 2020. – С. 1749-1755.

[14] Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. URL: https://arxiv.org/abs/1810.04805

[15] Sadana A. et al. NSIT@ NLP4IF-2019: Propaganda detection from news articles using transfer learning // Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda. – 2019. – С. 143-147.

[16] Vlad G. A. et al. Sentence-level propaganda detection in news articles with transfer learning and BERT-BiLSTM-capsule model // Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda. – 2019. – С. 148-154.

[17] Oliinyk V. A. et al. Propaganda detection in text data based on NLP and machine learning // CEUR Workshop Proceedings. – 2020. – Vol. 2631. – P. 132-144.

[18] Машечкин И.В., Петровский М.И., Царёв Д.В. Методы вычисления релевантности фрагментов текста на основе тематических моделей в задаче автоматического аннотирования // Вычислительные методы и программирование. – 2013. – Т. 14, № 1. – С. 91–102.

[19] Воронцов К.В., Потапенко А.А. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование. – 2012. – Т. 4, № 4. – С. 693–706.

[20] Пархоменко П.А., Григорьев А.А, Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов // Труды ИСП РАН. – 2017. – Т. 29 (2). – С. 161-200. DOI: 10.15514/ISPRAS-2017-29(2)-6.

[21] Yakunin K., Ionescu G.M., Murzakhmetov S., Mussabayev R., Filatova O., Mukhamediev R. Propaganda Identification Using Topic Modelling // Procedia Computer Science. 2020. Vol. 178. P. 205–212. https://doi.org/10.1016/j.procs.2020.11.022

[22] Vorontsov K. et al. Bigartm: Open source library for regularized multimodal topic modeling of large collections // International Conference on Analysis of Images, Social Networks and Texts. – Springer, Cham, 2015. – С. 370–381.

[23] Blei D.M., Ng A.Y., Jordan M.I. Latent dirichlet allocation // Journal of machine Learning research. – 2003. – Т. 3. – No Jan. – P. 993–1022.

[24] Jelodar H. et al. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey // Multimedia Tools and Applications. – 2018. – С. 1–43.

[25] Mimno D., Wallach H., Talley Ed., Leenders M. & McCallum A. Optimizing Semantic Coherence in Topic Models // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. – 2011. – P. 262–272.

[26] Barakhnin V. B., et al.: Methods to identify the destructive information // Journal of Physics. 1405(1), 012004. – 2019.

[27] Mukhamediev R.I., Mustakayev R., Yakunin K., Kiseleva S., Gopejenko V. Multi-Criteria Spatial Decision Making Support system for Renewable Energy Development in Kazakhstan // IEEE Access. 2019. 7, 122275-122288.

[28] Mukhamediev R. I. et al. Classification of Negative Information on Socially Significant Topics in Mass Media //Symmetry. – 2020. – Т. 12. – №. 12. – С. 1945.

[29] Zhu X., Goldberg A. B. Introduction to semi-supervised learning // Synthesis lectures on artificial intelligence and machine learning. – 2009. Vol. 3(1). – P. 1-130.

[30] Bradley A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997. – Vol. 30 (7). – 1145–1159. URL: https://doi.org/10.1016/s0031-3203(96)00142-2/.

[31] Akobeng A. K. Understanding diagnostic tests 3: receiver operating characteristic curves // Wiley Online Library, 21–Mar–2007. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1651-2227.2006.00178.x.

[32] Якунин К.О., Мусабаев Р.Р., Елис М.С., Мухамедиев Р.И. Тема энергетики в новостных публикациях // Возобновляемые источники энергии. Материалы Всероссийской научной конференции и XII Молодежной школы с международным участием. 24-25 ноября 2020 г. М.: Наука, 2020. С. 451-456.



**Мухамедиев Равиль Ильгизович**, докт. инженерных. наук, профессор Satbayev University, (https://official.satbayev.university/en/teachers/ravil-muhamedyev), email: r.mukhamediev@satbayev.university, elibrary.ru: authorid=1123472, ORCID: orcidID=0000-0002-3727-043X

**Филатова Ольга Георгиевна**, канд. философ. наук, доцент кафедры «Связи с общественностью в политике и государственном управлении» Санкт-Петербургского государственного университета, Санкт-Петербург (https://spbu.ru), email: o.filatova@spbu.ru, elibrary.ru: authorid=625633, ORCID: orcidID=0000-0001-9568-1002

**Якунин Кирилл Олегович,** магистр технических наук, докторант Satbayev University - email: yakunin.k@mail.ru elibrary.ru: authorid=1099914, ORCID: orcidID= 0000-0002-7378-9212

# Identification of Propaganda Documents in the News Text Corpora

R. Mukhamediev, O. Filatova, K. Yakunin

***Abstract* — The article demonstrates the possibilities of using topic modeling to identify propaganda in the media. In modern conditions of increasing information confrontation between countries, propaganda and counter-propaganda come to the forefront, since states need to protect their citizens from various informational threats, to ensure their safety, which is a necessary condition for the further development of the state. To achieve this research projects are necessary to test methods for identifying propaganda. One of such projects, focused on the use of artificial intelligence systems in various applied research areas at the intersection of machine learning, natural language processing and social studies, is presented in the article. The described approach for identifying such a semantically fuzzy phenomenon as propaganda is proposed for the first time. The following definition for political propaganda is suggested - a coordinated, systematic informational influence of the subject of propaganda on target audiences to achieve political goals and promote political ideas.**

**The proposed method includes four main stages: formation of corpus sections, calculation of a thematic model of an overall corpus, calculation of imbalance estimates of corpuses for each topic; extrapolation of the imbalance estimates results to all documents. The method was cross-checked on a subsample of 1000 news marked by an expert and showed a fairly high classification result. Harmonic measure score (F1-Score) varies from 0.72 to 0.94 depending on the selected threshold.**

***Keywords* — Propaganda; natural language processing; topic modeling; text classification; mass media analysis**

## REFERENCES

[1] Filatova O.G. Propaganda in the era of bots, trolls and fake-news: theoretical approaches and applied research // Strategic communications in business and politics. - 2018 .-- Vol. 1 (4). - S.86-94. (In Russ.)

[2] Barakhnin V.B., Muhamedyev R.I. , Mussabaev R.R., Kozhemyakina O.Yu., Issayeva A., Kuchin Ya.I., Murzakhmetov S.B., Yakunin K.O. Methods to identify the destructive information // Journal of Physics: Conf. Series. – 2019. – V. 1117. –10 p. URL: http://dx.doi.org/10.1088/1742-6596/1117/1/012001.

[3] Muhamedyev R. Machine learning methods: An overview // Computer Modelling & New Technologies. – 2015. – Vol. 19 (6). – С. 14-29.

[4] Korencÿic D., Ristov, S., Sÿnajder, J. Document-based topic coherence measures for news media text // Expert Systems with Applications. – 2018. – Vol. 114. – P. 357-373.

[5] Neuendorf K. A. The content analysis guidebook. Sage. – 2016.

[6] Steinberger J., Ebrahim M., Ehrmann M., Hurriyetoglu A., Kabadjov M., Lenkova P., Steinberger R., Tanev H., VGŸzquez S., Zavarella V. Creating sentiment dictionaries via triangulation // Decision Support Systems. – 2012. – Vol. 53 (4). – P. 689-694.

[7] Clerwall C. Enter the robot journalist: Users' perceptions of automated content // Journalism Practice. –2014. – Vol. 8. – P. 519–531.

[8] Popescu O., Strapparava C. Natural Language Processing meets Journalism // Proceedings of the 2017 EMNLP Workshop. Copenhagen, Denmark: Association for Computational Linguistics. – 2017.

[9] Hirschberg J., Manning C. D. Advances in natural language processing // Science. – 2015. – Vol. 349 (6245). – P. 261–266.

[10] Barrón-Cedeno A. et al. Proppy: A system to unmask propaganda in online news //Proceedings of the AAAI Conference on Artificial Intelligence. – 2019. – Т. 33. – №. 01. – С. 9847-9848.

[11] Barrón-Cedeno A., Jaradat I., Da San Martino G., Nakov P. Proppy: Organizing the news based on their propagandistic content // Information Processing & Management. 2019. Vol. 56 (5). – P. 1849-1864.

[12] Da San Martino G., Yu S., Barrón-Cedeno A., Petrov R., Nakov P. Fine-grained analysis of propaganda in news article . In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). – 2019. – P. 5640–5650.

[13] Altiti O., Abdullah M., Obiedat R. JUST at SemEval-2020 Task 11: Detecting Propaganda Techniques Using BERT Pre-trained Model // Proceedings of the Fourteenth Workshop on Semantic Evaluation. – 2020. – С. 1749-1755.

[14] Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. URL: https://arxiv.org/abs/1810.04805

[15] Sadana A. et al. NSIT@ NLP4IF-2019: Propaganda detection from news articles using transfer learning // Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda. – 2019. – С. 143-147.

[16] Vlad G. A. et al. Sentence-level propaganda detection in news articles with transfer learning and BERT-BiLSTM-capsule model // Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda. – 2019. – С. 148-154.

[17] Oliinyk V. A. et al. Propaganda detection in text data based on NLP and machine learning // CEUR Workshop Proceedings. – 2020. – Vol. 2631. – P. 132-144.

[18] Mashechkin I.V, Petrovsky M.I, Tsarev D.V. Methods for calculating the relevance of text fragments based on thematic models in the problem of automatic annotation // Computational methods and programming. - 2013. – Vol. 14 (1). - P. 91-102. (In Russ.)

[19] Vorontsov K.V., Potapenko A.A. Regularization, robustness and sparsity of probabilistic topic models // Computer Research and Modeling. 2012. Vol. 14 (4). P. 693-706. (In Russ.)

[20] Parhomenko P.A., Grigorev A.A., Astrakhantsev N.A. A survey and an experimental comparison of methods for text clustering: application to scientific articles // Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS). 2017. Vol. 29(2). P. 161-200. (In Russ.) https://doi.org/10.15514/ISPRAS-2017-29(2)-6.

[21] Yakunin K., Ionescu G.M., Murzakhmetov S., Mussabayev R., Filatova O., Mukhamediev R. Propaganda Identification Using Topic Modelling // Procedia Computer Science. 2020. Vol. 178. P. 205–212. https://doi.org/10.1016/j.procs.2020.11.022

[22] Vorontsov K. et al. Bigartm: Open source library for regularized multimodal topic modeling of large collections // International Conference on Analysis of Images, Social Networks and Texts. – Springer, Cham, 2015. – С. 370–381.

[23] Blei D.M., Ng A.Y., Jordan M.I. Latent dirichlet allocation // Journal of machine Learning research. – 2003. – Т. 3. – No Jan. – P. 993–1022.

[24] Jelodar H. et al. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey // Multimedia Tools and Applications. – 2018. – С. 1–43.

[25] Mimno D., Wallach H., Talley Ed., Leenders M. & McCallum A. Optimizing Semantic Coherence in Topic Models // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. – 2011. – P. 262–272.

[26] Barakhnin V. B., et al.: Methods to identify the destructive information // Journal of Physics. 1405(1), 012004. – 2019.

[27] Mukhamediev R.I., Mustakayev R., Yakunin K., Kiseleva S., Gopejenko V. Multi-Criteria Spatial Decision Making Support system for Renewable Energy Development in Kazakhstan // IEEE Access. 2019. 7, 122275-122288.

[28] Mukhamediev R. I. et al. Classification of Negative Information on Socially Significant Topics in Mass Media //Symmetry. – 2020. – Т. 12. – №. 12. – С. 1945.

[29] Zhu X., Goldberg A. B. Introduction to semi-supervised learning // Synthesis lectures on artificial intelligence and machine learning. – 2009. Vol. 3(1). – P. 1-130.

[30] Bradley A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997. – Vol. 30 (7). – 1145–1159. URL: https://doi.org/10.1016/s0031-3203(96)00142-2/.

[31] Akobeng A. K. Understanding diagnostic tests 3: receiver operating characteristic curves // Wiley Online Library, 21–Mar–2007. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1651-2227.2006.00178.x.

[32] Yakunin K.O, Mussabayev R.R., Eylis M.S., Mukhamediev R.I. Energy topic in news publications // Renewable energy sources. Proceedings of the All-russian scientific conference and the XIII youth school with international participation. 24–25 november, 2020. Moscow, 2020. P. 451-456. (In Russ.)

**Muhamedyev Ravil Ilgizovich**, Doctor of eng. sci., professor Satbayev University (https://official.satbayev.university/en/teachers/ravil-muhamedyev-), email: r.mukhamediev@satbayev.university, elibrary.ru: authorid=1123472, ORCID: orcidID=0000-0002-3727-043X
**Filatova Olga Georgievna**, Phd., Associate Professor, Department of Public Relations in Politics and Public Administration, St. Petersburg State University, St. Petersburg (https://spbu.ru), email: o.filatova@spbu.ru, elibrary.ru: authorid = 625633 , ORCID: orcidID = 0000-0001-9568-1002.
**Yakunin Kirill Olegovich,** master of technical sciences, PhD student at Satbayev University - email: yakunin.k@mail.ru elibrary.ru: authorid=1099914, ORCID: orcidID= 0000-0002-7378-9212