Министерство образования и науки Республики Казахстан

Акционерное общество «Информационно-аналитический центр»

«УТВЕРЖДАЮ»

И.о. президента
АО «Информационно-
аналитический центр»

_____ М.К. Атанаева

«_10_» октября 2018 г.

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

по научно-технической программе: 1) № BR05236839 «Разработка информационных технологий и систем для стимулирования устойчивого развития личности как одна из основ развития цифрового Казахстана»
(промежуточный)

в рамках Договора №7 от 23 мая 2018 года
на оказание услуг по выполнению научно-исследовательских работ

Астана 2018

СПИСОК ИСПОЛНИТЕЛЕЙ

Атанаева Мираим Кажмухамбетовна	заместитель президента
Булдыбаев Тимур Керимбекович	заместитель директора Департамента по работе со средствами массовой информации и населением
Оспанова Улжан Абаевна	менеджер Департамента образовательной статистики и сопровождения информационных систем
Акоева Инесса Георгиевна	менеджер Департамента аналитики, мониторинга и оценки образования
Нурумов Кайдар Серикович	проектный менеджер Департамента международных сопоставительных исследований
Аманғазы Мағжан Заңғарұлы	заместитель президента
Тиштыбаев Азат Ашенович	директор Департамента образовательной статистики и сопровождения информационных систем
Увалиева Жанар Саметовна	заместитель директора Департамента образовательной статистики и сопровождения информационных систем
Багланова Гульназ Асылбековна	главный аналитик Департамента образовательной статистики и сопровождения информационных систем
Смагулова Айжан Акболатовна	главный аналитик Департамента образовательной статистики и сопровождения информационных систем
Айтжанова Эльмира Мергенбаевна	главный аналитик Департамента образовательной статистики и сопровождения информационных систем
Кусиденова Астай Аскаровна	главный аналитик Департамента образовательной статистики и сопровождения информационных систем
Капсалыкова Айдана Талгатовна	менеджер Департамента образовательной статистики и сопровождения информационных систем
Утенова Сымбат Сагидуллаевна	специалист Департамента образовательной статистики и сопровождения информационных систем
Мамбетов Нариман Мухтарович	Инженер-системотехник Департамента образовательной статистики и сопровождения информационных систем
Байгулова Айгуль Данияровна	заместитель директора Департамента международных сопоставительных исследований

Черкашина Надежда Ивановна	главный аналитик Департамента международных сопоставительных исследований
Картпаев Берикбол Айткожевич	главный аналитик Департамента международных сопоставительных исследований
Зуева Әйгерім Еркинкызы	менеджер Департамента международных сопоставительных исследований
Культуманов Темирлан Канатович	специалист Департамента международных сопоставительных исследований
Омарова Ботакоз Серикбаевна	специалист Департамента международных сопоставительных исследований
Ногайбаева Гульбаршин Аскеновна	директор Департамента аналитики, мониторинга и оценки образования
Ахметжанова Аида Амантаевна	заместитель директора Департамента аналитики, мониторинга и оценки образования
Карбаева Галия Галимжановна	главный аналитик Департамента аналитики, мониторинга и оценки образования

РЕФЕРАТ

Отчет на 96 страницах. В отчете содержится 6 рисунков, 10 таблиц. Состоит из 15 частей.

Ключевые слова: средства массовой информации, информативные признаки, медиа-текст, резонансность, тональность, достоверность, медиа-вовлеченность, объективность, комплексная оценка значимости, разметка, корпус.

Работа выполняется на основании технического задания по научно-технической программе № BR05236839 «Разработка информационных технологий и систем для стимулирования устойчивого развития личности как одна из основ развития цифрового Казахстана» в рамках проектно-целевого финансирования.

Объектом исследования являются открытые электронные текстовые информационные источники новостного характера и их контент.

Цель исследования: реализация услуг по научно-технической программе.

Для достижения цели и решения задач исследования на каждом этапе работ применялись различные методы и подходы, такие как: социологический опрос, экспертный опрос, метод теоретического анализа и обобщения научной и аналитической литературы, аддитивный метод построения индексов, общенаучные методы наблюдения, систематизации, обобщения, статистический метод, лингвистические методы контекстуального, дискурсивного, интерпретативного, прагматического и компонентного анализа, интроспекция и др.

Научная новизна работы заключается в прикладном эмпирическом характере исследования. Впервые основное внимание было уделено информативным критериям оценки публикаций новостного характера и их влиянию на социум. Данная работа является первой в Казахстане, в которой комбинированно применены методы количественного исследования и статистического моделирования для получения оценки значимости предложенных информативных критериев для анализа влияния публикаций новостного характера на социум. Также научная новизна исследования заключается в том, что впервые был сформирован корпус казахстанских публикаций новостного характера и произведена их разметка по критериям достоверности, объективности, резонансности, социальной значимости, политизированности, наличия манипулятивных техник, казахстанского содержания и принадлежности к сфере в разрезе тематики.

По результатам проведенных работ проведено социологическое исследование по оценке влияния открытых информационных источников (электронных СМИ) на социум Казахстана. Опросом были охвачены 3200 респондентов из 16 регионов Казахстана.

Выработаны информативные признаки, разработаны методологические подходы по проведению комплекса работ, разработаны 3 инструкции и др. Проведены комплексная оценка значимости используемых информативных признаков и экспериментальное подтверждение. Сформирован сбалансированный, представительный и размеченный корпус публицистических текстов общим объемом 5 210 текстов из 5 источников. Осуществлена лингвистическая разметка корпуса по 2000 текстам, по результатам которой сформированы словари и формализованы правила по 3 признакам.

Результаты исследования будут применены для разработки автоматизированной системы оценки влияния открытых текстовых информационных источников на социум. Итоги исследования также будут полезны для применения в области политологии, социологии, журналистики и информационных технологий.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 11
1. РАЗРАБОТКА МЕТОДИКИ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ НА ОСНОВЕ АНАЛИЗА ПУБЛИКУЕМОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ ПРИМЕНИТЕЛЬНО К СИСТЕМЕ СОЦИАЛЬНОГО ДОВЕРИЯ 14
2. ИССЛЕДОВАНИЕ СУЩЕСТВУЮЩИХ ИНФОРМАТИВНЫХ ПРИЗНАКОВ С ЦЕЛЬЮ ОПРЕДЕЛЕНИЯ ПРИЕМЛЕМЫХ КРИТЕРИЕВ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ ПРИМЕНИТЕЛЬНО К ССД 18
3. ИССЛЕДОВАНИЕ КОМПЛЕКСНОЙ ОЦЕНКИ ЗНАЧИМОСТИ ИСПОЛЬЗУЕМЫХ ИНФОРМАТИВНЫХ КРИТЕРИЕВ 33
4. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ, НАПРАВЛЕННЫХ НА ЭКСПЕРИМЕНТАЛЬНОЕ ПОДТВЕРЖДЕНИЕ АДЕКВАТНОСТИ ПОЛУЧЕННОЙ КОМПЛЕКСНОЙ ОЦЕНКИ ЗНАЧИМОСТИ ИСПОЛЬЗУЕМЫХ ИНФОРМАТИВНЫХ КРИТЕРИЕВ ДЛЯ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ НА ОСНОВЕ АНАЛИЗА ПУБЛИКУЕМОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ 38
5. ИССЛЕДОВАНИЕ ОСОБЕННОСТЕЙ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ В УСЛОВИЯХ РЕСПУБЛИКИ КАЗАХСТАН ПРИМЕНИТЕЛЬНО К ССД 45
7. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО ФОРМИРОВАНИЮ СБАЛАНСИРОВАННОГО, ПРЕДСТАВИТЕЛЬНОГО, РАЗМЕЧЕННОГО И ПРОКЛАССИФИЦИРОВАННОГО КОРПУСА ПУБЛИЦИСТИЧЕСКИХ ТЕКСТОВ: ТЕКСТОВ ОФИЦИАЛЬНЫХ СМИ, ЖУРНАЛЬНЫХ ПУБЛИКАЦИЙ, НОВОСТНЫХ ПОРТАЛОВ И ДР. 69
8. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО АНАЛИЗУ ФАКТИЧЕСКИХ ХАРАКТЕРИСТИК СФОРМИРОВАННОГО КОРПУСА (СОСТАВ, СТРУКТУРА, ПРЕДСТАВИТЕЛЬНОСТЬ, СБАЛАНСИРОВАННОСТЬ, УРОВНИ РАЗМЕТКИ И КЛАССИФИКАЦИИ, ИХ КОРРЕКТНОСТЬ И ПОЛНОТА И ДР.) 71
9.РАЗРАБОТКА МЕТОДОВ ФОРМАЛИЗАЦИИ ПРАВИЛ И СЛОВАРЕЙ ДЛЯ РАСПОЗНАВАНИЯ В РАЗМЕЧЕННОМ КОРПУСЕ ИНФОРМАТИВНЫХ ПРИЗНАКОВ И ОПРЕДЕЛЕНИЯ СТЕПЕНИ ИХ ВЫРАЖЕННОСТИ 73

10.ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО АНАЛИЗУ И ВЫЯВЛЕНИЮ ЗАКОНОМЕРНОСТЕЙ, ПРИСУТСТВУЮЩИХ В СФОРМИРОВАННОМ КОРПУСЕ, С ЦЕЛЬЮ ФОРМАЛИЗАЦИИ ПРАВИЛ И ФОРМИРОВАНИЯ СЛОВАРЕЙ ДЛЯ РАСПОЗНАВАНИЯ В РАЗМЕЧЕННОМ КОРПУСЕ ИНФОРМАТИВНЫХ ПРИЗНАКОВ И ОПРЕДЕЛЕНИЯ СТЕПЕНИ ИХ ВЫРАЖЕННОСТИ НА ОСНОВЕ РАЗРАБОТАННЫХ МЕТОДОВ.....81
11. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО ЭКСПЕРИМЕНТАЛЬНОМУ ИССЛЕДОВАНИЮ ПОЛНОТЫ И КОРРЕКТНОСТИ СФОРМИРОВАННОГО МНОЖЕСТВА ПРАВИЛ И СЛОВАРЕЙ ДЛЯ РАСПОЗНАВАНИЯ В РАЗМЕЧЕННОМ КОРПУСЕ ИНФОРМАТИВНЫХ ПРИЗНАКОВ И ОПРЕДЕЛЕНИЯ СТЕПЕНИ ИХ ВЫРАЖЕННОСТИ 84
12. РАЗРАБОТКА МЕТОДИКИ ОПРЕДЕЛЕНИЯ СТЕПЕНЕЙ ВЕРОЯТНОЙ ДОСТОВЕРНОСТИ ОТДЕЛЬНОЙ ПУБЛИКАЦИИ, ВЕРОЯТНОЙ ДОСТОВЕРНОСТИ ПУБЛИКАЦИЙ ОПРЕДЕЛЁННОГО АВТОРА, ВЕРОЯТНОЙ ДОСТОВЕРНОСТИ ПУБЛИКАЦИЙ ОПРЕДЕЛЁННОГО ПЕЧАТНОГО СМИ И КРИТЕРИИ ИХ ВЕРИФИКАЦИИ.....85
13.РАЗРАБОТКА ИНСТРУКЦИИ ДЛЯ ЭКСПЕРТНОГО АНАЛИЗА И КЛАССИФИКАЦИИ ДОСТОВЕРНОСТИ ПОДОЗРИТЕЛЬНЫХ ПУБЛИКАЦИЙ 88
ЗАКЛЮЧЕНИЕ.....89
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 92
СПИСОК ОПУБЛИКОВАННЫХ ПУБЛИКАЦИЙ ПО ТЕМЕ ИССЛЕДОВАНИЯ 94
ПРИЛОЖЕНИЯ 95

ОПРЕДЕЛЕНИЯ

Средство массовой информации	периодическое печатное издание, теле-, радиоканал, кинодокументалистика, аудиовизуальная запись и иная форма периодического или непрерывного публичного распространения массовой информации, включая Интернет-ресурсы. [18]
Массовая информация	предназначенные для неограниченного круга лиц печатные, аудиовизуальные и иные сообщения и материалы. [18]
Уполномоченный орган в области средств массовой информации	государственный орган, осуществляющий государственное регулирование в области средств массовой информации. [18]
Журналист (представитель средства массовой информации)	физическое лицо, осуществляющее деятельность по сбору, обработке и подготовке сообщений и материалов для средств массовой информации на основе трудовых или иных договорных отношений. [18]
Официальное сообщение	информация, предоставляемая и (или) распространяемая через СМИ обладателем информации, установленным в соответствии с Законом Республики Казахстан «О доступе к информации». [18]
Публикация	текст, опубликованный в каком-либо издании. [21]
Медиа-текст	текст массовой коммуникации как информационный продукт средств массовой информации. [22]
Достоверность	позволяет оценить достоверность публикуемой информации на основе наличия ссылок в публикациях на источник, авторитетности издания, на котором размещены публикации, освещение одного и того же события различными изданиями, проверки аутентичности фото и видео, которое подтверждает событие, соответствие заголовка публикации содержанию медиа-текста.
Резонансность	позволяет оценить масштаб распространения публикации новостного характера и определить основной характер отношения аудитории к публикации на основе показателей медиа-охвата, количества репостов, шейров, комментариев и тональности комментариев.

Объективность	позволяет оценить объективность подачи фактов, информации в публикации.
Тональность	позволяет оценить эмоциональный тон подачи информации, фактов в публикации, который автор транслирует на аудиторию. Существует два основных метода автоматического определения тональности статистический метод и метод, основанный на словарях и правилах.
Медиа-вовлеченность	позволяет оценить поведение читателя на определенном веб-ресурсе на основе анализа глубины и длительности просмотра публикации и частоты возращения на сайт.
Информативный признак	достаточное условие для принадлежности объекта некоторому классу.
Фейк	от английского «fake» – фальшивый, поддельный. Это ложные новости или сообщения, которые генерируются в онлайн-пространстве. [23]
Корпус	структурированная система собрания текстов на некотором языке в электронной форме, связанных между собой определенными параметрами. [24]

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

СМИ	Средства массовой информации
PR	Public Relations, или связь с общественностью
НЛП	Нейролингвистическое программирование
Концепция	Концепция информационной безопасности Казахстана до 2016 года
Кодекс	Кодекс этики журналиста Республики Казахстан
НПА	Нормативно-правовые акты
ЗРК	Закон Республики Казахстан
РК	Республика Казахстан
ООН	Организация Объединённых Наций
ВОЗ	Всемирная организация здравоохранения
МВФ	Международный валютный фонд
ОЭСР	Организация экономического сотрудничества и развития
ССД	Система социального доверия

ВВЕДЕНИЕ

В целях реализации услуг по научно-технической программе № BR05236839 «Разработка информационных технологий и систем для стимулирования устойчивого развития личности как одна из основ развития цифрового Казахстана» проведён комплекс работ.

Для выполнения научно-технической программы были поставлены следующие задачи:

1. Разработка методики для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации применительно к системе социального доверия.

2. Исследование существующих информативных признаков с целью определения приемлемых критериев оценки влияния открытых текстовых информационных источников на социум;

3. Исследование комплексной оценки значимости используемых информативных критериев.

4. Экспериментальное подтверждение адекватности полученной комплексной оценки значимости используемых информативных критериев для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации.

5. Исследование особенностей оценки влияния открытых текстовых информационных источников на социум в условиях Республики Казахстан применительно к ССД.

6. Комплексное исследование, а также экспертный анализ необходимого и достаточного состава формируемого текстового корпуса, его структуры, представительности, сбалансированности, требуемых уровней разметки и классификаций, их однозначности; способов сбора, хранения и представления текстового корпуса; методов организации работы и взаимного контроля задействованных экспертов при его формировании.

7. Проведение комплекса работ по формированию сбалансированного, представительного, размеченного и проклассифицированного корпуса публицистических текстов: текстов официальных СМИ, новостных порталов.

8. Проведение комплекса работ по анализу фактических характеристик сформированного корпуса (состав, структура, представительность, сбалансированность, уровни разметки и классификации, их корректность и полнота и др.).

9. Разработка методов формализации правил и словарей для распознавания в размеченном корпусе информативных признаков и определения степени их выраженности.

10. Проведение комплекса работ по анализу и выявлению закономерностей, присутствующих в сформированном корпусе, с целью формализации правил и формирования словарей для распознавания в размеченном корпусе информативных признаков и определения степени их выраженности на основе разработанных методов.

11. Проведение комплекса работ по экспериментальному исследованию полноты и корректности сформированного множества правил и словарей для распознавания в размеченном корпусе информативных признаков и определения степени их выраженности.

12. Разработка методики определения степеней вероятной достоверности отдельной публикации, вероятной достоверности публикаций определённого автора, вероятной достоверности публикаций определённого печатного СМИ и критерии их верификации.

13. Разработка инструкции для экспертного анализа и классификации достоверности подозрительных публикаций.

Для решения поставленных задач на каждом этапе работ применялись различные методы и подходы, такие как: социологический опрос, экспертный опрос, метод теоретического анализа и обобщения научной и аналитической литературы, аддитивный метод построения индексов, общенаучные методы наблюдения, систематизации, обобщения, статистический метод, лингвистические методы контекстуального, дискурсивного, интерпретативного, прагматического и компонентного анализа, интроспекция и др.

Научная новизна работы заключается в прикладном эмпирическом характере исследования. Впервые основное внимание было уделено информативным критериям оценки публикаций новостного характера и их влиянию на социум. Данная работа является первой в Казахстане, в которой комбинированно применены методы количественного исследования и статистического моделирования в целях получения определения степени значимости предложенных информативных критериев для оценки влияния публикаций новостного характера на социум. Также научная новизна исследования заключается в том, что впервые был сформирован корпус казахстанских публикаций новостного характера и произведена их разметка по критериям достоверности, объективности, резонансности, социальной значимости, политизированности, наличия манипулятивных техник, казахстанского содержания и принадлежности к сфере в разрезе тематики.

Проведенные работы в рамках реализации НТП являются одним из первых подготовительных этапов работ по выработке методики оценки влияния открытых текстовых информационных источников на социум.

По результатам проведенных работ выработаны информативные признаки, разработаны методологические подходы по проведению комплекса работ, разработаны 3 инструкции и др. Проведена комплексная оценка значимости используемых информативных признаков и их экспериментальное подтверждение. Проведено социологическое исследование по оценке влияния открытых информационных источников (электронных СМИ) на социум Казахстана. Опросом были охвачены 3200 респондентов из 16 регионов Казахстана. Сформирован сбалансированный, представительный и размеченный корпус публицистических текстов общим объемом 5 210 текстов из 5 источников. Проведена лингвистическая разметка 2000 текстов, по результатам которой сформированы словари и формализованы правила по 3 признакам.

1. РАЗРАБОТКА МЕТОДИКИ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ НА ОСНОВЕ АНАЛИЗА ПУБЛИКУЕМОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ ПРИМЕНИТЕЛЬНО К СИСТЕМЕ СОЦИАЛЬНОГО ДОВЕРИЯ

Актуальность разработки методики

Разработка методики оценки влияния открытых информационных источников на социум с применением автоматизированной информационной системы приобретает особую актуальность в мире.

Это в первую очередь обусловлено возрастающей ролью СМИ оказывать непосредственное влияние на построение общественных убеждений и взглядов. Также СМИ используют различные виды психологического влияния, а зачастую и их комбинации для трансляции определенной позиции по той или иной ситуации.

Согласно исследованию Edelman Trust Barometer (Барометр доверия), проведенному в 28 странах мира и охватившему 33 тыс. респондентов, почти семь из десяти опрошенных респондентов обеспокоены тем, что фейковые новости и дезинформация используются сегодня как оружие. 63% респондентов утверждают, что среднестатистический человек не знает, как отделить хорошую журналистику от слухов или лжи.

По данным социологического исследования населения Казахстана, только 17,3% респондентов считают, что казахстанские Интернет-источники СМИ при освещении события подают информацию достоверно. При этом ½ часть респондентов (50,5%) отмечают скорее достоверность. Только 46,6% населения Казахстана считают деятельность казахстанских Интернет-источников СМИ прозрачной. 58% опрошенных респондентов считают, что СМИ оказывают влияние на видение ситуации в стране и являются инструментом формирования общественного мнения (60,3%). Основными последствиями необъективного отражения информации в информационных Интернет-источниках СМИ, по мнению населения, могут стать потеря доверия (11,7%), проведение митингов и акций протестов (7,7%), хаос и паника среди населения (6,5%) и необъективное восприятие информации людьми (4,9%).

Основными преимуществами разрабатываемой методики оценки являются:

- групповые (сводные) и индивидуальные индексы варьируются в интервале от 0 до 1;
- предложена целостная и сбалансированная система показателей влияния открытых текстовых информационных источников на социум;
- обеспечена возможность получения многоуровневой системы оценок (по

новостным порталам в целом, по отдельному новостному порталу, по отдельной публикации) и ранжирования в зависимости от поставленных задач анализа и уровня детализации;

- оценка позволяет учитывать динамику качества текстовой информации за определенный период и сравнивать с различными значениями показателей;
- оценка позволяет получить как интегральное значение влияния открытых текстовых информационных источников, так и групповые значения по отдельным критериям оценки.

Общие положения разрабатываемой методики

Цель оценки: определение уровня влияния открытых информационных источников на социум на основе измеряемых показателей с применением автоматизированной информационной системы.

Основные задачи:

- определение уровня влияния медиа-текстов казахстанских СМИ на основе критериев, характеризующих факторы медиа-воздействия;
- определение степени достоверности публикуемой информации новостного характера в казахстанских СМИ;
- генерация аналитики, статистических данных на основе оценки;
- мониторинг медиа-потребления.

Объект оценки: публикации казахстанских информационных новостных Интернет-источников.

Предмет оценки: достоверность, тональность, резонансность, объективность публикаций информационных новостных Интернет-порталов и медиа-вовлеченность аудитории.

Этапы разработки методики

Разработка методики оценки влияния открытых текстовых информационных источников на социум на основе анализа публикуемой текстовой информации включает несколько этапов, где на каждом этапе используется определенный метод и подход.

Первый этап - исследование существующих информативных признаков для определения приемлемых критериев оценки влияния открытых текстовых информационных источников на социум.

Второй этап - определение характера влияния открытых информационных источников на различные социальные группы населения Казахстана.

Третий этап - определение критериев и показателей оценки влияния открытых текстовых информационных источников на социум, выработанных на основе информативных признаков.

Четвертый этап - проведение комплексной оценки значимости используемых информативных критериев оценки и их показателей.

Пятый этап - экспертная оценка (разметка) сформированного корпуса публицистических текстов с применением информативных критериев и показателей.

Шестой этап - формализация правил и формирование словарей по информативным признакам, которые определяются лингвистическими методами.

Определение критериев оценки

Результаты исследования существующих информативных признаков влияния открытых текстовых информационных источников на социум позволили определить существующие информативные признаки оценки публикации новостного характера (см. раздел 2). Выбранные информативные признаки были сгруппированы в 5 информативных критериев, которые наиболее приемлемы для оценки: *достоверность, резонансность, тональность, объективность и медиа-вовлеченность.* Для оценки влияния открытых текстовых информационных источников по предложенным критериям были выработаны 22 показателя.

Для определения значимости выбранных информативных критериев и показателей оценки проведена комплексная оценка значимости (см. раздел 3 и 4). По результатам комплексной оценки значимости информативные критерии публикации, такие как достоверность, объективность и тональность, оказывают статистически значимое влияние на количество просмотров.

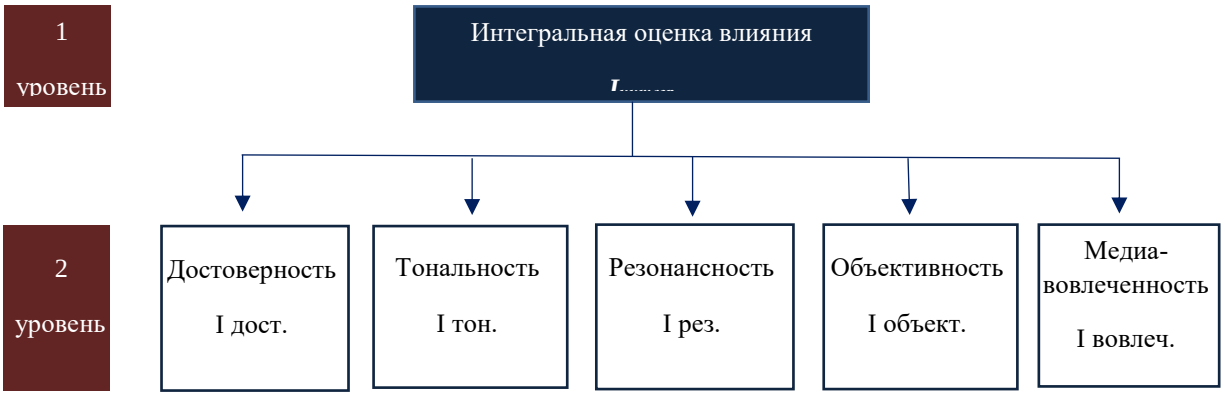
Для распознавания информативных признаков в публикациях и выработки методики их расчета сформирован сбалансированный, представительный корпус публицистических текстов, размеченного и проклассифицированного по части параметров. Для формирования корпуса была проведена 2-х этапная систематическая кластерная выборка.

Разметка корпуса текстов проведена двумя подходами. Первый подход – отнесение публикации к определенному классу и параметру. Второй– глубинная разметка публикаций для формализации лингвистических правил и формирования словарей по параметрам.

Оценка влияния открытых информационных источников на социум

Разрабатываемая система оценки влияния открытых текстовых информационных источников представляет собой совокупность оценочных процедур и многоуровневую систему показателей, построенную по принципам декомпозиции и агрегирования.

Рисунок 1 - Декомпозиция показателей оценки влияния открытых текстовых информационных источников на социум



Как следует из рисунка 1, на самом верхнем (первом) уровне осуществляется общая (интегральная) оценка влияния. Декомпозиция интегрального индикатора влияния осуществляется по пяти содержательным критериям оценки: достоверность, тональность, резонансность, объективность и медиа-вовлеченность.

Для каждого критерия оценки рассчитываются соответствующие групповые индексы, которые будут агрегированы в общий (интегральный) индекс влияния.

Важной задачей при расчете комплексных оценок является обоснование весовых коэффициентов, отражающих значимость того или иного показателя (или группы показателей). Весовые коэффициенты задаются на основе экспертных оценок.

Предлагаемая система процедур оценки и показателей включает в себя следующие срезы:

- по критериям оценки всех новостных порталов: достоверность, тональность, резонансность, объективность и медиа-вовлеченность;
- по критериям оценки отдельного новостного портала;
- по критериям оценки отдельной публикации.

2. ИССЛЕДОВАНИЕ СУЩЕСТВУЮЩИХ ИНФОРМАТИВНЫХ ПРИЗНАКОВ С ЦЕЛЬЮ ОПРЕДЕЛЕНИЯ ПРИЕМЛЕМЫХ КРИТЕРИЕВ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ ПРИМЕНИТЕЛЬНО К ССД

Цель исследования: изучение ключевых аспектов влияния открытых текстовых информационных источников на социум, анализ существующих методик оценки медиа-текстов и достоверности публикаций для выработки информативных критериев для последующей автоматизации процесса проведения оценки.

Объектом исследования являются существующие информативные признаки влияния открытых текстовых информационных источников на социум.

Методология исследования. Использован метод теоретического анализа и обобщения научной и аналитической литературы.

В исследовании изучены такие вопросы, как роль СМИ в формировании общественного мнения, международный опыт и государственное регулирование СМИ, опыт использования «Больших данных» для мониторинга и анализа массива информации, изучение манипулятивных механизмов влияния СМИ на социум, методы контент-анализа и сентимент-анализа содержания медиа-текстов, опыт применения специализированных программ для анализа контента и анализ существующих словарей.

Научная новизна работы заключается в выработке информативных признаков для возможного использования автоматизированного подхода в оценке влияния открытых текстовых информационных источников на социум с минимальной вовлеченностью экспертов.

Гипотезы исследования:

- Гипотеза 1. Определение информативных признаков находится в сильной зависимости от целей исследования.
- Гипотеза 2. Повсеместно наблюдается рост влияния и распространения электронных источников информации открытого доступа.
- Гипотеза 3. Резонансность публикации является одним из ключевых критериев оценки ее влияния на социум.
- Гипотеза 4. Автоматизированное определение уровня тональности на основе тональных словарей позволяет выявить потенциально опасное содержание и искажение фактов в публикации для влияния на мнение читателей.
- Гипотеза 5. Объективность публикуемой информации играет одну из ключевых ролей в процессе формирования общей картины доверия к электронным СМИ.
- Гипотеза 6. Достоверность публикации возможно определить путем проверки репутации и авторитетности источника информации, аутентичности фото и

видео, которые подтверждают событие, а также перекрестной проверки освещения события на других информационных порталах СМИ.

В соответствии с целью и гипотезами исследования были определены следующие задачи:

1. Изучение роли СМИ в формировании общественного мнения.
2. Анализ международного и казахстанского законодательства в области регулирования деятельности СМИ.
3. Изучение казахстанского опыта проведения фактчекинга публикаций.
4. Изучение опыта использования «Больших данных» (Big data) для мониторинга и анализа данных.
5. Изучение методов контент и сентимент-анализа медиа-текстов.
6. Определение основных инструментов, техник и механизмов воздействия СМИ на общественное мнение.
7. Анализ существующих методик оценки влияния СМИ на социум и определение наиболее приемлемых подходов к разработке методики оценки влияния открытых текстовых информационных источников на социум.
8. Анализ подходов к определению достоверности публикации новостного характера.

Роль СМИ в формировании общественного мнения

СМИ играют центральную роль в информировании общественности и, соответственно, в формировании мировоззрения личности и общественного поведения. При формировании представлений личности о мире, помимо эмпирического опыта, лишь два источника для получения информации являются основными: (1) окружение человека - его семья, друзья, коллеги, знакомые и другие люди, находящиеся в непосредственном контакте с ним; (2) общедоступные источники информации, такие как книги, журналы, Интернет, СМИ и другие. Если общедоступные источники информации справочного характера, такие как энциклопедии, учебники, книги и др., помогают сформировать фундаментальные знания, то СМИ являются основным источником новостей, которые ориентируют общество в отношении того, что происходит в мире, стране в настоящее время используя при этом различные виды психологического влияния, а зачастую и их комбинации для трансляции определенной позиции в отношении той или иной ситуации. [1]

Вопрос достоверности публикаций находится в фокусе внимания многих исследователей. Так, по данным исследования Массачусетского технологического института (США), было установлено, что 1% ложных новостей, получивших наибольшее

распространение в Twitter, распространяются среди аудитории от 1000 до 100 000 людей, тогда как масштаб распространения достоверных сообщений не превышают 1000 человек. Ложные данные распространяются значительно дальше, быстрее, глубже и шире, чем правда во всех категориях информации, а эффекты ложных политических новостей являются более ярко выраженными, чем эффект недостоверных новостей на тему терроризма, природных катаклизмов, науки, городских легенд или финансовой информации [2].

Согласно исследованию Edelman Trust Barometer (Барометр доверия), проведенному в 28 странах мира и охватившему 33 тыс. респондентов, в 2018 году по сравнению с 2017 годом отмечается снижение доверия к социальным платформам и увеличение доверия к традиционной журналистике и Интернет СМИ. Почти семь из десяти охваченных исследованием респондентов обеспокоены тем, что фейковые новости и дезинформация используются сегодня как оружие. 63% респондентов утверждают, что среднестатистический человек не знает, как отделить хорошую журналистику от слухов или лжи [3].

Агентство GlobalWebIndex исследовало количество времени, затрачиваемого людьми на просмотр тех или иных видов СМИ в 34 странах мира. Результаты исследования свидетельствуют о преобладании цифровых источников информации над традиционными СМИ, где основное время приходится на просмотр телевидения [4].

По данным социологического исследования населения Казахстана, только 17,3% респондентов считают, что казахстанские Интернет-источники СМИ при освещении события подают информацию достоверно. При этом ½ часть респондентов (50,5%) отмечают скорее достоверность. Только 46,6% населения Казахстана считают деятельность казахстанских Интернет-источников СМИ прозрачной. 58% опрошенных респондентов считают, что СМИ оказывают влияние на видение ситуации в стране и являются инструментом формирования общественного мнения (60,3%). Основными последствиями необъективного отражения информации в информационных Интернет источниках СМИ, по мнению населения, могут стать потеря доверия (11,7%), проведение митингов и акций протестов (7,7%), хаос и паника среди населения (6,5%) и необъективное восприятие информации людьми (4,9%) [5].

В связи с этим анализ медиа-контента - того, о чем нам рассказывают и о чем умалчивают, - является одним из главных вопросов, требующих повышенного внимания.

Законодательное регулирование СМИ

Анализ международного опыта в сфере регулирования деятельности СМИ показал, что правовые системы, регулирующие их деятельность в различных странах, неодинаковы.

В одних странах права и обязанности журналистов имеют четкое и детализированное закрепление на законодательном уровне, в других - основаны на традициях и принципах общего права. В некоторых странах прессу обязывают печатать лишь «проверенные факты» или «истину». Законодательство Казахстана в области СМИ обязывает журналиста проверять достоверность получаемой информации, обращаясь при этом к компетентным специалистам или органам.

Одним из инструментов саморегулирования деятельности СМИ являются этические стандарты профессии. Большинство профессиональных объединений журналистов или отдельные издания принимают кодексы журналисткой этики. В некоторых государствах издания и индивидуальные журналисты подчиняются данному кодексу. В других странах этические нормы журналистов нашли отражение в национальных законодательствах. К примеру, в США разные издания придерживаются собственных правил журналистской этики. Некоторые кодексы носят обязательный характер, и нарушение того или иного положения может привести к увольнению работодателем или исключению из профессионального объединения журналистов. Но большинство кодексов этики вместо этого предлагают добровольные руководящие принципы, помогающие журналистам принимать морально и профессионально обоснованные решения. Тем самым кодексы поощряют большую подотчетность читателям и зрителям [6].

В большинстве кодексов этических норм главным принципом является поиск истины. В Казахстане Кодекс этики журналиста (далее – Кодекс) придерживается принципов объективности, непредвзятости и честности. Вместе с тем Кодекс является необязательным к соблюдению и является инструментом самодисциплины и морально-этическим ориентиром, который способствует достижению доверия и уважения к журналистам и средствам массовой информации [7].

Таким образом, регулирование деятельности СМИ едины в нормах обеспечения объективности, опубликования информации соответствующей действительности, в поиске истины или достоверности, стремление быть подотчётными. В большинстве кодексов этических норм журналиста главным принципом является достоверность и объективность публикуемой информации.

Оценка влияния СМИ на социум

Поскольку значимость СМИ в построении мировоззрения человека и их роль в формировании общественного сознания не вызывает сомнений, оценка влияния СМИ на социум является одной из популярнейших областей прикладных исследований. На

сегодняшний момент фокус таких исследований сместился с оценки влияния телевидения на оценку влияния социальных сетей вследствие роста популярности последних.

Итоги исследования, проведенного маркетинговой компанией MediaKix в 2016 году и посвященного тому, сколько времени поводят пользователи в социальных сетях, показали, что электронные СМИ и социальные сети вышли на второе место, уступив только телевидению. При этом в 2015 году в США пользователи потратили больше времени на просмотр социальных приложений, чем на телевидение.

Для большинства населения Казахстана (59,8%) основными источниками получения информации новостного характера являются электронные источники: социальные сети/блоггеры (30,3%), а также новостные Интернет-сайты (29,5%) [5].

Несмотря на неугасающий интерес научного сообщества к оценке влияния СМИ, данная область характеризуется наличием множества различных подходов и теорий. Единые, общепринятые концепции и методики отсутствуют. Также теоретические наработки оценки влияния СМИ на социум многочисленны, неструктурированы и значительно отличаются друг от друга.

Методология оценки влияния СМИ на социум включает широко используемые методы социологических исследований, контент-анализа и ряд количественных показателей. К наиболее распространенным методам можно отнести социологические опросы, экспертные оценки, дискурс-анализ, контент-анализ, кейсовый метод, графематический и синтаксический анализ.

Манипулятивные механизмы влияния СМИ

СМИ используют различные виды психологического влияния, а зачастую и их комбинации для трансляции определенной позиции в отношении той или иной ситуации.

По данным социологического исследования населения Казахстана, большинство респондентов считают, что казахстанские информационные Интернет-сайты используются как инструменты для дискредитации определенных лиц (43,4%), формирования положительного образа (55,1%). При этом 52,7% населения полагают, что казахстанские Интернет-источники СМИ используются для освещения событий и (или) деятельности в той или иной сфере под определенным углом, выгодным определенному кругу лиц [5]. Эти данные позволяют заключить, что, по мнению казахстанцев, СМИ являются инструментом формирования и манипуляции общественным сознанием.

Изученные механизмы влияния СМИ на социум представляют собой не полный набор инструментов, которые используют СМИ, для формирования картины мира у общественности. Главным информативным признаком для оценки влияния в данном

аспекте можно считать наличие манипулятивных техник, среди которых может быть генерализация, политизированность, искажение информации и др.

Контент-анализ СМИ

В основу исследования и разработки методики оценки влияния СМИ на социум легли исследования по контент-анализу СМИ. Для этого были рассмотрены и обобщены основные методы контент-анализа их преимущества и недостатки, а также существующие программные инструменты для его проведения.

Как показали результаты анализа, для проведения контент-анализа необходимо обеспечить достаточное количество репрезентативной выборки, при которой бы тексты поддавались анализу по изначально заданным параметрам, в зависимости от целей исследования. При незначительном количестве текстов надежность принципа формализации снижается. Необходимо, чтобы исследуемое содержание позволило задать однозначное правило для надежного фиксирования нужных характеристик (принцип формализации), а также чтобы интересующие исследователя элементы содержания встречались с достаточной частотой (принцип статистической значимости). [8]

Процесс проведения контент-анализа сводится к определению темы, формированию задачи, разметке категорий, выбору единиц анализа. При верной разметке основные признаки статей (проблема, уровень напряженности, причины, решения и др.) выстраиваются в четкую структуру, рассмотрев которую можно сделать прогноз или заключение об уровне достоверности, резонансности, объективности и воздействия на читателя предоставляемой информации.

При формировании категорий для проведения контент-анализа необходимо выявление уникальных информативных признаков, по которым будет происходить четкое распределение текста. Данные признаки должны в полной мере соответствовать поставленным целям и задачам исследования. Категории контент-анализа должны охватывать все части содержания медиа-текста, быть взаимоисключающими (одни и те же части не должны принадлежать различным категориям), надежными (четкие критерии распределения текста по категориям) и уместными (соответствовать поставленной задаче и исследуемому содержанию).

Метод контент-анализа содержания публикаций имеет ряд серьёзных преимуществ перед другими методами и в тоже время ряд недостатков, как теоретического, так и прикладного характера. Основными преимуществами контент-анализа являются: (1) четко сформулированные параметры анализа, схемы кодирования и процедура выборки, что обеспечивает воспроизводимость проводимого эксперимента; (2) предоставление

информации об изменениях установленных единиц счета с течением времени (элемент лонгитюдного анализа); (3) объективный метод анализа текстов при корректном использовании. К основным недостаткам можно отнести: (1) расход значительного времени на проведение анализа; (2) высокая степень субъективизма при применении качественного метода контент-анализа;(3) частая ориентация на упрощение результатов, так как опирается на простой подсчёт слов; (4) непростота в применении компьютерных технологий и автоматизации исследования. [9]

Контент-анализ способен определить уместность или представленность различных эмоциональных и содержательных элементов авторов публикаций и участников коммуникаций, их предрассудков, намерений, идентифицировать ошибки и недосказанности. Метод контент-анализа может быть использован для изучения любого по содержанию и форме текстовой информации и быть применен в различных областях или сферах. Также контент-анализ тесно связан с социо- и психолингвистикой, он играет ключевую интегральную роль в системах разработки искусственного интеллекта.

Информативные признаки оценки влияния

Для анализа влияния открытых текстовых информационных источников на социум используются разные информативные признаки, на основе которых разрабатываются показатели оценки. Различные медиа, сайты и исследователи используют собственные показатели и способы их расчета, и, соответственно, наблюдается значительный диапазон вариаций информативных признаков для оценки.

Учитывая прямую взаимосвязь показателей и информативных признаков, авторами проделана работа по определению информативных признаков на основе анализа показателей оценки, применяемых на практике в рамках мирового опыта оценки влияния СМИ. Изучены широко применяемые показатели оценки влияния СМИ, таких как медиа-охват, индекс вовлеченности, индекс отказа, темп роста аудитории сообщества, unlikes rate, тональность упоминаний публикации и комментариев.

Проведенный комплексный анализ теоретических выкладок и практических наработок в области влияния СМИ на социум показал, что на сегодняшний день СМИ располагают широким арсеналом различных методов и техник воздействия на аудиторию и активно их используют. Эффективное оперирование такими техниками обусловливает роль СМИ в качестве одного из самых значимых факторов при формировании мировоззрения личности.

Несмотря на многочисленность показателей, позволяющих оценить определенные аспекты влияния СМИ, общая структура анализа влияния, как и сущность, характер и

методы расчета количественных показателей, и определение информативных признаков зависят от целей исследования. Так один и тот же показатель может рассчитываться по-разному, в зависимости от того, какой именно аспект медиа-импакта или медиа-эффекта выступает объектом оценки медиа-воздействия, а также в зависимости от того, каким путем и какие именно исходные данные были собраны для оценки.

По итогам их анализа, авторами были определены две группы информативных признаков, которые было бы целесообразно использовать при оценке влияния открытых информационных источников на социум. К информативным признакам первой группы, характеризующим критерий оценки, определенный как «Резонансность», относятся следующие признаки:

– наличие определенного количества читателей;
– заинтересованность читателей;
– реакция читателей.

Для оценки влияния СМИ на социум с помощью автоматизированной информационной системы предлагается использовать следующие показатели на основе определенных информативных признаков. Данные показатели являются количественным выражением информативных признаков и позволяют более развернуто оценить публикации по следующим признакам.

1. Медиа-охват. Данный показатель позволяет оценить, сколько человек просмотрело публикацию и, соответственно, в курсе опубликованной информации новостного характера. Характеризует масштабность ее распространения. Показатель рассчитывается как сумма количества просмотров публикации с одного IP-адреса.

2. Количество шейров. Данный показатель позволяет оценить, сколько человек поделилось данной публикацией в аккаунтах социальных сетей или своих аккаунтах на страницах веб-ресурсов. Показатель позволяет измерить то, что читатель согласен с точкой зрения автора и считает, что новость является важной и ее необходимо распространить. Показатель может рассчитываться как сумма количества шейров публикации к общему количеству просмотров.

3. Количество комментариев. Данный показатель измеряет интерес новости у аудитории и побуждает читателей комментировать. Показатель может рассчитываться как сумма количества комментариев к одной публикации к общему количеству просмотров данной публикации.

4. Показатель «тональность комментариев» определяется в диапазоне: негативная, позитивная, нейтральная. Рассчитывается с применением метода лингвистического анализа, когда программа регистрирует слова-маркеры негативного, позитивного или

нейтрального отношения отдельных читателей. Еще одним вариантом определения тональности информационной системой является использование технологии машинного обучения DeepLearning.

Показатели, выработанные на основе информативных признаков критерия «Резонансность», характеризуют особенности оценки медиа-охвата, позволяют определить масштабность распространения публикаций новостного характера, определить основной характер отношения аудитории к публикации.

Следующая группа информативных признаков относится к критерию «Медиа-вовлеченность», который описывает поведение читателя на определенном веб-ресурсе. Эта группа информативных признаков не может применятся для оценки отдельной публикации. Причиной такого разделения является невозможность применения метрик медиа-вовлеченности к единице анализа «публикация» из-за их сущности. Однако эта группа информативных признаков позволяет оценивать лояльность и заинтересованность аудитории в самом веб-ресурсе и выступает условным индикатором доверия аудитории к нему и потенциала его влияния на аудиторию. Данный аспект влияния не может быть исключен из комплексной оценки, поскольку, согласно проведенному авторами социологическому опросу населения, наибольшее количество респондентов (32%) при определении уровня своего доверия к новостным сообщениям самым важным фактором считают сам источник информации и его репутацию [5].

Таким образом, определены информативные признаки, относящиеся к критерию медиа-вовлеченность:

- лояльность аудитории вэб-ресурса;
- наличие у аудитории заинтересованности в публикациях конкретного веб-ресурса, ее читательские предпочтения.

Количественным выражением этих информативных признаков могут являться показатели, измеряющие:

1. Глубину просмотра. Показатель рассчитывается как отношение количества сессий с просмотром более N страниц к общему количеству сессий.

2. Длительность просмотра. Показатель рассчитывается как отношение количества сессий, длящихся более N минут, к общему количеству сессий.

3. Частоту возращения на сайт. Показатель рассчитывается как отношение количества сессий с просмотром более N страниц за N период к общему количеству посещений сайта.

Информативные признаки для определения достоверности информации

Данные анализа различных литературных источников определения достоверности информации новостного характера показывают, что подходы и признаки определения достоверности публикаций едины. Вместе с тем большинство признаков определения достоверности сведений, информации, данных, а также публикации основаны на экспертных оценках.

Достоверность текстовой информации новостного характера определяется с применением комбинированного подхода – экспертного анализа публикаций и с использованием информационных систем.

По результатам анализа подходов к определению достоверности публикации новостного характера определены следующие информативные признаки достоверности отдельной публикации:

1. Наличие ссылки в публикации на компетентный источник:

- информации государственного органа РК;
- информации официального Интернет-ресурса государственного органа РК;
- информации, опубликованной организациями квазигосударственного сектора РК;
- информации, опубликованной национальными компаниями РК;
- официальной информации информационных агентств, электронных СМИ и иных аналогичных компаний с высоким уровнем репутации;
- данных аналитических отчетов, публикаций, исследований, подготовленных авторитетными и признанными международными организациями (ООН, ВОЗ, ЮНЕСКО, МВФ, Всемирный банк, Европейский банк реконструкции и развития, Интерпол и другие ресурсы государственных ведомств других стран);
- признанных международных статистических баз данных (Институт статистики ЮНЕСКО, Евростат, статистическая информация ОЭСР, Всемирного банка и др.);
- информации профессиональных экспертных площадок (консалтинговые, рейтинговые платформы), признанные в своей области, на которые опираются представители авторитетных международных организаций;
- результатов научных исследований, подготовленными научными организациями, вузами и др.;
- ответ на запрос в государственные органы и международные организации.

2. Ссылка на первоисточник информации.

3. Наличие автора публикации.

4. Рейтинг/репутация издания (сайты новостных и информационных агентств, электронных СМИ, иных аналогичных сайтов), на котором размещена публикация.

5. Освещение одного и того же события различными изданиями: кроссчекинг, наличие дискурса с другими публикациями.

6. Аутентичность фото и видео, которые подтверждают событие.

7. Соответствие заголовка публикации содержанию публикации.

Информативные признаки для определения тональности и объективности публикации

Определение тональности публикации является одним из механизмов мониторинга медиа-текстов на предмет оценки уровня влияния на мнение читателей, становясь одним из ключевых элементов при проведении контент-анализа.

Все три вариации проведения сентимент-анализа: статистический, использование словарей и правил, а также смешанный анализ - подразумевают выявление в тексте определенных слов и словосочетаний, которые и являются информативными признаками для отнесения публикации к той или иной категории в разрезе тональности.

Существенным препятствием в адекватном анализе тональности текста является наличие в тексте иронии и сарказма, их правильное определение может существенно повысить точность результатов. При использовании в тексте элементов иронии и сарказма значение и общая тональность текста может приобрести обратное прямому толкованию текста значение. Так, по данным исследования Высшей школы экономики, некоторыми признаками наличия иронии/сарказма в тексте могут выступать усиленные восклицания: избыточное использование восклицательных или вопросительных знаков, наличие междометий, маркеры ирреалиса, а также их сочетание [10].

Результаты изучения методов контент- и сентимент-анализа медиа-текстов позволили определить следующие информативные признаки по критерию «Тональность»:

- наличие в тексте слов/словосочетаний с ярко выраженной негативной окраской;
- наличие в тексте ненормативной лексики;
- наличие иронии и сарказма;
- наличие в тексте слов/словосочетаний с ярко выраженной позитивной окраской;
- наличие в тексте похвалы.

Объективность информации в целом рассматривается как отсутствие предвзятости, применительно к СМИ объективность является одним из основных принципов профессиональной журналистской этики, оценка которой включает в себя четыре составляющих (по Д. Брюэру). Это сбалансированное освещение тем и взглядов, отражение

разных мнений, изучение противоположных точек зрения и проверка того, что ни одна из сторон не оказалась недостаточно отраженной в материале.

Объективность является одним из неотъемлемых элементов этического кодекса журналистов и ЗРК «О средствах массовой информации» и дает возможность оценить непредвзятость автора публикации к рассматриваемым событиям. Объективность является социально значимым свойством журналистики, при котором автор точно и всесторонне излагает факты, подтверждает информацию компетентными мнениями и не манипулирует информацией, навязывая тем самым мнение какой-либо из сторон-участниц события. Как показали результаты социологического исследования, большинство респондентов (65%) частично или полностью согласно с мнением, что, пользуясь информацией только из казахстанских Интернет-источников, можно составить объективное представление о событиях, происходящих в стране.

Исходя из изученной информации были выработаны следующие информативные признаки для работы с классификацией медиа-текстов в разрезе объективности:

1. Наличие в медиа-тексте выраженного личного мнения автора публикации.

При этом само по себе наличие выраженного мнения не является основанием для отнесения текста к классу «необъективный», однако слишком частое употребление предложений с ярким выражением личного мнения автора в сочетании с отсутствием/малым количеством фактов, а также наличие в тексте признаков сарказма/иронии, агитации/антиагитации, может указывать на необъективность представленной информации.

2. Наличие манипуляторных техник в медиа-тексте:
   - степень генерализации (слабо выраженная, сильно выраженная, отсутствует);
   - политизированность;
   - наличие призыва к действию;
   - искажение информации;
   - наличие фактов в статье.

Первые четыре указывают на необъективность, последний индикатор является показателем объективности.

Выбранные информативные признаки для оценки влияния открытых текстовых информационных источников на социум можно объединить в 5 информативных критериев, которые наиболее приемлемы для исследования: *достоверность, резонансность, тональность, объективность и медиа-вовлеченность*.

Критерий «достоверность» позволяет оценить достоверность публикуемой информации на основе наличия ссылок в публикациях на источник, авторитетности

издания, на котором размещены публикации, освещения одного и того же события различными изданиями, проверки аутентичности фото и видео, подтверждающие событие, соответствие заголовка публикации содержанию медиа-текста.

Критерий «резонансность» позволяет оценить масштаб распространения публикации новостного характера и определить основной характер отношения аудитории к публикации на основе показателей медиа-охвата, количества репостов, шейров, комментариев и тональности комментариев.

Критерий «тональность» позволяет оценить эмоциональный тон подачи информации, фактов в публикации, который автор транслирует на аудиторию. Существует два основных метода автоматического определения тональности статистический метод и метод, основанный на словарях и правилах.

Критерий «объективность» позволяет оценить объективность подачи фактов, информации в публикации.

Критерий «медиа-вовлеченность» позволяет оценить поведение читателя на определенном веб-ресурсе на основе анализа глубины и длительности просмотра публикации и частоты возращения на сайт.

На основе информативных признаков определены следующие показатели оценки, которые объедены по вышеуказанным критериям.

I. Показатели по критерию «Достоверность»:

1. Наличие/отсутствие ссылки в публикации на компетентный источник:
   1.1. информации государственного органа РК;
   1.2. информации официального Интернет-ресурса государственного органа РК;
   1.3. информации, опубликованной организациями квазигосударственного сектора;
   1.4. информации опубликованной национальными компаниями РК;
   1.5. официальная информация информационных агентств, электронных СМИ и иных аналогичных компаний с высоким уровнем репутации;
   1.6. данные аналитических отчетов, публикаций, исследований, подготовленных авторитетными и признанными международными организациями (ООН, ВОЗ, ЮНЕСКО, МВФ, Всемирный банк, Европейский банк реконструкции и развития, Интерпол и другие ресурсы государственных ведомств других стран);
   1.7. признанные международные статистические базы данных (Институт статистики ЮНЕСКО, Евростат, статистическая информация ОЭСР, Всемирного банка и др.);

1.8. профессиональные экспертные площадки (консалтинговые, рейтинговые платформы, информация), признанные в своей области, на которые опираются представители авторитетных международных организаций;

1.9. результаты научных исследований, подготовленными признанными научными организациями, вузами и др.;

1.10. ответы на запросы в государственные органы и международные организации.

2. Ссылка на первоисточник
3. Освещение одного и того же события различными изданиями: кросс-чекинг, наличие дискурса с другими статьями. (Регистрируемые ИС параметры: Субъект (Предмет обсуждения) /Место/ Действие/Время).
4. Аутентичность фото, которое подтверждает событие.
5. Аутентичность видео, которое подтверждает событие.
6. Наличие факта последующей правки в уже опубликованной статье.
7. Соответствие заголовка публикации содержанию медиа текста.
8. Наличие/отсутствие автора.
9. Наличие/отсутствие проверяемых фактов в статье.
10. Репутация издания (сайты новостных и информационных агентств, электронных СМИ, иных аналогичных сайтов), на которых опубликована информация.

II. Показатели по критерию «Резонансность»:

1. Медиа-охват (количество просмотров).
2. Количество шейров информации.
3. Количество комментариев публикации.
4. Тональность комментариев:

– негативный;
– нейтральный;
– позитивный.

III. Показатель по критерию «Тональность»:

1. Тональность публикации:

– негативный;
– нейтральный;
– позитивный.

IV. Показатели по критерию «Объективность»:

1. Наличие в медиа-тексте выраженного личного мнения автора публикации.

2. Наличие манипуляторных техник в медиа-тексте.
3. Степень политизированности публикации.
4. Степень генерализации:

– слабовыраженная;
– сильно выраженная;
– нет.

V. Показатели по критерию «Медиа-вовлеченность»:

1. Глубина просмотра публикации.
2. Длительность просмотра публикации аудиторией.
3. Частота возращений аудитории на Интернет-портал СМИ.

Таким образом, по результатам проведенного исследования были получены следующие результаты:

1. Изучена роль СМИ в формировании общественного мнения.
2. Рассмотрены международное и казахстанское законодательство в сфере регулирования деятельности СМИ для выработки основы для критериев оценки.
3. Определены инструменты, техники и механизмы воздействия на социум, используемые СМИ.
4. Определены и исследованы существующие информативные признаки медиа-текстов.
5. Выработаны критерии и показатели оценки влияния открытых текстовых информационных источников на социум на основе информативных признаков для последующего применения в автоматизированной информационной системе оценки влияния СМИ на социум.

3. ИССЛЕДОВАНИЕ КОМПЛЕКСНОЙ ОЦЕНКИ ЗНАЧИМОСТИ ИСПОЛЬЗУЕМЫХ ИНФОРМАТИВНЫХ КРИТЕРИЕВ

Результаты исследования существующих информативных признаков влияния открытых текстовых информационных источников на социум. позволили определить приемлемые критерии оценки. Методика их определения и выбора основана на комбинированном подходе – экспертный метод и статистический метод оценки.

Выбранные информативные признаки были сгруппированы в 5 информативных критериев, которые наиболее приемлемы для оценки: достоверность, резонансность, тональность, объективность и медиа-вовлеченность. Для оценки влияния открытых текстовых информационных источников по предложенным критериям были выработано 22 показателя.

Важным этапом определения значимости предложенных информативных критериев оценки и их показателей является проведение комплексной оценки значимости. В этой связи *целью исследования* являлось определение приемлемого подхода к проведению комплексной оценки значимости, проведение выборки публикаций и формирование корпуса текстов и их разметка.

Объектом исследования являются информативные критерии оценки влияния открытых текстовых информационных источников на социум.

В ходе исследования был использован аддитивный метод к построению индексов.

Научная новизна работы заключается в прикладном эмпирическом характере исследования. Впервые основное внимание было уделено информативным критериям оценки публикаций новостного характера и их влиянию на социум. Данная работа является первой в Казахстане, в которой комбинированно применены методы количественного исследования и статистического моделирования для получения оценки значимости предложенных информативных критериев для оценки влияния публикаций новостного характера на социум.

Проведение выборки публикаций и определение показателей для комплексной оценки значимости

Для выработки возможных расчетов по распознаванию показателей оценки сформирован корпус текстов новостного характера по 5 информационным новостным Интернет-источникам: Central Asia Monitor, Tengrinews, Радио Азаттык, Zakon.kz, Казахстан 2.0.

Формирование корпуса проведено с применением 2-х этапной систематической кластерной выборки [11]. Для проведения комплексной оценки значимости основным

критерием отбора корпуса служило наличие таких данных по критерию резонасности, как количество просмотров, количество комментариев, количество шейров. В этой связи базой выборки для построения индексов выступили публикации, опубликованные на информационных новостных Интернет-порталах: Tengrinews.kz и Central Asia Monitor.

Для проведения процедуры комплексной оценки значимости был проанализирован корпус из 500 публикаций с наибольшим количеством просмотров. Данные публикации были просмотрены свыше 20 млн раз. К оценке публикации были привлечены 2 эксперта в области политологии и анализа данных.

Для получения данных с целью построения статистико-математической модели проведена экспертная оценка отобранных публикаций. Для эффективной организации экспертного анализа и получения объективных результатов была разработана инструкция, по оценке публикаций для проведения комплексной оценки значимости информативных критериев. Инструкция состоит из 4 разделов: определение достоверности публикации, определение резонансности публикации, определение тональности публикации, определение объективности публикации.

Экспертная оценка 500 публикаций проведена по следующим 16 показателям 4-х критериев оценки:

1. Критерий «Достоверность»

Показатели:

a. наличие/отсутствие ссылки в публикации на компетентный источник;
b. ссылка на первоисточник;
c. освещение одного и того же события различными изданиями: кросс-чекинг, наличие дискурса с другими статьями;
d. соответствие заголовка публикации содержанию медиа текста4
e. наличие автора в статье;
f. наличие проверяемых фактов в статье (Субъект (Предмет обсуждения) /Место/ Действие/Время);
g. репутация издания (сайты новостных и информационных агентств, электронных СМИ, иных аналогичных сайтов), на котором опубликована информация.

2. Критерий «Резонансность»

Показатели:

a. количество просмотров публикации;
b. количество репостов/шейров информации;
c. количество комментариев публикации;

d. тональность комментариев.

3. Критерий «Тональность»

Показатель: Тональность публикаций

4. Критерий «Объективность»

Показатели:

a. наличие в медиа-тексте выраженного личного мнения автора публикации;
b. наличие манипуляторных техник в медиа-тексте;
c. степень политизированности публикации;
d. степень генерализации.

Обоснование выбора подходов к построению индекса

Для проведения комплексной оценки значимости было предложено два подхода к построению индекса.

Первый подход – аддитивный. При таком построении переменным, составляющим индекс, присваиваются статистические веса, которые в последующем складываются для составления индекса. Как статистическая переменная, такой индекс представляет собой непрерывную переменную, значения которой варьируются в промежутке от [-100:100].

Второй подход – комплексное построение индекса с помощью техник уменьшения размерности [12] и факторного анализа [13].

Методика уменьшения размерности, а именно анализ главных компонент показывает, какие переменные имеют наибольшее собственное значение (eigenvalue) или же по-другому объясняющие наибольшую порцию дисперсии данных. По существу, методика позволяет отбросить переменные не «нужные» и оставить только «нужные». Далее на основе оставшихся переменных проводится факторный анализ, группирующий их в отдельные факторы, которые в дальнейшем можно трансформировать в индексы интереса.

Фактически два подхода представляют собой техники двух противоположных ответвлений машинного обучения: обучения без учителя (unsupervisedlearning) и обучения с учителем (supervisedlearning) (Ramasubramanian&Singh, 2017) [14].

К классу алгоритмов машинного обучения с учителем относятся методы регрессионного анализа. Структура данных должна содержать четко обозначенную зависимую переменную. При наличии такой переменной алгоритм выстраивает цепочку независимых переменных для выявления взаимозависимостей, вычисления прогноза или же установления статистически значимых переменных. В свою очередь к классу алгоритмов машинного обучения без учителя относятся методы кластеризации,

уменьшения размерности, нейронные сети. При данном алгоритме структура данных не имеет четко обозначенных переменных и обучение происходит, опираясь на измерение дистанции между каждой строчкой в множестве данных. Далее методики группируют данные согласно дистанции и образуют из них гомогенные массивы.

Для построения индекса можно использовать оба подхода. Однако в этом случае структура данных содержит в себе такой критерий оценки, как резонансность, в которую входят непрерывные переменные: (1) количество просмотров, (2) количество лайков, (3) количество комментариев, (4) количество шейров.

Вышеупомянутые переменные имеют четкое обозначение и могут служить в качестве зависимых переменных в статистической модели. Однако лишь «количество просмотров» обладает полнотой данных во всех медиа-ресурсах, что оставляет ее в качестве единственного кандидата на роль подходящей зависимой переменной. Другим моментом, который выступает в пользу построения индекса, согласно аддитивному подходу, является наличие четко обозначенных теоретически, информативных критериев публикации. Только лишь при отсутствии таковых наиболее адекватным шагом было бы использование второго подхода.

Таким образом, структура данных и наличие теоретических концептов позволяют использовать аддитивный подход к построению индексов, цель которых адекватно измерить предложенные критерии оценки влияния публикации новостного характера на социум.

При аддитивном построении составным частям индекса присваивается вес по следующей формуле:

$$W_i = \frac{1}{\text{кол}-\text{во переменных}} \quad (1)$$

где: i - индекс

Если составная часть представляет собой бинарную переменную (0-1), то вес остается согласно вышеприведённой формуле, если же часть представляет собой значение более 1, то тогда внутри части вес рассчитывается по формуле:

$$W_{ik=} \frac{w_i}{\text{максимальная величина переменной}} \quad (2)$$

где: k – переменная

В итоге полученный вес складывается для построения переменной, которая находится в максимальных пределах [-1:1]. Далее с целью построения индекса, переменная умножается на 100.

Статистическая модель

В связи с тем, что используется алгоритм обучения с учителем (supervisedlearning), целесообразно использовать регрессию в качестве основной статистической модели. Зависимая переменная «количество просмотров» имеет непрерывный тип данных, вследствие этого можно использовать модель обычной множественной регрессии.

Независимые переменные включили в себя построенные индексы:

– индекс достоверности публикации;
– индекс тональности публикации;
– индекс объективности публикации.

Таким образом, общая модель приняла следующий функциональный вид:

$$y(\text{кол} - \text{во просмотров}) = \beta_0 + \beta_1 \text{Индекс достоверности} + \beta_2 \text{индекс тональности} + \beta_3 \text{индекс обьективности} + \varepsilon \quad (3)$$

где,

β_0 – пересечение, а остальные бета коэффициенты характеризуют наклонную,

ε – случайная ошибка.

Также дополнительно можно посмотреть на статистическое взаимодействие переменных индекса тональности и индекса объективности. Статистические взаимодействия позволяют увидеть одновременное влияние двух переменных на зависимую переменную.

Таким образом, частная модель которая включает в себя статистическое взаимодействие индекса тональности и индекса объективности выглядит следующим образом:

$$y(\text{кол} - \text{во просмотров}) = \beta_0 + \beta_1 \text{Индекс достоверности} + \beta_2 \text{индекс тональности} + \beta_3 \text{индекс обьективности} + \beta_2 \text{индекс тональности} * \beta_3 \text{индекс обьективности} + \varepsilon \quad (4)$$

4. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ, НАПРАВЛЕННЫХ НА ЭКСПЕРИМЕНТАЛЬНОЕ ПОДТВЕРЖДЕНИЕ АДЕКВАТНОСТИ ПОЛУЧЕННОЙ КОМПЛЕКСНОЙ ОЦЕНКИ ЗНАЧИМОСТИ ИСПОЛЬЗУЕМЫХ ИНФОРМАТИВНЫХ КРИТЕРИЕВ ДЛЯ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ НА ОСНОВЕ АНАЛИЗА ПУБЛИКУЕМОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

Одним из главных элементов определения приемлемых критериев оценки влияния открытых текстовых информационных источников на социум является определение степени их значимости.

По результатам экспертной оценки корпуса публикаций новостного характера была проведена процедура комплексной оценки значимости с построением аддитивных индексов. Также проведена процедура техники бутстрэп.

В ходе данной процедуры были построены три индекса:

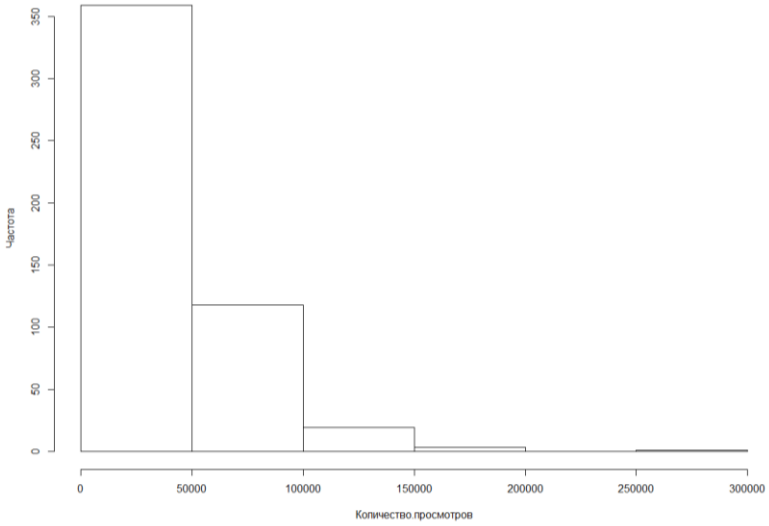
– индекс объективности;
– индекс достоверности;
– индекс тональности.

По результатам множественной регрессионной модели оценки значимости установлен статистически значимый эффект между влияниями индексов достоверности, тональности и объективности на переменную «количество просмотров». Так как статистические тесты показали наличие автокорреляции и гетероскедастичности остатков, для вычисления стандартных ошибок были использованы робастные стандартные ошибки Нью-Уэста, которые хоть и увеличили коридор отклонения средней величины, тем не менее все же показали высокий уровень статистической значимости коэффициентов при каждом из трех использованных индексов. Также R – квадрат модели показал умеренно хороший коэффициент детерминации (0.44), что указывает на то, что модель довольно неплохо отразила изучаемый феномен. Результаты были проверены посредством техники бутстрэпинг, которая показала незначительное отклонение от результатов модели.

Построение статистико-математической модели

В качестве основной статистико-математической модели была использована множественная регрессия, в которой параметры интереса были оценены с помощью метода наименьших квадратов. В роли зависимой переменной был использован показатель «количество просмотров». Независимые переменные включили в себя: индекс достоверности публикации, индекс тональности, а также индекс объективности. Следует отметить, что при детальном рассмотрении зависимая переменная показала экспоненциальное распределение (рисунок 2).

Рисунок 2. Статистическое распределение переменной «количество просмотров»



Для более точного вычисления была использована логарифмическая трансформация переменной «количество просмотров», таким образом, общая модель приняла следующий функциональный вид:

$$\log(y) = \beta_0 + \beta_1 \text{Индекс достоверности} + \beta_2 \text{индекс тональности} + \beta_3 \text{индекс обьективности} + \varepsilon \quad (5)$$

Также дополнительно было создано статистическое взаимодействие переменных индекса тональности и индекса объективности. Статистические взаимодействия позволяют увидеть одновременное влияние двух переменных на зависимую переменную. Таким образом, частная модель включила в себя статистическое взаимодействие индекса тональности и индекса объективности:

$$log(y) = \beta_0 + \beta_1 \text{Индекс достоверности} + \beta_2 \text{индекс тональности} + \beta_3 \text{индекс обьективности} + \beta_2 \text{индекс тональности} * \beta_3 \text{индекс обьективности} + \varepsilon \quad (6)$$

Так как составная часть индекса резонансности была использована в качестве зависимой переменной, переменная тональность комментариев была объединена с переменной тональность публикации для создания индекса тональности. Для непосредственного вычисления построенной модели был использован язык статистического программирования – R.

Полученные результаты

В таблице 1 представлены результаты вычислений. Если обратить внимание на значение-р, то можно увидеть, что все независимые переменные показывают значимые результаты (значение-р меньше порогового уровня ошибки 5%). Так как зависимая

переменная прошла логарифмическую трансформацию, то при интерпретации результатов необходимо использовать следующую формулу $\exp(\beta) - 1$.

Так, к примеру, увеличение индекса достоверности на один процент при прочих равных условиях ведет к повышению просмотров на 2%, увеличение индекса тональности на один процент ведет к уменьшению количества просмотров на 1/3 процента, увеличение индекса объективности на 1 процент ведет к сокращению просмотров на 2,3 процента.

Следует отметить, что из трех результатов тональность показывает наименьшее влияние на количество просмотров, однако это скорее всего связано с меньшим количеством критериев, по которым был построен индекс.

Таблица 1. Результаты общей модели множественной регрессии

	параметр оценки	стандартная ошибка	t-критерий	значение-p
пересечение	9.5805	0.1919	49.918	2e-16***
индекс достоверности	0.0193	0.0031	6.126	1.84e-09***
индекс тональности	-0.0026	0.0010	-2.514	0.0122*
индекс объективности	-0.0230	0.0014	-15.563	2e-16***
R-квадрат – 0.4274		Множественный R-квадрат -0.424		

Также представлены результаты второй модели со статистическим взаимодействием между индексом объективности и индексом тональности (таблица 2).

Таблица 2. Результаты частной модели множественной регрессии

	параметр оценки	стандартная ошибка	t-критерий	значение-p
пересечение	9.586e+00	1.918e-01	49.970	2e-16***
индекс достоверности	1.950e-02	3.160e-03	6.172	1.41e-09***
индекс тональности	-1.586e-03	1.373e-03	-1.155	0.249
индекс объективности	-2.428e-02	1.753e-03	-13.849	2e-16***
Объективность* тональность	-5.307e-05	4.085e-05	-1.299	0.194
R-квадрат – 0.4294		Множественный R-квадрат -0.4248		

Из таблицы 2 видно, что коэффициент взаимодействия не имеет статистической значимости, более того переменная взаимодействия не повлияла на улучшение модели и повлекла за собой ухудшение коэффициента тональности, в добавок к этому введение новой переменной не повлекло за собой улучшение величины R-квадрат. В то же время достоверность и объективность остались на прежнем уровне значений-p. Вследствие выше сказанного в дальнейшем анализе использовалась общая модель множественной регрессии.

Несмотря на высокие показатели значимости коэффициентов необходимо отметить, что модель имеет две серьезные проблемы, а именно:

1) автокорреляция остатков;

2) гетероскедастичность остатков.

Данные выводы были сделаны в результате проверки на соответствующие статистические тесты.

Таблица 3. Результаты тестов на гетероскедастичность и автокорреляцию

Тест Дарбина-Уотсона на наличие автокорреляции случайных ошибок	Значение-р (<2.2e-16)
Тест Бройша-Пагана на наличие гетероскедастичности	Значение-р (<0.0046)

В обоих тестах значение-р достаточно мало, чтобы сделать вывод в пользу наличия автокорреляции и гетероскедастичности.

Для устранения данных проблем существует два варианта.

Первый связан с использованием новых данных, в особенности, данных зависимой переменной.

Второй предполагает оставить все как есть, однако расширить интервалы доверия вокруг параметров интереса. В данном случае был выбран второй вариант, в частности, использованы робастные (состоятельные, крепкие) стандартные ошибки робастных стандартных ошибок (HAC), предложенные Нью и Уэстом (1987). [15] Результаты вычислений представлены в таблице 4.

Таблица 4. Результаты общей модели множественной регрессии со стандартными ошибками HAC

	параметр оценки	стандартная ошибка HAC
пересечение	9.5805	0.35100
индекс достоверности	0.0193	0.00538
индекс тональности	-0.0026	0.001125
индекс объективности	-0.0230	0.003320

Из таблицы видно, что после использования процедуры робастные стандартные ошибки HAC увеличили интервалы доверия почти в два раза во всех коэффициентах кроме индекса тональности.

Альтернативное построение модели с помощью техники бутстрэп

Так как истинное статистическое распределение популяции неизвестно, а распределение в имеющейся выборке принимает экспоненциальный характер, была проведена процедура техники бутстрэп.

Бутстрэп представляет собой непараметрическую технику вычисления распределения выборки и параметров интереса, без традиционных допущений об истинной форме статистического распределения популяции (Fox, 2002) [16].

Техническое исполнение процедуры проводится посредством выборки с замещением наблюдений не из популяции, а из имеющейся выборки. Данная процедура производится некоторое количество раз (чем больше, тем лучше) и в дальнейшем позволяет получить все необходимые параметры интереса. Тем не менее в данном случае интерес представляют собой не статистические параметры популяции, а параметры регрессионной модели, что в некоторой степени модифицирует стандартную технику построения бутсрэп.

Для того, чтобы правильно построить коэффициенты бета и их стандартные модели, техника бутстрэп была проведена непосредственно на самих регрессиях методом перевыборки случайных-х.

Формально метод может быть описан следующим образом. Представим независимую переменную и зависимые переменные как матрицу $M_{n\times k}$ = $(Y_{ij}, X_{i1}, X_{i2}, X_{i3})$, где i=1…..n. Далее производится выборка некоторого количества наблюдений O из матрицы $M_{n\times k}$, так что $0 < O < n$, после этого производится вычисление регрессионной модели из которой коэффициенты бета и их стандартные ошибки сохраняются в матрицу $M_{p\times c}$, где p – количество итераций, а c – число коэффицентов бета и их сопутствующих стандартных ошибок. Таким образом, вышеописанная процедура была проведена 1000 раз и получена матрица L размером 1000×8.

В таблице 5 приведены сравнительные результаты параметров, полученных после процедуры бутстрэп и обычных вычисленных параметров модели.

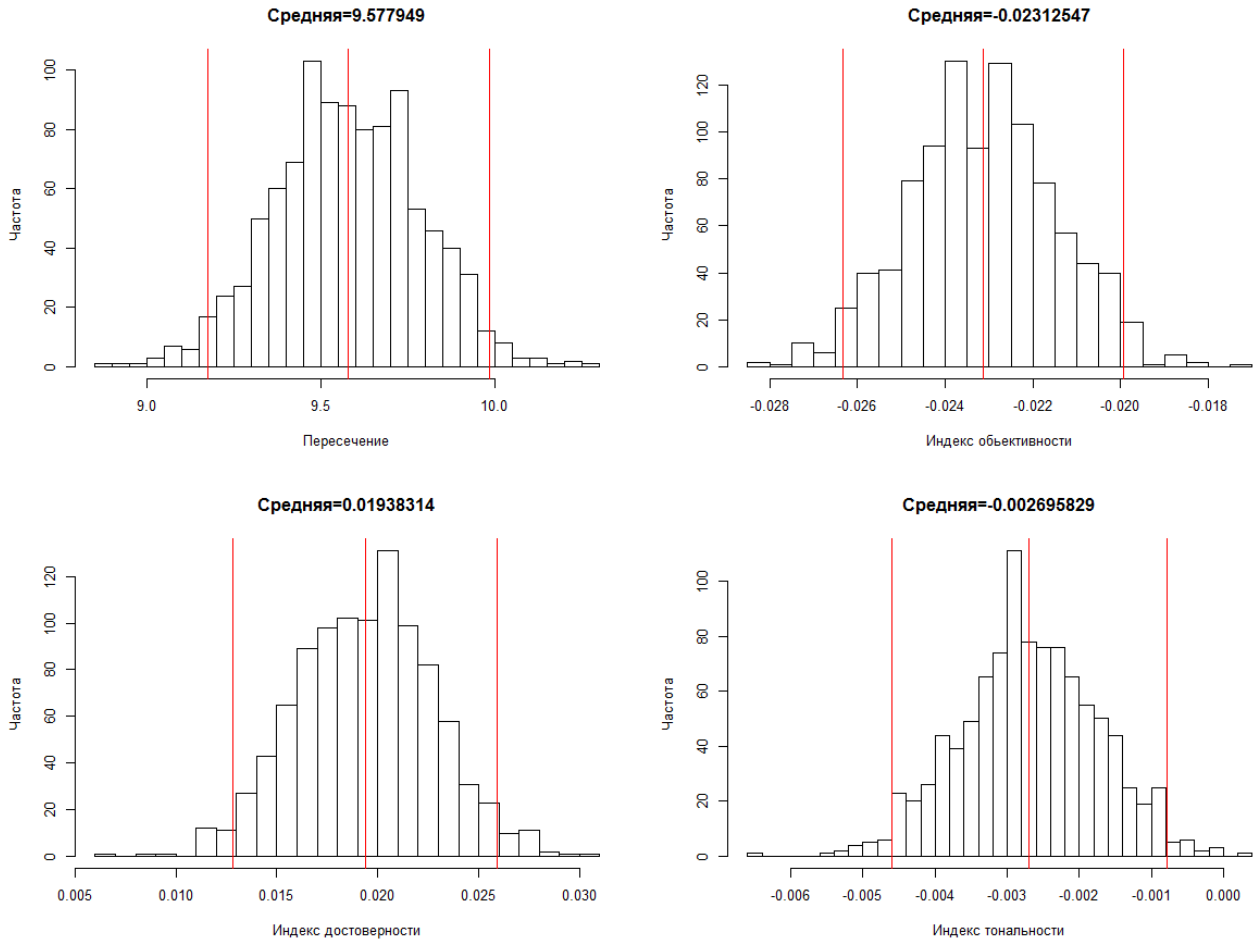
Таблица 5. Результаты процедуры «бутстрэп»

	параметр оценки (обычная модель)	стандартная ошибка HAC (обычная модель)	параметр оценки (бутстрэп)	стандартная ошибка HAC (бутстрэп)
пересечение	9.5805	0.35100	9.5779	0.20186
индекс достоверности	0.0193	0.00538	0.0193	0.00327
индекс тональности	-0.0026	0.001125	-0.0026	0.000954
индекс объективности	-0.0230	0.003320	-0.0231	0.001597

Из таблицы видно, что процедура бутстрэп значительно сократила значения стандартных ошибок. Также необходимо отметить, что смещение результатов

незначительно в сравнении с обычной моделью. Ниже представлен рисунок параметров бутстрэп с вычисленными средними величинами и интервалами доверия.

Рисунок 3. Результаты процедуры «бутстрэп»



Графики показывают, что интервалы доверия, построенного с помощью процедуры бутстрэп, находятся в рамках области статистического распределения, не далее, чем в два стандартных отклонения от средней величины.

Таким образом, по результатам экспериментального подтверждения комплексной оценки значимости можно сделать следующие выводы.

Информативные критерии публикации, такие как достоверность, объективность и тональность, оказывают статистически значимое влияние на количество просмотров. Так, к примеру, уменьшение индексов объективности и тональности положительно влияет на количество просмотров, в свою очередь, увеличение индекса достоверности увеличивает количество просмотров.

Данные результаты были получены посредством статистической модели множественной линейной регрессии. Помимо статистически значимых результатов модель показала хороший коэффициент детерминации (R квадрат=0.44).

Однако необходимо отметить что в ходе анализа выяснилась довольно проблемная структура данных, в особенности моменты, связанные с автокорреляцией и гетероскедастичностью остатков. Хотя лучшим выходом было бы собрать новые данные или изменить спецификацию модели, тем не менее, учитывая, что основная цель - проверить значимость информационных критериев, данные проблемы были исправлены посредством использования HAC, а также непараметрической техники бутстрэп.

Результаты статистического моделирования показывают, что предложенные информативные критерии публикации могут в дальнейшем служить каркасом для создания автоматизированной модели машинного обучения.

По поводу потенциальных улучшений следует сказать, что необходимо увеличить число переменных внутри информативных критериев. Данный шаг позволит получить более точный и валидный индекс при построении более полной модели.

Также довольно неожиданным результатом стало создание расширенного индекса тональности, а именно включение в него переменной – тональность комментариев. Хотя индекс и показал наименьшее влияние на количество просмотров, тем не менее результаты оказались статистически значимыми как в модели, так и после процедуры бутстрэп.

5. ИССЛЕДОВАНИЕ ОСОБЕННОСТЕЙ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ В УСЛОВИЯХ РЕСПУБЛИКИ КАЗАХСТАН ПРИМЕНИТЕЛЬНО К ССД

Целью исследования является изучение существующего опыта мониторинга СМИ в Казахстане и определение характера влияния открытых информационных источников на различные социальные группы населения Казахстана.

Объектом исследования выступила оценка влияния СМИ на социум Казахстана.

Метод проведения работы – анализ законодательства Республики Казахстан и существующих практик мониторинга СМИ, а также социологический опрос населения.

Исследование охватывает такие вопросы, как законодательное регулирование в области мониторинга СМИ, опыт мониторинга открытых текстовых информационных источников в Казахстане и влияние казахстанских информационных новостных Интернет-источников на социум.

В целях определения характера влияния открытых информационных источников (электронных СМИ) на социум проведено социологическое исследование по оценке влияния открытых информационных источников на социум Казахстана. В опросе приняло участие 3 200 респондентов из 16 регионов Казахстана (14 областей и 2 города республиканского значения). Исследованием были охвачены все социальные группы населения страны в разрезе гендерной принадлежности *(52,5% мужчин и 47,5% женщин)*, возрастных групп *(15-19, 20-24, 25-34, 35-44, 45-54, 55-64, 65 лет и старше)*, национальностей *(61,8% казахи, 26,8% русские и 11,4% другие национальности)*, языка общения, семейного положения, типа поселения *(56,7% город и 43,3% село)*, по типу занятости *(28,8% наемные работники частного сектора, 17,5% работники госсектора, 13,1% пенсионеры, 4,6% предприниматели и др.)* и материального благосостояния.

Для исследования особенностей оценки влияния открытых текстовых информационных источников были выдвинуты следующие гипотезы:

1. В общей структуре потребителей электронных СМИ новостного характера в Казахстане преобладает экономически активное население.
2. Население считает, что СМИ влияют на их видение ситуации, событий, явлений.
3. Население считает, что СМИ объективно отражают казахстанскую реальность.
4. Большинство населения считает, что средства массовой информации используются как инструмент для освещения событий, сферы и др. под определенным углом, выгодным для определенного круга лиц.
5. Большинство населения доверяет новостным Интернет-порталам.

6. Оперативность и доступность новостной информации в Интернете не всегда подкрепляется ее достоверностью, что оказывает влияние на доверие к Интернет-источникам.
7. Большинство электронных средств массовой информации зачастую жертвуют точностью, чтобы первыми осветить событие для привлечения аудитории.
8. Большинство населения не может отличить качественно подготовленный журналистский материал от информации, основанной на слухах и лживых данных.

В соответствии с целью и гипотезами исследования были определены следующие задачи:

1. Исследование особенностей законодательного регулирования в области мониторинга СМИ;
2. Анализ опыта мониторинга открытых текстовых информационных источников в Казахстане;
3. Проведение социологического исследования населения по оценке влияния открытых информационных источников (электронных СМИ) на социум.

Ряд выводов и полученных результатов обладают новизной. Новизна проведенного исследования заключается, прежде всего, в определении социально-демографического портрета среднестатистического читателя информационных порталов и поведенческих особенностей населения при медиа-потреблении информации. Получены результаты оценки мнения населения об уровне объективности и достоверности информации в освещении событий электронными СМИ в Казахстане и уровень доверия к казахстанским новостным информационным источникам. Построены индексы доверия каналам распространения информации, доверия информационным Интернет-сайтам, объективности информационных Интернет-источников и их прозрачности в разрезе регионов Казахстана.

Законодательное регулирование в области мониторинга СМИ

Законодательство Казахстана обязывает СМИ придерживаться принципов объективности, законности, достоверности и уважения частной жизни, чести и достоинства человека и гражданина.

В Казахстане уполномоченным органом в области СМИ является Министерство информации и коммуникаций Республики Казахстан (далее – МИК РК). При обнаружении в медиа-пространстве информации, способной вызвать негативную реакцию у общества, МИК РК обязан своевременно ставить в известность первого руководителя государственного органа [17].

Согласно ЗРК «О средствах массовой информации», государственное регулирование в области СМИ осуществляется путем правового обеспечения, государственного контроля за соблюдением казахстанского законодательства о СМИ [18]. К компетенции МИК РК относится проведение мониторинга средств массовой информации. Мониторинг СМИ осуществляется на предмет соблюдения казахстанского законодательства. Мониторинг осуществляется в соответствии с Правилами проведения мониторинга средств массовой информации. Данные правила регулируют порядок проведения мониторинга СМИ [19]. Мониторинг СМИ осуществляется с использованием информационной системы «Автоматизированный мониторинг национального информационного пространства» (при ее наличии).

Мониторинг доступности и свод информации, запрещенной к распространению по сетям телекоммуникаций на территории Казахстана, осуществляется организацией, выполняющей работы по техническому и методическому обеспечению мониторинга, с использованием аппаратно-программных средств.

Опыт мониторинга открытых текстовых информационных источников в Казахстане

Существующая практика и проведённые исследования в сфере СМИ охватывают только определённые направления СМИ. В частности, государственный мониторинг осуществляется на предмет соблюдения законодательства в сфере СМИ. Систематический мониторинг и оценка влияния СМИ в Казахстане проводится РГП «Центром анализа информации» на предмет соблюдения казахстанского законодательства. Основная деятельность по мониторингу и оценке влияния СМИ, как правило, ведется в разрезе одного определенного аспекта, например, оценка эффективности политической информационной кампании, деятельности организации и т.д. Целью такой оценки чаще всего выступает продвижение нововведений и популяризация государственной политики.

Консалтинговые услуги по проведению мониторинга и контент анализа СМИ направлены только для оценивания отношения к деятельности организации, персонам, брендам и предназначены в большей степени для продвижения нововведений, услуг и товаров.

По информации базы данных АО «Национальный центр государственной научно-технической экспертизы», в период с 1995 по 2016 годы в Казахстане зарегистрировано 27 научно-исследовательских работ, косвенно связанных с темой исследования. Вместе с тем, проведенные НИР малочисленны, охватывают только определенные области изучения роли СМИ и лишь косвенно затрагивают проблематику комплексной оценки медиа-воздействия и влияния открытых тестовых информационных источников на социум.

Основной методологией оценки влияния СМИ на социум выступает экспертный подход. Хотя организации, занимающиеся мониторингом и оценкой влияния СМИ, располагают определенными набором вспомогательных инструментов, например, программа транскрибирования текста из видеоролика и получение данных в текстовом формате, вместе с тем отсутствуют инструменты автоматизированного систематического мониторинга и оценки, особенно в части контент-анализа.

Автоматизированные средства мониторинга СМИ достаточно широко применяются на этапе сбора данных, однако при анализе данных автоматизированные инструменты применяются либо редко, либо в коммерческих целях. В области оценки медиа-воздействия исследования, кодировка и анализ данных в основном проводятся экспертами вручную. В условиях постоянного увеличения количества информации, вопрос автоматизации будет приобретать все большую актуальность.

Результаты социологического исследования по оценке влияния открытых информационных источников (электронных СМИ) на социум Казахстана

Методология социологического исследования

Цель: определение характера влияния открытых информационных источников (электронных СМИ) на социум.

Задачи опроса:

1. определить основные каналы получения населением информации новостного характера;

2. определить социальный портрет среднестатистического читателя электронных СМИ;

3. выявить уровень медиа-потребления населением СМИ, в том числе электронных;

4. оценить уровень прозрачности электронных СМИ по мнению населения;

5. определить уровень доверия населения к СМИ, в том числе электронных;

6. выявить по мнению населения уровень достоверности освещения событий казахстанскими СМИ, в том числе электронных;

7. Определить уровень влияния электронных СМИ на общественное мнение.

Метод опроса: анкетный опрос населения (face-to-face).

Целевая аудитория опроса: население Республики Казахстан от 15 лет и старше.

Рабочие гипотезы опроса:

1. В общей структуре потребителей электронных СМИ новостного характера в Казахстане преобладает экономически активное население.
2. Население считает, что СМИ влияют на их видение ситуации, событий, явлений.

3. Население считает, что СМИ объективно отражают казахстанскую реальность.
4. Большинство населения считает, что средства массовой информации используются как инструмент для освещения событий, сферы и др. под определенным углом, выгодным для определенного круга лиц.
5. Большинство населения доверяет новостным Интернет-порталам.
6. Оперативность и доступность новостной информации в Интернете не всегда подкрепляется ее достоверностью, что оказывает влияние на доверие к Интернет-источникам.
7. Большинство электронных средств массовой информации зачастую жертвуют точностью, чтобы первыми осветить событие для привлечения аудитории.
8. Большинство населения не может отличить качественно подготовленный журналистский материал от журналистского материала, основанного на слухах и лживой информации.

Параметры опроса:

– География опроса: 14 областей и 2 города республиканского значения.
– Выборка опроса: репрезентативная населению Казахстана (15 лет и старше), равно наполненная по областям и 2 городам республиканского значения, квота на пол и возраст по определенным категориям населения (без квоты на пол и возраст).
– Объем выборки: 3 200 респондентов из расчета по 200 респондентов в каждом из 16 регионов (таблица 6).

Объем выборочной совокупности n=3200 удовлетворяет условиям доверительной вероятности 95% с доверительным интервалом ±1,73%.

Объем выборочной совокупности по регионам n=200 удовлетворяет условиям доверительной вероятности 95% с доверительным интервалом ±6,93%.

Таблица 6. Распределение выборки по регионам и местности проживания

№	Регионы	Генеральная совокупность			Выборочная совокупность		
		15 лет и старше	город	село	15 лет и старше	город	село
1.	Акмолинская	562 724	265 960	296 764	200	96	104
2.	Актюбинская	614 002	387 463	226 539	200	126	74
3.	Алматинская	1 388 296	332 703	1 055 593	200	48	152
4.	Атырауская	410 424	195 893	214 531	200	95	105
5.	Западно-Казахстанская	480 360	244 471	235 889	200	102	98
6.	Жамбылская	747 626	306 666	440 960	200	82	118
7.	Қарагандинская	1 060 927	847 986	212 941	200	160	40
8.	Қостанайская	702 907	376 482	326 425	200	107	93

№	Регионы	Генеральная совокупность			Выборочная совокупность		
		15 лет и старше	город	село	15 лет и старше	город	село
9.	Қызылординская	519 813	227 410	292 403	200	87	113
10.	Мангистауская	423 170	184 148	239 022	200	87	113
11.	Южно-Казахстанская	1 850 560	846 206	1 004 354	200	91	109
12.	Павлодарская	588 825	420 565	168 260	200	143	57
13.	Северо-Казахстанская	446 705	200 477	246 228	200	90	110
14.	Восточно-Казахстанская	1 080 259	660 210	420 049	200	122	78
15.	г.Астана	704 246	704 246	0	200	200	-
16.	г.Алматы	1 374 944	1 374 944	0	200	200	-
ВСЕГО		**12 955 788**	**7 575 830**	**5 379 958**	**3 200**	**1 843**	**1 836**

Социально-демографическая структура участников исследования

Качественный состав респондентов по социально-демографическим характеристикам показывает, что опросом охвачены все социальные группы населения Казахстана в разрезе гендерной принадлежности *(52,5% мужчин и 47,5% женщин)*, возрастных групп*(15-19, 20-24, 25-34, 35-44, 45-54, 55-64, 65 лет и старше)*, национальностей *(61,8% казахи, 26,8% русские и 11,4% другие национальности)*, языка общения, семейного положения, типа поселения *(56,7% город и 43,3% село)*, по типу занятости *(28,8% наемные работники частного сектора, 17,5% работники госсектора, 13,1% пенсионеры, 4,6% предприниматели и др.)* и материального благосостояния.

Анализ источников потребления информации новостного характера и социальный потрет среднестатистического читателя

Результаты исследования показали, что предпочтения участников опроса при выборе основного источника получения информации почти с одинаковыми показателями разделились на три ведущих сегмента: социальные сети, блогеры (30,3%), телевидение (29,8%) и новостные Интернет-сайты (29,5%). Доля респондентов, предпочитающих получать информацию из печатных изданий, составила 6,2%. Другие каналы информации являются менее значимыми.

Для молодежи источником информации в первую очередь в сравнении с другими возрастными категориями выступают социальные сети и блогеры. Исследование выявило зависимость пользования социальными сетями от возраста: чем моложе респондент, тем чаще для него социальные сети выступают в качестве основного источника информации. И, наоборот, чем старше респондент, тем реже он использует сетевые платформы (социальные сети, блоги) как источник новостной информации. Реже всего в социальных

сетях проводят время сельские жители и респонденты старшего возраста, чаще всего – учащиеся (15-19 лет) и городские жители.

Противоположная картина складывается в отношении телевидения. Частота просмотра телевизора прямо пропорциональна возрасту респондента, т.е. чем старше человек, тем чаще он смотрит телевизор, и, наоборот, чем моложе, тем реже.

Интернет-сайтами зачастую пользуются респонденты в возрасте 25-34 и 35-44 лет, имеющие высшее образование.

Предварительная гипотеза исследования по поводу структуры потребителей электронных СМИ новостного характера частично подтверждается за исключением социальных сетей. Среди потребителей почти всех видов электронных СМИ преобладает экономически активное население. Потребителями социальных сетей, в первую очередь, является молодежь в возрасте до 24 лет и школьники.

Больше половины респондентов (59%) делают акцент на то, что при потреблении новостей они хотят только быть в курсе и ограничиваются одним, максимум двумя новостями о событии. Только каждый пятый респондент стремится использовать расширенный список источников информации при потреблении новостей, чтобы детально узнать о происходящем событии.

По результатам опроса, лидерами как по популярности, так и по доли регулярной аудитории является новостной портал *nur.kz* и интернет-портал *mail.ru.* 61,8% пользователей Интернета отметили, что они знают сайт nur.kz. Доля регулярной аудитории данного сайта (тех, кто регулярно пользуется сайтом) составляет 51,3%. На третьем и четвертом месте разместились новостные порталы *Tengrinews.kz* и *Zakon.kz.* Эти сайты знают каждый третий-четвертый Интернет-пользователь (32,8% и 28%). На долю постоянной аудитории данных сайтов (тех, кто регулярно пользуется сайтом) приходится 24,1% и 16% соответственно. Популярность сайтов *365 info.kz*, *Азаттық радиосы* и *Караван* варьируется на уровне 10%. Доля ежедневной аудитории данных сайтов составляет 4-5% (таблица 7).

Таблица 7 - Скажите, какие казахстанские информационные/новостные/аналитические Интернет-сайты Вы знаете/Назовите все Интернет-сайты, которыми Вы пользовались хотя бы однажды/Назовите те Интернет-сайты из перечисленных, которыми Вы пользуетесь регулярно? (% от общего числа опрошенных)

Название источника	Знание		Пользование хотя бы однажды	Пользование регулярно
	Знание без подсказки	Знание с подсказкой		
nur.kz	**61,8%**	**27,9%**	**54,7%**	**51,3%**
tengrinews.kz	**32,8%**	**29,4%**	**33,2%**	**24,1%**

Название источника	Знание		Пользование хотя бы однажды	Пользование регулярно
	Знание без подсказки	Знание с подсказкой		
camonitor.kz	5,3%	10,7%	5,8%	2,6%
mail.ru	**57,3%**	**30,3%**	**52,8%**	**52,3%**
zakon.kz	**28,0%**	**36,3%**	**34,8%**	**16,0%**
vlast.kz	6,3%	16,4%	7,8%	3,1%
forbes.kz	8,2%	15,8%	8,3%	4,6%
azattyq.org	6,8%	10,7%	5,0%	3,7%
caravan.kz	9,0%	15,9%	9,1%	4,0%
bnews.kz	6,8%	16,7%	8,9%	2,9%
kapital.kz	4,9%	12,4%	5,2%	1,5%
365info.kz	9,9%	18,5%	12,6%	4,8%
time.kz	6,0%	18,7%	8,7%	2,8%
rikatv.kz	2,5%	2,0%	2,3%	2,3%
Затрудняюсь ответить	4,3%	8,7%	5,1%	6,0%

Социально-демографический портрет активных пользователей новостных сайтов приведен в таблице 8.

Таблица 8 - Социально-демографический портрет активных пользователей информационных сайтов (% от числа пользующихся сайтом регулярно)

	Mail.ru (52%)	**Nur.kz (51%)**	**Tengrinews.kz (24%)**	**Zakon.kz (16%)**	**365info.kz (5%)**	**Азаттық радиосы (5%)**
Пол	Женщины	Не имеет значения	Не имеет значения	Мужчины	Не имеет значения	Не имеет значения
Возраст	15-34	20-34	25-34	35-44; 45-54	55-64; 25-34	45-54; 20-24
Этнос	Другие этнические группы	Казахи	Казахи	Казахи и русские	Казахи и др. этнические группы	Казахи
Образование	Неполное среднее и высшее	Высшее	Высшее	Высшее	Высшее	Среднее общее
Материальное положение	Средний и ниже среднего уровень дохода	Выше среднего уровень дохода	Высокий уровень дохода	Средний и выше среднего уровень дохода	Высокий уровень дохода	Выше среднего уровень дохода
Язык общения	На русском	На казахском; на казахском и русском	На казахском и русском	На казахском и русском	На казахском и русском	На казахском
Местность проживания	Село	Не имеет значения	Город	Город	Город	Село

Достоверность в освещении событий и уровень доверия населения к казахстанским новостным информационным источникам

Результаты опроса показывают, что телевидение продолжает оставаться наиболее доверительным каналом распространения информации. Более того, этот источник чаще всего выбирали как наиболее достоверный, правдивый (61%).

Также по результатам опроса была подтверждена гипотеза о том, что в целом новостным Интернет-порталам, как каналам информации, доверяет большинство казахстанцев: 30% сказали, что доверяют, а 53% скорее доверяют. Не доверяют Интернет-порталам 12% опрошенных. Тем не менее по достоверности в восприятии населения Интернет-сайты занимают вторую позицию (35%), после телевидения. Меньше всего доверия вызывают частные разговоры и социальные сети. Этим каналам в той или иной степени не доверяют 20% и 30% от общего числа опрошенных соответственно.

Для сравнительного анализа уровня доверия каналам распространения информации рассчитан Индекс доверия каналам. При расчете Индекса был использован подход расчета индексов для равнения групп [20]. Индекс по уровню доверия каждому каналу распространения информации рассчитан по следующей формуле:

$$I_c = (n_+ + 0{,}5 * n_{1\pm} - 0{,}5 * n_{2\pm} - n_-)/(n_+ + n_{1\pm} + n_{2\pm} + n_- + n_0), \quad (7)$$

где:

n_+ - число доверяющих,

$n_{1\pm}$ - число скорее доверяющих,

$n_{2\pm}$ - число скорее не доверяющих,

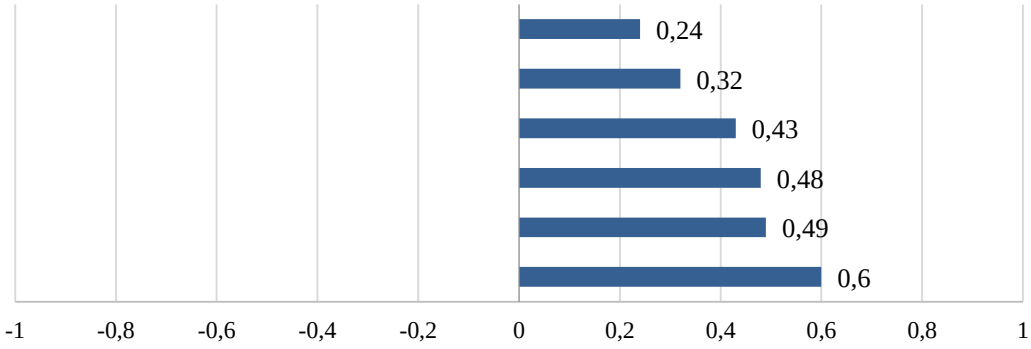
n_- - число не доверяющих,

n_0 - число затруднившихся ответить.

По результатам расчёта, индексы могут принимать значения от 1 до -1, где 1 – абсолютное доверие, а -1 – абсолютная недоверие.

Наибольший индекс доверия отмечается у телевидения, наименьшее доверие наблюдается к социальным сетям (рисунок 4).

Рисунок 4. Индексы доверия каналам распространения информации



	Индекс доверия телевидению	Индекс доверия интернет-сайтам	Индекс доверия печатным изданиям	Индекс доверия радио	Индекс доверия частным беседам	Индекс доверия социальным сетям
Индекс	0,6	0,49	0,48	0,43	0,32	0,24

Социально-демографический анализ утвердительных ответов («да, доверяю») на вопрос о доверии каналам распространения информации показал, что новые медиа в большей степени пользуются доверием в молодежных группах до 24 лет. Старшее поколение доверяет традиционным СМИ: телевидению, газетам и радио (таблица 9).

Таблица 9. Основные социально-демографические характеристики респондентов, выбравших вариант «да, доверяю» на вопрос о доверии каналам распространения информации

Характе-ристики	Социальные сети	Интернет-сайты	Газеты	ТВ	Радио	Разговоры
пол	все	все	все	женщины	все	все
возраст (лет)	15-19; 20-24	15-19; 20-24	65 лет и старше	55-64; 65 лет и старше	20-24; 65 лет и старше	15-19
этнос	казахи	казахи	казахи	Казахи; другие этносы	казахи	все
образование	Неполное среднее; среднее	все	все	все	все	Неполное среднее
достаток	Низкий; ниже среднего; высокий	Ниже среднего	Выше среднего	Ниже среднего; выше среднего	Выше среднего	Низкий
местность	село	село	село	село	село	село
регион	Костанайская, ЮКО, Г.Алматы Алматинская, Мангистауская Кызылординская	ЮКО, Костанайская, Акмолинская, Алматинская, ЗКО, Г.Астана	Атырауская, Костанайская, ЗКО, Акмолинская, ЮКО, Алматинская	ЮКО, Атырауская, Г.Алматы, Акмолинская, Костанайская,	Атырауская, Алматинская, Акмолинская, Костанайская	Акмолинская ЮКО, ВКО

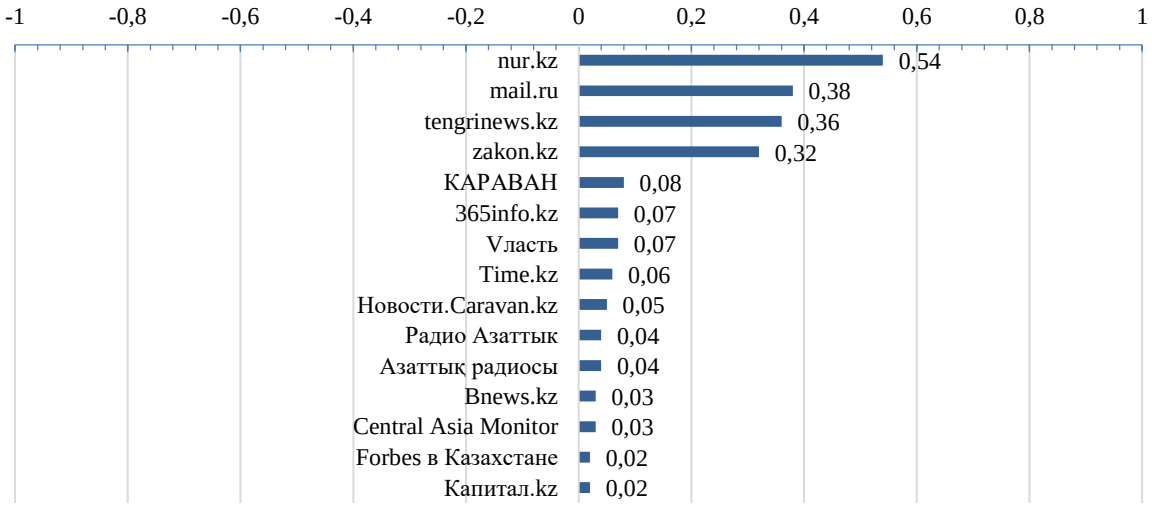
Важную роль в формировании доверительного отношения к информации играет отношение к источникам, которые ее распространяют. К основным факторам доверия респонденты отнесли:

– канал распространения информации (ТВ, газеты, Интернет-сайты);
– статус источника информации (республиканский/областной/городской);
– наличие в сообщении аргументированных доводов, экспертной оценки.

По результатам опроса установлено, что оперативность и доступность передаваемой информации не являются залогом доверия полученным сообщениям. Это подтверждает предположение о том, что доверие информации в первую очередь основывается на оценке ее достоверности, а не оперативности. 39% поверят сообщению от официального лица, находящегося на месте происшествия; каждый четвертый (29%) поверит информации, которую распространят посредством телевидения или радио, а не мессенджеров или социальных сетей.

Для сравнительного анализа уровня доверия каналам распространения информации введен индекс доверия информационным Интернет-сайтам. Наибольшим доверием среди населения пользуется новостной Интернет портал nur.kz (рисунок 5).

Рисунок5.Индексы доверия информационным Интернет-сайтам



Новостные Интернет-порталы пользуются большим доверием со стороны широкой аудитории, чем сайты, позиционирующие себя как информационно-аналитические. Очевидна разница в аудиториях этих сайтов. Новостным порталам, таким как mail.ru или nur.kz, чаще доверяют молодые, еще социально не адаптированные респонденты. Таким же сайтам, как Власть или Forbes, в Казахстане больше доверяют состоявшиеся респонденты, с образованием и достатком (таблица 10).

Таблица 10 - Основные социально-демографические характеристики респондентов, выбравших вариант «да, доверяю» на вопрос о доверии каналам распространения информации, % от общего числа опрошенных

Характе-ристики	**mail.ru**	**nur.kz**	**tengrinews.kz**	**zakon.kz**	**Власть**	**Forbes в Казахстане**
пол	все	все	все	все	мужчины	все
возраст (лет)	15-19; 20-24	15-19; 20-24	20-24, 25-34	20-24	35-44	20-24
этнос	Другие этносы	все	казахи	все	все	все
образование	Неполное среднее; среднее	Высшее, неоконченное высшее, среднее специальное	Высшее, неоконченное высшее	Высшее, неоконченное высшее	Высшее, неоконченное высшее	Высшее, неоконченное высшее
достаток	Ниже среднего; средний	Ниже среднего	Ниже среднего, выше среднего	Средний; Высокий	Высокий	Высокий
местность	село	село	все	все	село	село

регион	Костанайская, ЮКО, Акмолинская	ЮКО, Костанайская, Акмолинская, Алматинская, г.Астана	Мангистауская, Костанайская, Акмолинская, г.Астана Алматинская	Алматинская, г.Астана, г.Алматы, Акмолинская, Костанайская	Акмолинская, г.Алматы	Акмолинская, Мангистауская, Костанайская,

Интернет-сайты, как канал информации, пользуются однозначным доверием у 30% респондентов. И, тем не менее, каждый второй (53%) сказал, что доверяет Интернет-порталам с оговоркой «скорее», а 12% отметили, что в той или иной степени не доверяют.

Одной из причин невысокого доверия информационным порталам выступает отсутствие практик пользования этими сайтами, на основе которых мог бы сформироваться положительный или отрицательный опыт пользования новостными Интернет-сайтами. В этих условиях важным становится понимание того, каким противовесом обладают официально зарегистрированные СМИ по отношению к социальным сетям.

По результатам опроса, подавляющее большинство казахстанцев (76,3%) больше доверяет сообщениям, которые публикуются на сайтах, официально зарегистрированных СМИ, чем сообщениям, которые разные люди и блогеры пишут в социальных сетях (Фейсбук, Твиттер, ВКонтакте и т.п.).

По результатам опроса, в уровне доверия казахстанской журналистике прослеживается та же линия тренда, что и в отношении социальных сетей. Разница в уровне доверия журналистике лишь на 9 пунктов выше уровня доверия социальным сетям (суммарный показатель 52,2% против 43,3%).

По результатам опроса, одна из причин недоверия информации кроется в распространении ложных новостей, которые по оценкам респондентов являются частым явлением не только в социальных сетях, но и на сайтах, официально зарегистрированных СМИ. При этом доля тех, кто считает, что в социальных сетях материалы, основанные на слухах и лживой информации, публикуются в 1,5 больше, чем аналогичный показатель по зарегистрированным СМИ (38% и 21% соответственно).

По результатам опроса подтвердилась гипотеза, что большинство населения не может отличить качественно подготовленный журналистский материал от журналистского материала, основанного на слухах и лживой информации (59%). Более того, критериев, которые аудитория использует для распознавания фейковых новостей, крайне мало, и они не всегда состоятельны.

В вопросе необходимости госрегулирования публикуемой информации в социальных сетях в общественном мнении нет устойчивого мнения: 44% поддерживают, 25% выступают против, 31% затруднившихся, т.е. чаша весов может в любой момент качнуться в другую сторону.

Основными аргументами за необходимость введения госрегулирования со стороны участников опроса выступили следующие:

- гарантия получения правдивой, достоверной информации (31%);
- обеспечение функции контроля со стороны государства за распространением информации (27%).

А аргументами против стали: нарушение прав граждан на свободу слова (29%), нарушение демократических принципов, угроза диктатуры (23%) и сокращение доступа к информации (8%).

Надо отметить, что если госрегулирование вызывает неоднозначные оценки, то получить инструмент для оценки достоверности публикации хотело бы большинство опрошенных (67%). Идея о введении меры для оценки доверия журналистам и редакторам вызвала меньшую поддержку (63%).

С предложениями критериев достоверности выступили около 50% опрошенных, но их качество, как и процентная наполняемость являются незначительными. Наиболее часто упоминаемыми критериями стали такие, как наличие аргументации, фактажа в сообщении (13%), сравнение с информацией из других источников (7%), наличие реакции, резонанса в обществе (7%).

Объективность освещения событий казахстанскими новостными информационными источниками

В ходе опроса респондентам было предложено оценить, насколько точно казахстанские новостные Интернет-сайты освещают события, происходящие в стране. Результаты опроса показывают, что ни одно из сфер не получило оценку выше 4 баллов (по 5-балльной шкале). Общая оценка уровня освещения информации равна 3,36 балла. Заметно, что положительные оценки больше свойственны сельским жителям (3,46 балла); респондентам с невысокими доходами (3,45 балла) и казахам (3,4 балла).

Оценка качества освещения в СМИ важнейших направлений жизни общества показала, что выше всех оценивается деятельность по направлению «чрезвычайные ситуации» (3,56 балла). Самую низкую оценку из направлений получила тема трудовых отношений и занятости – 3,17 балла.

Согласно результатам исследования, большинству респондентов достаточно той информации о деятельности государственных органов, которую они получают из СМИ. На недостаточность такого рода информации пожаловался каждый четвертый опрошенный. 67% респондентов считают, что информация о деятельности государственных органов отображается достоверно, правдиво. Каждый четвертый респондент отметил, что информация о деятельности государственных органов, необъективная.

Результаты опроса выявили недостаточно высокий уровень доверия СМИ. Только 46% респондентов сообщили, что деятельность казахстанских Интернет-источников является прозрачной. Треть опрошенных не согласна с мнением, что казахстанские Интернет-источники ведут прозрачную деятельность.

Большинство респондентов (68%) считает, что казахстанские интернет-источники подают достоверную и правдивую информацию при освещении событий в стране. При этом 23% опрошенных выразили недоверие казахстанским Интернет-источникам.

Как показывает анализ полученных данных, к главным недостаткам отечественных Интернет СМИ относятся избыток информации по не интересующим темам (28,8%), сокрытие определенных фактов и событий (27,8%) и недостаток информации по интересующим темам (27,6%). Кроме этого, отсутствие всесторонней информации по тем или иным событиям волнует 26,1% респондентов. Каждый пятый респондент обратил внимание на то, что в новостях из Интернет-источников встречается предвзятое толкование фактов и недостоверная информация. Дополнительный анализ результатов показывает, что казахоязычная Интернет-аудитория страдает от недостатка информации по интересующим темам, отсутствия всесторонней информации по событиям в стране и недостатка оперативности получения информации из Интернет-источников.

В ходе опроса большинство (65%) респондентов частично или полностью было согласно с мнением, что, пользуясь информацией только из казахстанских Интернет-источников, можно составить объективное представление о событиях, происходящих в стране. Таким образом, можно говорит о подтверждении гипотезы о том, что население считает, что СМИ объективно отражают казахстанскую реальность.

результаты опроса частично подтвердили гипотезу о том, что большинство электронных СМИ зачастую жертвуют точностью, чтобы первыми осветить событие для привлечения аудитории. Почти половина участников опроса согласна с утверждением, что на практике у казахстанских интернет-СМИ оперативность подачи информации превалирует над ее точностью и достоверностью. Высокий уровень поддержки со стороны участников опроса получили суждения о том, что Интернет-СМИ информируют общественность, казахстанские новостные интернет-сайты озабочены привлечением большой аудитории.

Частичное подтверждение по результатам исследования получила гипотеза о том, что большинство казахстанцев склонно думать, что информационные Интернет-сайты используются в качестве инструмента освещения событий, сфер под углом, выгодным для достижения определенных целей. Доля тех, кто поддерживает данный тезис, варьируется от 40 до50%, что свидетельствует о преобладании мнения, но не о его доминировании. Чаще

респонденты соглашаются с тем, что Интернет-сайты используются, чтобы освещать события под углом, который выгоден определенному кругу лиц (53%), а также чтобы формировать положительный образ того или иного субъекта (55%). Чуть реже поддержку находят утверждения о том, что Интернет-сайты занимаются отвлечением аудитории от серьезных проблем (47%) или навязыванием определенного мнения читателям (45%). Еще реже, что с помощью информационных Интернет-сайтов распространяется информация, дискредитирующая какого-то субъекта (43%). Соглашательскую позицию чаще всего занимают респонденты в возрасте 20-24 лет и 25-34 лет, имеющие высшее образование. При этом почти значительная часть – от 25 до 35% – считает, что информационные Интернет-сайты не являются инструментами для манипуляционных действий с общественным мнением.

Влияние информационных Интернет-источников на социум

В рамках изучения рефлексивной коммуникации респондентов в электронных СМИ был выявлен ряд особенностей. Так, почти равное число опрошенных (48% и 50%) читают и не читают комментарии к публикациям. На интерес к комментариям оказывают влияние в основном две характеристики респондентов: возраст и образование. Так, в молодых группах количество читающих комментарии выше, чем в более старших, и, наоборот. Чем выше уровень образования у респондентов, тем чаще они читают комментарии.

Большинство опрошенных практически не пишет комментарии к текстам в Интернет-источниках (57% – никогда, 22% – изредка). Только пятая часть респондентов призналась, что осуществляет подобную активность (5% – часто, 15% – иногда).

Исследование не выявило доминирующей причины того, почему респонденты пишут комментарии к публикациям. При этом можно заметить, что негативные причины (несогласие с автором публикации и комментариями) более распространены в качестве причины комментария, чем положительные.

Большинство опрошенных (57%) наиболее привлекательным аспектом распространения информации назвало то, что в Интернете есть возможность не только ознакомиться с новостями, но и выяснить, как те или иные события воспринимаются другими пользователями. Несколько менее привлекательным, но достаточно весомым (44%) является такой аспект, как возможность посредством Интернета узнать информацию от непосредственных участников событий.

Способность Интернета помочь избежать ответственность и обеспечивать для пользователей анонимность и даже определенную безнаказанность является привлекательной стороной для 16% и 14% опрошенных соответственно. Чаще других

данные аспекты отмечали респонденты в возрасте 15-19 лет; мужчины; казахи; со средним и средним специальным образованием.

Большинство респондентов (60%) практически никогда и редко делятся новостями, мнениями людей или иными информационными сообщениями, пересылают их через социальные сети. Только 15% опрошенных можно отнести к активным Интернет-пользователям, которые делятся или пересылают разного рода информацию через социальные сети. Чаще это опрошенные в возрасте от 15 до 34 лет; имеющие неполное среднее образование (школьники); казахи; представители крайних по доходам, наименее и наиболее обеспеченных, групп.

Исследование не выявило самого главного мотива, из-за которого основная часть опрошенных совершила бы определенное ответное действие на публикацию в Интернете. При этом треть респондентов (37%) хотела бы поделиться информацией, которую посчитали важной и интересной для широкого круга читателей; четверть опрошенных (25%) были обеспокоены содержанием публикации и в ответ на это проявили активность; пятая часть (20%) была не согласна с точкой зрения автора публикации и поэтому поделилась информацией. Только 7% респондентов последовали бы призывам авторов публикаций.

Подтвердилась гипотеза о том, что, по мнению населения, СМИ влияют на их видение ситуации, событий, явлений. По итогам исследования, большинство опрошенных респондентов (58%) признали наличие влияния СМИ на их видение ситуации в стране. Вместе с тем, каждый третий (33%) отрицает такое влияние. Таким образом, можно выделить три условные группы респондентов: «поддающиеся влиянию СМИ», «неподдающиеся влиянию СМИ», «затрудняющиеся с оценкой влияния СМИ», различающиеся по своим социально-демографическим характеристикам.

60% опрошенных согласны с утверждением, что казахстанские Интернет-источники являются инструментом формирования общественного мнения. Каждый четвертый (27%) не согласен с данным утверждением. Вместе с тем, на число согласившихся с мнением о наличии влияния отечественных Интернет-источников, оказывают влияние в большей степени социальные характеристики респондентов (материальное положение, язык общения, уровень образования), а не демографические (возраст).

Вместе с тем характер и интенсивность данного влияния с точки зрения потребления Интернет-источников, как показал опрос, не являются очевидными. Так, только35% опрошенных часто и очень часто обсуждает в своем окружении увиденное/услышанное/прочитанное в Интернет-источниках. Другие 35% опрошенных делают это редко и очень редко. При этом 1/3 опрошенных затруднились с ответом.

Опрошенные казахстанцы не склонны по итогам публикаций в СМИ совершать какие-либо действия: 53% респондентов сказали, что очень редко и редко предпринимают определенные действия, еще 33% затруднились ответить на вопрос. Лишь 14% респондентов, отвечая на вопрос, сказали, что часто и очень часто совершают какие-либо действия по итогам публикаций в СМИ. При этом самой важной чертой этой группы является возраст: положительно ответили на этот вопрос респонденты группы 15-24 лет. В то же время место жительства, пол, этнический признак респондентов не оказывают существенного влияния на ответы.

Большинство опрошенного населения (52%) не смогли или не захотели ответить на поставленный вопрос о последствиях необъективного отражения информации в Интернет-источниках. В то же время полученные ответы от 48% участников исследования показывают, что чаще всего ими прогнозируются прямые (на уровне вызов-ответ) негативно-эмоциональные и протестные реакции населения (потеря доверия, митинги и акции протеста, хаос и паника, непредсказуемые и печальные, скандалы). В то же время рациональных, эмоционально неокрашенных ответов сравнительно мало.

Таким образом по результатам проведенного исследования поставленные цели и задачи были полностью достигнуты. В ходе социологического исследования были получены следующие результаты:

1. Определены основные каналы получения населением информации новостного характера.
2. Определен социально-демографический портрет среднестатистического читателя информационных порталов.
3. Определены поведенческие особенности населения при медиа-потреблении информации.
4. Получены результаты оценки мнения населения об уровне объективности и достоверности информации в освещении событий в стране и уровень доверия к казахстанским новостным информационным источникам.
5. Рассмотрены значимые аспекты влияния электронных СМИ на общественное мнение.
6. По данным социологического исследования построены индексы доверия каналам распространения информации, индекс доверия информационным Интернет-сайтам, индекс объективности информационных Интернет-источников и индекс прозрачности информационных Интернет-источников.

6.КОМПЛЕКСНОЕ ИССЛЕДОВАНИЕ, А ТАКЖЕ ЭКСПЕРТНЫЙ АНАЛИЗ НЕОБХОДИМОГО И ДОСТАТОЧНОГО СОСТАВА ФОРМИРУЕМОГО ТЕКСТОВОГО КОРПУСА, ЕГО СТРУКТУРЫ, ПРЕДСТАВИТЕЛЬНОСТИ, СБАЛАНСИРОВАННОСТИ, ТРЕБУЕМЫХ УРОВНЕЙ РАЗМЕТКИ И КЛАССИФИКАЦИЙ, ИХ ОДНОЗНАЧНОСТИ; СПОСОБОВ СБОРА, ХРАНЕНИЯ И ПРЕДСТАВЛЕНИЯ ТЕКСТОВОГО КОРПУСА; МЕТОДОВ ОРГАНИЗАЦИИ РАБОТЫ И ВЗАИМНОГО КОНТРОЛЯ ЗАДЕЙСТВОВАННЫХ ЭКСПЕРТОВ ПРИ ЕГО ФОРМИРОВАНИИ

Целью исследования является изучение международного опыта и методологии формирования текстовых корпусов для выработки рекомендаций по составлению представительного, сбалансированного, размеченного корпуса текстов. В ходе исследования был использован метод теоретического анализа и обобщения научной и аналитической литературы, экспертный и социологический опросы.

Научная новизна работы заключается в том, что впервые аспекты формирования корпусов и выработка рекомендаций изучается с точки зрения применимости для автоматизации оценки влияния казахстанских электронных СМИ на социум.

Объектом исследования выступают международный опыт и методологии формирования текстовых корпусов, мнение экспертов относительно исследуемой области.

Методология исследования включает теоретический анализ и обобщение научной и аналитической литературы, экспертный и социологический опросы. Результаты проведенного комплексного исследования представлены в данном отчете.

Были изучены и проанализированы ранее проводившиеся исследования, в ходе которых производилась компиляция и разметка текстовых корпусов. Был проведен социологический опрос населения Казахстана, в ходе которого были выявлены основные каналы получения информации новостного характера населением, определен социально-демографический портрет среднестатистического представителя аудитории информационных новостных порталов. В ходе данного социологического исследования был также обоснован выбор электронных источников и их контента в качестве основного объекта исследования всего проекта.

Кроме этого в целях определения необходимого и достаточного состава формируемого корпуса был проведен экспертный опрос, получены рекомендации ведущих казахстанских и российских экспертов и описаны их результаты.

Были определены классы, параметры и виды разметки корпуса, первичная (метаданные), экспертная и лингвистическая.

Исходя из поставленной цели основными задачами исследования являются:

1. изучение и анализ исследований, проведенных на основе текстовых корпусов;

2. определение основных каналов получения населением информации новостного характера и составление социально-демографического портрета среднестатистического читателя информационных порталов;
3. изучение мнений экспертов по определению необходимого и достаточного состава формируемого корпуса;
4. определение параметров разметки публикаций и их классификаций;
5. определение способа сбора данных для формирования корпуса.

Изучение и анализ исследований, проведенных на основе текстовых корпусов

Несмотря на существующее в настоящее время многообразие готовых многомилионных текстовых корпусов, не все они соответствуют требованиям настоящего исследования.

Структура текстов корпуса зависит от поставленных исследованием задач. Тексты в корпус выбираются в соответствии с проблемной областью, т.е. областью, интересующей исследователя. Одним из требований предъявляемым к составу и структуре корпуса является требование полноты. Публикации, отобранные в корпус для анализа, должны характеризоваться репрезентативностью, т.е. корпус должен отражать все свойства исследуемой области.

Корпус должен быть сбалансированным. При сбалансированности в корпус включаются тексты, представляющие разные модусы дискурса, разнообразные по жанрам, стилям, тематике, классам и т.п. При разработке корпуса устанавливаются пропорции, в которых должны быть представлены тексты разных жанров, стилей и т.п. Такие корпусы имеют фиксированный объем, и процедура пополнения таких корпусов происходит согласно строгой и четко выработанной процедуре отбора.

Объем корпуса на данном этапе исследований должен составлять не менее 2000 публикаций. Для решения различных лингвистических задач необходимо, чтобы тексты содержали дополнительную лингвистическую и металингвистическую информацию – разметку.

Опираясь на проведенный анализ текстовых корпусов и релевантных исследований в данной области, особенно в казахстанском сегменте, было принято решение о необходимости формирования собственного корпуса текстовых статей новостного характера. Сформированный тестовый корпус текстовой информации новостного характера является сбалансированной, репрезентативной структурой с присвоенной метаразметкой, а также с разметкой по 9 информативным признакам.

Определение источников для формирования корпуса

По результатам проведенного социологического исследования по оценке влияния открытых информационных источников (электронных СМИ) на социум в качестве источника для формирования корпуса публицистических тестов были определены информационные новостные Интернет-порталы.

Такой выбор обусловлен тем фактом, что электронные источники являются основными источниками получения информации новостного характера для основной части населения (59,8% респондентов). Дополнительным фактором, повлиявшим на такой выбор, является доверие, которым пользуются электронные ресурсы у населения. Так по результатам социологического опроса, было выявлено, что среди открытых текстовых информационных источников, Интернет-сайты занимают второе место по индексу доверия населения с незначительным отрывом от печатных изданий, уступив лишь телевидению, которое не является текстовым источником информации.

Для объективного отображения исследуемого влияния электронных СМИ на социум и учета интересов различных слоев социума в выборку корпуса включены наиболее популярные, средне и менее популярные среди населения Казахстана новостные Интернет-порталы, определенные по результатам сравнительного анализа существующих рейтингов новостных порталов и результатам социологического опроса.

По версии Интернет-статистики, приоритетные места рейтинга занимают Tengrinews.kz, Nur.kz и Zakon.kz. По результатам социологического опроса, наиболее посещаемые населением новостные порталы являются Nur.kz и Mail.ru. Далее следуют Tengrinews.kz и Zakon.kz. Популярность сайтов 365info.kz и Азаттық радиосы варьируется на уровне 5%. [5]

Результаты экспертного опроса для определения необходимого и достаточного состава формируемого корпуса

По результатам анализа различных мнений экспертов основными требованиями обеспечения представительности и сбалансированности корпуса публикаций были определены:

- наличие в корпусе публикаций, подготовленных различными изданиями по характеру политической направленности;
- включение в корпус информационных порталов, специализирующихся на предоставлении ежедневных новостей о событиях в стране;
- отбор в корпус публикаций СМИ, пользующихся разным уровнем популярности среди населения;

– исключение из выборки узкоспециализированных СМИ.

Для обеспечения сбалансированности формируемого корпуса были определены следующие параметры:

– публикации информационных порталов СМИ (наиболее популярные среди населения и менее популярные);
– определение анализируемого периода публикаций (с августа 2017 года по август 2018 год);
– наличие публикаций, которые соответствуют анализируемым классам;
– периодичность выборки каждые n дней через n-ный интервал для отслеживания дискурса (3 последовательных дня).

Для снижения уровня предвзятости в оценке публикаций к проведению разметки будут привлечены работники в сфере массовой коммуникации, практикующие политологи, филологи, юристы, редакторы, аналитики в различных сферах.

В целях формирования единого понимания, обеспечения единства мнения в оценке публикации и валидности результатов экспертами было рекомендовано разработать инструкцию.

Для распознавания информативных признаков в сформированном корпусе, по мнениям экспертов, необходимо провести глубинную разметку с формированием словарей по классам и параметрам, по которым необходима лингвистическая разметка.

Определение параметров разметки публикации и их классификация

При формировании корпуса публицистических текстов необходимо осуществить первоначальную разметку публикаций. Результаты изучения опыта в проведении разметок корпуса позволяют сформулировать следующие выводы.

При формировании корпуса осуществить первоначальную разметку публикаций по следующим метаданным:

– автор (журналист);
– источник;
– объем текста;
– количество просмотров;
– количество комментариев;
– количество лайков, голосов, реакций;
– количество «Share»;
– URL картинок, фото;
– URL видео;

– дата просмотра (Referral date).

В целях определения уровней разметки публикаций и их возможных классификаций необходимо, прежде всего, определить классы и параметры, по которым тексты будут размечены. Также важным аспектом является определение подходов к разметке.

Согласно первому подходу для разработки машинного обучения отнесения публикации к определенному классу и параметру необходима разметка, где эксперты относят текст к определенному классу и параметру. Далее компьютер обучается угадывать, какие тексты эксперты отнесли к определённому классу и параметру.

Второй подход – глубинная разметка публикаций для формализации лингвистических правил и формирования словарей по параметрам. Для этого необходимо определить перечень классов и параметров, по которым необходима лингвистическая разметка.

Для объективной разметки корпуса и минимизации субъективизма в оценках необходимо определить способы верификации данных при обобщении экспертных оценок.

Для получения качественных результатов разметки публикации необходимо привлечение квалифицированных экспертов. Привлекаемые к разметке должны отвечать ряду требований к их знаниям и квалификации:

– иметь соответствующее профильное образование;
– обладать аналитическим мышлением;
– иметь определенный уровень экспертизы и знаний в своей области специализации.

К экспертному анализу были привлечены практикующие журналисты, специалисты в области политологии, государственного управления, аналитики в различных сферах. Главными критериями при отборе экспертов выступили область профессиональной деятельности, специальность, профессиональные навыки и опыт работы.

Для эффективной организации экспертного анализа и получения объективных результатов необходима разработка экспертно-аналитической рамки анализа и инструкции.

По результатам консультаций с ведущими казахстанскими и российскими экспертами, а также из целей и задач исследования будет произведена разметка корпуса по 3 видам:

– первичная разметка, заключающаяся в сборе метаданных;
– экспертная разметка – разметка, выполняемая экспертами в соответствующих предметных областях для определения принадлежности публикации к выработанным классам и выявления соответствия критериям и параметрам разметки;

– лингвистическая разметка – глубинная разметка текста лингвистами, направленная на выявления лингвистических средств выражения и определения в тексте параметров разметки с формированием словарей и формализацией правил.

Способы сбора публикаций

Существуют различные способы формирования корпуса.

1. *Граббинг* – это сканирование каких-либо ресурсов для получения нужной информации. В применении к Интернету это чаще всего сканирование сайтов и скачивание с них информации.
2. *Парсер* – это программа для автоматизации процесса парсинга, то есть обработки информации по определенному алгоритму.

Иногда термины «граббер» и «парсер» употребляют как слова-синонимы. Однако основная цель применения парсера другая. Такой программный продукт сделан для того, чтобы расшифровывать добытый граббером код.

Применяя заданные параметры, парсер выполняет анализ полученных данных, а затем после обработки выдает результат.

Такие программы, как граббер и парсер используются для решения следующих задач:

– актуализация данных при размещении на своем сайте сервисов, предполагающих регулярные обновления (например, курсы валют или прогноз погоды);
– обновление страниц (Указанные программы являются быстрым способом получения свежего контента интересующей тематики, правда не очень честным);
– сбор графических изображений в существенных объемах без необходимости сохранять каждый отдельно взятый элемент вручную;
– сбор требуемых ссылок, имеющих нужные параметры (например, для конкретного домена и др.);
– объединение данных из многих источников на одной странице (Примером может служить новостной портал, на котором размещаются свежие материалы из разных источников со ссылкой на оригинал. Пользователи в данном случае получают возможность ознакомиться со всеми последними новостями в одном месте);

3. *«ручное» целевое формирование корпуса* (Формирование корпуса производится вручную с привлечением большого количества специалистов. Для формирования корпуса заранее определяется выборка и определенное количество статей);
4. *Приобретение готового корпуса* (Необходимо, чтобы корпус соответствовал целям и задачам исследования).

Для проведения работ по выработке механизмов по распознаванию информативных признаков в корпусе публицистических текстов и их апробации будет сформирован тренировочный корпус. Формирование корпуса для целей настоящего исследования будет осуществлён «ручным» способом. При этом необходимо определить период, за который будет осуществляться сбор данных таким образом, чтобы он позволял проследить тренды информационного пространства. Также необходимо тщательно подойти к выбору ресурсов, с которых предполагается производить сбор публикаций. Список ресурсов, вошедших в выборку корпуса, должен включать источники разнообразные по стилю, политической направленности, редакционной политике и по уровню популярности среди населения.

7. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО ФОРМИРОВАНИЮ СБАЛАНСИРОВАННОГО, ПРЕДСТАВИТЕЛЬНОГО, РАЗМЕЧЕННОГО И ПРОКЛАССИФИЦИРОВАННОГО КОРПУСА ПУБЛИЦИСТИЧЕСКИХ ТЕКСТОВ: ТЕКСТОВ ОФИЦИАЛЬНЫХ СМИ, ЖУРНАЛЬНЫХ ПУБЛИКАЦИЙ, НОВОСТНЫХ ПОРТАЛОВ И ДР.

На основе анализа рекомендаций экспертов и международного опыта формирования корпусов были сформулированы и описаны требования к корпусу применительно для целей проекта согласно специфике этапа его реализации. Основными требованиями к корпусу стали:

- соблюдение принципа представительности (репрезентативности);
- соблюдение принципа сбалансированности;
- обеспечение объема не менее 2000 публикаций;
- включение в выборку корпуса публикаций информационных порталов СМИ с разным уровнем популярности среди населения;
- определение анализируемого периода публикаций (с августа 2017 года по август 2018 год);
- наличие публикаций, которые соответствуют анализируемым классам;
- периодичность выборки каждые n дней через n-ный интервал для отслеживания дискурса (3 последовательных дня) и др.

Были определены основные параметры и классы разметки корпуса в рамках всей методики: 6 классов, тематическая кластеризация по 18 сферам, 22 информативных признака, 12 параметров лингвистической разметки.

Была проведена работа по сбору данных, в результате которой в выборку были включены собранные вручную 5210 публикаций из 5 источников (tengrinews.kz, zakon.kz, azattyq.org, camonitor.kz, kz.expert) за однолетний период с 1 августа 2017 по 31 июля 2018 по определенным датам (3 последовательных дня с *n*-ным шагом) с первичным сбором мета-данных (первичная разметка). Также был разработан методологический подход и проведены мероприятия по организации работы с экспертами и лингвистами, привлеченными к разметке, позволяющие обеспечить надежность полученных данных (триангуляция данных, разработка инструкций и разъяснительные собрания экспертов). Были определены параметры разметки в рамках методики в целом и параметры, соответствующие данному этапу исследования.

Была проведена первичная разметка 5210 публикаций по мета-данным (17 категорий данных), результаты которой представлены на Диске 1. Важно отметить, что архитектура

сайтов, отобранных в выборку, не всегда предусматривает возможность сбора и наличие всех вышеуказанных параметров мета-разметки. В случае, где сбор данных не возможен по объективным причинам (например, на сайте нет возможности комментировать публикацию или нет данных о количестве просмотров), были собраны только те данные, которые находятся в открытом доступе. Соответственно, данные мета-разметки неоднородны, и часть необходимых данных отсутствует. Так полные данные на этапе мета-разметки по всем 5 источникам удалось собрать по параметру «Объем текста» и «Имя автора публикации».

Была проведена экспертная разметка по 5210 публикациям (9 категорий: достоверность, объективность, социальная значимость, потенциальная резонансность, тональность, наличие манипулятивных техник, политизированность, казахстанское содержание, принадлежность к сферам по тематике). Для верификации данных и повышения надежности ее результатов методом триангуляции была проведена оценка сходимости между кодировщиками. Для данной оценки использован онлайн-калькулятор https://nlp-ml.io/jg/software/ira/. Для оценки согласия кодировки между экспертами был использован фактический процент согласия A_{obs} (сходимость кодировщиков) по каждой статье. В целом кодировка характеризуется довольно высоким результатом согласия по всем переменным. Промежуточные результаты экспертной разметки по 9 параметрам представлены на Диске 2. В настоящий момент ведется работа по окончательному определению 9 параметров экспертной оценки с учетом триангуляции данных.

Также была проведена глубинная лингвистическая разметка с выделением фрагментов текста, формированием словарей и формализацией правил определения параметров в тексте по 2000 публикациям (3 категории: тональность, политизированность, наличие манипуляторных техник). Результаты анализа корпуса текстов по критериям *«Тональность», «Политизированность»* и *«Наличие манипулятивных техник»* описаны в разделе 9 и представлены на Диске 3.

8. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО АНАЛИЗУ ФАКТИЧЕСКИХ ХАРАКТЕРИСТИК СФОРМИРОВАННОГО КОРПУСА (СОСТАВ, СТРУКТУРА, ПРЕДСТАВИТЕЛЬНОСТЬ, СБАЛАНСИРОВАННОСТЬ, УРОВНИ РАЗМЕТКИ И КЛАССИФИКАЦИИ, ИХ КОРРЕКТНОСТЬ И ПОЛНОТА И ДР.)

В ходе проведенного комплекса работ по формированию сбалансированного, представительного, размеченного и проклассифицированного корпуса публицистических текстов электронных новостных порталов был сформирован корпус, включающий 5210 публикаций из 5 новостных порталов: tenginews.kz (36,4%), zakon.kz (38,4%), azattyq.org (7,3%), camonitor.kz (11,9%), kz.expert (6,2%). В этой связи целью данной работы является изучение и анализ фактических характеристик сформированного корпуса публицистических текстов для автоматизированной оценки влияния отрытых электронных информационных источников на социум.

Все требования и рекомендации, выдвигаемые к корпусу для обеспечения сбалансированности и репрезентативности, были соблюдены. Была проведена мета-разметка (5210 публикаций), экспертная разметка (5210 публикаций) и лингвистическая разметка (2000 публикаций) корпуса. Мета-разметка корпуса проводилась по 17 параметрам.

Анализ фактических характеристик корпуса показал, что данные мета-разметки неоднородны, и часть необходимых данных отсутствует. Так полные данные на этапе мета-разметки по всем 5 источникам удалось собрать по параметру «Объем текста» и «Имя автора публикации». Это обусловлено тем, что архитектура сайтов, отобранных в выборку, не всегда предусматривает возможность сбора и наличие всех отобранных параметров мета-разметки (например, на сайте нет возможности комментировать публикацию или нет данных о количестве просмотров).

Объем текстов, включенных в корпус из разных источников, значительно отличается по количеству знаков, что объясняется редакторской политикой ресурса. Среднее количество знаков в одной публикации на tengrinews.kz и zakon.kz, ресурсах, специализирующихся в основном на предоставлении ежедневных новостных сводок и кратких информационных сообщений, в среднем в 2,5 раза меньше, чем в публикациях camonitor.kz, контент которого включает и информационные сообщения, и аналитику; в среднем в 3-4 раза меньше, чем в публикациях сайтов azattyq.org и kz.expert, контент которых составляет в основном аналитические и художественно-публицистические статьи.

Поскольку процесс экспертной разметки на момент предоставления настоящего отчета находится на этапе обработки результатов и дальнейшее окончательное определение

параметров разметки с учетом специфики процедуры триангуляции данных находится на стадии завершения, в работе приведены промежуточные результаты экспертной оценки по первой группе экспертов.

9. РАЗРАБОТКА МЕТОДОВ ФОРМАЛИЗАЦИИ ПРАВИЛ И СЛОВАРЕЙ ДЛЯ РАСПОЗНАВАНИЯ В РАЗМЕЧЕННОМ КОРПУСЕ ИНФОРМАТИВНЫХ ПРИЗНАКОВ И ОПРЕДЕЛЕНИЯ СТЕПЕНИ ИХ ВЫРАЖЕННОСТИ

Успешное решение задач по автоматизации оценки лингвистического корпуса из СМИ требует системной работы по созданию словарного массива, отбору языковых единиц и способов (правил) их употребления, в совокупности способных обучить машину распознавать тексты предвзято-негативного/позитивного и подозрительного характера.

В связи с этим *целью данной работы* является формирование словарей и разработка методов формализации правил для распознавания определенных информативных признаков в сформированном корпусе текстов.

Объектом анализа стали 2 000 текстов из 5 информационных новостных Интернет источников: «KazakhSTAN 2.0», «Central Asia Monitor», «Zakon.kz», «Радио Азаттык», «Tengrinews.kz».

Для достижения цели были поставлены и решены следующие задачи:

1. установлен перечень релевантных информативных признаков, позволяющих сформировать корпус слов и словосочетаний, необходимый для распознавания и оценки текста;
2. разработана инструкция проведения глубинного анализа текстовой информации в СМИ в контексте цели исследования;
3. проведен глубинный анализ сформированного корпуса публикаций СМИ;
4. размечен текстовый корпус по обозначенным информативным признакам;
5. сформирован словарь на базе анализируемых текстов;
6. формализованы правила для распознавания определенных информативных признаков в корпусе текстов.

В процессе исследования и работы с текстами использовались общенаучные методы наблюдения, систематизации, обобщения, статистический метод, лингвистические методы контекстуального, дискурсивного, интерпретативного, прагматического и компонентного анализа, интроспекция.

В процессе работы по формированию корпуса и экспертной консультации были определены информативные признаки для формирования словарей и формализации правил.

Формирование словарей и формализация правил для распознавания признаков в заданном лингвистическом корпусе осуществлялось по следующим параметрам:

1. Тональность: позитивный, нейтральный, негативный.

Основной практической задачей являлось определение позитивной, негативной, нейтральной оценки слов или словосочетаний с включением их в словарь.

Включение данного признака объясняется его универсальностью для любого текста (т.е. сопутствию каждому тексту) и отражает знак оценки (хорошо, плохо, нормально) автора текста к описываемым событиям.

2. Политизированность.

Анализ данного признака предполагает определение экспертами перечня слов или словосочетаний, придающих тексту политизированную окраску.

Под политизированностью текста мы понимаем придание политической окраски событиям, напрямую не относящихся к государственному строю, системе государственного управления, властным структурам. Политизированные тексты требуют пристального внимания и относятся к той группе текстов, распознавание которых востребовано в первую очередь, поэтому анализ данного признака включен в первый этап исследования.

3. Предвзятость.

Определение предвзятости мнения автора обозначается через перечень соответствующих слов или словосочетаний.

Под предвзятостью подразумевается предубежденное необъективное мнение, которое складывается заранее. Следует отметить, что включение этого признака в перечень анализируемых вызвано необходимостью обучить машину своевременно распознать и нейтрализовать вред предвзятых публикаций, оказываемый развитию общества.

4. Социальная значимость тематики.

Данный признак раскрывается через определение перечня социально значимых слов или словосочетаний.

С целью расширения словаря обозначения социально значимых объектов также предполагается составление синонимических рядов. Релевантность и необходимость данного признака для исследования объясняется свойством социально значимых текстов вызывать широкий общественный резонанс, затрагивать интересы общества и/или вызвать дестабилизацию общества.

5. Казахстанское содержание.

Включение данного признака в круг исследуемых обусловлено необходимостью обучить машины определять принадлежность содержания текста к казахстанским реалиям и отделять такие материалы от информационного корпуса других государств.

6. Определение наличия манипулятивных техник.

Анализ данного признака предполагает определение наличия манипулятивных техник и их характеристику. Также формируется перечень слов или словосочетаний, которые используются в качестве манипуляций (подмена понятий, повторение некоторых слов, призыв к действию и т.д.).

7. Формирование онтологий источников, на данные которых ссылаются в тексте.

Наличие или отсутствие ссылки на источник, вид источника является отчетливым сигналом, указывающим на достоверность или недостоверность излагаемых в публикации фактов.

8. Формирование онтологий, ключевых слов или словосочетаний, позволяющих определить принадлежность текста к сфере.

Анализ и результаты данного признака позволяют ускорить обучение машины по определению принадлежности текстов к определенной сфере.

9. Определение наличия проверяемых фактов в статье: субъект, предмет обсуждения, место, действие, время.

Сочетания и закономерности в связке субъект-предмет-место-действие. Анализ данного признака предполагает составление словаря ключевой тематики обсуждения в тексте, который будет способствовать наиболее эффективному распознаванию машиной темы и предмета, обозначенных в информационном материале.

10. Соответствие заголовка публикации содержанию медиа-текста.

Признак соответствие/несоответствие заголовка содержанию часто сопутствует политизированным и предвзятым текстам, в связи с чем анализ текста по этому параметру является актуальным и высоко информативным.

11. Определение ключевых слов текста.

Анализ данного признака позволяет выявить основную мысль текста.

12. Призыв к действию.

Данный признак также относится к числу наиболее актуальных, требующих оперативного распознавания машиной параметров, поскольку в контексте негативной тональности обладает потенциальной угрозой общественному порядку и безопасной жизнедеятельности человека.

На данном этапе глубинный анализ проведен по 3 информативным признакам: тональность, политизированность и наличие манипулятивных техник.

Разработка экспертно-аналитической рамки (инструкции)

С целью проведения объективной разметки, формирования словарей и формализации правил была разработана инструкция. Инструкция включает в себя цель,

задачи проекта, перечень информативных признаков, являющихся предметом исследования, подробный алгоритм обработки текстового материала. Экспертно-аналитическая рамка содержит характеристику 12 признаков и схему их анализа.

Для обеспечения понимания экспертами лингвистической разметки и исключения расхождений в толковании понятийного аппарата в инструкции приведена детализированная информация по каждому из классов и параметров разметки. По каждому из классов приведено определение, раскрывающее суть классификации.

По каждому признаку дано указание зафиксировать номер публикации и скопировать его с заголовком, произвести выделение слов или фрагментов (словосочетаний, устойчивых выражений, слов-предложений и т.д.) текста определенным цветом. В инструкции представлены различные таблицы (в зависимости от содержания признака) по составлению словаря. В целях формализации правил в инструкции сформулировано соответствующее задание.

Для обеспечения качества лингвистической разметки к глубинному анализу были привлечены эксперты-лингвисты, имеющие значительный опыт в реализации научных исследований в сфере современной лингвистики (дискурсивного (текстового) анализа, прагма-, этно и психолингвистических исследований и т.д.).

Результаты глубинного анализа текстов для формализации правил по тональности

В результате анализа корпуса текстов по критерию «Тональность» был сформирован лингвистический словарь. Было выявлено 15 076 словоупотреблений (слов и словосочетаний) с негативной и позитивной семантикой, из них 9 791 однословное словоупотребление и 5 285 словосочетаний. На базе данных словоупотреблений был составлен словарь маркированной лексики, содержащий 2 892 лексемы (1 928 - с негативной семантикой, 964 - с позитивной семантикой).

В процессе анализа прежде всего учитывалось отношение автора к событиям.

Статистический подсчет по видам тональности в 2000 предложенных текстах показал явное преобладание текстов нейтрального характера – 1 559. Было выявлено 352 текста с негативной тональностью. Наименьшее количество текстов (89) определено как тексты с позитивной тональностью. Данная статистика обусловлена общей тенденцией авторов публикаций стремиться объективно и бесстрастно излагать информацию, что и отразилось в преобладании текстов с нейтральной тональностью.

В ходе анализа текстов по признаку «Тональность» были выявлены закономерности и произведена формализация правил.

Негативная тональность текста формируется в результате:

– использования в тексте слов с ярко выраженной негативной тональностью;
– частотного использования частицы *не* в качестве основного отрицания того, что обозначается словом с нейтральной или позитивной семантикой, тональностью, к которому она относится по смыслу (***не*** *находят своего решения,* ***не*** *может получить вестей*);
– частотного использования слов с приставкой *не-* со значением полной противоположности того, что выражает слово с нейтральной или позитивной семантикой без этой приставки (***не****удовлетворительное состояние национальной экономики*);
– использования глагольного слова *нет* со значением отрицания того, что названо словом с нейтральной или позитивной семантикой (***нет*** *доступа к политической арене*);
– сознательного использования в пределах речевого целого единиц разных стилей;
– использование просторечных и разговорных слов для наименования, характеристики объекта речи;
– использования слов в переносном значении, имеющем помету «разговорное», в одной языковой конструкции с объектом речи «первые лица государства», «президент», «правительство»;
– использования кавычек в словах с нейтральной или позитивной тональностью, употребленных в прямом значении, с целью создания значения с противоположными признаками;
– использования кавычек в словах с нейтральной или позитивной тональностью, употребленных в прямом значении, называющих и характеризующих президента;
– использования антонимических пар в одной языковой конструкции с объектом речи «президент», «правительство»;
– использования стилистической фигуры лексического повтора для придания негативной экспрессивности;
– использования ложно-риторических вопросов с эмоциональными оттенками сомнения, несогласия;
– употребления частицы «ни» и повторяющегося союза «ни… ни» перед глаголами с отрицанием (включающими частицу «не») для усиления / акцентирования отрицания;

– использования сегментированных (в частности, парцеллированных) конструкций.

Позитивная тональность текста формируется в результате:

– использования слов и словосочетаний с позитивной тональностью;
– отсутствия в тексте слов с ярко выраженной негативной тональностью;
– использования частицы «не» при словах с негативной тональностью;

Нейтральная тональность текста формируется в результате:

– отсутствия в тексте слов с ярко выраженной негативной и позитивной тональностью;
– сочетания в пределах одной речевой конструкции слов с позитивной и негативной семантикой.

Результаты глубинного анализа текстов для формализации правил по политизированности

Прежде всего, необходимо отметить, что в представленном для анализа лингвистическом корпусе тексты, обладающие признаком «Политизированность», встретились 12 раз, таким образом, они составляют менее 1% из всего проанализированного материала (2 000 текстов). Однако для дальнейших этапов исследования этот результат является весьма полезным, поскольку будет способствовать пересмотру подходов по формированию экспериментального корпуса текстов.

В результате анализа корпуса текстов по признаку «Политизированность» сформирован лингвистический словарь, в который вошли: 156 словоупотреблений (84 словоупотребления слов и 72 словосочетания). На базе данных словоупотреблений был составлен словарь маркированной лексики, содержащий 37 лексем.

Формализованы правила:

– наличие в основном тексте статьи политических тем и микротем, не полностью относящихся к заявленной в названии статьи теме;
– неполное совпадение ключевых слов темы, заявленной в названии статьи (в лиде), и ключевых слов основного текста статьи;
– использование слов и словосочетаний с негативной тональностью в речевых конструкциях, относящихся к политическим темам.

Результаты глубинного анализа текстов для формализации правил по манипулятивным признакам

Результаты глубинного анализа корпуса позволили определить 283 текста (что составляет менее 15% от всего проанализированного корпуса), содержащих манипулятивные техники.

В результате анализа корпуса текстов по признаку «Наличие манипулятивных техник» сформирован лингвистический словарь, в который вошли: 3 797 словоупотреблений (1 995 словоупотреблений слов и 1 802 словосочетания). На базе данных словоупотреблений был составлен словарь маркированной лексики, содержащий 430 лексем.

Правила формализации для манипулятивных техник:

- использование слов и словосочетаний с негативной семантикой;
- повтор языковых единиц с негативной оценкой;
- создание иронии путем использования слов в переносном значении, имеющем помету «разговорное», в одной языковой конструкции с объектом речи «первые лица государства», «президент», «правительство»; путем использования и обыгрывания языковых штампов;
- создание противопоставлений с использованием антонимических пар в одной языковой конструкции с объектом речи «президент», «правительство»;
- частотное использование частицы не в качестве основного отрицания того, что обозначается словом с нейтральной или позитивной семантикой, тональностью, к которому она относится по смыслу;
- использование кавычек в словах с позитивной семантикой, употребленных в прямом значении, называющих и характеризующих президента, правительство.

По итогам работы экспертов-лингвистов по глубинному анализу и детальной разметке медиа-текстов получены следующие результаты:

1. Сформирован словарь из 2 892 слов (15 076 словоупотреблений) по информативному признаку тональности.
2. Сформирован словарь из 37 слов (156 словоупотреблений) по информативному признаку политизированности.
3. Сформирован словарь из 430 слов (3797 словоупотреблений) по информативному признаку наличия манипулятивных техник.
4. Сформулированы правила формализации по определению информативных признаков тональности, политизированности, наличия манипулятивных техник.

Следующим и заключительным этапом работы является глубинный анализ 9 признаков, заявленных в плане исследования.

Как показали результаты работы первого этапа исследования, в рамках следующего этапа должно уделяться большое значение созданию формализованных правил для употребления каждой отдельной словоформы, имеющей негативную или позитивную тональность.

Такая работа предполагает анализ не отдельных текстов (что само по себе является также весьма важным), а отдельных лексем в общем корпусе текстов (2 000), а также, возможно, искусственное моделирование конструкций, построенных по языковым законам на базе существующих в языке моделей. Например, анализ слова *развитие* в сочетании с другими словами, анализ частиц *не* в сочетании с глаголами с позитивной и негативной тональностью, с прилагательными, местоимениями, существительными, сочетание *не* с другими частями речи, повтор и т.д. Если эксперт в процессе анализа частицы *не* на базе всего корпуса текстов (2000) не обнаружил некоторых конструкций либо значений, он вправе обратиться к существующим грамматикам, словарям, в которых возможны перечисления не выявленных в анализируемых текстах конструкций.

10. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО АНАЛИЗУ И ВЫЯВЛЕНИЮ ЗАКОНОМЕРНОСТЕЙ, ПРИСУТСТВУЮЩИХ В СФОРМИРОВАННОМ КОРПУСЕ, С ЦЕЛЬЮ ФОРМАЛИЗАЦИИ ПРАВИЛ И ФОРМИРОВАНИЯ СЛОВАРЕЙ ДЛЯ РАСПОЗНАВАНИЯ В РАЗМЕЧЕННОМ КОРПУСЕ ИНФОРМАТИВНЫХ ПРИЗНАКОВ И ОПРЕДЕЛЕНИЯ СТЕПЕНИ ИХ ВЫРАЖЕННОСТИ НА ОСНОВЕ РАЗРАБОТАННЫХ МЕТОДОВ

Проведен комплекс подготовительных работ для определения и выявления закономерностей в сформированном корпусе публицистических текстов с целью формализации правил.

Комплекс подготовительных работ включал следующие этапы:

1. Определение необходимого и достаточного состава формируемого корпуса текстов.
2. Определение параметров и уровней разметки публикаций и их классификация.
3. Формирование сбалансированного и представительного корпуса публицистических текстов.
4. Проведение разметки корпуса текстов.

Разметка корпуса публицистических текстов проводится с применением двух подходов. Первый подход – отнесение публикации к определенному классу и параметру. Второй – глубинная разметка публикаций для формализации лингвистических правил и формирования словарей по параметрам.

К экспертному анализу корпуса с применением первого подхода были привлечены практикующие журналисты, специалисты в области политологии, государственного управления, аналитики в различных сферах. Разметка с применением первого подхода осуществлена по 9 параметрам:

1. Достоверный, недостоверный, трудно определяется;
2. Объективный, предвзятый, трудно определяется;
3. Социально значимый, социально не значимый, трудно определяется;
4. Потенциально резонансный, потенциально не резонансный, трудно определяется;
5. Тональность: позитивный, нейтральный, негативный (необходимо оценить весь текст и каждый абзац).
6. Наличие манипуляторных техник, отсутствие манипуляторных техник, трудно определить
7. Политизированный, не политизированный, трудно определяется;
8. Казахстанское содержание, не казахстанское содержание;

9. Принадлежность к сферам.

Результаты разметки по первому подходу описаны в разделе 9.2 и 9.3.

Глубинная разметка публикаций для формализации лингвистических правил и формирования словарей проведена по 12 параметрам:

1. Тональность: позитивный, нейтральный, негативный.
2. Политизированность.
3. Предвзятость.
4. Социальная значимость тематики.
5. Казахстанское содержание.
6. Определение наличия манипулятивных техник.
7. Формирование онтологий источников, на данные которых ссылаются в тексте.
8. Формирование онтологий, ключевых слов или словосочетаний, позволяющих определить принадлежность текста к сфере.
9. Определение наличия проверяемых фактов в статье: субъект, предмет обсуждения, место, действие, время.
10. Соответствие заголовка публикации содержанию медиа-текста.
11. Определение ключевых слов текста.
12. Призыв к действию.

На данном этапе глубинный анализ проведен по 3 информативным признакам: тональность, политизированность и наличие манипулятивных техник.

По результатам анализа корпуса по тональности публикации установлены следующие закономерности.

Негативная тональность. Сознательное использование в пределах речевого целого единиц разных стилей. Использование просторечных и разговорных слов для наименования, характеристики объекта речи. Использования ложно-риторических вопросов с эмоциональными оттенками сомнения, несогласия

По параметру «наличие манипулятивных техник» установлены следующие закономерности.

Среди видов манипуляций наиболее частотным является:

– попытка вызвать у читателей негативные эмоции и оценки;
– убеждение (аргументация);
– дискредитация.

Среди используемых методов воздействия наиболее частотными являются:

– метод упрощения (позволяет внушить главную мысль аудитории в краткой и лаконичной форме утверждения);

– метод повторения (придает фактам дополнительный вес, призывает аудиторию акцентировать внимание на наиболее важных для инициатора моментах).

На основе полученных результатов глубинного анализа текстов по 12 информативным признакам будут разработаны методы формализации правил. Результаты разметки по первому подходу описаны в разделе 9.

Таким образом, проведен комплекс подготовительных работ для выявления закономерностей, присутствующих в сформированном корпусе. Полученные результаты выявленных закономерностей позволят более точно формализовать правила по каждому из выбранных информативных признаков. Также полученные результаты позволят провести экспериментальную работу по определению корректности разработанных методов формализации правил.

11. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО ЭКСПЕРИМЕНТАЛЬНОМУ ИССЛЕДОВАНИЮ ПОЛНОТЫ И КОРРЕКТНОСТИ СФОРМИРОВАННОГО МНОЖЕСТВА ПРАВИЛ И СЛОВАРЕЙ ДЛЯ РАСПОЗНАВАНИЯ В РАЗМЕЧЕННОМ КОРПУСЕ ИНФОРМАТИВНЫХ ПРИЗНАКОВ И ОПРЕДЕЛЕНИЯ СТЕПЕНИ ИХ ВЫРАЖЕННОСТИ

Проведен комплекс подготовительных работ по проведению экспериментальных работ по исследованию полноты и корректности сформированного множества правил и словарей.

Для проведения экспериментального исследования полноты и корректности сформированного множества правил и словарей проводится глубинный анализ 2000 текстов и формирование словарей по 12 параметрам (см. раздел 9). На данном этапе глубинный анализ проведен по 3 информативным признакам: тональность, политизированность и наличие манипулятивных техник. На основе полученных результатов глубинного анализа текстов по 12 информативным признакам будут разработаны методы формализации правил.

Следующим этапом работ будет проведение экспериментального исследования полноты и корректности сформированного множества правил и словарей для распознавания в размеченном корпусе информативных признаков с применением информационных технологий.

Экспериментальное исследование планируется провести в несколько этапов.

На первом этапе будет создана экспериментальная группа экспертов. Экспертам будут предоставлены тексты, по которым были формализованы правила и сформированы словари по обозначенным признакам. Эксперты должны будут оценить тексты и отнести к определенному классу или параметру, применяя предоставленные правила и словари.

На втором этапе экспертам будут предоставлены новые тексты. Применяя предоставленные правила и словари, эксперты должны оценить тексты и отнести к определенному классу или параметру. Далее проводится анализ сходимости в результатах.

На третьем этапе формализованные правила и сформированные словари экспериментально подтверждаются с применением информационных систем. Далее проводится анализ результатов оценки компьютером и экспертами.

Таким образом, полученные результаты экспериментального подтверждения позволят определить корректность сформированного множества правил и словарей.

12. РАЗРАБОТКА МЕТОДИКИ ОПРЕДЕЛЕНИЯ СТЕПЕНЕЙ ВЕРОЯТНОЙ ДОСТОВЕРНОСТИ ОТДЕЛЬНОЙ ПУБЛИКАЦИИ, ВЕРОЯТНОЙ ДОСТОВЕРНОСТИ ПУБЛИКАЦИЙ ОПРЕДЕЛЁННОГО АВТОРА, ВЕРОЯТНОЙ ДОСТОВЕРНОСТИ ПУБЛИКАЦИЙ ОПРЕДЕЛЁННОГО ПЕЧАТНОГО СМИ И КРИТЕРИИ ИХ ВЕРИФИКАЦИИ

Актуальность разработки методики

Вопрос достоверности публикаций находится в фокусе внимания многих исследователей. Так, по данным исследования Массачусетского технологического института (США), было установлено, что 1% ложных новостей, получивших наибольшее распространение в Twitter, распространяются среди аудитории от 1000 до 100 000 людей, тогда как масштаб распространения достоверных сообщений не превышают 1000 человек. Ложные данные распространяются значительно дальше, быстрее, глубже и шире, чем правда во всех категориях информации, а эффекты ложных политических новостей являются более ярко выраженными, чем эффект недостоверных новостей на тему терроризма, природных катаклизмов, науки, городских легенд или финансовой информации.

В законодательствах различных стран одним из основных принципов средств массовой информации является достоверность. Законодательство Казахстана в области СМИ обязывает журналиста проверять достоверность получаемой информации, обращаясь при этом в компетентные органы или специалистам.

Данные анализа различных источников выявления и проверки достоверности информации показывают, что подходы и критерии к определению достоверности публикаций едины. Вместе с тем большинство методик определения достоверности сведений, информации, данных, а также публикации основаны на экспертных оценках. Проверка достоверности текстовой информации проводится в комбинации с применением информационных систем и экспертных мнений.

В этой связи, учитывая озабоченность мирового сообщества к производству и распространению недостоверной «фейковой» информации с использованием информационных технологий, разработка методики определения степени вероятной достоверности публикации является актуальной.

Общие положения разрабатываемой методики

Цель оценки - определение степени вероятной достоверности публикаций в рамках оценки уровня влияния открытых информационных источников на социум на основе измеряемых показателей с применением автоматизированной информационной системы.

Задачи оценки:

- определение показателей определения степени вероятной достоверности публикаций;
- подтверждение степени значимости критерия достоверность в рамках оценки влияния СМИ (открытых информационных источников на социум);
- формирование и разметка корпуса текстов на основе источников информации для определения степени вероятной достоверности публикаций.

Объект оценки: публикации казахстанских информационных новостных порталов.

Предмет оценки: достоверность публикаций информационных новостных Интернет-порталов.

Основные этапы разработки методики:

1. Сбор и обобщение данных на основе изучения международного опыта исследований в данной области.
2. Выработка подходов к определению показателей к информативному признаку «достоверность».
3. Проведение социологического опроса для сбора данных для определения показателей достоверности.
4. Комплексная оценка показателя и формирование индекса достоверности.
5. Формирование корпуса публикаций для апробации показателей информативного признака «достоверность»:
   - выбор источников медиа-текстов;
   - экспертный опрос для определения необходимого и достаточного состава формируемого корпуса;
   - формирование сбалансированного, репрезентативного корпуса публикаций;
   - экспертная разметка корпуса по признаку достоверность.
6. Выявление закономерностей и правил по информативному признаку «достоверность» на основе экспертной оценки.
7. Выгрузка размеченного корпуса текстов для машинного обучения.

Определение показателей оценки достоверности публикации

Результаты исследования существующих информативных признаков влияния открытых текстовых информационных источников на социум позволили определить существующие информативные признаки определения достоверности публикации новостного характера (см. раздел 2). Выработанный на основе информативных признаков достоверности публикации критерий «достоверность» позволяет оценить достоверность публикуемой информации на основе наличия ссылок в публикациях на источник,

авторитетности издания, на котором размещены публикации, освещение одного и того же события различными изданиями, проверки аутентичности фото и видео, которые подтверждают событие, соответствие заголовка публикации содержанию медиа-текста.

Выявление закономерностей и правил по информативному признаку «достоверность» на основе экспертной оценки.

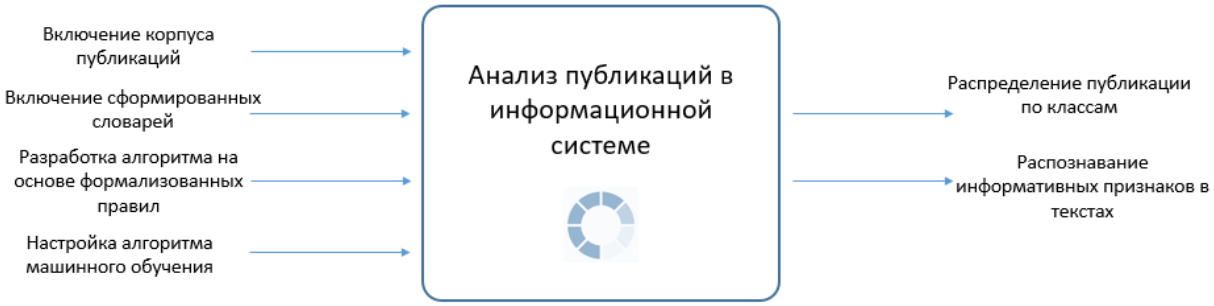
Для распознавания информативных признаков достоверности публикации и выработки методики их расчета сформирован сбалансированный, представительный корпус публицистических текстов.

Проведена разметка публикации по достоверности с привлечением практикующих журналистов, политологов, аналитиков и других узкоспециализированных специалистов. Для получения объективных результатов оценки разработана инструкция (см. раздел 13).

Разметка корпуса текстов проведена двумя подходами. Первый подход – оценка публикации по трем параметрам достоверности: достоверный, не достоверный и трудно определяется. Второй – глубинная разметка публикаций для формализации лингвистических правил и формирования словарей по компетентным источникам, на которые ссылаются в тексте.

Заключительным этапом работы является экспериментальное исследование полноты и корректности сформированного словаря и правил для распознавания классов и информативных признаков в корпусе (достоверность). Исходя из целей и задач, экспериментальное исследование проводится с использованием информационной системы по следующей схеме (рисунок 6).

Рисунок 6 – Экспериментальное исследование определения классов и информативных признаков



13.РАЗРАБОТКА ИНСТРУКЦИИ ДЛЯ ЭКСПЕРТНОГО АНАЛИЗА И КЛАССИФИКАЦИИ ДОСТОВЕРНОСТИ ПОДОЗРИТЕЛЬНЫХ ПУБЛИКАЦИЙ

В целях обеспечения понимания методологии и для исключения расхождений в толковании понятийного аппарата разработана инструкция для экспертного анализа и классификации достоверности подозрительных публикаций. В инструкцию включена информация, позволяющая обеспечить формирование единого мнения о целях и задачах экспертного анализа. Инструкция составлена лаконично, четко структурирована и содержит конкретные и детализированные вопросы.

Дополнительно перед началом работы с экспертами, выполняющими разметку, классификацию и анализ в соответствии с инструкцией, предусмотрено проведение установочных собраний с описанием сути исследования, методов и ожидаемых результатов их работы, а также с разъяснением возникающих у экспертов вопросов. Кроме этого, планируется проведение обучения работы с программным обеспечением, позволяющим произвести определение согласованности кодировщиков - https://nlp-ml.io/jg/software/ira/.

Инструкция составлена для двух последовательных этапов классификации и анализа достоверности подозрительных публикаций:

1) разметка публикаций и классификация подозрительных публикаций по классу «Достоверность» с тремя возможными вариантами классификации: «достоверная», «недостоверная», «трудно определяется» ;
2) экспертный анализ результатов разметки подозрительных публикаций в соответствии с классификацией по параметру «Достоверность» с двумя классами, каждый из которых включает 6 и 8 подклассов соответственно.

В обеих частях инструкции содержатся пошаговые указания для экспертов, позволяющие обеспечить максимальную точность исполнения заданий и направленные на получение однозначных и репрезентативных данных результатов классификации и анализа.

ЗАКЛЮЧЕНИЕ

В ходе проведения исследования получены следующие результаты.

1. Проведено исследование существующих информативных признаков с целью определения приемлемых критериев оценки влияния открытых текстовых информационных источников на социум.

В ходе исследования изучена роль СМИ в формировании общественного мнения, проанализированы международное и казахстанское законодательство в сфере регулирования деятельности СМИ с целью выработки основы для критериев оценки, определены инструменты, техники и механизмы воздействия на социум, используемые СМИ, определены и исследованы существующие информативные признаки медиа-текстов, выработаны критерии и показатели оценки влияния открытых текстовых информационных источников на социум на основе информативных признаков для последующего применения в автоматизированной информационной системе оценки влияния СМИ на социум.

2. Проведено исследование комплексной оценки значимости используемых информативных критериев. Также проведен комплекс работ, направленных на экспериментальное подтверждение адекватности полученной комплексной оценки значимости используемых информативных критериев, для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации.

По результатам исследования были построены три индекса: индекс объективности, индекс достоверности и индекс тональности. По результатам множественной регрессионной модели оценки значимости установлен статистически значимый эффект между влияниями индексов достоверности, тональности и объективности на переменную «количество просмотров». Информативные критерии публикации, такие как достоверность, объективность и тональность, оказывают статистически значимое влияние на количество просмотров. Так, к примеру, уменьшение индексов объективности и тональности положительно влияет на количество просмотров, в свою очередь, увеличение индекса достоверности увеличивает количество просмотров. Данные результаты были получены посредством статистической модели множественной линейной регрессии. Помимо статистически значимых результатов, модель показала хороший коэффициент детерминации (R квадрат=0.44).

3. Проведено исследование особенностей оценки влияния открытых текстовых информационных источников на социум в условиях Республики Казахстан применительно к ССД. По итогам исследования получены следующие результаты:

1) Проведен анализ казахстанского законодательства в области мониторинга СМИ.

2) Изучен опыт проведения мониторинга СМИ и проанализированы проведенные НИР в области СМИ в Казахстане.
3) Определены основные каналы получения населением информации новостного характера.
4) Определен социально-демографический портрет среднестатистического читателя информационных порталов.
5) Определены поведенческие особенности населения при медиа-потреблении информации.
6) Получены результаты оценки мнения населения об уровне объективности и достоверности информации в освещении событий в стране и уровень доверия к казахстанским новостным информационным источникам.
7) Рассмотрены значимые аспекты влияния электронных СМИ на общественное мнение.
8) По данным социологического исследования построены индекс доверия каналам распространения информации, индекс доверия информационным Интернет-сайтам, индекс объективности информационных Интернет-источников и индекс прозрачности информационных Интернет-источников.

4. Разработана методика для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации применительно к системе социального доверия.

5. Проведено комплексное исследование, а также экспертный анализ необходимого и достаточного состава формируемого текстового корпуса, его структуры, представительности, сбалансированности, требуемых уровней разметки и классификаций, их однозначности; способов сбора, хранения и представления текстового корпуса; методов организации работы и взаимного контроля задействованных экспертов при его формировании.

В результате комплекса работ были изучены и проанализированы ранее проводившиеся исследования, в ходе которых производилась компиляция и разметка текстовых корпусов. В целях определения необходимого и достаточного состава формируемого корпуса был проведен экспертный опрос, получены рекомендации ведущих казахстанских и российских экспертов и описаны их результаты. Были определены классы, параметры и виды разметки корпуса: первичная (метаданные), экспертная и лингвистическая.

6. Проведен комплекс работ по формированию сбалансированного, представительного, размеченного и проклассифицированного корпуса публицистических текстов: текстов официальных СМИ, журнальных публикаций, новостных порталов и др.

В ходе работ был сформирован представительный и сбалансированный корпус публицистических. В корпус вошли 5210 публикаций из 5 источников (tengrinews.kz, zakon.kz, azattyq.org, camonitor.kz, kz.expert) за однолетний период с 1 августа 2017 по 31 июля 2018 по определенным датам (3 последовательных дня с *n*-ным шагом) с первичным сбором мета-данных (первичная разметка). Также был разработан методологический подход и проведены мероприятия по организации работы с экспертами и лингвистами, привлеченными к разметке, позволяющие обеспечить надежность полученных данных (триангуляция данных, разработка инструкций и разъяснительные собрания экспертов). Была проведена первичная разметка 5210 публикаций по мета-данным (17 категорий данных).

8. Проведен комплекс работ по анализу фактических характеристик сформированного корпуса (состав, структура, представительность, сбалансированность, уровни разметки и классификации, их корректность и полнота и др.).

9. По итогам работы экспертов-лингвистов по глубинному анализу и детальной разметке медиатекстов получены следующие результаты:

1) сформирован словарь из 2892 слов (15 076 словоупотреблений) по информативному признаку тональности;
2) сформирован словарь из 37 слов (156 словоупотреблений) по информативному признаку политизированности;
3) сформирован словарь из 430 слов (3797 словоупотреблений) по информативному признаку наличия манипулятивных техник;
4) сформулированы правила формализации по определению информативных признаков тональности, политизированности, наличия манипулятивных техник.

10. Разработана методика определения степеней вероятной достоверности отдельной публикации.

11.Разработана инструкция для экспертного анализа и классификации достоверности подозрительных публикаций.

Все поставленные задачи были выполнены в полном объеме согласно технической спецификации и календарному плану. По завершению данного этапа исследования дальнейшая работа по реализации НТП будет продолжена, полученные результаты могут применяться в социологии, политологии, журналистике и др.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Flegel, R. C., Chaffee, S. H. (1971). Influences of editors, readers, and personal opinions on reporters. Journalism Quarterly, 48(4), 645–651.
2. Vosoughi, S., Roy, D., Aral, S. The spread of true and false news online // Science. – 2018. – Vol. 359, pp. 1146-1151.
3. Глобальное исследование «Барометр доверия Эдельман 2018» [Электронный ресурс]. -2018. - URL: http://cms.edelman.com/sites/default/files/2018-02/2018_Edelman_Trust_Barometer_Global_Report_FEB.pdf (дата обращения: 29.07.2018).
4. Digital vs. Traditional Media Consumption. Analyzing time devoted to online and traditional forms of media at a global level, as well as by age and across countries. INSIGHT REPORT SUMMARY | Q1 2017
5. Результаты социологического опроса по оценке влияния открытых информационных источников (электронных СМИ) на социум / АО «Информационно-аналитический центр» МОН РК; - Астана, 2018.
6. Бюро международных информационных систем, Государственный департамент США, Законодательство и СМИ, с. 72.
7. Кодекс этики журналиста Республики Казахстан [Электронный ресурс] // URL: https://www.inform.kz/ru/kodeks-etiki-zhurnalista-respubliki-kazahstan-polnyy-tekst_a2506438(дата обращения: 29.07.2018).
8. Семёнова А.В., Корсунская М.В. Контент-анализ СМИ: проблемы и опыт применения. М. – 2010.
9. Bryman A., Bell E. Business Research Methods. Oxford, 2003.
10. Рыжакова Е.В. Автоматическое определение иронии и сарказма в тексте, 2014.
11. Для более полной информации о кластерные выборки см. Groves, M.R., Floyd, J.F., Couper, M.P., Lepkowski, J.M, Singer, E & Tourangeau R. (2009). Survey Methodology (2nd Ed.). New Jersey: John Wiley & Sons
12. Jolliffe, I. T. (2002). Principal Component Analysis. (2nd Ed.). New-York: Springer.
13. Pages, J. (2015). Multiple Factor Analysis by Example Using R. Boca-Raton: Taylor & Francis Group.
14. Ramasubramanian K. & Singh A. (2017). Machine Learning Using R. New-York: Apress.
15. Newey, W.K. & West, K.D. (1987). A Simple Positive-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. “Econometrica”, 55, 703-708.

16. Fox, J. (2002). Bootstrapping Regression Models. Retrieve from Stanford University website: http://statweb.stanford.edu/~tibs/sta305files/FoxOnBootingRegInR.pdf
17. Приказ Министра информации и коммуникаций Республики Казахстан от 23 февраля 2018 года № 50 «Об утверждении Типового положения об уполномоченном лице (подразделении) по взаимодействию со средствами массовой информации».
18. Закон Республики Казахстан «О средствах массовой информации» от 23 июля 1999 года № 451-I.
19. Приказ и.о. Министра по инвестициям и развитию Республики Казахстан от 26 января 2016 года № 73.
20. Татарова Г. Методология анализа данных в социологии, М.1999
21. Определение Публикации [Электронный ресурс] // URL: http://tolkslovar.ru/p24649.html (дата обращения: 10.08.2018).
22. Красноярова О.В. Медийный текст: его особенности и виды // Известия ИГЭА. – 2010. - №3 (71). – сс. 177-180.
23. Русский язык как государственный [Электронный ресурс] // URL: http://rus-gos.spbu.ru/index.php/dictionary/index/1 (дата обращения: 23.07.2018).
24. О национальном корпусе русского языка. [Электронный ресурс] URL: http://www.ruscorpora.ru/corpora-structure.html (дата обращения: 30.09.2018)/

СПИСОК ОПУБЛИКОВАННЫХ ПУБЛИКАЦИЙ ПО ТЕМЕ ИССЛЕДОВАНИЯ

№	Наименование статьи	Авторы	Журнал	Статус
1.	ОПРЕДЕЛЕНИЕ ИНДИКАТОРОВ ОЦЕНКИ ДОСТОВЕРНОСТИ ПУБЛИКАЦИЙ В СМИ ДЛЯ РАЗРАБОТКИ МЕТОДИКИ ОПРЕДЕЛЕНИЯ СТЕПЕНИ ВЕРОЯТНОЙ ДОСТОВЕРНОСТИ ОТДЕЛЬНОЙ ПУБЛИКАЦИИ	М.К. Атанаева, У.А. Оспанова, Т.К. Булдыбаев, И.Г. Акоева, К.С. Нурумов	Вестник КазНПУ, Серия филологическая	Принято к публикации
2.	ИСПОЛЬЗОВАНИЕ ЛИНГВИСТИЧЕСКИХ ПОДХОДОВ В ОПРЕДЕЛЕНИИ КЛАССОВ И ИНФОРМАТИВНЫХ ПРИЗНАКОВ В ТЕКСТАХ	М.К. Атанаева, У.А. Оспанова, Т.К. Булдыбаев, И.Г. Акоева, К.С. Нурумов	Вестник КазНПУ, Серия филологическая	Принято к публикации
3.	КОМПЛЕКСНАЯ ОЦЕНКА ЗНАЧИМОСТИ ИНФОРМАТИВНЫХ ПРИЗНАКОВ ДЛЯ ОЦЕНКИ ВЛИЯНИЯ НА СОЦИУМ НА ОСНОВЕ АНАЛИЗА ПУБЛИКУЕМОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ	К.С. Нурумов, М.К. Атанаева, Т.К. Булдыбаев, У.А. Оспанова, И.Г. Акоева	Вестник Евразийского национального университета имени Л.Н. Гумилева, Серия Журналистика	Принято к публикации
4.	ИССЛЕДОВАНИЕ ИНФОРМАТИВНЫХ ПРИЗНАКОВ ПУБЛИКАЦИЙ В ЭЛЕКТРОННЫХ СМИ С ТОЧКИ ЗРЕНИЯ НАСЕЛЕНИЯ КАЗАХСТАНА	Оспанова У.А., Атанаева М.К., Булдыбаев Т.К., Акоева И.Г.	«Вестник КазГосЖенПУ», №3(75), 2018	Принято к публикации
5.	ИНСТРУКЦИЯ ДЛЯ ЭКСПЕРТНОГО АНАЛИЗА И КЛАССИФИКАЦИИ ДОСТОВЕРНОСТИ ПОДОЗРИТЕЛЬНЫХ ПУБЛИКАЦИЙ НОВОСТНОГО ХАРАКТЕРА	М. Атанаева, Т. Булдыбаев, У. Оспанова, И. Акоева	«Вестник КазГосЖенПУ», №3(75), 2018	Принято к публикации
6.	Identification of indicators of mass media publications reliability assessment for developing a methodology for determining the level of probable reliability of a separate publication	M. Atanayeva, T. Buldybayev, U. Ospanova, I. Akoyeva	Journal of International Communication	На стадии рецензирования
7.	Determination of priority informative features of reliability and bias of electronic media from the point of view of the population of Kazakhstan	M. Atanayeva, T. Buldybayev, U. Ospanova, I. Akoyeva	Canadian Journal of Communication	На стадии рецензирования

ПРИЛОЖЕНИЯ

1. Разработка методики для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации применительно к системе социального доверия

2. Исследование существующих информативных признаков с целью определения приемлемых критериев оценки влияния открытых текстовых информационных источников на социум применительно к ССД

3. Исследование комплексной оценки значимости используемых информативных критериев

4. Проведение комплекса работ, направленных на экспериментальное подтверждение адекватности полученной комплексной оценки значимости используемых информативных критериев для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации

5. Исследование особенностей оценки влияния открытых текстовых информационных источников на социум в условиях Республики Казахстан применительно к ССД

6. Комплексное исследование, а также экспертный анализ необходимого и достаточного состава формируемого текстового корпуса, его структуры, представительности, сбалансированности, требуемых уровней разметки и классификаций, их однозначности; способов сбора, хранения и представления текстового корпуса; методов организации работы и взаимного контроля задействованных экспертов при его формировании

7. Проведение комплекса работ по формированию сбалансированного, представительного, размеченного и проклассифицированного корпуса публицистических текстов: текстов официальных СМИ, журнальных публикаций, новостных порталов и др.

8. Проведение комплекса работ по анализу фактических характеристик сформированного корпуса (состав, структура, представительность, сбалансированность, уровни разметки и классификации, их корректность и полнота и др.)

9. Разработка методов формализации правил и словарей для распознавания в размеченном корпусе информативных признаков и определения степени их выраженности

10. Проведение комплекса работ по анализу и выявлению закономерностей, присутствующих в сформированном корпусе, с целью формализации правил и формирования словарей для распознавания в размеченном корпусе информативных признаков и определения степени их выраженности на основе разработанных методов

11. Проведение комплекса работ по экспериментальному исследованию полноты и корректности сформированного множества правил и словарей для распознавания в размеченном корпусе информативных признаков и определения степени их выраженности

12. Разработка методики определения степеней вероятной достоверности отдельной публикации, вероятной достоверности публикаций определённого автора, вероятной достоверности публикаций определённого печатного СМИ и критерии их верификации

13. Разработка инструкции для экспертного анализа и классификации достоверности подозрительных публикаций