Министерство образования и науки Республики Казахстан

Акционерное общество «Информационно-аналитический центр»

«УТВЕРЖДАЮ»

Заместитель президента
АО «Информационно-
аналитический центр»

_____ М. К. Атанаева

«11» октября 2019 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

Нур-Султан 2019

СПИСОК ИСПОЛНИТЕЛЕЙ

Координатор исследования	(подпись, дата)	М. К. Атанаева, (введение, заключение)
Руководитель исследования	(подпись, дата)	Т. К. Булдыбаев, (введение, разделы 1-12, заключение)
Основные исполнители исследования	(подпись, дата)	У. А. Оспанова, (разделы 2, 4, 6, 8, 10, 11, 12)
	(подпись, дата)	И. Г. Акоева, (разделы 1, 2, 7, 8, 11, 12)
	(подпись, дата)	К. С. Нурумов, (разделы 1, 2, 3, 4, 7, 8, 9, 10)
	(подпись, дата)	М. А. Баймаханбетов, (разделы 1, 2, 3, 7, 8, 11, 12)
	(подпись, дата)	М. Т. Шакенова, (разделы 6, 8, 10)
	(подпись, дата)	А. О. Пилипенко, (разделы 2, 4, 5)
	(подпись, дата)	Э. М. Айтжанова, (разделы 11, 12)
	(подпись, дата)	Г. Г. Карбаева, (раздел 11)
Соисполнители исследования	(подпись, дата)	А. А. Тиштыбаев, (раздел 11)
	(подпись, дата)	Ж. С. Увалиева, (раздел 11)
	(подпись, дата)	А. А. Смагулова, (раздел 11)
	(подпись, дата)	А. А. Кусиденова, (раздел 11)
	(подпись, дата)	А. Т. Капсалыкова, (раздел 11)
	(подпись, дата)	С. С. Утенова, (раздел 11)
	(подпись, дата)	С. К. Мухамеджанова, (раздел 11)
	(подпись, дата)	А. А. Абылхаиров, (раздел 11)

2

РЕФЕРАТ

Количество страниц в отчете: 106. В отчете содержится 12 рисунков, 35 таблиц, состоит из 15 частей.

Ключевые слова: новостные тексты, информативные признаки, тональность, объективность, манипулятивность, разметка, корпус.

Работа выполняется в рамках Договора №7 на оказание услуг по выполнению научно-исследовательских работ в рамках научно-технической программы №BR05236839 «Разработка информационных технологий и систем для стимулирования устойчивого развития личности как одна из основ развития цифрового Казахстана».

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

Цель исследования: выработка методических основ для проектирования информационной системы оценки влияния открытых текстовых информационных источников на социум на основе публикуемой информации применительно к системе социального доверия.

В исследовании применялись различные подходы: машинное обучение, методы статистического анализа, лингвистические методы контекстуального, дискурсивного, интерпретативного, прагматического и компонентного анализа, интроспекция.

Научная новизна исследования заключается в том, что данная работа является первой в Казахстане по выработке комплексного анализа казахстанского медиапространства с применением комбинированных методов машинного обучения и лексико-словарного подхода. Также в рамках исследования сделана попытка улучшить существующие подходы распознавания тональности текста, и выработать подходы к определению объективности и манипулятивности новостного текста.

По результатам проведенных работ в 2019 году получены промежуточные результаты. В частности, разработана методика оценки влияния открытых текстовых информационных источников на социум на основе публикуемой информации. Сформированы 27 словарей по информативным признакам тональности, политизированности, манипулятивности, призыва к действию, предвзятости, именованных сущностей, тематических областей. Получены результаты экспериментального подтверждения сформированных словарей и правил для распознавания признака «тональности» с использованием алгоритмов машинного обучения и определения лексико-словарным подходом полярность тональности и степень объективности новостного текста.

Сформированы корпуса новостных текстов, описаны их фактические характеристики и выявлены статистические и текстовые закономерности.

Результаты исследования могут быть применены для проектирования информационной системы оценки влияния открытых текстовых информационных источников на социум на основе анализа публикуемой информации.

Полученные результаты исследования также будут полезны для применения в области социальных наук, журналистики и информационных технологий.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 8
1. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО АНАЛИЗУ ЭФФЕКТИВНОСТИ РАЗРАБОТАННОЙ МЕТОДИКИ ПРИМЕНИТЕЛЬНО К ССД ДЛЯ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ НА ОСНОВЕ АНАЛИЗА ПУБЛИКУЕМОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ 11
2. НА ОСНОВЕ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ АНАЛИЗА ЭФФЕКТИВНОСТИ И ВЫЯВЛЕННЫХ НЕДОСТАТКОВ ДОРАБОТКА И УСОВЕРШЕНСТВОВАНИЕ МЕТОДИКИ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ НА ОСНОВЕ АНАЛИЗА ПУБЛИКУЕМОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ ПРИМЕНИТЕЛЬНО К ССД 21
3. РАЗРАБОТКА МЕТОДИКИ РАСЧЁТА ИНФОРМАТИВНЫХ КРИТЕРИЕВ ДЛЯ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ ПРИМЕНИТЕЛЬНО К ССД (ОПИСАНИЕ, ЦЕЛЬ, ИСТОЧНИКИ ДАННЫХ, МЕТОДЫ И ФОРМУЛЫ РАСЧЕТА, НОРМАЛИЗАЦИЯ И ОБРАБОТКА ИСПОЛЬЗУЕМЫХ ДАННЫХ) 49
4. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО ИНФОРМАЦИОННОМУ, АНАЛИТИЧЕСКОМУ И ЭКСПЕРТНОМУ СОПРОВОЖДЕНИЮ ПРОЦЕССА ПРОЕКТИРОВАНИЯ И РАЗРАБОТКИ ИНФОРМАЦИОННОЙ СИСТЕМЫ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ ПРИМЕНИТЕЛЬНО К ССД 52
5. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО МОНИТОРИНГУ ОТКЛОНЕНИЙ В РАБОТЕ РАЗРАБОТАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ ПРИМЕНИТЕЛЬНО К ССД 55
6. РАЗРАБОТКА МЕТОДОВ ФОРМАЛИЗАЦИИ ПРАВИЛ И СЛОВАРЕЙ ДЛЯ РАСПОЗНАВАНИЯ В РАЗМЕЧЕННОМ КОРПУСЕ ИНФОРМАТИВНЫХ ПРИЗНАКОВ И ОПРЕДЕЛЕНИЯ СТЕПЕНИ ИХ ВЫРАЖЕННОСТИ 56
7. 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 68

8. 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 73
9. РАЗРАБОТКА МЕТОДИКИ ОПРЕДЕЛЕНИЯ СТЕПЕНЕЙ ВЕРОЯТНОЙ ДОСТОВЕРНОСТИ ОТДЕЛЬНОЙ ПУБЛИКАЦИИ, ВЕРОЯТНОЙ ДОСТОВЕРНОСТИ ПУБЛИКАЦИЙ ОПРЕДЕЛЁННОГО АВТОРА, ВЕРОЯТНОЙ ДОСТОВЕРНОСТИ ПУБЛИКАЦИЙ ОПРЕДЕЛЁННОГО ПЕЧАТНОГО СМИ И КРИТЕРИИ ИХ ВЕРИФИКАЦИИ 76
10. РАЗРАБОТКА ИНСТРУКЦИИ ДЛЯ ЭКСПЕРТНОГО АНАЛИЗА И КЛАССИФИКАЦИИ ДОСТОВЕРНОСТИ ПОДОЗРИТЕЛЬНЫХ ПУБЛИКАЦИЙ 80
11. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО ФОРМИРОВАНИЮ ДОПОЛНИТЕЛЬНОГО СБАЛАНСИРОВАННОГО, ПРЕДСТАВИТЕЛЬНОГО, РАЗМЕЧЕННОГО И ПРОКЛАССИФИЦИРОВАННОГО КОРПУСА ПУБЛИЦИСТИЧЕСКИХ ТЕКСТОВ: ТЕКСТОВ ОФИЦИАЛЬНЫХ СМИ, ЖУРНАЛЬНЫХ ПУБЛИКАЦИЙ, НОВОСТНЫХ ПОРТАЛОВ И ДР. 81
12. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО АНАЛИЗУ ФАКТИЧЕСКИХ ХАРАКТЕРИСТИК СФОРМИРОВАННОГО КОРПУСА (СОСТАВ, СТРУКТУРА, ПРЕДСТАВИТЕЛЬНОСТЬ, СБАЛАНСИРОВАННОСТЬ, УРОВНИ РАЗМЕТКИ И КЛАССИФИКАЦИИ, ИХ КОРРЕКТНОСТЬ И ПОЛНОТА; И ДР.) 93
ЗАКЛЮЧЕНИЕ 97
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 99
ПРИЛОЖЕНИЯ 106

ОПРЕДЕЛЕНИЯ

Тональность	позволяет оценить эмоциональный тон подачи информации, фактов в публикации, который автор транслирует на аудиторию.
Объективность	информация, которая отвечает критериям достоверности, и рассматривается с точки зрения ее подачи в объективном виде без субъективной точки зрения автора.
Манипулятивность	совокупность, определенные комбинации речевых стратегий и тактик, создаваемые с целью неявного воздействия на общественное сознание и выраженных определенными языковыми индикаторами.
Корпус	структурированная система собрания текстов на некотором языке в электронной форме, связанных между собой определенными параметрами [1].

ВВЕДЕНИЕ

По данным Worldometers[1] ежедневно в мире публикуется свыше 200 млн. газет[2], 3 млн. блогов[3], отправляются свыше 310 млн. твитов[4]. Количество интернет пользователей превысило 4,4 млрд человек. С усилением цифровизации всех сфер жизнедеятельности, информация становится легкодоступной для проведения исследований, использования в коммерческих целях и др. В тоже время данная тенденция создает возможности для распространения необъективной, фейковой, негативной, дезинформирующей информации, которые влияют на доверие ко всей новостной экосистеме [2] и влияют на отдельные категории социума [3]. В связи со стремительным увеличением объемов информации возрастает потребность в автоматическом анализе больших объемов новостных текстовых данных.

В целях разработки методических основ информационной системы оценки влияния открытых текстовых информационных источников на социум на основе публикуемой информации применительно к системе социального доверия (ССД) проведен комплекс взаимосвязанных работ. Комплекс работ включал следующие виды работ: (1) доработка и усовершенствование методики оценки влияния открытых информационных источников на социум; (2) разработка методики расчёта информативных критериев; (3) формализацию правил и составление словарей; (4) экспериментальное исследование полноты и корректности сформированных словарей и лингвистических правил; (5) формирование корпуса новостной текстовой информации и (6) его разметку, (7) анализ и выявление закономерностей в корпусе новостных текстов; (8) разработку методики определения степени объективности новостной текстовой информации.

Научная новизна исследования заключается в том, что данная работа является первой в Казахстане по выработке комплексного анализа казахстанского медиапространства с применением комбинированных методов машинного обучения и лексико-словарного подхода. Также в рамках исследования сделана попытка улучшить существующие подходы распознавания тональности текста, и выработать подходы к определению объективности и манипулятивности новостного текста.

По результатам исследования получены: (1) алгоритм машинного обучения по распознаванию тональности новостной текстовой информации с использованием новых предикторов; (2) алгоритм определения степени объективности и (3) полярности

---

[1] https://www.worldometers.info/
[2] Всемирная ассоциация газет и новостных изданий (WAN-IFRA)
[3] Technorati on State of the Blogosphere reports and elaboration of other data, WordPress Stats - Automattic Inc.
[4] Internet Live Stats (InternetLiveStats.com)

тональности новостных текстов с использованием лексико-словарного подхода; (4) предложен подход определения манипулятивных техник с использованием лексико-словарного подхода; (5) решение проблемы несбалансированности размеченного корпуса текстовых данных; (6) сформированы 27 словаря по информативным признакам тональности, политизированности, манипулятивности, призыва к действию, предвзятости, именованных сущностей, тематических областей; (7) получены результаты экспериментального подтверждения сформированных словарей и правил для распознавания признака «тональности» с использованием алгоритмов машинного обучения и определения лексико-словарным подходом полярность тональности и степень объективности новостного текста; (8) сформированы корпуса новостных текстов, описаны их фактические характеристики и выявлены статистические и текстовые закономерности.

В соответствии с календарным планом в рамках данного исследования на 2019 отчетный год запланирована реализация следующих задач:

1) проведение комплекса работ по анализу эффективности разработанной методики применительно к ССД для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации;

2) на основе полученных результатов анализа эффективности и выявленных недостатков, доработка и усовершенствование методики оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации применительно к ССД;

3) разработка методики расчёта информативных критериев для оценки влияния открытых текстовых информационных источников на социум применительно к ССД (описание, цель, источники данных, методы и формулы расчета, нормализация и обработка используемых данных);

4) проведение комплекса работ по информационному, аналитическому и экспертному сопровождению процесса проектирования и разработки информационной системы оценки влияния открытых текстовых информационных источников на социум применительно к ССД;

5) проведение комплекса работ по мониторингу отклонений в работе разработанной информационной системы оценки влияния открытых текстовых информационных источников на социум применительно к ССД;

6) разработка методов формализации правил и словарей для распознавания в размеченном корпусе информативных признаков и определения степени их выраженности;

7) 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;

8) 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;

9) разработка методики определения степеней вероятной достоверности отдельной публикации, вероятной достоверности публикаций определённого автора, вероятной достоверности публикаций определённого печатного СМИ и критерии их верификации;

10) разработка инструкции для экспертного анализа и классификации достоверности подозрительных публикаций;

11) проведение комплекса работ по формированию дополнительного сбалансированного, представительного, размеченного и проклассифицированного корпуса публицистических текстов: текстов официальных СМИ, журнальных публикаций, новостных порталов и др.;

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

Все поставленные задачи выполнены в полном объёме и их текущее состояние отражено в данном промежуточном отчёте.

1. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО АНАЛИЗУ ЭФФЕКТИВНОСТИ РАЗРАБОТАННОЙ МЕТОДИКИ ПРИМЕНИТЕЛЬНО К ССД ДЛЯ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ НА ОСНОВЕ АНАЛИЗА ПУБЛИКУЕМОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

1.1 Краткое описание работ по разработке методики

В результате проведения комплекса работ в 2018 году разработана методика оценки влияния открытых информационных источников на социум на основе анализа публикуемой тестовой информации применительно к системе социального доверия (далее – Методика). Комплекс работ включал следующие мероприятия: (1) социологическое исследование населения по оценке влияния открытых информационных источников (электронных СМИ) на социум Казахстана; (2) теоретический анализ и обобщение научной и аналитической литературы по анализу новостных текстов; (3) статистический анализ и моделирование для получения оценки значимости информативных критериев; (4) формирование корпуса новостных текстов; (5) разметка корпуса текстов экспертами и лингвистами.

По итогам социологического исследования определены (1) основные каналы получения населением информации новостного характера; (2) социально-демографический портрет среднестатистического читателя информационных порталов; (3) поведенческие особенности при медиа-потреблении; (4) мнения населения об уровне объективности и достоверности информации в освещении событий электронными СМИ; (5) уровень доверия к казахстанским новостным информационным источникам.

Обзор научной и аналитической литературы позволил определить косвенные информативные признаки, которые, возможно применить в анализе новостных текстов. Было предложено 16 информативных признаков, объединенных в 4 информативных критерия: (1) достоверность; (2) объективность или предвзятость; (3) тональность; (4) резонансность (популярность) публикации. В частности, (1) наличие ссылки в публикации на компетентный источник; (2) ссылка на первоисточник; (3) соответствие заголовка содержанию публикации; (4) освещение одного и того же события различными изданиями: кросс-чекинг, наличие дискурса с другими статьями; (5) наличие автора в статье; (6) наличие проверяемых фактов в статье; (7) репутация издания, на котором опубликована информация; (8) наличие манипулятивных техник; (9) политизированность публикации; (10) наличие в медиа-тексте выраженного личного мнения автора публикации; (11) генереализация текста; (12) тональность (позитивная/негативная); (13) количество просмотров, (14) шейров, (15) комментариев к публикации; (16) тональность комментариев (позитивная/негативная).

Для определения значимости предложенных информативных признаков и критериев проведена комплексная оценка их значимости. Процедура комплексной оценки значимости включает: (1) формирование выборки корпуса публикаций, (2) построение индексов для измерения информативных признаков, (3) построение статистической модели и (4) модели машинного обучения. Для формирования корпуса новостных текстов была проведена двухэтапная систематическая кластерная выборка. Для определения значимости информативных признаков была проведена экспертная разметка корпуса новостных публикаций. Разметка осуществлена по 16 информативным признакам. Для построения индексов был использован аддитивный подход. Данный подход был выбран в связи с тем, что структура данных и наличие теоретических концептов позволяют использовать аддитивный подход к построению индексов, цель которых адекватно измерить предложенные критерии оценки влияния публикации на социум. В ходе проведения данной процедуры построены три индекса: индекс объективности, индекс достоверности, индекс тональности. Полученные результаты статистического моделирования показали, что предложенные информативные критерии публикации могут в дальнейшем служить каркасом для создания автоматизированной модели машинного обучения. Полученные результаты были экспериментально подтверждены посредством техники бутстрэпинг, которая показала незначительное отклонение от результатов модели.

Таким образом в предложенной Методике определены информативные признаки и критерии, проведена оценка их значимости с позиции резонансности (популярности), выработана схема оценки текстовой информации и предложен подход распознавания информативных признаков в текстах.

1.2 Результаты экспериментального подтверждения эффективности разработанной Методики

Экспериментальное подтверждение значимости влияния информативных признаков на резонансность проведено по результатам экспертной и лингвистической разметки малого корпуса новостных текстов из 5211 публикаций по 5 новостным интернет-порталам.

Разметка корпуса была проведена по 5 признакам: достоверный/недостоверный, объективный/предвзятый; тональность (позитивная, нейтральная или негативная); наличие/отсутствие манипулятивных техник; политизированное/неполитизированное содержание текста. Для обеспечения объективности, верификации данных и повышения надежности результатов экспертной разметки при обработке и анализе результатов экспертной разметки применялся *метод триангуляции данных.* Для оценки результатов разметки использовано ПО Inter-Rater Agreement, учитывающее каппу Лайта и Флейса, альфу Криппендорфа с функцией поддержки недостающих данных, находящееся в

открытом онлайн доступе [4]. В ходе обработки данных с помощью ПО Inter-Rater Agreement происходила фильтрация и отбраковка мнений и определялась принадлежность публикации к классу и параметру. В целом кодировка характеризуется довольно высоким результатом согласия по всем переменным. В ходе проведенного анализа, результаты которого приведены ниже в Таблицах 1 и 2, было установлено, что триангуляция данных повышает точность машинного обучения на основании триангулированных данных для алгоритма К-ближайших соседей (K-Nearest Neighbors, K-NN) и Наивный Байес (Naïve Bayes, NB).

Таблица 1 – Метрики результативности машинного обучения методом K-NN на основе триангулированных и нетриангулированных данных

K-NN - Нетриангулированные данные				**K-NN ML Триангулированные данные**			
accuracy=	0.7771			**accuracy**=	0.816		
precision (neg)=	0.73	recall (neg)=	0.96	precision (neg)=	0.78	recall (neg)=	0.99
precision (neut)=	0.94	recall (neut)=	0.38	precision (neut)=	0.99	recall (neut)=	0.46
precision (pos)=	0.77	recall (pos)=	0.99	precision (pos)=	0.79	recall (pos)=	1.00
precision (macro avg)=	0.82	**recall (macro avg)=**	0.78	**precision (macro avg)=**	0.85	**recall (macro avg)=**	0.82
F1 score	0.75			**F1 score**	0.79		

Таблица 2 – Метрики результативности машинного обучения методом NB на основе триангулированных и нетриангулированных данных

NB ML Нетриангулированные данные				**NB ML Триангулированные данные**			
accuracy=	0.4237			**accuracy**=	0.4734		
precision (neg)=	0.48	recall (neg)=	0.21	precision (neg)=	0.62	recall (neg)=	0.33
precision (neut)=	0.37	recall (neut)=	0.17	precision (neut)=	0.37	recall (neut)=	0.21
precision (pos)=	0.42	recall (pos)=	0.88	precision (pos)=	0.46	recall (pos)=	0.89
precision (macro avg)=	0.42	**recall (macro avg)=**	0.42	**precision (macro avg)=**	0.49	**recall (macro avg)=**	0.47
F1 score	0.37			**F1 score**	0.43		

Из вышеприведенных результатов видно, что триангуляция данных улучшает точность машинного обучения на 4% для метода K-NN и на 5% для метода NB. Однако дальнейшая работа показала, что в различных условиях, как, к примеру, в условиях исключения предварительного проведения процедуры ADASYN, триангуляция данных

позволяет увеличить точность машинного обучения (accuracy) на 11,15% и 7,5%, а F-меры (F1 score) на 3% и 6,6% для алгоритмов K-NN и NB соответственно.

Также в ходе дальнейшей работы над проектом было установлено, что эффект улучшения результативности машинного обучения за счет триангуляции данных сохраняется при различных условиях, при различных алгоритмах машинного обучения: XGBoost, Multilayer Perceptron, Random Forest. При этом, чем ниже исходная точность машинного обучения, тем сильнее положительный эффект триангуляции данных.

По результатам лингвистической разметки текстов сформированы словари и словосочетания по 4 признакам, использованным в процессе оценки (таблица 3).

Таблица 3 – Результаты лингвистической разметки

Словари	**Количество**	
	слов	**словосочетаний**
Тональность	3291	6955
Политизированность	62	238
Манипулятивные техники	887	1479
Компетентные источники		1167

По итогам анализа разметки публикаций Источника 1 (Kz.expert) наблюдается наибольшее количество недостоверных, предвзятых, политизированных новостей с присутствием манипулятивных техник. В Источнике 2 (Tengrinews) и 3 (Zakon) наблюдается низкое количество подозрительных публикаций (рисунки 1, 2).

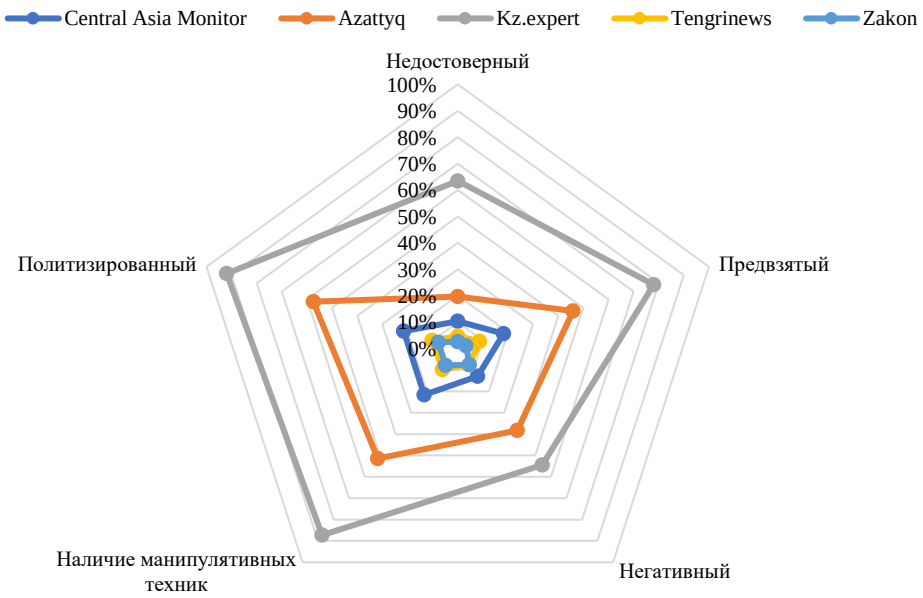

Рисунок 1. Доля публикаций в разбивке по признакам от общего количества новостей по конкретному порталу

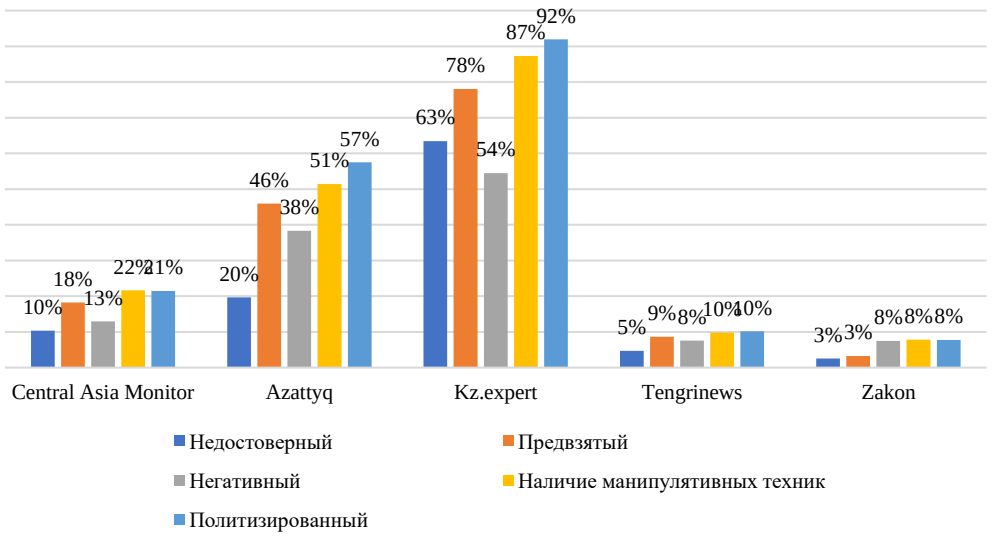

Рисунок 2. Доля публикаций в разбивке по признакам от общего количества новостей по конкретному порталу

Результаты экспериментального подтверждения позволили определить влияние отобранных информативных признаков на признаки резонансности (популярности).

Для определения взаимозависимостей между признаками и показателями резонансности *(просмотры, шейры, комментарии)* построена матрица корреляции (рисунок 3).

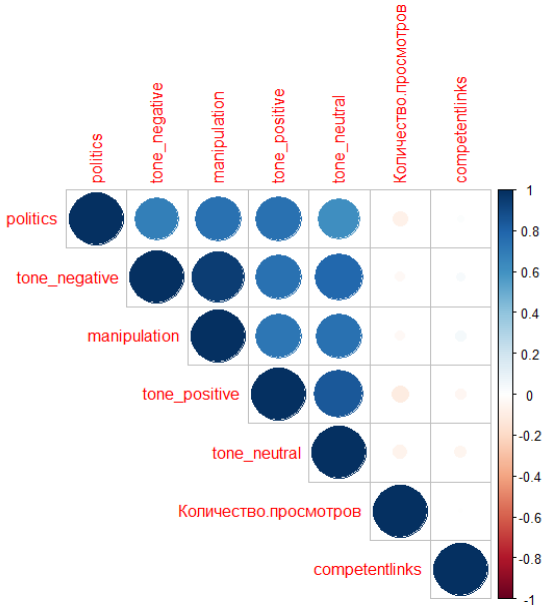

Рисунок 3. Матрица корреляции

Корреляционная матрица (n=2204) показывает, что переменная «количество просмотров» имеет незначительную корреляцию с переменными «политизированность», «тональность» и «наличие компетентных источников в публикациях».

Результаты корреляции:

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.2473009 0.0493633 -25.268 < 2e-16 ***
tone_positive -0.0317086 0.0026947 -11.767 < 2e-16 ***
tone_neutral -0.0025930 0.0010811 -2.398 0.016466 *
manipulation 0.0285773 0.0028012 10.202 < 2e-16 ***
competentlinks 0.1049667 0.0251786 4.169 3.06e-05 ***
Объем.текста 0.0004658 0.0001299 3.586 0.000336 ***

Коэффициенты модели (n=5143) показывают положительное влияние на резонансность таких переменных как манипулятивные техники, наличие компетентных источников и объём текста. С их повышением увеличивается вероятность того, что публикация окажется резонансной. В то же время повышение частоты слов с нейтральной и позитивной тональностью ведет к понижению вероятности резонансности статьи.

При прочих равных условиях:

- повышение позитивной тональности ведет к уменьшению вероятности нахождения резонансной статьи;
- повышение нейтральной тональности ведет к понижению вероятности резонансной статьи;
- повышение манипулятивных техник ведет к повышению вероятности резонансной статьи;
- повышение наличия компетентных источников ведет к увеличению вероятности резонансной статьи;
- увеличение объёма текста ведет к увеличению вероятности нахождения резонансной статьи.

Факторный анализ, проведенный на переменных «тональность», «манипулятивные техники», «политизированность», «объём текста», показал результаты, указанные в таблице 4. Для удобства были оставлены лишь показатели со значением выше 0,6.

Таблица 4 – Результаты факторного анализа

	Factor1	**Factor2**
объем текста	0.887	
тон. негативная		0.814
тон. позитивная	0.764	
тон. нейтральная	0.903	
наличие манипулятивных техник		0.914
политизированность		0.640
	Factor 1	Factor 2
кумулятивная дисперсия	0.461	0.876

Из анализа видно, что такие переменные как «негативная тональность», «манипулятивные техники» и «политизированность», составляют один фактор. Второй фактор составляют такие переменные, как «объем текста», «позитивная» и «нейтральная» тональность. Оба фактора объясняют 87% всех данных.

На основе полученных факторов построена модель множественной регрессии для определения влияния признаков на резонансность публикации. В качестве независимых переменных были использованы факторы, построенные в ходе исследовательского факторного анализа, указанные в таблице 4. В качестве зависимой переменной было взято «количество просмотров». Полученные факторы были сохранены как регрессионные очки и использованы как две независимые переменные в модели множественной регрессии. Также в качестве независимых переменных был использован аддитивный индекс, состоящий из таких переменных, как «достоверность», «объективность», «социальная значимость».

Результаты множественной регрессии:

```
Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.144923 0.116722 1.242 0.214
 index -0.181201 0.143806 -1.260 0.208
 scale(Factor1) -0.094898 0.020100 -4.721 2.47e-06 ***
 scale(Factor2) 0.007435 0.019928 0.373 0.709
---
Residual standard error: 0.9961 on 2504 degrees of freedom
Multiple R-squared: 0.009028,	Adjusted R-squared: 0.00784
F-statistic: 7.604 on 3 and 2504 DF, p-value: 4.635e-05
```

Результаты множественной регрессии показывают, что стандартизированный фактор 1, состоящий из таких переменных, как «объем текста», «позитивная» и «нейтральная» тональность, - демонстрирует статистически значимое влияние на стандартизированное количество просмотров. С повышением фактора на один юнит, количество просмотров падает на -0.094898 стандартных значений. Таким образом, модель хоть и имеет низкий R квадрат, тем не менее показывает, что статистически значимое влияние наблюдается и может быть использовано для построения модели машинного обучения.

Для установления зависимости объёма публикации от информативных критериев был также использован факторный анализ, который показал, что если убрать объём публикации, то одного фактора достаточно для группировки таких переменных как позитивная, негативная и нейтральная тональность, наличие манипулятивных техник и политизированность статьи (таблица 5).

Таблица 5 – Результаты факторного анализа

	Фактор
тон. негативная	0.967
тон. позитивная	0.845
тон. нейтральная	0.718
наличие манипулятивных техник	0.987
политизированность	0.867
	Фактор 1
кумулятивная дисперсия	0.77

Полученный фактор объясняет 77% всех данных. Фактор был использован в модели множественной регрессии, наряду с переменной индекс, состоящей из переменных «достоверность», «объективность», «социальная значимость».

```
Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.43660 0.04834 9.031 <2e-16 ***
 index -0.55659 0.06065 -9.177 <2e-16 ***
 Factor1 0.76207 0.00883 86.301 <2e-16 ***
---
Residual standard error: 0.6195 on 5208 degrees of freedom
Multiple R-squared: 0.6164,	Adjusted R-squared: 0.6162
F-statistic: 4184 on 2 and 5208 DF, p-value: < 2.2e-16
```

Модель показывает статистически значимые результаты влияния переменной индекс и фактор 1 на объём текста публикации. Так, увеличение индекса на 1 юнит ведет к уменьшению объёма текста на -0.55659 стандартных значений, увеличение фактора 1 на один юнит ведет к увеличению объёма публикации на 0.76207 стандартных значений.

Таким образом, модель показала статистически значимое влияние переменных при довольно высоком R квадрате 0,61. Полученные результаты могут быть использованы для построения модели машинного обучения.

1.3 Выводы и дальнейшая работа

Исходя из вышеизложенного, различные модели позволили определить информативные признаки, влияющие на резонансность (популярность) публикации. Полученные результаты возможно использовать для построения модели машинного обучения по распознаванию информативных признаков в новостных текстах. Также по результатам исследования установлено, что выявление признаков с использованием словарей обеспечивает большую надежность результатов. Таким образом, тестирование выбранных информативных признаков в различных моделях каждый раз подтверждает их значимость. Вместе с тем, полученные результаты анализа по некоторым признакам необходимо подтвердить на более расширенном количестве размеченных публикаций.

В результате проведенной работы можно выделить следующие мероприятия по улучшению методики оценки влияния текстовой информации на социум.

1) Дальнейшее уточнение информативных признаков публикаций. Необходима дальнейшая работа по детализации понятий информативных признаков, а также их распознавания в тексте. Имеющиеся наработки не позволяют однозначно и в полной мере определить наличие/отсутствие информативного(ых) признака(ов) и на основе такого определения принять решение в отношении того, относится та или иная публикация к классу по данному признаку. Этот вывод был сделан в результате анализа корпуса, размеченного экспертами по 5211 публикациям. В тех случаях, когда эксперт затруднялся классифицировать публикацию по информативному признаку, он прибегал к субъективному суждению и оценке публикации, что негативно сказывалось на результатах машинного обучения на основе такой разметки.

Бинарные переменные не всегда отражают действительную характеристику публикации. Так, к примеру, если в публикации в равной степени присутствуют слова из словарей негативной и позитивной тональности, однозначная классификация статьи, по одной из этих категорий, потребует либо выставления промежуточных подклассов, либо выставления пороговых значений. По некоторым из выделенных признаков, более точным представляется измерение степени присутствия (по нормализованной шкале от 0% до 100%) того или иного информативного признака в публикации. Более того, к каждому информативному признаку необходим индивидуальный подход.

В ходе консультаций с экспертами авторы исследования пришли к заключению о необходимости исключения следующих признаков либо по причине отсутствия данных, либо по высокой стоимости реализации: (1) ссылка на первоисточник; аутентичность (2) фото и (3) видео, которое подтверждает событие; (4) наличие факта последующей правки в уже опубликованной статье; (5) репутация издания (сайты новостных и информационных агентств, электронных СМИ, иных аналогичных сайтов), на которых опубликована информация; (6) степень генерализации; (7) глубина просмотра публикации; (8) длительность просмотра публикации аудиторией; (9) частота возвращений аудитории на интернет портал СМИ.

2) Экспертная разметка публикаций. Распределение данных, полученных в результате экспертной разметки корпуса публикаций по информативным признакам, получилось крайне неравномерным. Так, к примеру, около 89% всех публикаций были оценены как объективные, около 85% всех публикаций не имеют манипулятивных техник. Неравномерность распределения данных прямым образом сказывается на качестве и точности процедур машинного обучения. Алгоритмы будут с успехом определять класс информативных признаков, имеющих большую представленность в размеченном корпусе, в то время как признаки, имеющие низкую представленность, алгоритм определить не

сможет. Необходимо отметить, что такие результаты разметки представляют собой объективное отражение реальности, где основная масса новостей генерируется крупными новостными агрегаторами, которые стремятся к объективной и нейтральной подаче информации. Также новостные сайты учитывают особенности казахстанского законодательства в области регулирования СМИ и соответствующую ответственность, предусматриваемую за клевету и распространение ложной информации. Возможно, при значительном расширении размеченного корпуса с помощью экспертного подхода удалось бы обучить алгоритм определять все классы, однако данный шаг был бы неэффективен с точки зрения финансовых и временных ресурсов. В дополнение необходимо принимать во внимание трудности с определением и классификацией публикаций по информативным признакам на основе имеющейся методики. В этой связи необходимо увеличить количество подозрительных публикаций (недостоверных, политизированных, негативных, позитивных и др.) в составе корпуса.

3) Создание словарей по тематическим направлениям и информативным признакам. Подключение к анализу корпуса словарей, составленных лингвистами, повысило эффективность распознавания информативных признаков. Логистическая регрессия показала статистически значимое влияние количественных переменных, представленных словарями. Требуется дальнейшая работа со словарями, в особенности в области формализации лингвистических правил и шаблонов и нахождения их в тексте. Также необходимо провести комплекс работ по экспериментальному подтверждению сформированных словарей. После анализа эффективности сформированных словарей необходимо выработать алгоритм распознавания информативных признаков в тексте с использованием словарного подхода и машинного обучения и производить дальнейшую разметку на их основе.

2. НА ОСНОВЕ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ АНАЛИЗА ЭФФЕКТИВНОСТИ И ВЫЯВЛЕННЫХ НЕДОСТАТКОВ ДОРАБОТКА И УСОВЕРШЕНСТВОВАНИЕ МЕТОДИКИ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ НА ОСНОВЕ АНАЛИЗА ПУБЛИКУЕМОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ ПРИМЕНИТЕЛЬНО К ССД

По результатам анализа предложенной методики оценки влияния открытых информационных источников на социум на основе публикуемой текстовой информации проведен комплекс работ по доработке и усовершенствованию методики.

В комплекс работ входит: (1) формализация правил и составление словарей, (2) экспериментальное исследование полноты и корректности сформированных словарей и лингвистических правил, (3) формирование корпуса новостной текстовой информации и (4) его разметка, (5) анализ и выявление закономерностей в корпусе, (6) решение задачи несбалансированности размеченного корпуса, (7) проведение тематического моделирования, (8) разработка алгоритмов распознавания информативных признаков и (9) составление методики их расчета.

2.1. Определение доверия населения к новостным каналам информации

По результатам социологического исследования, проведенного в 2018 году, для 59,8% населения Казахстана основными источниками получения новостной информации являются интернет-источники (30,3% социальные сети и блогеры и 29,5% новостные интернет-порталы). 1/3 респондентов (29,8%) предпочитают получать информацию посредством телевидения.

Согласно Индексу доверия каналам распространения информации наибольшим доверием у населения пользуются телевидение, интернет-порталы и печатные издания. Среди текстовых информационных источников новостные интернет-порталы занимают первое место по индексу доверия населения с незначительным отрывом от печатных изданий, уступив лишь телевидению, которое не является текстовым источником информации (таблица 6).

Таблица 6 - Индексы доверия каналам распространения информации

	Индекс	**Ранг**
Индекс доверия телевидению	0,60	1
Индекс доверия интернет-сайтам	0,49	2
Индекс доверия печатным изданиям	0,48	3
Индекс доверия радио	0,43	4
Индекс доверия частным беседам	0,32	5
Индекс доверия социальным сетям	0,24	6

Большинство казахстанцев (76%) 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.). Выявлено, что, чем старше респондент, тем больше он доверяет официальным СМИ. И наоборот, чем моложе респондент, тем чаще он доверяет мнению знакомых и друзей из социальных сетей. Одной 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 чаще в 1,5 раза, чем в зарегистрированных СМИ (38% и 21% соответственно).

2.2 Описание структуры оценки

Методика представляет собой комплекс взаимосвязанных процедур (1) сбора текстовой информации, (2) тематического моделирования, (3) анализа текстовых данных по распознаванию информативных признаков и (4) выдачи аналитики по результатам анализа.

На рисунке 4 представлена общая концептуальная схема оценки новостных текстов.

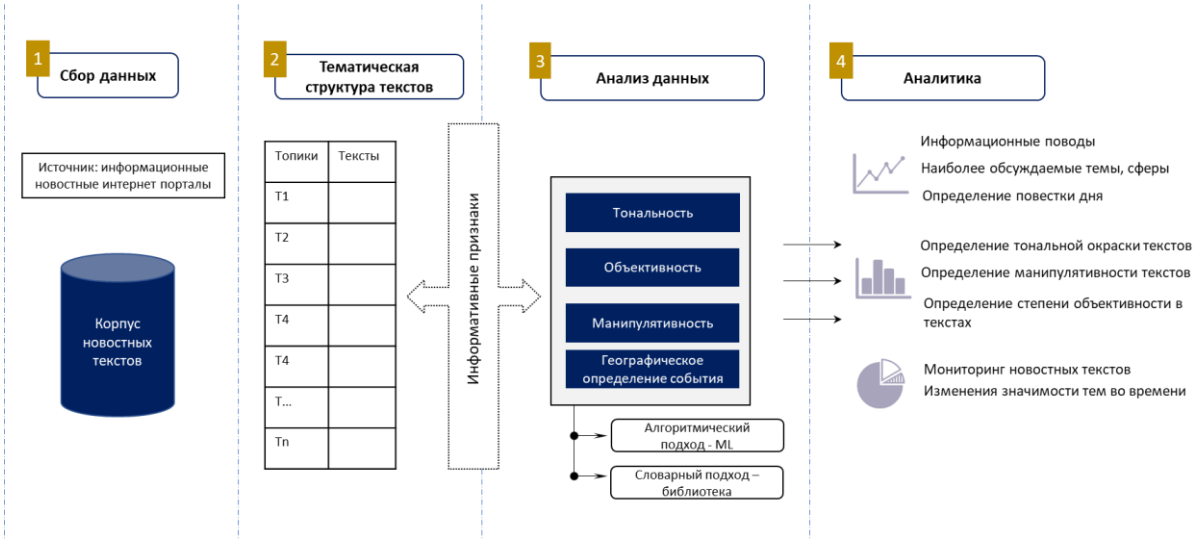

Рисунок 4 – Концептуальная схема оценки новостных текстов

На первом этапе осуществляется сбор новостных текстов из открытых текстовых казахстанских информационных интернет-источников с применением автоматических способов извлечения текстов или путем загрузки готовых коллекций (корпусов) данных.

На следующем этапе осуществляется *тематическое моделирование*. Большинство моделей разрабатываются на основе латентного размещения Дирихле LDA. Модель латентного размещения Дирихле (Latent Dirichlet Allocation) была предложена Дэвидом Блеем в 2003 году. [5] Это современный активно развивающийся вероятностный

инструментарий, который применяется в задачах анализа данных [6]. Для решения данной задачи тематические модели оказались полезными инструментами [7].

В контексте новостных средств массовой информации тематическое моделирование, существуют два основных типичных варианта его применения: анализ повестки дня СМИ и анализ новостных фреймов. Само тематическое моделирование, как процесс, является одним из инструментов автоматизированного анализа корпуса новостных текстов и оценки их влияния на социум [8].

Тематическое моделирование при анализе новостной информации может быть применено для различных задач, как в исследовательском анализе в области социальных наук, так и при решении практических задач в государственном и частном секторах. Результаты тематического моделирования могут быть полезны при определении повестки дня СМИ, информационных поводов и изменения значимости тем во времени [7].

На следующем этапе осуществляется распознавание информативных признаков тональности, объективности, манипулятивности в новостных текстах. Распознавание проводится с применением (1) алгоритмического и (2) словарного подходов. В частности, применяются алгоритмы машинного обучения такие как Random Forest, SVM, XGboost и KNN. При втором подходе используются сформированные экспертами-лингвистами словари по информативным признакам.

На завершающем этапе осуществляется визуализация результатов анализа данных с приведением аналитики. Результаты анализа могут определить информационные поводы, тематические сферы, наиболее обсуждаемые тематики, тональную окраску новостей, степень объективности и манипулятивности текстов.

2.3 Решение проблем несбалансированности размеченного корпуса новостных текстов

В процессе распознавания тональности новостных текстов исследователи зачастую сталкиваются с проблемой несбалансированности разметки корпусов. Наблюдаемый перекос в разметке текстов может объясняться отражением объективной реальности, когда преобладание нейтральных публикаций обусловлено принципами журналистской этики. В свою очередь, преобладание публикаций с негативной тональностью над позитивной можно объяснить их большей популярностью у аудитории и способностью к вирусному распространению, как было доказано в исследовании Хансена и соавторов [9]. Таким образом, для решения вопроса классификации текстов в несбалансированных корпусах, предлагается рассмотреть возможность применения метода ADASYN.

Существует множество метрик оценки результативности машинного обучения, такие как confusion matrix (матрица ошибок), precision-recall (повышенная точность-полнота), F1

score (F-мера) и другие. При работе с несбалансированными данными показатели точности не соответствуют результативности машинного обучения при решении задачи классификации миноритарных классов из-за неравномерного распределения данных.

Несбалансированность корпусов характерна не только для корпуса, собранного в ходе данного исследования, но и для аналогичных корпусов текстов различного происхождения. К примеру, корпус новостных публикаций на русском языке, предоставляемый в открытом доступе, который был собран в ходе проекта Linis-crowd, реализованного Лабораторией интернет-исследований Высшей школы экономики (Россия, Санкт-Петербург) [10]. В данном корпусе соотношение позитивных, негативных и нейтральных публикаций составляет 9,8%, 38%, и 52,2% соответственно, что также указывает на его несбалансированность по признаку «тональность». Схожая тенденция наблюдается в зарубежных корпусах новостных текстов доступных в открытой базе массивов данных на различных иностранных языках kaggle.com. Необходимо учитывать, что несбалансированность корпуса характерна прежде всего для необработанных корпусов. Как правило, для целей машинного обучения корпуса проходят предварительную чистку, в ходе которой часть текстов, способная стать причиной несбалансированности удаляется.

При работе с несбалансированными массивами данных, преобладание мажоритарных классов оказывает значительное влияние на результаты машинного обучения. При этом точность (accuracy) машинного обучения становится ненадежной метрикой его результативности. Так, расчет точности машинного обучения на несбалансированных данных зачастую дает приемлемый и даже высокий уровень (от 80% и выше), но при этом другие метрики результативности машинного обучения, такие как матрица ошибок (confusion matrix) [11] повышенная точность (presicion), полнота (recall), [12], и F-мера [13] (F1 score), демонстрируют противоположную картину. Неравномерное распределение данных приводит к неспособности алгоритмов распознавать и классифицировать миноритарные классы текстов.

Наука о данных выработала несколько стратегий для нивелирования эффекта, оказываемого несбалансированностью, на результаты машинного обучения. Основная стратегия заключается в изменении баланса классов. Наиболее широко применимым подходом для решения указанной проблемы, в силу своей практичности и эффективности, является применение различных стратегий балансирования выборки – сэмплинг (sampling), которые можно разделить на две группы: случайные и специальные. Сэмплинг используется для изменения распределения классов и называется восстановлением равновесия (rebalancing) с целью получения более сбалансированного обучающего множества [14].

Восстановление баланса классов может проходить тремя путями. В первом случае удаляют некоторое количество примеров мажоритарного класса андерсэмплинг (undersampling), во втором – увеличивают количество примеров миноритарного оверсэмплинг (oversampling) [15]. В первой ситуации случайным образом выбирается n записей миноритарного класса и их полностью копируют, во второй – удаляют k записей мажоритарного класса. Третий подход составление сбалансированных подмножеств из исходного массива данных, к которым относится бустинг (boosting) [16].

В данном случае причиной отказа от андерсэмплинга стал размера корпуса, который для целей машинного обучения является небольшим. Составление сбалансированных подмножеств в нашем случае не дало существенного улучшения точности и других метрик результативности машинного обучения. Исходя из этого, было принято решение использовать оверсэмплинг.

Среди множества методов и техник оверсэмплинга, одной из сравнительно новых групп методов является генерирование синтетических данных (synthetic data generation). Генерирование синтетических данных является техникой схожей с бустингом, при которой машина искусственно генерирует данные на основе исходного массива. Среди методов генерирования синтетических данных можно отметить технику синтетического оверсэмплинга миноритарных классов SMOTE (Synthetic minority oversampling technique), введенную Чавла и соавторами [17]. SMOTE предусматривает случайную интерполяцию пар ближайших соседей в миноритарном классе, что приводит к увеличению размера подмножества, составляющего миноритарный класс. Введение техники SMOTE стало импульсом для разработки новых техник, основанных на совершенствовании данного метода. К ним относятся такие методы как SMOTEBoost [18], Borderline-SMOTE [19], RWO-sampling [20], DataBoost-IM [21]. SMOTE также является предшественником метода ADASYN, разработанного Хе и соавторами [22]. Отличие метода от исходной техники SMOTE состоит в том, что ADASYN генерирует данные адаптивным способом, при котором количество генерируемых данных происходит на основе весов, определяемых в зависимости от распределения данных и уровня трудности обучения на миноритарных классах.

В данном случае использование ADASYN рекомендуется ввиду того, что при небольшом размере корпуса, он позволяет проводить машинное обучение без удаления основной части элементов выборки (модель становиться более устойчивой к вероятности ошибки), вместо того, чтобы использовать генерирование синтетических данных для простого увеличения размерности выборки.

Ключевой идеей алгоритма ADASYN является использование плотности распределения Γ в качестве критерия для автоматического принятия решения о необходимом количестве для каждого элемента миноритарного подмножества путем адаптивного изменения весов различных элементов миноритарного подмножества, что в итоге компенсирует неравномерное распределение (несбалансированность) данных.

Для подтверждения эффективности данного способа решения проблемы несбалансированности было проведено экспериментальное исследование. Эксперимент проводился на вышеуказанном корпусе. В ходе исследования на исходном корпусе были получены такие метрики результативности машинного обучения алгоритмами K-ближайших соседей (K-nearest neighbors, K-NN) и наивный Байес (Naïve Bayes, NB) как accuracy, precision-recall, F1 score.

Результаты машинного обучения до процедуры ADASYN для алгоритмов K-NN и Naïve Bayes представлены ниже в таблицах 7 и 8 соответственно.

Таблица 7 - Метрики результативности машинного обучения с применением алгоритма K-NN до процедуры ADASYN

K-NN ML			
accuracy	78%		
precision (neg)=	43%	recall (neg)=	24%
precision (neut)=	81%	recall (neut)=	94%
precision (pos)=	25%	recall (pos)=	4%
precision (macro avg)=	50%	**recall (macro avg)=**	41%
F1 score	42%		

Таблица 8 - Метрики результативности машинного обучения с применением алгоритма Naïve Bayes до процедуры ADASYN

NB ML			
accuracy	79%		
precision (neg)=	46%	recall (neg)=	33%
precision (neut)=	82%	recall (neut)=	94%
precision (pos)=	0%	recall (pos)=	0%
precision (macro avg)=	43%	**recall (macro avg)=**	42%
F1 score	42%		

Из таблиц выше видно значительное расхождение между метрикой accuracy и метриками precision-recall, F1 score. При этом если точность машинного обучения accuracy составила почти 80% для обоих алгоритмов, то остальные метрики показывают, что машинное обучение результативно только при определении класса публикаций «нейтральные». Класс «позитивные» либо не распознается машиной вовсе, либо в очень

редких случаях. В случаях с классом «негативные» машина правильно определяет 20-40% случаев.

Матрицы ошибок на тестовом подмножестве (30%) для алгоритмов K-NN и Naïve Bayes представлены ниже в таблицах 9 и 10 соответственно.

Таблица 9 - Матрица ошибок с применением алгоритма K-NN до процедуры ADASYN

		Предсказанный класс		
		негативный	**нейтральный**	**позитивный**
Действительный класс	негативный	33	105	1
	нейтральный	40	777	8
	позитивный	3	73	3

Таблица 10 - Матрица ошибок с применением алгоритма Naïve Bayes до процедуры ADASYN

		Предсказанный класс		
		негативный	**нейтральный**	**позитивный**
Действительный класс	негативный	38	55	6
	нейтральный	150	654	98
	позитивный	0	0	0

Вышеприведенные таблицы подтверждают наблюдение о неспособности машины правильно определять миноритарные классы. По алгоритму Naïve Bayes машинное обучение абсолютно не справилось с задачей определения публикаций, относящихся к классу «позитивные».

Далее была проведена процедура ADASYN, позволившая сбалансировать корпус. Характеристики корпуса после процедуры ADASYN представлены ниже в таблице 11.

Таблица 11 - Разметка корпуса по признаку «тональность» после ADASYN

Класс	Количество	%
Негативные	3333	33,24
Нейтральные	3295	32,86
Позитивные	3400	33,91
Итого:	10028	100

Затем были вычислены метрики результативности машинного обучения на корпусе, расширенном за счет процедуры ADASYN для алгоритмов K-NN и NB, представленные в таблицах 12 и 13 соответственно.

Таблица 12 - Метрики результативности машинного обучения с применением алгоритма K-NN после процедуры ADASYN

K-NN ML (ADASYN)			
accuracy	52%		
precision (neg)=	23%	recall (neg)=	31%

precision (neut)=	84%	recall (neut)=	65%
precision (pos)=	15%	recall (pos)=	22%
precision (macro avg)=	40%	**recall (macro avg)=**	48%
F1 score	39%		

Таблица 13 - Метрики результативности машинного обучения с применением алгоритма Naïve Bayes после процедуры ADASYN

NB ML (ADASYN)			
accuracy	33%		
precision (neg)=	39%	recall (neg)=	50%
precision (neut)=	86%	recall (neut)=	24%
precision (pos)=	11%	recall (pos)=	87%
precision (macro avg)=	45%	**recall (macro avg)=**	34%
F1 score	34%		

Если до процедуры ADASYN наблюдалось ярко выраженное несоответствие между показателями accuracy и такими метриками, как presicion-recall, F1 score так разница между accuracy и F1 составила 36% и 37% для алгоритмов K-NN и NB соответственно, то после процедуры ADASYN эта разница существенно сократилась, составив 13% и 1% для алгоритмов K-NN и NB соответственно.

Матрицы ошибок на тестовом подмножестве (30%) для алгоритмов K-NN и Naïve Bayes представлены ниже в таблицах 14 и 15 соответственно.

Таблица 14 - Матрица ошибок с применением алгоритма K-NN после процедуры ADASYN

		Предсказанный класс		
		негативный	**нейтральный**	**позитивный**
Действительный класс	негативный	67	52	20
	нейтральный	214	437	174
	позитивный	12	33	34

Таблица 15 - Матрица ошибок с применением алгоритма Naïve Bayes после процедуры ADASYN

		Предсказанный класс		
		негативный	**нейтральный**	**позитивный**
Действительный класс	негативный	69	31	39
	нейтральный	100	202	523
	позитивный	7	3	69

Вышеприведенные таблицы подтверждают, что с проведением процедуры уравновешивания классов ADASYN машина смогла лучше определять миноритарные классы.

Необходимо отметить, что хотя после процедуры ADASYN наблюдается резкое снижение точности машинного обучения (accuracy): на 26% и 46% для алгоритмов K-NN и NB соответственно, точность машинного обучения (accuracy) после процедуры ADASYN гораздо более соответствует фактической результативности машинного обучения, чем до данной процедуры.

Таким образом, было получено экспериментальное подтверждение эффективности метода ADASYN для решения проблемы несбалансированности корпуса по признаку «тональность», хотя точность машинного обучения (accuracy) осталась низкой после процедуры.

2.4 Определение тональности новостных текстов

Определение тональности текста или настроений находится в фокусе многих исследователей ввиду ее практической и научной значимости. Существует несколько подходов к распознаванию тональности текста. Первый, заключается в применении алгоритмов машинного обучения для классификации текстов по категориям положительная или отрицательная. Второй - в применении лексико-ориентированного подхода с использованием размеченных словарей слов.

Изучение специфики новостной области показывает, что определенные тематические области, как, например, происшествия и преступления, по своей сути связаны с негативным тоном. Наряду с этим специфические особенности контента, такие как манипулятивность, тесно связаны с апеллированием к эмоциям аудитории. Все это привело к предположению, что тематическая область и другие особенности контента публикации могут быть релевантными характеристиками для анализа тональности, способствующими повышению точности классификации.

Новый набор признаков включает в себя политизированные, манипулятивные, предвзятые слова, слова, призывающие к действию, а также 17 словарей из разных тематических областей (происшествия, образование, здравоохранение, оборона и др.). Для проверки предположения используются четыре популярных алгоритма машинного обучения с учителем, и проводится три эксперимента по классификации классов на трех размеченных корпусах новостных текстов. В результате предложенные наборы функций показали последовательное улучшение распознавания тональности на трех корпусах.

1. Обзор научной литературы

Одна из первых попыток распознавания тональности текстовых данных предлагает использование бинарной классификации обзоров фильмов [23]. Используя униграммы, биграммы и положение слов в тексте исследователи Панг и Ли применили стандартные алгоритмы машинного обучения для классификации обзоров фильмов на положительную

или отрицательную категории. Последующие попытки были направлены на повышение точности классификации, а также применении более сложных алгоритмов машинного обучения. Например, также они извлекли только те части документов, где содержалось субъективное мнение авторов, для моделирования полярности тональности, тем самым повысив точность определения тональности с 82,8% до 86,4% [24]. Другая попытка сместила фокус определения тональности с бинарной классификации на мультиклассовую классификацию обзоров фильмов с использованием пятибалльных рейтинговых шкал [25]. Достигнув хороших результатов на уровне документа, последующие исследования переключились на выявление тональности на уровне предложений. Например, МакДональд и соавторы предложили совместную классификацию документов и предложений, которая позволила снизить погрешность определения [26]. Икеда и соавторы использовали сдвиги полярности слов в предложениях, что позволило включить контекстную информацию в алгоритмы определения тональности [27]. Метод улучшил точность, особенно на уровне предложения с использованием небольшого количества обучающих данных. Маас и соавторы предложили комбинацию методов обучения с учителем и без учителя для определения тональности на уровне документ-слово [28]. Авторы использовали вероятностную модель представления слов в документах, а затем использовали логистическую регрессию для определения тональности слов. Ся и соавторы сосредоточились на проблеме распознавания изменения полярности [29]. В частности, они разработали трехэтапную процедуру, которая, во-первых, делит документы на суб-предложения и статистически выявляет изменения полярности, во-вторых, устраняет отрицания и, в-третьих, использует различные комбинации взвешенных классификаторов для классификации тональности. Следует отметить, что почти все вышеупомянутые исследования в основном опирались на традиционные признаки, полученные с помощью «мешка слов» и N-граммы. Последующие попытки расширили число возможных признаков и включили обучающие наборы признаков частеречной разметки. Такие признаки, как количество глаголов, существительных, местоимений, прилагательных, предлогов и местоимений, позволили достичь точности классификации в 90%. Более поздние исследования в области анализа тональности сместились к включению методов глубокого обучения, дальнейшей классификации на уровне предложений, а также тестирования наиболее эффективных подмножеств из существующих наборов признаков [30, 31, 32, 33, 34].

Параллельно с машинным обучением в научной литературе разрабатывался лексико-ориентированный подход. Вместо того чтобы полагаться на строгие алгоритмы машинного обучения и стандартные функции, сторонники лексического подхода широко

использовали размеченные словари слов. Например, Терни и Литтман использовали General Inquirer Lexicon для проверки результатов алгоритма обучения семантической ориентации слов [35]. Денеке, используя SentiWordNet, предложил методологию классификации полярности в многоязычной среде [36]. Динг и соавторы проанализировали контекстно-зависимые слова и частные лингвистические конструкции, которые трудно обнаружить с помощью алгоритмов обучения [37]. Табоада и соавторы разработали новую меру, измеряющую семантическую ориентацию субъективности и тональности в текстах [38]. Другие авторы коллективы ученых использовали комбинированные методы лексики и машинного обучения для создания более эффективной комбинированной процедуры, которая могла бы работать лучше, чем любой из двух подходов по отдельности [39, 40]. Кроме того, несмотря на существование надежных источников лексиконов (SentiWordNet, SentiNet, LIWC), ученые разработали новые модели, такие как VADER, которые достигают точности классификации в 0,96 (F1) [41]. Модель не только относительно точно определяет тональность новостных текстов, но также может быть использована и в других областях.

Тема распознавания тональности не нова, и предыдущие исследования достигли впечатляющих результатов в точности алгоритмов машинного обучения и словарей на основе лексики. Однако при работе с интернет СМИ важно использовать как тематическую область, так и информационный контент. В связи с этим предлагается использовать словари по 17 тематическим сферам и содержательными информативными признаками, такими как политизированность, предвзятость, призыв к действию, манипулятивные слова и тональность.

Хотя предыдущие исследования продемонстрировали возможность и эффективность комбинированных моделей тональности и текстовых тематик, тем не менее, данные попытки в основном были сфокусированы на включение алгоритмов распознавания тональности в рамках тематического моделирования. Насколько известно, нет ни одного исследования, которое бы явно основывало классификацию тональности на тематических областях в области интернет СМИ или социальных исследований. С точки зрения классификации, основанной на тематическом моделировании наиболее близкими работами, являются попытка провести анализ тональности на основе тематического моделирования LDA для прогнозирования фондового рынка [42], а также статья Кольцовой и соавторов, где авторы использовали тематическое моделирование LDA в процессе компиляции словарей тональности [10]. С точки зрения эксперимента с существующими функциями, самое близкое исследование было проведено в 2014 г., где авторы эмпирически проверили обширный набор признаков, собранных на основе предыдущих исследований. Однако вышеупомянутые исследователи также упустили из виду важность тематических

областей и содержательных информативных признаков для классификации тональности. Таким образом, предлагается проверить эффективность данных признаков с использованием четырех алгоритмов машинного обучения: Random Forest, Support Vector Machines, KNN и XGboost.

2. Признаки

Ченло и Лосада в своем эмпирическом исследовании классификации тональности и субъективности попытались собрать полный набор различных существующих признаков. Предлагаемый набор включает в себя словари, части речи, синтаксические структуры, позиционные признаки, лексические признаки, длину знаков, концепты и признаки теории риторической структуры. Авторы проверили их особенности для того, чтобы получить субъективность и полярность на уровне предложений. Группа лексических признаков включала в себя количество положительных и негативных слов, соотношение субъективных и отрицательных терминов, а также вопросительные и восклицательные предложения. В целом представляется разумным следовать группировке признаков, предложенной Ченло и Лосада, однако необходимо их расширение/исключение некоторых избыточных групп. Поскольку анализ концентрируется на уровне текста, а не на уровне предложения, мы не включаем позиционные признаки, а также признаки, полученные с помощью риторическо- структурной теории (RST features). Кроме того, необходимо исключить синтаксические структуры, потому что они в основном используются для анализа мнений из текстов. Большая доля предикторов получена с использованием словарного подхода, вычисляющего встречаемость слов, входящих в определенный словарь для каждого документа. Для составления словарей тематических сфер информативных признаков были привлечены эксперты в области лингвистики русского языка. Словари были составлены на основе углубленного лингвистического анализа 2000 новостных текстов, отобранных методом случайной выборки. Таким образом, конечные наборы признаков выглядят следующим образом:

2.1) Словарные признаки

Эта группа объединяет N-граммы, рассчитанные на основе встречаемости униграмм (1), биграмм (2) и количества уникальных слов (3). Другие 4 предиктора в этой группе были получены во время процедуры text-mining с помощью программного обеспечения Statistica[5], которое представляет собой автоматическую обработку текстовой информации, приводящую к цифровому представлению документов путем определения нескольких

[5] http://statsoft.ru

наиболее важных измерений (называемых «концептами»), которые имеют наибольшую вариативность между словами и документами – 4 концепта (4-7).

2.2) Лексические признаки

Большинство лексико-семантических признаков также были получены с помощью подхода «мешок слов» и включают количество точек (8), запятых (9), восклицаний (10), вопросов (11), многоточий (12), кавычек (13), скобок (14), тире (15) и дефисов (16). Два других предиктора, подпадающих под эту категорию, были получены с помощью словарного подхода и представляют собой встречаемость слов из словаря тональности в документах – позитивные (17) и негативные (18) слова. Словарь тональности включал 979 позитивных и 2112 негативных слов.

2.3) Признаки объема документа

Предикторы этой группы включают символы (19), слова (20) и предложения (21). Количество символов было рассчитано с учетом пробелов.

2.4) Части речи (PoS tagging)

Эта группа объединяет количество прилагательных (22), предлогов (23)[6], наречий (24), вспомогательных слов (25), координирующих союзов (26), определителей (27), междометий (28), существительных (29), чисел, включая цифры (30), частиц (31), местоимений (32), имен собственных (33), знаков препинаний (34), подчинительных союзов (35), глаголов (36). Для проведения PoS tagging была использована библиотека UDPipe [43], свободно доступная для языка статистического программирования R.

2.5) Тематические области

Тематические области были вычислены на основе встречаемости слов из словарей тематических сфер. Они включали словари по таким темам, как административно-территориальные вопросы (123 слова), экономика и финансы (238 слов), внешняя политика (95 слов), гражданское общество (188 слов), здравоохранение (141 слово), индустриально-инфраструктурное развитие и технологии (115 слов), оборона и безопасность (77 слов), образование и наука (160 слов), правительство и парламент (97 слов), происшествия и преступления (326 слов), сельское хозяйство (61 слово), социальное обеспечение (28 слов), спорт, культура и развлечения (243 слова), труд (25 слов), энергетика (70 слов), экология и туризм (40 слов), чрезвычайные ситуации (53 слова). Перечень тематических областей был составлен на основе анализа тематических тегов, заданных новостными сайтами, с последующим проведением экспертных консультаций (7 экспертов в области средств

[6] https://universaldependencies.org/u/pos/ADP.html

массовой информации). Эта группа предикторов объединяет 17 соответствующих признаков (37-50), рассчитанных по встречаемости слов из этих словарей.

2.6) Информативные признаки контента документа

Данные характеристики были измерены с помощью 4 предикторов: политизированные (51) и манипулятивные (52) слова, слова, призывающие к действию (53) и предвзятые слова (54), рассчитанных на основе встречаемости слов, включенных в соответствующие словари, в соответствии с вышеупомянутым углубленным лингвистическим анализом.

Выбор этих признаков обосновывается обширным обзором литературы, в которой установлены взаимосвязи между тональностью и фейковыми (ложными) новостями [44, 45, 46] и между фейком и политизированностью [47, 48, 49] (косвенные связи между политизированностью и тональностью); взаимосвязь между тональностью и манипулятивностью установлена в ходе вышеупомянутого лингвистического анализа и проявлена в значительном количестве научных работ в области когнитивной психологии и журналистики [50, 51, 52, 53, 54, 55]. Призыв к действию как вид манипуляции был также выделен в отдельный признак. Включение предвзятости в группу информативных признаков также основывалось на литературном обозрении [56, 57, 58, 59, 60].

Таким образом, полученные предикторы были сгруппированы в 3 группы признаков, где первый набор признаков включает словарь, лексику, признаки объема документа и функции PoS tagging; группа признаков два - набор признаков 1 и тематические области; и третья группа признаков объединяет предыдущие два, а также информативные признаки, как показано в таблице 16. Такая группировка позволяет проследить улучшение классификации тональности с использованием тематических областей и особенностей контента текстов.

Таблица 16. Группы признаков

<table>
<tr><td rowspan="14">Корпус 1</td><td rowspan="14">Корпус 2</td><td rowspan="14">Корпус 3</td><td colspan="4">Словарь</td></tr>
<tr><td>Униграмы (1)</td><td colspan="2">Биграмы (2)</td><td>Уникальные слова (3)</td></tr>
<tr><td>Концепты (4-7)</td><td colspan="2"></td><td></td></tr>
<tr><td colspan="4">Лексические признаки</td></tr>
<tr><td>Точки (8)</td><td>Вопр. знаки (11)</td><td>Скобки (14)</td><td>Позитивные слова (17)</td></tr>
<tr><td>Запятые (9)</td><td>Многоточия (12)</td><td>Тире (15)</td><td>Негативные слова (18)</td></tr>
<tr><td>Воскл. знаки (10)</td><td>Кавычки (13)</td><td>Дефисы (16)</td><td>-</td></tr>
<tr><td colspan="4">Признаки длины</td></tr>
<tr><td>Кол-во слов (19)</td><td colspan="2">Кол-во предложений (20)</td><td>Кол-во знаков (21)</td></tr>
<tr><td colspan="4">Разметка частей речи</td></tr>
<tr><td>Прилагательные (22)</td><td>Соединительные союзы (26)</td><td>Числительные (30)</td><td>Знаки пунктуации (34)</td></tr>
<tr><td>Предлоги (23)</td><td>Определения (27)</td><td>Частицы (31)</td><td>Подчинительные союзы (35)</td></tr>
<tr><td>Наречия (24)</td><td>Междометия (28)</td><td>Местоимения (32)</td><td>Глаголы (36)</td></tr>
</table>

			Вспом-ые слова (25)	Существительные (29)	Имена собственные (33)	-
X			*Тематические области*			
			Административные и территориальные вопросы (34)	Индустриально-инфраструктурное развитие и технологии (39)	Сельское хозяйство (44)	Экология и туризм (49)
			Экономика и финансы (35)	Оборона и безопасность (40)	Социальное обеспечение (45)	Чрезвычайные происшествия (50)
			Внешняя политика (36)	Образование и наука (41)	Спорт, культура и развлечения (46)	-
			Гражданское общество (37)	Правительство и парламент (42)	Труд (47)	-
			Здоровье и здравоохранение (38)	Преступления и происшествия (43)	Энергетика (48)	-
	X		*Содержательные информативные признаки*			
			Политизированность (51)	Манипуляция (52)	Призыв к действию (53)	Предвзятость (54)

3. Данные

Для проведения эксперимента было использовано 3 корпуса новостных текстов. Корпус 1, включающий 5211 новостей, был собран с 5 казахстанских новостных сайтов, отличающихся по уровню популярности, стилю представления информации и целевой аудитории. Оценка тональности проводилась с использованием Google Cloud Natural Language API [61]. Корпус 2 был взят из ресурса Kaggle (https://www.kaggle.com/), содержащий 8260 новостей, собранных из различных онлайн-социальных сетей [62]. Корпус 3, включил в себя 32292 новости, размеченные алгоритмом, предложенным в работе Кольцовой [10].

Поскольку тональность Google Sentiment имеет множество категорий в диапазоне [-1:1], необходимо было перекодировать значения в более простые категории, такие как положительная, отрицательная и нейтральная. В противном случае, - перед исследователем встает задача мульти классификации или регрессии. Можно заметить, что интенсивность не обеспечивает точных пороговых значений для классификации, а скорее указывает на то, что значения, близкие к нулю, единице или минус одному, выражают более нейтральную, более негативную или более позитивную тональность соответственно. В начале предполагалось использовать порог 0.1, однако итоговая оценка тональности показала несбалансированность позитивных (775), негативных (868) и нейтральных (3567) категорий. Для того чтобы избежать проблемы несбалансированности был выбран более строгий порог 0,0. Кроме того, оценка тональности в корпусе 3 была упрощена до трех категорий, поскольку она имела диапазон от -2 до +2. В таблице 17 приведено количество позитивных, нейтральных и негативных новостей во всех 3 корпусах.

Таблица 17 - Корпуса новостных текстов

Корпус	Категория	Количество
1) 5211 публикаций	Позитивная	1436
	Нейтральная	2068
	Негативная	1707
2) 8260 публикаций	Позитивная	2794
	Нейтральная	4032
	Негативная	1434
3) 32292 публикаций	Позитивная	3180
	Нейтральная	16856
	Негативная	12256

Из таблицы видно, что два корпуса показывают дисбаланс в сторону нейтральной категории. Первый корпус имеет лучшую представленность негативных и позитивных текстов, хотя нейтральная категория по-прежнему имеет большее число. Такой дисбаланс может быть результатом неоднозначной оценки тональности экспертами и проблем перевода, связанных с использованием Google Sentiment API. Однако мы склонны думать, что несбалансированность просто отражает истинную картину тональности публикуемых новостей. Таким образом, новостные порталы изобилуют публикациями с простой фактологической информацией и имеют гораздо меньше публикаций с позитивной и негативной тональностью.

4. Методы

Чтобы проверить эффективность наших тематических областей и содержательных информативных признаков, были использованы четыре популярных алгоритма машинного обучения с учителем: Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) и XGBoost (XGB). Из-за проблемы дисбаланса логично придерживаться разных стратегий классификации. Для RF и SVM был применен подход «один против всех» и тональность перекодирована в бинарные категории, такие как «позитивная против всех», «негативная против всех» и «нейтральная против всех». Учитывая, что подход «один против всех» дает еще более несбалансированные результаты, эксперимент был дополнен такими алгоритмами KNN и XGboost, используя все три класса без их перекодирования. Все результаты были получены с помощью 5-кратной перекрестной проверки.

5. Эксперимент

Бинарная классификация. Эксперимент начинается с применения алгоритмов Random Forest (случайный лес) и SVM (метод опорных векторов) на корпусе 1 (5211 текстов). Процедуры проводились отдельно по каждому набору признаков. Первый набор состоял из словаря, частей речи, длины и признаков лексики, таким образом, общим количеством 33 признака. Второй набор добавил тематические области и, следовательно, увеличил набор на 17 переменных. Третий набор включал все остальные содержательные

информативные признаки с конечным количеством 55 предикторов. Для расчета результатов была использована 5-кратная перекрестная валидация и тестировались различные числа деревьев для проведения Random Forest, а также линейные, полиномиальные и радиальные ядра при проведении алгоритма SVM. В таблице 18 приведены отдельные занятия и соответствующие результаты процедур машинного обучения. Третий столбец содержит параметры для алгоритмов SVM. В таблице оставлены только те параметры, которые дали лучшие результаты точности (accuracy), F1 score, precision и recall.

Таблица 18 – Эксперимент на Корпусе 1

Class	Features	model	accuracy	F1	Precision	Recall
Pos vs all (позитивные против всех)	Набор признаков 1	RF	81.09	87.80	82.82	93.41
		SVM (poly, Degree=3, Scale=3, C=0.25	82.82	88.70	85.10	92.62
	Набор признаков 2	RF	82.05	88.20	83.01	94.07
		SVM (lin, C=1)	82.72	88.38	84.88	92.19
	Набор признаков 3	RF	84.35	89.78	86.05	93.84
		SVM (poly, Degree=1, Scale=0.1, C=1)	84.83	90.00	86.84	93.97
Neg vs all (негативные против всех)	Набор признаков 1	RF	78.11	84.92	80.65	89.66
		SVM (lin, C=1)	79.27	86.06	79.97	93.15
	Набор признаков 2	RF	78.50	85.16	81.08	89.67
		SVM (lin, C=1)	78.11	85.13	79.92	91.07
	Набор признаков 3	RF	79.84	85.88	82.85	89.10
		SVM (poly, Degree=1, Scale=0.1, C=1)	80.99	86.90	82.53	91.75
Neut vs all (нейтральные против всех)	Набор признаков 1	RF	63.62	73.95	64.50	86.63
		SVM (rad, Sigma=0.2155, C=1)	62.85	73.43	63.99	86.15
	Набор признаков 2	RF	65.45	74.96	65.97	86.79
		SVM (poly, Degree=3, Scale=0.1, C=1)	63.05	72.44	65.20	81.48
	Набор признаков 3	RF	68.13	76.88	71.04	83.76
		SVM (rad, Sigma=0.079, C=1)	68.42	77.16	71.09	84.37

В целом, как Random Forest, так и SVM показали, что информативные признаки и тематические области улучшили эффективность классификации во всех классах. В частности, Pos vs all показал последовательное увеличение точности на 3-4%, F1 на 2-3%, precision 4-5%, однако recall показал лишь незначительное улучшение. В случае негативных против всех алгоритмы сигнализируют меньшее улучшение, примерно каждый параметр на 2%. Наибольший прирост наблюдается в случае Neut vs all, например, SVM сигнализировал о повышении точности на 5,74%, F1 на 3,73%, precision увеличился на 7,1%, однако recall несколько снизился с 86,63% до 84,37%. Можно заметить, что наилучший прирост достигается при сочетании тематических областей и содержательных информативных признаков в группе признаков 3.

Таблица 19 – Эксперимент на Корпусе 2

Class	Features	model	accuracy	F1	Precision	Recall
Pos vs all (позитивные против всех)	Набор признаков 1	RF	71.54	80.20	74.49	86.86
		SVM (rad, Sigma=0.2758 C=1	69.00	79.42	70.97	90.14
	Набор признаков 2	RF	75.12	79.42	77.90	93.53
		SVM (rad, Sigma=0.1354, C=1)	73.42	81.71	75.28	89.34
	Набор признаков 3	RF	74.45	81.85	77.58	86.62
		SVM (rad, Sigma=0.1162, C=1)	74.39	82.14	76.55	88.61
Neg vs all (Негативные против всех)	Набор признаков 1	RF	83.53	90.67	84.00	98.51
		SVM (rad, Sigma=0.2762, C=0.5	81.59	89.77	81.85	99.40
	Набор признаков 2	RF	84.62	91.24	84.92	98.58
		SVM (rad, Sigma=0.1252, C=1	83.47	90.69	83.64	99.03
	Набор признаков 3	RF	85.53	91.66	86.73	97.18
		SVM (rad, Sigma=12.12, C=1)	84.50	91.22	84.99	98.44
Neut vs all (нейтральные против всех)	Набор признаков 1	RF	60.35	62.63	61.06	64.28
		SVM (rad, Sigma=0.2938, C=1	58.77	61.84	59.29	64.63
	Набор признаков 2	RF	64.52	66.39	65.05	67.79
		SVM (rad, Sigma=0.1354, C=1)	63.80	66.88	63.44	70.72

	Набор признаков 3	RF	64.83	66.85	65.18	68.61
		SVM (rad, Sigma=0.1212, C=1)	64.52	67.40	64.19	70.96

Эксперимент на Корпусе 2 показал значительные улучшения в классе Pos vs all (таблица 19). Например, SVM увеличил точность на 5,39% и F1 на 2,72%, в то же время, хотя точность улучшилась на 5.58%, recall сигнализировал о снижении на 1.7%. В отличие от позитивной категории негативная категория практически не улучшилась при Random Forest с увеличением параметров менее чем на 1%, в то время как SVM достигла более лучших результатов в точности, F1, precision и recall. Лучшие результаты были получены в нейтральной категории, что свидетельствовало об улучшении параметров как в Random Forest, так и в SVM, с последующим увеличением точности на 5,75% и увеличением F1 на 5,56%. Как и в первом наборе данных, сочетание тематических областей и содержательных специфических информативных признаков повысило классификационную мощность алгоритмов.

Таблица 20 – Эксперимент на Корпусе 3

type	Features	model	accuracy	F1	Precision	Recall
Pos vs all (позитивные против всех)	Набор признаков 1	RF	88.66	93.91	91.02	96.99
		SVM (rad, Sigma=0.0598 C=1)	90.21	94.85	90.24	99.96
	Набор признаков 2	RF	89.23	94.23	91.27	97.39
		SVM (rad, Sigma=0.0312, C=1	90.27	94.88	90.27	1
	Набор признаков 3	RF	89.23	94.23	91.28	97.37
		SVM (rad, Sigma=0.0271, C=0.25)	90.27	94.88	90.27	1
Neg vs all (негативные против всех)	Набор признаков 1	RF	66.28	73.55	71.36	75.88
		SVM (rad, Sigma=0.0596, C=1)	65.12	75.46	66.75	86.79
	Набор признаков 2	RF	67.18	75.84	70.13	82.57
		SVM (rad, Sigma=0.0309, C=1)	67.45	76.24	69.99	83.71
	Набор признаков 3	RF	67.48	75.75	70.81	81.43
		SVM (rad, Sigma=0.0269, C=1)	67.80	76.29	70.56	83.04
		RF	61.41	59.66	59.66	59.66

Neut vs all (нейтральные против всех)	Набор признаков 1	SVM (rad, Sigma=0.0607, C=1)	60.32	52.85	61.23	46.48
	Набор признаков 2	RF	61.93	60.83	60.50	61.16
		SVM (rad, Sigma=0.0312, C=0.5)	61.31	55.99	62.19	50.91
	Набор признаков 3	RF	62.49	61.91	60.79	63.08
		SVM (rad, Sigma=0.0271, C=1)	62.23	58.56	62.32	55.23

В отличие от предыдущих двух корпусов третий корпус показал наименьшие улучшения как в обоих алгоритмах, так и во всех классах (таблица 20). Позитивная категория при Random Forest показала увеличение всех параметров менее чем на 1%, такая же картина наблюдается и при SVM. Негативная категория, в свою очередь, хотя и сигнализировала об улучшении точности при SVM, тем не менее, показала лучшее увеличение F1 при Random Forest. Лучший прирост был, достигнут в нейтральной категории, где оба алгоритма показали последовательные улучшения по всем параметрам.

В целом первая часть эксперимента продемонстрировала эффективность предложенных тематических направлений и содержательных информативных признаков. Хотя включение тематических областей улучшило классификационную способность алгоритмов, тем не менее, именно сочетание двух наборов признаков показало последовательное увеличение во всех трех корпусах. Интересно отметить, что алгоритмы сигнализировали лучшие результаты классификации в обучении нейтральной категории. Мы считаем, что такие результаты обусловлены высокой несбалансированностью размеченных корпусов. Те же рассуждения могут быть применены для объяснения высоких результатов положительной категории в первом корпусе и высоких результатов отрицательной категории во втором корпусе.

Трехклассовая классификация. Как и в предыдущей части эксперимента, были использованы три набора признаков отдельно. Сразу можно заметить, как три класса классификации понижают показатели во всех трех корпусах по сравнению с результатами двоичной классификации. Поскольку фокус направлен на эффективность предлагаемых функций, необходимо обратить внимание на последовательное повышение точности и F1 во всех корпусах и моделях. Например, алгоритм KNN улучшил точность, F1 и precision на 7%, в то время как recall увеличился на 8%. Те же результаты можно наблюдать и во втором наборе данных, где показатели сигнализировали еще больший рост. В частности, KNN и XGboost продемонстрировали повышение точности и F1 на 5% и 11% соответственно.

Третий корпус показал наименьший прирост точности и F1, обеспечивая результаты, аналогичные бинарной классификации (таблицы 21-23).

Таблица 21 – Эксперимент на Корпусе 1

Features	Model	Accuracy	F1	Recall	Precision
Feature set 1	XGboost	58.37	0.59	0.58	0.59
	KNN	45.20	0.45	0.45	0.46
Feature set 2	XGboost	61.12	0.61	0.61	0.61
	KNN	49.36	0.49	0.49	0.50
Feature set 3	XGboost	62.59	0.63	0.63	0.63
	KNN	52.74	0.53	0.53	0.53

Таблица 22 – Эксперимент на Корпусе 2

Features	Model	Accuracy	F1	Recall	Precision
Feature set 1	XGboost	57.46	0.56	0.57	0.59
	KNN	40.23	0.41	0.40	0.41
Feature set 2	XGboost	60.89	0.60	0.61	0.51
	KNN	50.20	0.50	0.50	0.51
Feature set 3	XGboost	62.24	0.62	0.62	0.63
	KNN	51.69	0.52	0.52	0.52

Таблица 23 – Эксперимент на Корпусе 3

Features	Model	Accuracy	F1	Recall	Precision
Feature set 1	XGboost	56.82	0.52	0.57	0.54
	KNN	51.20	0.49	0.51	0.49
Feature set 2	XGboost	58.74	0.55	0.59	0.56
	KNN	52.25	0.50	0.52	0.50
Feature set 3	XGboost	60.14	0.57	0.60	0.57
	KNN	53.16	0.51	0.53	0.51

Таким образом, три класса классификации подтвердили эффективность предложенных наборов признаков. Текстовый анализ средств массовой информации может быть улучшен путем включения информации о текстовых информативных особенностях и тематических областях. При объединении они могут значительно улучшить классификационную мощность алгоритмов распознавания тональности. Мы также считаем, что это сочетание функций может быть применено в многоязычной структуре и особенно в английском языке.

II. Лексический подход к разметке корпуса

Процедура измерения

Помимо выше предложенной методики был проведен комплекс работ по разметке корпусов с помощью сформированных словарей. Алгоритм разметки не требует машинного обучения, однако зависит от качества словаря позитивных и негативных слов.

Полярность тональности. Перед проведением алгоритма разметки была применена процедура распознавания частей речи (PoS tagging) на тестовом корпусе из 35 645 текстов.

Процедура PoS tagging несколько проблематична для русского языка, так как существует не так много надежных библиотек и открытых программных решений обученных словарей. Вследствие этого предпочтение было отдано в пользу библиотеки Udpipe [43] свободно доступной для статистического программного обеспечения C++ и R.

После процедуры PoS tagging тексты были размечены с помощью словаря тональности. Цель состояла в том, чтобы получить индикатор тональности в диапазоне [-1:1], который затем можно было бы использовать как непрерывную или категориальную переменную. В случае категориальной переменной можно либо использовать бинарную переменную, либо назначить несколько категорий, разделив диапазон на 11 групп, причем средняя группа является нейтральной категорией. Полярность документа — это нормализованная сумма полярностей на уровне предложений, тогда как полярность на уровне предложений была рассчитана с использованием отрицателей, усилителей и уменьшителей (таблица 24).

Таблица 24 - Пример конструкций и слов с негативной семантикой: отрицатели, усилители, уменьшители

№	Модели конструкций	примеры
Отрицатели		
1	Не + глагол с положительной семантикой	Не читать, не заступаться, не знал
2	Не + глагол с негативной семантикой	Не убивать, не обманывать
3	Не + модальный глагол + глагол	Не мог выполнить (обещанное), не сумел доказать
Усилители		
4	Слишком + слово с негативной семантикой	Слишком неграмотный
5	Очень + слово с негативной семантикой	Очень плохой
6	Крайне + слово с негативной семантикой	Крайне недружелюбный
Уменьшители		
7	Не + очень + слово	Не очень хороший
8	Почти + не + глагол	Почти не знает
9	Практически + не + глагол	Практически не соответствует

Отрицатели использовались, чтобы изменить полярность слов, если отрицатель был найден до и после 4 слов тональности, он автоматически изменит знак слова тональности с +1 до -1. Рассмотрим следующее простое предложение:

Погода была плохой (без частицы-отрицателя тональность предложения (-1)
Погода была не плохой (с частицей отрицателем, тональность предложения (+1)

Не считая отрицателя «нет», можно ошибочно считать второе предложение позитивным, но присутствие отрицателя явно меняет полярность тональности. Таким же образом были использованы усилители и уменьшители, чтобы максимально точно определить интенсивность тональности.

Погода была весьма неплохой (с увеличителем очень итоговая тональность (+1.5)

Алгоритм очень точно определяет усилитель и присваивает вес 0,5[7], точно так же алгоритм идентифицирует слова, которые уменьшают полярность слов в каждом предложении. Таким образом, результирующая полярность, полученная на уровне предложения, суммировалась на уровне документа в диапазоне [-1: 1], где значение ближе к -1 интерпретируется как публикация с более негативной тональностью, значение ближе к 0 - как публикация с более нейтральной тональностью, а значение ближе к 1 означает публикацию с более положительной тональностью.

2.5. Определение манипулятивности новостных текстов

В современном мире СМИ оказывают серьезнейшее влияние на общество, его граждан, на их политические взгляды и убеждения. Зачастую способы влияния СМИ могут представлять опасность для социума и способствовать социальной нестабильности общества. В связи с этим возникает необходимость выявления механизмов манипуляции СМИ в современном мире.

В современной лингвистической науке явления манипуляции и манипулятивности языка рассматриваются с различных аспектов: когнитивного, прагмалингвистического, дискурсивного, психологического и лингвоюридического подходов. В связи с этим учеными выработан целый ряд определений манипуляции (манипулятивности) как характерной черты медиатекстов [63, 64]. Перечислим некоторые из них.

Наиболее активно в научной литературе манипуляция рассматривается с позиции дискурс-анализа. Под манипуляцией исследователь понимает аргументированный дискурс, в рамках которого власть или политическая сила, выступающая в качестве манипулятора, пытается укрепить свои позиции и воздействовать на реципиента (читателя), находящегося в роли жертвы, не способной противостоять манипуляции в силу определенных факторов контекстуального принуждения [65].

В аспекте когнитивно-прагматических исследований манипуляция рассматривается как определенно выстроенный контекст, при котором из целого ряда возможных пресуппозиций читатель выбирает только один важный и актуальный манипулятору вариант. Остальные пресуппозиции блокируются в сознании читателя в рамках манипулятивно выстроенного контекста [66].

Наиболее изучен феномен манипуляции в психологии. Одной из активно цитируемой учеными дефиницией манипуляции является определение, предложенное Е.Л. Доценко, в котором манипуляция представлена как психологическое воздействие в целях

[7] Параметр присвоения весов может меняться для увеличения влияния частиц уменьшителей или увеличителей

получения одностороннего выигрыша в результате скрытого побуждения кого-либо к действиям, выгодным манипулятору [67].

Расширяя рамки манипуляции за пределы психологического воздействия, В.М. Герасимов и другие рассмаривают ее в качестве определенного метода социального и психологического контроля» [68].

С лингвистической точки зрения при изучении манипуляции представляет большой интерес отбор и использование языковых средств, с помощью которых осуществляется скрытое манипулятивное воздействие на читателя, после чего им воспринимается полученная информация как объективная [69].

На материале лингвоюридических текстов И. Беляева рассматривает манипуляцию как определенный прагматический акт, в котором специально затемнена интенция адресата, в результате которой достигается манипулятивное воздействие на читателя и расширяется «иллюзорная субъективная реальность», воспринимаемая как объективная информация [70].

Исследователь А. Колмогорова, обобщая множество определений манипуляции, отмечает, что манипуляция представляет собой наложение определенных контекстуальных ограничений, в результате чего читатель движется по выстроенной манипулятором траектории посредством использования специфических вербальных маркеров [63].

Другими словами, манипулятивность медиатекста означает наличие в нем определенных лингвистических и экстралингвистических маркеров, с помощью которых осуществляется манипуляция мнением читателя, в результате которой читатель получает только часть информации, выгодную автору, часть информации может быть затемнена, а оказываемое на читателя воздействие воспринимается как объективная информация.

В медиатексте могут применяться различные стратегии и тактики манипуляции.

Экстралингвистические факторы (момент, во время которого наносится информационный «удар», возраст, пол, социальный статус автора текста и объекта манипуляции, социально-экономическая и общественно-политическая обстановка в стране, информационный фон, дистанция между участниками коммуникации и т.д.), хотя и играют заметную роль в продвижении манипулятивного текста и цели, которую установил для себя автор текста, остаются вне поля собственно корпуса языковых единиц и в рамках этого проекта представляют скорее сугубо теоретический интерес и имеют отношение больше к углубленному изучению замысла и правильной подачи (в нужное время и в нужном месте, после определенных событий и т.д.) анализируемого типа текстов.

Главными показателями манипулятивности текста являются собственно лингвистические факторы, к которым относятся языковые единицы разных уровней:

фонетического, лексического, и грамматического (словообразование, морфология, синтаксис и пунктуация текста).

В контексте данного этапа проекта, направленного на разработку *автоматизированной* информационной системы, позволяющей опознавать тексты манипулятивного характера, нами сформулировано рабочее определение манипулятивности.

Под манипулятивностью текста мы понимаем совокупность, определенные комбинации речевых стратегий и тактик, создаваемых с целью неявного воздействия на общественное сознание и выраженных определенными языковыми индикаторами.

Исходя из определения манипулятивности, нами разработан алгоритм анализа медиатекстов, который представляет собой комплекс речевых (стратегии и тактики) и языковых индикаторов с помощью которых они реализуются (лексические единицы и грамматические конструкции).

Материалом для лингвистического анализа является 1000 текстов манипулятивного характера, которые были отобраны из 10 000 текстов путем их изучения и наблюдения над языковым материалом.

Ниже представлен перечень речевых и языковых индикаторов, которые вошли в схему анализа медиатекстов. Следует отметить, что собственно маркерами для обучения машины автоматическому распознаванию манипулятивных текстов являются только языковые индикаторы (слова, словосочетания, грамматические конструкции и т.д.), которые представлены под цифрой II. Речевые индикаторы необходимы для теоретических обобщений и расширения сферы исследования за счет включения в него экстралингвистических факторов, и в конечном итоге внесут вклад в изучение понимания природы манипуляции как междисциплинарного объекта научного исследования.

I. Речевые индикаторы: стратегии и тактики (извлекаются из текстов для теоретических обобщений и разработки методологии исследования)

С точки зрения *объекта речи* выделяются 2 стратегии манипуляции: стратегия дискредитации и стратегия позитивного продвижения.

Стратегия дискредитации направлена снижение авторитета, разоблачение объекта речи (государство, политика, реформы, политические деятели и др.), развенчание его позиции. Данная стратегия реализуется через использование следующих тактик манипуляции:

1. Тактика анализ-«минус» - основанное на фактах рассмотрение ситуации, предполагающее выражение отрицательного отношения к описываемому, а также к людям, их действиям и поступкам.

2. Тактика обвинения (безличного обвинения) - приписывание кому-либо какой-либо вины, признание виновным в чём-либо, а также раскрытие, обнаружение чьих-либо неблаговидных действий, намерений, качеств (обвинение, при котором не указываются виновники осуждаемых действий или поступков).

3. Тактика оскорбления - нанесение обиды, унижение, уязвление кого-либо, сопровождаемое экспликацией эмоционального составляющего компонента вместо приведения доказательств.

4. Тактика угрозы - запугивание, обещание причинить адресату неприятность, зло.

Стратегия позитивного продвижения объекта отражает стремление говорящего максимально увеличить значимость объекта речи. Данная стратегия реализуется через использование следующих тактик манипуляции:

1. Тактика анализ - «плюс» - основанное на фактах рассмотрение, разбор ситуации, предполагающий имплицитное выражение положительного отношения говорящего к описываемой ситуации.

2. Тактика презентации - представление кого-либо в привлекательном виде.

3. Тактика отвода критики - приведение с целью доказательства невиновности аргументов и/или фактов, с помощью которых можно объяснить (оправдать) какие-либо действия и поступки.

4. Тактика самооправдания - отрицание негативных суждений об объекте критики и его причастности к тому, чему даётся отрицательная оценка.

С точки зрения адресата речи (читателя) используются следующие тактики манипуляции:

1. Тактика призыва к действию (в данной стратегии возможна определенная градация: просьба, принуждение, приказ и др.).

2. Тактика обещания - добровольное обязательство сделать что-либо.

3. Тактика прогнозирования - предсказание, суждение о дальнейшем течении, развитии чего-либо на основании интерпретации различных имеющихся данных.

4. Тактика предупреждения - предостережение, предваряющее извещение о возможных событиях, действиях, ситуациях и т.п.

5. Тактика убеждения (аргументации).

6. Тактика формирования благосклонности (завоевание симпатии, расположение).

7. Тактика единения – использование инклюзивного «мы» для формирования чувства единения и солидарности.

8. Тактика политизации – придание политической окраски событиям, не связанным с политикой, а также политическим событиям, которым необъективно придают негативную окраску.

9. Тактика воздействия на сенсорные каналы.

II. Языковые индикаторы (извлекаются из текстов для машинной обработки).

Лексический уровень:

1. Прецедентные имена и сочетания, т. е. широко известные имена, географические наименования, которые вызывают у общественности стойкие ассоциации с событием, символом, набором определенных качеств или ценностей. (Методологический комментарий: Использование прецедентных имен является одним из распространенных инструментов манипуляции, способов задеть за живое, обратить внимание, сформировать определенное отношение к фактам, описываемым в тексте).

2. Слова и словосочетания, отражающие социально значимые объекты, явления и т.д., используемые с целью манипуляции, не относящиеся к данной тематике.

3. Слова с высоким индексом (1) негативной тональности. (Методологический комментарий: Индексация степени тональности вводится в алгоритм анализа медиа-текстов с целью разграничения, собственно, манипулятивных текстов от текстов критического характера. Как показали наблюдения за языковыми единицами в манипулятивных текстах, все такие тексты содержат слова с высоким индексом негативной (позитивной) тональности.

4. Слова со средним и невысоким (2) индексом негативной тональности.

5. Устойчивые выражения, метафоры, штампы с негативной тональностью. (Методологический комментарий: с помощью устойчивых выражений, в основе которых всегда присутствует метафора (образное сравнение), авторы текста оказывают влияние на сознание читателей. Особенно большой силой воздействия на сознание человека в связи с особенностью человеческого мышления наиболее ярко воспринимать негативные оценки, обладают устойчивые выражения с негативной тональностью.

6. Слова с негативной тональностью в заголовке статьи.

7. Слова с негативной тональностью в заголовке статьи, характеризующие страну, государство, правительство, социально значимые объекты, явления и т.д.

8. Слова и словосочетания, отражающие казахстанское содержание текста

9. Слова с негативной тональностью в тексте статьи, характеризующие Казахстан и относящиеся к нему объекты и т.д., (казахстанское содержание).

10. Слова и словосочетания, отражающие социально значимые объекты, явления и т.д.

11. Слова с высоким индексом негативной тональности в тексте статьи, характеризующие страну, государство, правительство, социально значимые объекты, явления и т.д.

12. Слова со средним и невысоким индексом негативной тональности в тексте статьи, характеризующие страну, государство, правительство, социально значимые объекты, явления и т.д.

13. Устойчивые выражения, штампы с негативной тональностью, характеризующие страну, государство, правительство, социально значимые объекты, явления и т.д.

14. Повтор слов с негативной тональностью.

15. Генерализация (обобщающие слова).

16. Визуальные предикаты и словосочетания.

17. Аудиальные предикаты и словосочетания.

18. Кинестетические предикаты и словосочетания.

Примечание: по завершении работы составляются сводные словари на уровне каждого индикатора.

Грамматический уровень:

1. Риторические вопросы.
2. Безличные конструкции.

3. РАЗРАБОТКА МЕТОДИКИ РАСЧЁТА ИНФОРМАТИВНЫХ КРИТЕРИЕВ ДЛЯ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ ПРИМЕНИТЕЛЬНО К ССД (ОПИСАНИЕ, ЦЕЛЬ, ИСТОЧНИКИ ДАННЫХ, МЕТОДЫ И ФОРМУЛЫ РАСЧЕТА, НОРМАЛИЗАЦИЯ И ОБРАБОТКА ИСПОЛЬЗУЕМЫХ ДАННЫХ)

3.1 Методика расчета информативного признака «тональность»

Для распознавания тональности новостного текста с использованием машинного обучения предложен новый набор информативных признаков, указанных в таблице 25.

Таблица 25 - Набор признаков для распознавания тональности текста

№	Информативный признак	Количество слов
1	Политизированность	30
2	Манипулятивность	652
3	Предвзятость	31
4	Призыв к действию	89
5	Административно-территориальные вопросы	123
6	Экономика, финансы	238
7	Внешняя политика	95
8	Гражданское общество	188
9	Здравоохранение	141
10	Индустриально-инфраструктурное развитие и технологии	115
11	Оборона и безопасность	77
12	Образование и наука	160
13	Правительство и Парламент	97
14	Правонарушение и происшествия	326
15	Сельское хозяйство	61
16	Социальное обеспечение	28
17	Спорт, культура и развлечения	243
18	Труд	25
19	Энергетика	70
20	Экология и туризм	40
21	Чрезвычайные происшествия	53

Данные признаки получены с использованием словарного подхода, вычисляющего встречаемость слов, входящих в определенный словарь для каждого документа.

Для проверки эффективности тематических областей и содержательных информативных признаков были использованы четыре популярных алгоритма машинного обучения с учителем: Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) и XGBoost (XGB). Из-за проблемы дисбаланса классов были выбраны несколько стратегий к решению задачи классификации. Для RF и SVM был применен подход «один против всех» (one-vs-all), при котором тональность была перекодирована в «позитивный против всех», «негативный против всех» и «нейтральный против всех» двоичных категорий. Учитывая, что подход «один против всех» дает еще более несбалансированные результаты, эксперимент был дополнен алгоритмами KNN и XGboost,

используя все три класса без предварительной перекодировки. Все результаты были получены с помощью 5-кратной перекрестной проверки.

Результаты эксперимента представлены в методике для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации применительно к системе социального доверия.

Программный код по распознаванию тональности представлен в приложении на диске.

3.2 Методика расчета информативного признака «объективность»

Для определения объективности новостного текста предложен алгоритм определения степени объективности. Объективность понимается как объективно поданная информация, соответствующая критериям достоверности без субъективной точки зрения автора.

Для определения степени объективности статьи используются сформированные лингвистические словари: (1) словарь персоналий; (2) компетентные источники; (3) используемые связи при ссылке на источник (таблица 26).

Таблица 26 – Словарь именованных сущностей

№ п/п	Словарь	Количество слов
1	Компетентные источники (должности и организации)	825
2	Персоналии	401
3	Используемые связи при ссылке на источник	88

Степень объективности рассчитывается по следующей формуле:

$$\text{CO} = \sum_{i=1}^{n} \frac{x_i}{y_i} \qquad (1)$$

где,

CO – степень объективности

x_i – количество объективных предложений

y_i – общее количество предложений

Алгоритм определения степени объективности строится на основе соответствий заранее определенным условиям и определяется на уровне предложений. Предложение считается объективным если оно отвечает, как минимум одному их следующих условий.

1. наличие слова связки и сайта (доверенные новостные сайты);

2. наличие слова связки и компетентного источника+ персоналии;

3. наличие слова связки и цифры (статистической информации).

Условия можно сложить чтобы получить не только бинарный показатель объективное-субъективное предложение, но и степень объективности как категориальный показатель, состоящий из максимум 3 категорий, где вес каждого условия равен 0.33.

Предложение считается субъективным если в нем не содержится ни одного из вышеупомянутых условий и содержатся местоимения и глаголы с личными окончаниями, указывающие на определенное лицо. Полученная объективность на уровне предложений может быть суммирована и представлена как отношение объективных к субъективным предложениям и нормализована на уровне публикации. Таким образом, степень объективности находится в интервале – [0;1], где значение ближе к 0 означает, что текст субъективен, ближе к 1 – объективен.

Экспериментальные результаты определения степени объективности новостных текстов представлена в методике определения степеней вероятной достоверности публикаций отдельной публикации.

Программный код по определению степени объективности представлен в приложении на диске.

4. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО ИНФОРМАЦИОННОМУ, АНАЛИТИЧЕСКОМУ И ЭКСПЕРТНОМУ СОПРОВОЖДЕНИЮ ПРОЦЕССА ПРОЕКТИРОВАНИЯ И РАЗРАБОТКИ ИНФОРМАЦИОННОЙ СИСТЕМЫ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ ПРИМЕНИТЕЛЬНО К ССД

Основной целью проектирования и разработки информационной системы оценки влияния открытых текстовых информационных источников на социум применительно к системе социального доверия является выработка эффективных алгоритмов анализа новостных текстов для определения тематической структуры медиапространства и распознавания информативных признаков в открытых текстовых данных с применением технологий машинного обучения и словарного подхода.

Задачами проектирования и разработки данной системы являются:

- сбор (парсинг) новостной текстовой информации в хранилище;
- тематическое моделирование корпуса новостных текстов;
- определение социально-значимых топиков (тематик);
- анализ текстов по распознаванию в них информативных признаков;
- разработка аналитических инструментов для визуализации результатов анализа.

В целом система позволит:

1. проводить тематический мониторинг медиа-пространства;
2. решить задачу классификации новостных текстов по тематических областям;
3. определять тональность, степень объективности и манипулятивности новостных текстов;
4. выявлять социально-значимые тематические области;
5. определять принадлежность новостных текстов определенному региону.

Данная информационная система может быть полезна для широкой аудитории: исследователей по различным областям науки, государственных и частных организаций, медиа компаний и других заинтересованных лиц.

Информационная система может предоставлять следующие возможности:

1) осуществить выборку новостных текстов по определенной тематической области;

2) определить тональность, степень объективности и манипулятивности интересующих новостных текстов или тематической области;

3) найти публикации, которые адресованы определённому региону;

4) использовать сформированные словари, корпуса текстов, алгоритмы для собственных исследований.

На первом этапе осуществляется сбор новостных текстов из открытых информационных интернет-источников. Формирование корпуса новостных текстов осуществляется с применением парсинга (автоматическое извлечение и сбор открытой информации) или путем использования готовых коллекций текстовых данных.

Модуль тематического моделирования

Модуль предназначен для распределения корпуса по топикам и последующего анализа списка сформированных топиков.

Латентное размещение Дирихле (LDA). Анализ текстов осуществляется с использованием методов машинного обучения и/или прописанного алгоритма распознавания информативных признаков. Как альтернатива система может обратиться к уже прописанным скриптам тематического моделирования и анализа текстов на языке python и/или R.

Входным набором данных является набор текстов, загруженных из файлового хранилища. Тематики представляются в виде групп наборов слов.

В рамках анализа полученные тематики должны сравниваться с имеющимися словарями тематических областей (к примеру, топик со словами «президент», «правительство», «государство» должен сравниваться со словарем государственных органов). Данный шаг необходим для автоматического определения названия топика. Далее внутри каждого направления проводится дополнительная процедура тематического моделирования, которая позволит детализировать имеющиеся топики (на согласовании). В дополнении в данном модуле возможно разместить график, отражающий частоту появления новостных материалов, связанных с определенным топиком, в разрезе временных периодов (например, месяцев).

Модуль анализа текстов

Данный модуль предназначен для классификации текстов на соответствие определенным информативным признакам: тональность, объективность, манипулятивность на основе имеющихся словарей.

Функционирование модуля планируется по следующей схеме: пользователь загружает коллекцию текстовых данных или данные, собранные парсером, который впоследствии анализируется на выявление степени присутствия вышеназванных признаков.

Распознавание тональности, объективности и манипулятивности осуществляется двумя подходами (1) алгоритмическим и (2) словарным. Алгоритмы распознавания признаков описаны в методике для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации применительно к ССД.

По результатам проведенного анализа выстраиваются графики и/или шкалы соответствия по каждому из признаков. Также модуль должен визуализировать присутствие ключевых слов из словарей в тексте для наглядного представления.

Выявление социально-значимых тематик и оценка влияния СМИ на социум

Тематики, имеющие социальную значимость, определяются, как правило, путем опроса экспертов, обладающих данными о состоянии общественного мнения, которые, в свою очередь, могут быть получены посредством социологических опросов населения.

С точки зрения определения социально значимых тематик наиболее результативным методом является экспертное оценивание. Эксперты, обладающие знаниями специфики деятельности СМИ и особенностями социальных взаимоотношений и убеждений социума, смогут внести значительный вклад в оценивание тематик, освещаемых в СМИ. Результаты тематического моделирования не предусматривают наименование топика. Использование экспертного анализа позволит: (1) дать названия полученным топикам; (2) определить топики, являющиеся значимыми для социума.

Для анализа интереса различных социальных групп к различным тематикам наиболее результативным методом является рандомизированное социологическое исследование в форме социологического опроса. Основной целью социологического исследования является определение интереса различных социальных групп к различным тематикам. Выборка должна быть произведена в соответствии с общепринятыми процедурами. Социологический опрос позволит связать тематики, освещаемые СМИ с социальными группами, у которых они вызывают интерес, изучить влияние текстовой информации в разрезе возрастных групп населения, в разрезе стратификации по принципу «город-село», регионов, гендерной стратификации, уровня образования и профессиональной деятельности респондентов.

С точки зрения оценки влияния СМИ на социум оптимальным подходом к исследованию может стать экспериментальное исследование, поскольку только такой дизайн исследования позволяет однозначно установить и доказать наличие причинно-следственных связей. Основной целью эксперимента является установление факта влияния информативных признаков (или их части), отобранных в ходе исследования, в качестве признаков, анализируемых в рамках информационной системы. Предположительно, такими признаками являются: тональность, манипулятивность, объективность. Данный метод позволит установить факт влияния СМИ на социум, подтвердить необходимость учета в рамках машинного анализа контента электронных СМИ самих отобранных информативных признаков, замерить степень влияния таких информационных признаков на социум. Более детальная стратегия экспериментального исследования будет разрабатываться и обсуждаться в ходе дальнейшего исследования.

5. ПРОВЕДЕНИЕ КОМПЛЕКСА РАБОТ ПО МОНИТОРИНГУ ОТКЛОНЕНИЙ В РАБОТЕ РАЗРАБОТАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ ПРИМЕНИТЕЛЬНО К ССД

Проведен мониторинг разработанной информационной системы Институтом информационных и вычислительных технологий. Система состоит из 3 модулей:

1. Модуль ETL. Модуль ETL включает блоки для сбора данных – скрапперы по определенным сайтам, управления процессами и предобработки данных (Airflow), тематического моделирования (Big Arthem), векторизации (Word2vec), отчетности и классификации;
2. Хранилище реализовано в двух модулях: PostgreSQL, которое является более структурированной системой хранения данных, требующих более длительного хранения, и Elasticsearch, использование которого позволяет обеспечить скорость работы ИС;
3. Модуль отображения. Он включает два подмодуля: Django – модуль пользовательского интерфейса с прототипом функционала для пользователя, и Kibana – модуль интерфейса администратора системы, для отображения работы системы в реальном времени.

На основе мониторинга и анализа прототипа разработанной информационной системы предлагаются следующие рекомендации:

1. Рассмотреть вопрос об авторских правах и праве на использование готового ПО и софтов для целей проекта на предмет необходимости покупки/получения лицензий от их разработчиков;
2. Рассмотреть возможность написания собственных кодов для использования дэшбордов, пользовательского интерфейса и для наличия возможности конфигурации функционала под специфику проекта;
3. Рассмотреть возможность включения разработанных в ходе проекта алгоритмов тематического моделирования LDA, машинного обучения распознавания тональности, определения полярности тональности и степени объективности новостных текстов лексико-словарным подходом.
4. Рассмотреть возможность создания собственной библиотеки с размещением сформированных в рамках исследования словарей и экспериментальных корпусов.

Работы по дальнейшему проектированию информационной системы оценки влияния открытых текстовых информационных источников на социум применительно к ССД будут продолжены.

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

Подготовительным этапом, обусловливающим решение задач по автоматизации оценки новостных текстов и распознаванию информативных признаков, является комплекс работ по созданию словарного массива, отбору языковых единиц и способов (правил) их употребления в новостных текстах. В связи с этим целью данной работы является разработка методов формализации правил и словарей для распознавания в корпусе новостных текстов информативных признаков и определения степени их выраженности.

Были определены 12 информативных признаков, для распознавания которых необходимо формирование словарей и формализация лингвистических шаблонов и правил.

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

Для достижения цели в текущем году были поставлены и решены следующие задачи:

1. Проведение глубинного анализа сформированного корпуса публикаций СМИ по 9 информативным признакам (3 признака в 2018 г., 9 – в 2019 г.).
2. Осуществление разметки текстового корпуса по обозначенным информативным признакам.
3. Формирование и систематизация словарей по определенным информативным признакам на базе анализируемых текстов.
4. Выявление закономерностей для формализации правил распознавания отдельных информативных признаков.
5. Получение экспертных заключений о реализуемости выявленных закономерностей в рамках проекта.
6. Описание методов формализации правил на основе лингвистического анализа текстов.

6.1 Описание предыдущего этапа работ

В ходе предыдущего этапа работ был проведен комплекс мероприятий, включающий следующее. *Первое,* определение информативных признаков для формирования словарей и формализации правил. *Второе,* разработка экспертно-аналитической рамки (инструкции). *Третье,* формирование пула экспертов. *Четвёртое,* глубинный анализ текстов для составления словарей и выявления закономерностей для формализации правил по трем признакам: 1) тональность: позитивная, нейтральная, негативная; 2) политизированность; 3) наличие манипулятивных техник.

В результате анализа корпуса текстов по критерию «тональность» был сформирован лингвистический словарь и формализован ряд правил. После обработки полученных словарей на базе данных словоупотреблений был составлен словарь маркированной лексики, содержащий 3091 лексему (2112 – с негативной семантикой, 979 - с позитивной семантикой). Словарь словосочетаний тональности включает в себя 3299 негативных языковых единиц и 2571 позитивных.

В результате анализа полученных данных по завершению работы группы лингвистов над корпусом текстов по признаку «политизированность», сформирован лингвистический словарь, в который вошли 30 слов и 136 словосочетаний.

По завершению анализа корпуса текстов по признаку «наличие манипулятивных техник» проведена чистка и коррекция данных, на основе которых сформирован лингвистический словарь, включающий 652 слова и 891 словосочетание.

6.2 Глубинный анализ текстов для составления словарей по признаку «социальная значимость»

В рамках заданного параметра исследовано 2000 текстов для выявления признаков социально значимых публикаций, т.е. публикаций, касающихся социально значимых для общества вопросов, способных вызвать широкий общественный резонанс, затрагивающих интересы общества и/или способных вызвать дестабилизацию общества.

Среди языковых единиц, маркирующих социально значимые публикации, были выделены наименования социальной инфраструктуры – комплекса объектов, предприятий, организаций, которые обеспечивают функциональную жизнедеятельность населения, формирование интеллектуально развитого индивида, общества. Также к данной группе языковых единиц относятся наименования должностей и персоналий, осуществляющих деятельность в социально значимых организациях. В отдельную группу были выделены языковые единицы, маркирующие социально значимые темы.

В результате анализа в текстах зафиксированы слова и словосочетания, сигнализирующие о социальной значимости текста и разработан соответствующий словарь, включающий 2598 языковых единиц. По завершении работы по чистке и коррекции данного словаря, количество языковых единиц составило 2426 слов и словосочетаний. Наибольшее количество слов и словосочетаний выявлено в группе «Темы» – 1196, далее по убывающей в группах «Организации» – 455, «Персоны» – 418, «Должности» – 325, «Объекты» – 32.

С целью увеличения потенциальных возможностей автоматизированного поиска социально значимых публикаций были составлены 213 синонимических рядов к социально маркированным языковым единицам. Как показали результаты исследования в рамках

данного признака, с помощью языковых средств успешно могут быть выявлены новостные тексты, затрагивающие социально значимые темы. В результате работы, языковые единицы были распределены по предложенным классам слов и достаточно четко дифференцированы.

6.3 Глубинный анализ текстов для составления словаря по признаку «казахстанское содержание»

В процессе анализа данного параметра объектом исследования стали слова и словосочетания, связанные с событиями, произошедшими на территории Казахстана либо связанными с деятельностью казахстанцев за рубежом. Предметом анализа стали казахстанские события/персоналии, наименования географических объектов, обычаев, традиций, предметов культуры, быта, специфических наименований должностей, территориально-административного деления, национальных брендов и т.д.

В результате анализа корпуса текстов по критерию «казахстанское содержание» была произведена разметка материала исследования и сформирован лингвистический словарь. По завершении чистки и коррекции полученных данных, в состав словаря включены 1398 слов и 5501 словосочетание. Сводная частеречная характеристика словаря, составленного в ходе анализа текстов в рамках параметра «казахстанское содержание», в подавляющем большинстве представлена именами существительными (1304 единицы), 92 прилагательными (без учета прилагательных в составных наименованиях организаций), 2 наречиями («по-казахски», «по-шымкентски»), отсутствуют глаголы, за исключением одного случая употребления глагола в составе наименования хэштега #НеМолчи.KZ.

Исследования в рамках данного признака показали высокую эффективность работы с языковыми средствами, отражающими казахстанское содержание, обусловленную наличием яркой линвокультурологической окраски анализируемой группы слов. Одним из важнейших и частотных признаков и показателей казахстанского содержания является наличие языковых средств, заимствованных из казахского языка, имеющих казахские (тюркские) корни.

6.4 Глубинный анализ текстов для составления словаря и формализации правил по признаку «ссылки на компетентные источники»

В рамках анализа данного признака в 2000 текстах по завершении чистки и коррекции словаря выявлены следующие лингвистические маркеры по признаку «Ссылки на компетентные источники»:

- используемые связи при ссылке на источник – 88 слов;
- компетентные источники – 836 (наименование организаций – 767, наименование должностей – 69);

– персоналии – 401;

– словосочетания со ссылками на компетентные источники – 1141.

На основе глубинного анализа текстов по данному критерию сформулированы правила и разработаны языковые модели, позволяющие определить в публикациях СМИ наличие ссылок на компетентные источники.

Как показали проведенные исследования, для наиболее достоверного определения в текстах наличия/отсутствия ссылок на компетентные источники необходим перечень наименований всех министерств, комитетов, департаментов, ФИО политических служащих Республики Казахстан. Данный перечень будет полезен и для анализа других параметров.

Исследования в рамках этого параметра показали эффективность использования языковых средств для определения наличия в тексте ссылок на компетентный источник. Свод формализованных правил, составленный словарь источников информации, перечень слов-маркеров (имен существительных, вводных слов и глаголов) и соответствующих словосочетаний, сигнализирующих о наличии ссылки на источник, достаточно четко ограничен от других лексико-семантических групп слов и типов словосочетаний. Это позволяет с высокой степенью вероятности выявить ссылки на источники путем последующей автоматизированной обработки.

6.5 Глубинный анализ текстов для составления словаря по признаку «принадлежность текста к определенной сфере»

В рамках исследования данного признака было проведено распределение тех или иных слов и словосочетаний в текстах по перечисленным в инструкции сферам. В результате сводного анализа полученных данных по итогам работы группы лингвистов с корпусом текстов по критерию «принадлежность текста к определенной сфере» был сформирован лингвистический словарь, позволяющий определять принадлежность текста к определенной сфере.

Словарь включает 2080 слов и 2218 словосочетаний, сформированных в рамках 17 тематических сфер: (1) «Правонарушение и происшествия» – 326; (2) «Спорт, культура и развлечения» – 243; (3) «Экономика, финансы» – 238; (4) «Гражданское общество» – 188; (5) «Образование и наука» – 160; (6) «Здравоохранение и здоровье» – 141; (7) «Административно-территориальные/коммунальные вопросы» – 123; (8) «Индустриально-инфраструктурное развитие и технологии» – 115; (9) «Внешняя политика» – 95; (10) «Правительство и Парламент» – 97; (11) «Оборона и безопасность» – 77; (12) «Энергетика» – 70; (13) «Сельское хозяйство» – 61; (14) «Чрезвычайные происшествия» – 53; (15) «Экология и туризм» – 40; (16) «Социальное обеспечение» – 28; (17) «Труд» – 25.

Исследования в рамках этого параметра показали дальнейшую целесообразность выявления языковых средств, сигнализирующих об отнесении текста к той или иной отрасли. Полученные результаты позволили сформировать достаточно четко дифференцированный словарь языковых единиц по разным сферам жизнедеятельности государства и вносят существенный вклад в лингвистическую базу по заданным кластерам для автоматизированной обработки текстов. Альтернативным, дополнительным путем создания такой лингвистической базы может служить работа с универсальным (без рассмотрения определенных медийных текстов) языком по созданию лексико-тематических групп по заданным сферам. Однако работа с конкретными текстами СМИ позволяет выявить (как в случае нашего исследования) наиболее частотные, актуальные и интересные для журналистов и населения сферы и может служить аналитической платформой для коррекции информационной политики в стране в целом.

6.6 Глубинный анализ текстов для составления словаря и формализации правил по признаку «наличие проверяемых фактов в статье» в разрезе «субъект/предмет–действие–место–время»

Данный параметр, включенный в структуру исследования, необходим для выявления наличия в тексте проверяемых фактов и обусловил их определение посредством вычленения из текстов языковых единиц, отражающих понятия «субъект/предмет», «время», «место», и «объект», в контексте которых произошли события основной линии сюжета. В рамках работы над параметром в соответствии с требованиями инструкции группой лингвистов была произведена текстовая разметка, составлен глоссарий слов и словосочетаний, а также разработаны правила формализации, способствующие автоматизированному определению наличествующих в текстах СМИ проверяемых фактов.

По результатам глубинного анализа сформирован словарь из 1202 слов и 4221 словосочетания. Наиболее частотными проверяемыми фактами являются «субъект/ предмет» – 1458 словосочетаний и «действие» – 525 слов. Выработаны правила формализации и семантико-грамматические модели для автоматизированного определения в текстах проверяемых фактов.

Исследования показали, что данный параметр является одним из самых лингвистически маркированных, т.е. позволяющих максимально формализовать средства выражения заданных позиций (объект, предмет, время, действие и т.д.) в «языковых координатах» как на лексическом, так и на грамматическом уровнях: часть речи, род, число, падеж, лексико-тематические группы слов и т.д. Для дальнейшего продолжения исследования в рамках данного параметра, безусловно целесообразного, более четкой

дифференциации требует определение понятий «субъект», «предмет» и «объект» в силу расхождения их трактовок в логическом и грамматическом регистрах.

6.7 Глубинный анализ текстов для составления словаря и формализации правил по признаку «соответствие заголовка публикации содержанию»

В рамках анализа данного критерия в 2000 текстах произведена разметка заголовков публикаций и ключевых слов и разработаны формализованные правила для автоматизированного определения соответствия/несоответствия заголовка публикации содержанию.

В результате исследовательской работы были выявлены три группы текстов: 1) заголовок, который соответствует содержанию; 2) заголовок, который частично соответствует содержанию; 3) заголовок, который не соответствует содержанию. Подавляющее большинство текстов было отнесено к первой группе, их количество составило 1813 единиц. Ко второй группе было отнесено 175 текстов (заголовок частично соответствует содержанию). Выявлено 12 текстов с заголовками, не соответствующими содержанию.

Результаты работы в рамках данного признака подтвердили выдвинутую в начале исследования гипотезу о возможности выявления соответствия/несоответствия заголовка содержанию текста посредством лингвистического анализа. Расширение базы исследования, т.е. анализ большего количества текстов на предмет заданного параметра, позволит на основе увеличения группы данных получить более точные и детальные языковые способы для автоматизированного определения соответствия/несоответствия заголовка содержанию текста.

6.8 Глубинный анализ текстов для составления словаря по признаку «определение ключевых слов текста»

В результате глубинного анализа текстов был составлен словарь из 4001 ключевых слов и 6409 ключевых словосочетаний. Общее количество выявленных языковых единиц составило 10410 единиц.

Выявление ключевых слов производилось с опорой на следующие позиции: (1) частотность употребления, (2) совпадение с заголовком, принадлежность к нейтральному стилю, (3) способность в комплексе с другими ключевыми словами отражать тематику и общий смысл текста. В результате анализа работы, проделанной группой лингвистов по поиску, разметке и составлению словаря ключевых слов, выявлена статистическая закономерность: количество ключевых слов в текстах составляет от 2 до 10 единиц. Следует отметить «подвижность» ключевых слов в тексте: они могут быть локализованы в любом абзаце текста.

Частотной оказалась следующая структура: в компактных новостях (коротких новостных сообщениях) зафиксировано совпадение наименования заголовка и ключевых слов, отсутствие ключевых слов в первоначальных абзацах и их использование в финальных абзацах текста. Также следует отметить, что совпадение заголовка и ключевых слов в большинстве случаев наблюдается в текстах с негативной тональностью. Другими словами, выведение авторами текста ключевого слова в заголовок усиливает эмоциональное воздействие на читателя, чаще всего негативное.

Исследование языковых средств текста в рамках данного параметра, для обучения машины автоматизированной обработке текстов на сегодняшний день является хорошо зарекомендовавшей себя и активно используемой практикой в информационно-поисковых системах, программах по мониторингу СМИ и др. Выявление ключевых слов в контексте нашего исследования подтвердило успешность такого опыта, показало «прозрачность» определения данной категории слов в текстах, позволило сформулировать некоторые закономерности их использования и является определенным вкладом в расширение базы ключевых слов для автоматизированной обработки текстов медийного характера. Однако, ввиду разнообразия тем статей медийного пространства, исследовательский интерес представляет не словарь ключевых слов, а закономерности и способы их выражения, которые можно вывести машинными средствами на основе лингвистической разметки.

6.9 Глубинный анализ текстов для составления словаря и формализации правил по признаку «предвзятость»

На основе глубинного анализа данного признака составлен перечень слов и словосочетаний, отражающих ярко выраженное личное мнение автора без приведения различных точек зрения. Из 2000 текстов после работы лингвистов, а также после сводной чистки и коррекции словаря было выявлено 32 слова и 423 словосочетания, сигнализирующих о наличии в тексте предвзятого отношения автора. Также на основании анализа текстов были сформулированы правила формализации для автоматизированного определения текстов с предвзятостью. В силу большого сдвига в сторону от собственно языковых явлений в пользу экстралингвистических факторов (необходимость учета возраста, пола, наличия неприязни или, наоборот, симпатий, этнической, конфессиональной принадлежности, политических взглядов и т.д.), маркеров, характеризующих предвзятые тексты, на данном этапе исследования определение предвзятости только с учетом языковых средств не представляется возможным, в связи с чем предлагается исключить данный признак.

6.10 Глубинный анализ текстов для составления словаря и правилам формализации по признаку «призыв к действию»

Результаты глубинного анализа корпуса позволили выявить 519 текстов (что составляет 26% от всего проанализированного корпуса), содержащих призывы к действию.

На основе анализа корпуса текстов по признаку «призыв к действию» сформирован лингвистический словарь, маркированной лексики, содержащий 89 слов и 1427 словосочетаний.

Основу словаря маркированной лексики в рамках данного параметра составляют глаголы – 69 слов. Также в эту группу вошли 9 существительных, 5 модальных слов, 6 прилагательных и причастий.

Закономерности, позволяющие формализовать правила для распознавания текстов, содержащих призывы к действию, включают:

1) использование слов с модальной семантикой, например, «нельзя», «можно», «нужно», «должен» и т.д.;

2) использование глаголов в форме повелительного наклонения, например, «требуйте», «возьмите», «отдайте», «верните»;

3) использование глаголов в инфинитиве «призвать», «требовать», «приказывать» и др.;

4) использование частицы не в качестве основного отрицания того, что обозначается словом с нейтральной или позитивной семантикой, к которому она относится по смыслу;

5) использование восклицательного знака.

Как показали проведенные исследования, данный признак достаточно эффективно исчисляем посредством анализа языковых средств и может быть успешно применен для обучения машины автоматизированному распознаванию текстов с призывами. С учетом современных языковых процессов в СМИ и особенно в социальных сетях и интернет-пространстве, которые характеризуются устойчивой тенденцией к изменению используемых языковых средств, форматов подачи информации для активного привлечения целевой аудитории в рамках поставленных авторами или их заказчиками общественно-политических задач, дальнейшее исследование данного признака на большем материале текстов является одним из важнейших направлений работы по обеспечению и укреплению информационной безопасности страны.

6.11 Чистка и коррекция словарей

По итогам глубинного анализа новостных текстов лингвистами по составлению словаря в рамках определенных параметров была проведена работа по подготовке полученных данных (языковых единиц) к экспериментальной машинной обработке.

Была произведена коррекция словарей путем:

- сокращения словосочетаний;
- удаления контекстуальных языковых единиц;
- удаления повторов и технических ошибок;
- удаления словоформ.

6.12 Экспертное обсуждение результатов работы по выявлению закономерностей для формализации правил

По завершении работы группы лингвистов, выявленные закономерности по 8 признакам из 12 были вынесены на обсуждение. В ходе экспертного обсуждения по выявленным закономерностям были даны комментарии и заключения относительно применимости и реализуемости результатов работы лингвистов, где заключение «реализуемо» означает наличие программных средств, позволяющих реализацию выявленной закономерности в рамках разрабатываемого продукта; заключение «сложно реализуемо» означает, что реализация выявленной закономерности требует построения сложных многоэтапных алгоритмов и временных затрат и, возможно, нецелесообразно для реализации с точки зрения целей и бюджета проекта; заключение «не реализуемо» означает отсутствие программных средств, позволяющих реализацию выявленной закономерности в рамках разрабатываемого продукта; заключение «требует обсуждений» означает необходимость получения дополнительных консультаций с экспертами для вынесения окончательного заключения.

Свод комментариев и заключений экспертной группы по выявленным группой лингвистов закономерностям для формализации правил представлен в отчете «Разработка методов формализации правил и словарей для распознавания в размеченном корпусе информативных признаков и определения степени их выраженности».

Выводы и дальнейшая работа

Таким образом, на основе сводного анализа и коррекции результатов работы группы лингвистов по исследованию и детальной разметке 2000 медиа-текстов, получены следующие результаты:

1. Сформирован словарь из 8961 языковых единиц (3091 слов и 5870 словосочетаний) по информативному признаку «тональность».
2. Сформирован словарь из 166 языковых единиц (30 слов и 136 словосочетаний) по информативному признаку «политизированность».
3. Сформирован словарь из 1543 языковых единиц (652 слова и 891 словосочетание) по информативному признаку «наличие манипулятивных техник».

4. Сформирован словарь из 2639 языковых единиц (по 6 параметрам) по информативному признаку «социальная значимость».
5. Сформирован словарь из 1398 слов и 5501 словосочетания по информативному признаку «казахстанское содержание».
6. Сформирован словарь компетентных источников (именованные сущности: организации, должности и персоналии), состоящий из 2378 слов и словосочетаний, 88 слов-связок, используемых при ссылке на источник.
7. Сформирован словарь из 1202 слов и 4221 словосочетания в рамках информативного признака «наличие проверяемых фактов в разрезе субъект/предмет – действие – место – время».
8. Сформирован словарь из 4001 слов и 6409 словосочетаний по информативному признаку «ключевые слова» в текстах.
9. Сформирован словарь по информативному признаку «предвзятость» – 32 слова и 423 словосочетания.
10. Сформирован словарь по информативному признаку «призыв к действию» – 89 слов и 1427 словосочетаний.
11. Выявлены закономерности для формализации правил по определению информативных признаков тональности, политизированности, наличия манипулятивных техник, ссылок на компетентные источники, наличия проверяемых фактов в разрезе «субъект/предмет – действие – место – время», соответствия заголовка публикации содержанию, предвзятости и призыва к действию.

Была произведена чистка, систематизация и внесены дополнения в вышеуказанные словари.

Также в ходе анализа корпуса была произведена разметка текстов с выделением текстовых фрагментов, включающих маркеры информативных признаков.

По результатам закономерностей, выявленных в ходе лингвистического анализа корпуса, были получены экспертные заключения об их реализуемости в рамках проекта от российских и казахстанских экспертов.

Также были описаны методы формализации правил на основе закономерностей, выявленных в ходе лингвистического анализа корпуса.

Полученные результаты исследования в целом подтвердили необходимость и возможность выявления заданных информативных признаков на языковом уровне для автоматизированного распознавания искомых параметров текста.

В число однозначно определяемых с помощью лингвистических средств информативных признаков вошли тональность, социальная значимость, казахстанское содержание, ссылки на компетентные источники и призыв к действию.

В то же время, как показали результаты работы данного этапа исследования, часть информативных признаков, отобранная для оценки новостных текстов, должна определяться и распознаваться с применением комбинированных средств – лингвистических способов и машинных алгоритмов. В число таких признаков входит политизированность, наличие манипулятивных техник, ключевые слова, предвзятость. В рамках этой работы необходимо искусственное моделирование конструкций, построенных по лингвистическим правилам и шаблонам на базе моделей и словоупотреблений, определенных в ходе глубинного анализа новостных текстов. 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.

Данная работа, на наш взгляд, крайне необходима, так как выделенные в результате анализа текстов лексемы в других контекстах могут иметь другие значения (ср.: *развитие страны* и *развитие преступности*).

Кроме того, для более успешной автоматизированной обработки текста на предмет выявления признака «политизированность» четкого определения требует само понятие «политизированность». В наиболее общем виде политизированный текст представляет собой описание событий, напрямую не относящихся к политике, однако приобретающих политизированную окраску из-за обилия политической лексики, призывов и наличия негативной тональности. Бытийной канвой политизированного текста является казахстанское содержание. Другими словами, для точного определения наличия признака политизированности в тексте необходимо рассматривать его в совокупности нескольких факторов и включать их данные в автоматизированную обработку.

В целях более эффективной работы по определению наличия казахстанского содержания текста необходимо включение в материал для автоматизированного распознавания, помимо извлеченных из 2000 текстов казахстанских персон, топонимов, предметов обихода и т.д., в силу ограниченности объемом проанализированного материала универсального перечня наименований географических объектов, известных персон, должностных лиц, наименований казахстанских государственных органов, учреждений, организаций и бизнес-структур. Подобные словари в разы повысят «узнаваемость» казахстанского содержания в тексте.

Несмотря на то, что глубинный анализ текстов в рамках параметров «политизированность», «наличие манипулятивных техник» и «предвзятость» выявил определенные языковые средства и материал для научных наблюдений и обобщений, составление формализованных правил для автоматизированного распознавания указанных признаков с помощью только языковых единиц не представляется возможным. Как показали исследования, на языковом уровне политизированность теста, наличие манипулятивных техник и предвзятость практически полностью совпадают с тональностью, в особенности с негативной.

Определение манипулятивных техник и предвзятости в силу двойственной природы их происхождения: прагматические характеристики (взаимоотношения между автором и лицом, о котором идет речь, социальные, гендерные, политические установки, интенции автора), с одной стороны, и собственно языковые средства, с другой, – возможно только с учетом комплекса «человеческих» факторов, которые пока еще не освоены искусственным интеллектом, т.е. недоступны для автоматизированного распознавания. Кроме того, само понятие «предвзятость» в самом широком его понимании трактуется в естественном языке как субъективность, которая, как известно, хотя бы в самой маленькой степени всегда все же присутствует в любом тексте.

В то же время автоматизированное распознавание вышеуказанных «проблемных» признаков текста может быть осуществлено при условии включения в методику исследования текстов, кроме собственно лингвистических параметров, факторов экстралингвистического толка, что, безусловно, должно являться предметом отдельного исследования и темой другого проекта.

7. 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

Для выявления закономерностей в сформированном корпусе из общего корпуса, содержащего 1 982 433 новостных текста, на основе стратифицированной выборки были отобраны 100 тысяч текстов (рисунок 5).

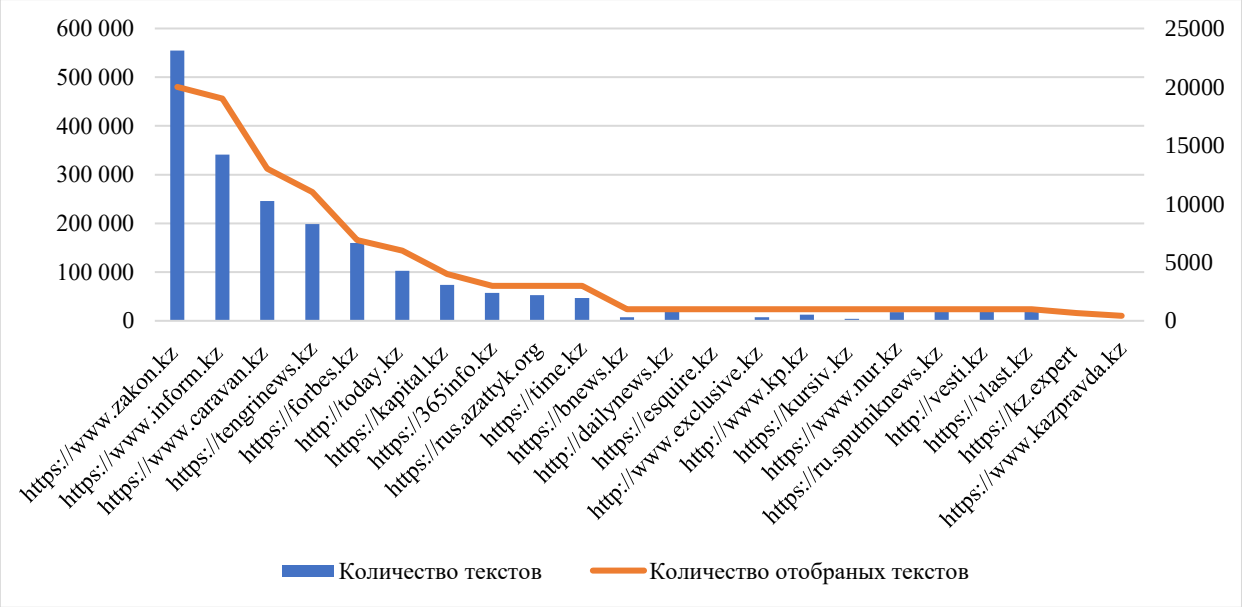

Рисунок 5 – Сравнение представленности новостных текстов в выборке и корпусе

Данный корпус размечен по 56 признакам с использованием сформированных словарей по информативным признакам, лексическим признакам, объему публикаций и частей речи.

Для каждого признака была высчитана мера несимметричности распределения. Значения коэффициента асимметрии по признакам строго больше нуля. Это означает, что у всех признаков большинство значений меньше среднего.

По результатам статистического анализа словарь негативных слов коррелируют с со словарями тематических областей: правонарушения и происшествия, манипулятивность. Словарь позитивных слов показывает корреляционную связь со словарем сферы «правительство и парламент». Словарь манипулятивных слов коррелирует с политизированными словами (рисунок 6).

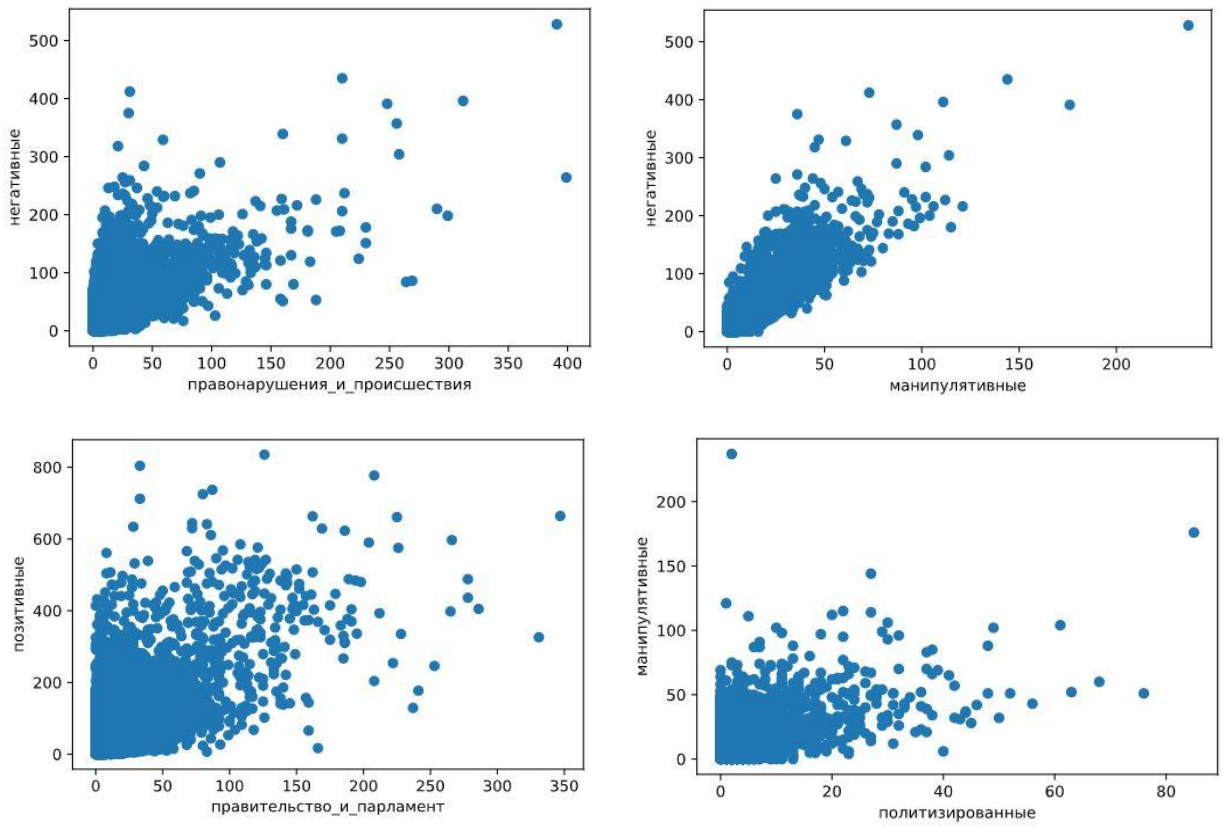

Рисунок 6 - Диаграмма рассеяния между информативными признаками

Процедура определения значимых признаков на основе непрерывных данных осуществлялась путем корреляционного анализа, в ходе которого осуществлялось вычисление коэффициента Пирсона. При определении связи определенной тематической области и отобранных информативных признаков было выявлено что:

- тематические области «гражданское общество», «образование и наука» и «правительство и парламент» входят в тройку наиболее значимых для информативного признака «позитивная тональность»;

- тематическая область «правонарушения и происшествия» является наиболее значимой для информативного признака «негативная тональность»;

- тематическая область «правительство и парламент», ожидаемо, значима для информативного признака «политизированность»;

- информативный признак «манипулятивность» наиболее часто проявляется в тематической области «правонарушение и происшествия», «гражданское общество» и «правительство и парламент».

Для определения текстовых закономерностей были выявлены наиболее часто встречающиеся в корпусе существительные, прилагательные, наречия и глаголы, ключевые связки слов.

Кроме этого, результат разметки частей речи позволит провести анализ взаимовстречаемости определенных слов и вывести последовательность связок слов уровне предложений (рисунок 7). Данный анализ проведен с использованием библиотеки UDpipe, свободно доступной для языка статистического программирования R.

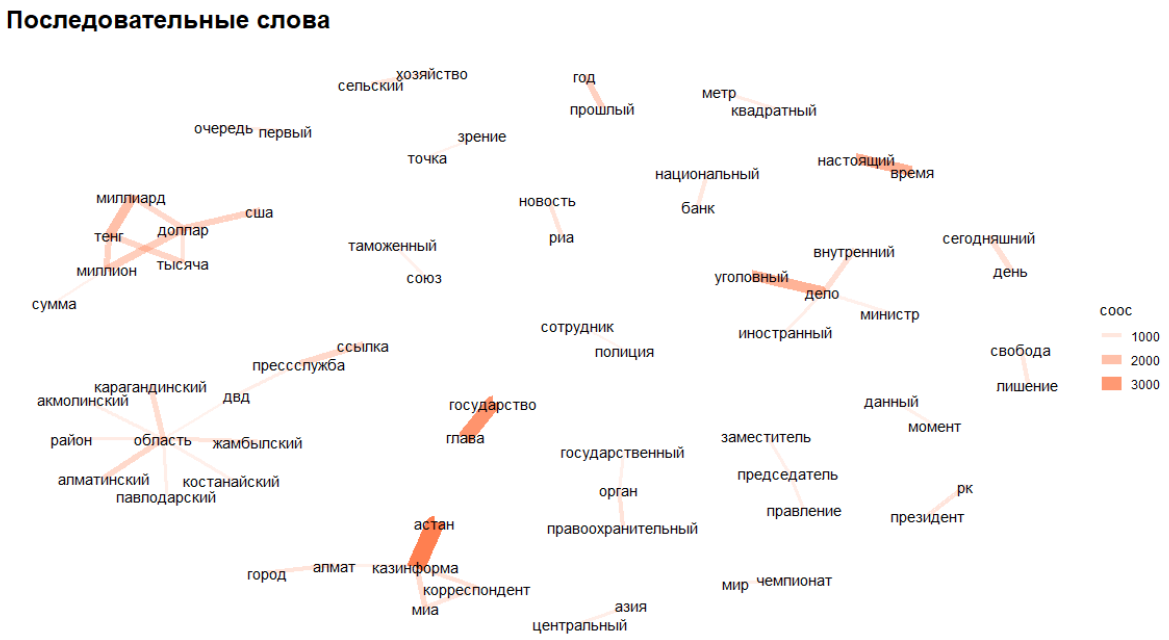

Рисунок 7 – Взаимовстречаемость слов в предложениях

На рисунке 7 жирными линиями выделены последовательности слов, встречающихся в анализируемом новостном корпусе более 3000 раз. Менее жирными выделены связки, встречающиеся более 2000 раз и более 1000 раз. Последовательности дают ясную картину взаимовстречаемости слов в предложениях.

Также поиск закономерностей в корпусе позволил выявить пересечение топиков, полученных в результате LDA с тематическими сферами, определёнными экспертным подходом, что потенциально в дальнейшем позволит производит категоризацию публикаций и топиков по тематическим сферам.

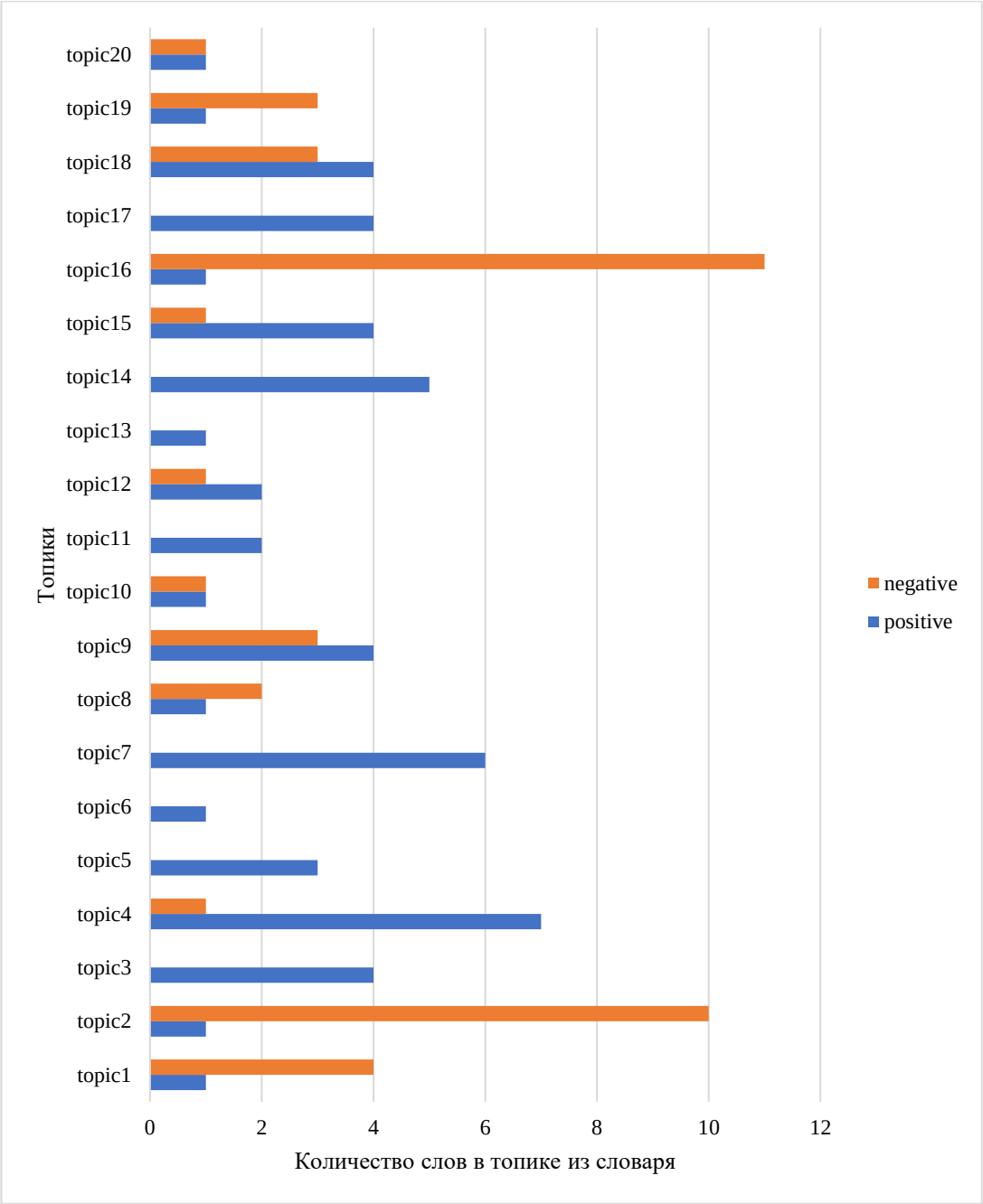

Рисунок 8 - Взаимовстречаемость слов из словаря тональности со словами в топике

Как видно из рисунка 8, наибольшее количество слов из словаря негативных слов встречается в топиках 2 и 16, позитивных слов в топиках 4, 7 и 14. В результате возможно предположить, что большинство текстов в данных топиках являются с негативной или позитивной тональностью.

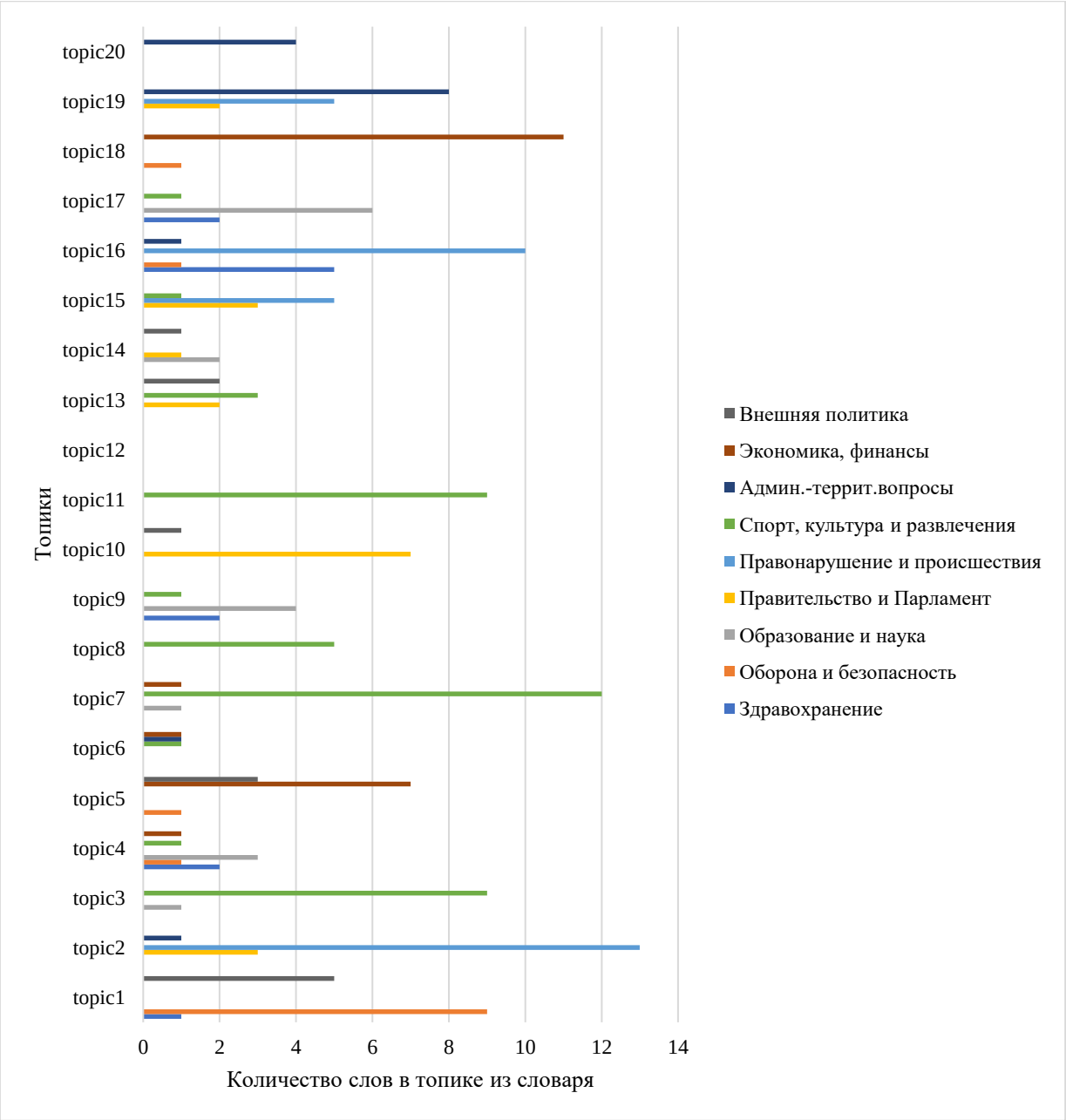

Рисунок 9 - Взаимовстречаемость слов из словаря тематических направлений со словами в топике

На рисунке 9 представлена взаимовстречаемость слов из словаря тематических направлений со словами в топике. По 9 из 17 тематических словарей взаимовстречаемость выше 5 слов, или 25%. Наибольшая взаимовстречаемость отмечается по топикам 2 со словарем «правонарушения и происшествия», 7 – «спорт, культура и развлечение», 18 – «экономика и финансы», 1 - «оборона и безопасность». В результате возможно предположить, что большинство публикаций в данных топиках представлены данными тематиками.

8. 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

С целью распознавания информативных признаков для оценки влияния открытых текстовых информационных источников на социум лингвистическим способом были сформированы словари слов и словосочетаний по 11 информативным признакам.

Сформированные лингвистами словари были расширены двумя способами.

1. Первый способ заключался в расширении слов имеющимися словоформами. Такое расширение лингвистических словарей признаков состояла из нескольких этапов. На первом этапе на базе сформированного корпуса, состоящего из 1,982 млн. новостных текстов был создан словарь, размер которого составляет 787,3 тыс. уникальных словоформ. В словарь включены уникальные словоформы, встречающиеся в корпусе более 2-х раз. На втором этапе осуществлялось автоматическое расширение лексических единиц, в словарях, сформированных лингвистами путем включения их словоформ. (таблица 27). Данный словарь возможно использовать на новостных текстах без проведения процедуры лемматизации.

Таблица 27 – Расширенные словари признаков словоформами

№ п/п	Словарь	Количество слов
1	Позитивные	6890
2	Негативные	10433
3	Политизированные слова	212
4	Манипулятивные слова	4620
5	Предвзятые слова	91
6	Слова, призывающие к действию	347
7	Слова-связки, используемые при ссылке на источник	410
	Слова по определенной сфере:	
8	Административно-территориальные вопросы	905
9	Экономика, финансы	1700
10	Внешняя политика	562
11	Гражданское общество	1244
12	Здравоохранение	780
13	Индустриально-инфраструктурное развитие и технологии	908
14	Оборона и безопасность	500
15	Образование и наука	1046
16	Правительство и Парламент	714
17	Правонарушение и происшествия	2726
18	Сельское хозяйство	327
19	Социальное обеспечение	176
20	Спорт, культура и развлечения	1802
21	Труд	194

22	Энергетика	436
23	Экология и туризм	294
24	Чрезвычайные происшествия	422

2. Второй способ заключался в расширении имеющихся базовых слов в сформированных лингвистами словарях ручным способом. В частности словарь тональности расширен синонимами и антонимами. [71] В словарь компетентных источников добавлены наименования государственных органов, организации квази-государственного сектора, высших учебных заведений. Словарь персоналий расширен государственными и политическими служащими, руководителями квази-государственного сектора, парламентариями и др. лицами. Кроме того, составлен список областей, городов и населенных пунктов с географическими координатами (таблица 28). Данный словарь возможно использовать после процедуры лемматизации текста.

Таблица 28 – Расширенные словари признаков ручным способом

№ п/п	Словарь	Количество слов
1	Позитивные	3475
2	Негативные	5580
3	Словарь слов с негативной семантикой: отрицатели, усилители, уменьшители	54
4	*Компетентные источники (организации и должности)*	826
5	*Используемые связки при ссылке на источник*	88
6	Персоналии	401
7	Наименование областей, городов, населенных пунктов Казахстана с координатами	16 745

Для проверки эффективности сформированных словарей и лингвистических правил проведены эксперименты по распознаванию информативного признака «тональность» с использованием словарей признаков с помощью алгоритмов машинного обучения и словарного подхода, а также определению степени объективности новостных текстов с использованием словарного подхода.

Для распознавания тональности предложен новый набор информативных признаков, полученных в результате формирования словарей, таких как политизированные, манипулятивные, предвзятые слова, слова, призывающие к действию, и слова-маркеры 17 тематических областей. Результаты эксперимента продемонстрировали эффективность предложенных тематических областей и содержательных информативных признаков при бинарной и трехклассовой классификации тональности. Хотя включение тематических областей улучшило классификационную способность алгоритмов, тем не менее именно

сочетание двух наборов признаков показало последовательное увеличение во всех трех корпусах.

С использованием словарного подхода предложен алгоритм определения полярности тональности. Полярность новостного текста находится в интервале – [-1;1], где значение ближе к -1 означает, что текст является более негативным, ближе к 1 – позитивным. Алгоритм полярности тональности использует сформированные словари: словарь (1) позитивных и (2) негативных слов; (3) словарь отрицателей, уменьшителей и увеличителей тональности. Алгоритм разметки не требует машинного обучения, однако зависит от качества словаря позитивных и негативных слов.

С использованием словарного подхода предложен алгоритм определения степени объективности. Степень объективности определяется следующими критериями: (1) наличие в новостном тексте статистической информации; (2) наличие в новостном тексте ссылки на компетентные источники или персоны. Для определения степени объективности статьи используются сформированные лингвистические словари: (1) словарь компетентных источников, (2) персоналий, (3) слова-связки, используемые при ссылке на источник.

Таким образом, в результате проведения комплекса работ сформированные словари и правила были экспериментально подтверждены для распознавания признака «тональности» с использованием алгоритмов машинного обучения, с использованием словарного подхода для определения полярности тональности, а также степени объективности новостного текста.

Сформированные словари представлены в приложении на диске.

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

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. Вместе с тем регулирование деятельности СМИ сходится в нормах обеспечения объективности, достоверности и подотчётности.

Два из международных принципов профессиональной журналистской этики, выдвинутых под эгидой ЮНЕСКО (1983 г.), посвящены объективности. Первый принцип касается прав граждан на достоверную информацию – *«Народ в целом и отдельные граждане имеют право видеть объективное отражение действительности путем получения достоверной и точной информации ...».* Второй принцип призывает журналистов правдиво и объективно освещать происходящие события – *«Первоочередная задача журналиста - обеспечить право граждан на достоверную и объективную информацию через правдивое освещение происходящих событий, когда факты освещаются в их собственном контексте, с указанием связей различных явлений и без искажений, будучи творчески обработаны журналистом. В этом случае общественность имеет возможность формирования картины реально происходящих событий через доступ к достоверной информации, в которой корни и природа событий, процесс развития и нынешнее состояние фактов отражены максимально объективно»*.[8]

В Законе РК «О средствах массовой информации» основными принципами деятельности СМИ определены: объективность, законность, достоверность и уважение частной жизни, чести, достоинства человека и гражданина.[9]

Вместе с тем определение достоверности новостного контента является довольно сложной задачей. Законодательства различных стран, в т.ч. и Казахстана обязывают журналистов публиковать информацию, соответствующую действительности, призывают к поиску истины и стремлению к достоверности. В связи с этим одной из важных задач

---

[8] International Principles of Professional Ethics in Journalism [Электронный ресурс] URL: https://accountablejournalism.org/ethics-codes/International-Principles (дата обращения: 21.05.2019).
[9] Закон Республики Казахстан «О средства массовой информации», от 23 июля 1999 года. № 451-I.

является определение информативных признаков, которые косвенно могут определить вероятную достоверность публикации.

Непроверенные и субъективные новости читатели могут воспринять с неправильной точки зрения. Наблюдается рост количества фейковых новостей. В свете скорости обмена информацией в Интернете, а также разнообразия новостей нелегко отличить правду от вымысла. [72].

Джанг в своей работе выделяет основные возможные способы борьбы с поддельными новостями: (1) повышение грамотности и когнитивных способностей отличать факты от вымысла, (2) корректирующий подход, который заключается в создании доступных процедур проверки фактов самой аудиторией, (3) обнаружение источников фейковых новостей и фильтрация информации из таких источников [73].

Таким образом информативный признак «объективность» рассматривается как достоверная информация, основанная на доказательной базе, проверяемых и неискажённых фактах, а также отражении разных мнений компетентных источников (персоналий, государственных органов, компаний и др.).

Описание подхода к определению объективности и полученные результаты

Информативный признак «объективность» определяется как информация, отвечающая критериям достоверности. Следует отметить, что данное понятие рассматривается с точки зрения подачи информации в объективном виде без субъективной точки зрения автора. Здесь важно понимать, что цель анализа объективности не в проверке истинности информации, но в проверке стиля подачи информации автором. В этом смысле распознавание фэйковых новостей хоть и близко концептуально, однако отличается по цели анализа.

В обоих случаях оценивается текст, однако в первом случае текст проверяется на критерии достоверности, в то время как во втором проверяется истинность информации в тексте. Вследствие вышесказанного объективная публикация понимается как публикация, отвечающая следующим критериям достоверности:

1) Наличие в новостном тексте ссылки на доверенный новостной сайт;

2) Наличие в новостном тексте ссылки на компетентные источники (государственные органы, местные исполнительные органы, национальные компании, организации образования, организации, медиакомпании и др.) + персоналии (политические или государственные служащие, депутаты, уполномоченные лица компаний или организаций и др.);

3) Наличие цифр (статистической информации).

Для определения степени объективности статьи используются сформированные лингвистические словари: (1) словарь персоналий; (2) компетентные источники; (3) используемые связи при ссылке на источник (таблица 29).

Таблица 29 – Словарь именованных сущностей

№ п/п	Словарь	Количество слов
1	Компетентные источники (должности и организации)	825
2	Персоналии	401
3	Используемые связи при ссылке на источник	88

Алгоритм определения степени объективности строится на основе соответствий заранее определенным условиям и определяется на уровне предложений. Предложение считается объективным если оно отвечает, как минимум одному их следующих условий:

1. наличие слова связки и сайта (доверенные новостные сайты);
2. наличие слова связки и компетентного источника+ персоналии;
3. наличие слова связки и цифры (статистической информации).

Условия можно сложить чтобы получить не только бинарный показатель объективное-субъективное предложение, но и степень объективности как категориальный показатель, состоящий из максимум 3 категорий, где вес каждого условия равен 0.33.

Предложение считается субъективным если в нем не содержится ни одного из вышеупомянутых условий и содержатся местоимения и глаголы с личными окончаниями, указывающие на определенное лицо. Полученная объективность на уровне предложений может быть суммирована и представлена как отношение объективных к субъективным предложениям и нормализована на уровне публикации. Таким образом, степень объективности находится в интервале – [0;1], где значение ближе к 0 означает, что текст субъективен, ближе к 1 – объективен.

Степень объективности рассчитывается по следующей формуле:

$$\text{CO} = \sum_{i=1}^{n} \frac{x_i}{y_i} \qquad (1)$$

где,
CO – степень объективности
x_i – количество объективных предложений
y_i – общее количество предложений

Ниже представлены экспериментальные результаты разметки на основе корпуса, состоящего из 5211 новостей по 5 новостным сайтам.

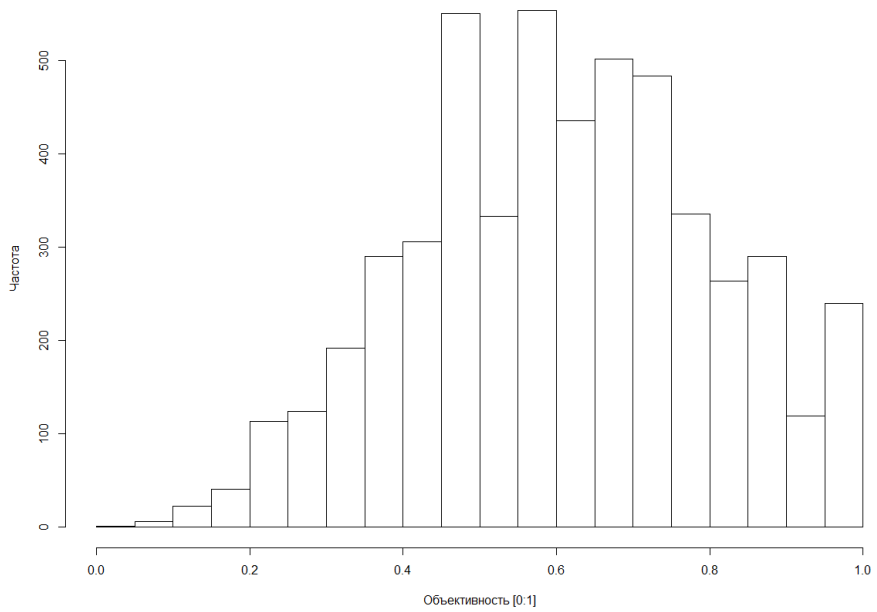

Рисунок 10 - Результаты разметки объективности текста

Из рисунка 10 видно, что основная часть публикаций в корпусе имеет степень объективности от 0.4 до 0.9. В то же время если возникнет необходимость, то пороговое значение объективности можно менять. К примеру, увеличить минимальное достаточное количество достоверных признаков в предложении с 1 на 2.

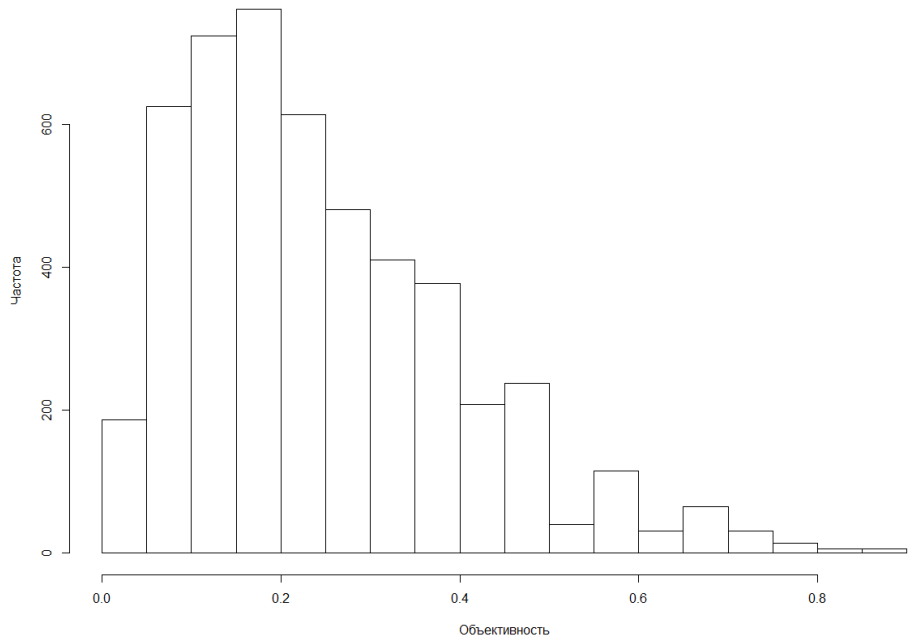

Рисунок 11 - Результаты разметки объективности текста

Экспериментальный корпус из 5211 публикаций, где минимальное количество критериев достоверности равно 2. Из рисунка 11 видно, что основная масса публикаций имеет пониженную степень объективности в диапазоне от 0.05 до 0.4.

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

В целях разработки подхода к определению и распознаванию манипулятивности в текстовых данных разработана инструкция для экспертного анализа и классификации по информативному признаку «манипулятивные техники».

В Инструкции на основании обзора научной литературы дано определение манипулятивности текста: «Под манипулятивностью текста понимается совокупность, определенные комбинации речевых стратегий и тактик, создаваемых с целью неявного воздействия на общественное сознание и выраженных определенными языковыми индикаторами».

Для количественного измерения концепта «манипулятивности текста» разработаны индикаторы:

1. Речевые индикаторы: стратегии и тактики (извлекаются из текстов для теоретических обобщений и разработки методологии исследования).
2. Языковые индикаторы (лексический уровень, грамматический уровень).

В Инструкции содержится схема анализа новостных текстов.

Материалом для лингвистического анализа являются 1000 текстов манипулятивного характера, которые были отобраны из 10 000 текстов путем их внимательного прочтения и наблюдения над языковым материалом.

В инструкции содержатся пошаговые указания для экспертов, позволяющие обеспечить максимальную точность в анализе и направленные на получение однозначных и репрезентативных данных результатов классификации и анализа.

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

Целью работы является формирование сбалансированного, представительного, размеченного и проклассифицированного корпуса для автоматизированной оценки влияния отрытых текстовых информационных источников на социум на основе публикуемой информации.

В рамках комплекса работ по формированию корпуса публицистических текстов: текстов официальных СМИ, журнальных публикаций, новостных порталов и др., были определены следующие задачи:

- проведение первичного отбора источников и сбор данных;
- определение требований к корпусу: сбалансированность, представительность, основные категории мета-разметки;
- формирование основного, малого и тестового корпусов;
- чистка корпусов и систематизация данных, разметка тестового корпуса.

11.1 Обзор научной литературы

При реализации задачи в рамках данного исследования, были учтены существующие практики работы с текстовыми корпусами в рамках научной литературы. Изученная информация была структурирована по основным характеристикам, необходимым для учета в дальнейшей работе. Принято во внимание, что текстовые корпуса, всегда представляют собой некую выборку, соответствующую определенным критериям и в определенной пропорции, согласуются с характеристиками и параметрами исследуемого феномена.

Структура текстов корпуса зависит от поставленных исследованием задач. В корпус тексты отбираются в соответствии с проблемной областью, т.е. областью, которой интересуется исследователь. В данном случае, поскольку объектом исследования являются электронные новостные интернет-источники и их контент, корпус должен включать *статьи электронных СМИ казахстанского медиапространства*.

Одним из самых важных требований, которое необходимо учитывать при формировании корпуса, является требование *представительности (репрезентативности)*. Под представительностью понимается способность корпуса отражать все свойства исследуемой области [74].

Дуглас Бибер, автор одной из наиболее цитируемых статей, посвященных представительности корпусов утверждает, что фокусирование внимания на размере

выборки является ошибочной стратегией для формирования корпусов и подчеркивает необходимость достижения максимально возможной представительности. Он также отмечает, что при формировании корпусов необходимо получить максимальную вариативность данных, описывающих исследуемый феномен для более глубокого его понимания [75].

Корпусная лингвистика опирается на корпус как на выборку большей популяции. Как и во всех других научных дисциплинах, эмпирические результаты, полученные на основе выборки, могут быть обобщены для большей популяции только в том случае, если выборка является репрезентативной для этой популяции.

Согласно Д. Биберу, представительность корпуса определяется «степенью, в которой выборка включает полный диапазон вариативности популяции» [75, стр. 244]. Представительность имеет решающее значение при формировании и проектировании корпусов, поскольку целью большинства исследований корпуса состоит в том, чтобы выявить количественные языковые модели в выборке корпуса и обобщить эти выводы для большей языковой группы.

Тем не менее, эмпирический опыт решения вопроса представительности представлен лишь в небольшой доле работ представителей научного сообщества, таких как Д. Бибер и С.Т. Грайс [76]. Такие исследователи как Г. Лич [77], Т. Варади [78], и С.Т. Грайс [79] в своих работах уделили внимание рекомендациям по формированию и обеспечению представительности корпусов. Также среди научного сообщества существует альтернативное мнение, заключающееся в том, что вопрос представительности корпуса не является существенным и обеспечивается за счет увеличения количества текстов в корпусе, т.е. объема выборки. Апологетами этой точки зрения являются Дж. Синклэр [80] и П. Хэнкс [81].

В работе МакЭнери и соавторов [82] описывается, что концептуально и методологически существует два основных типа репрезентативности в отношении двух параметров: исследуемая область и лингвистическая репрезентативность.

Репрезентативность корпуса в отношении исследуемой области представляет собой ту степень, в которой корпус содержит полный диапазон вариативности по типам тестов, характерным для исследуемой области. Репрезентативность в отношении исследуемой области является решающим аспектом для исследователя, и позволяет ему проецировать результаты, полученные на основе анализа одной конкретной выборки на всю исследуемую область.

Лингвистическая репрезентативность представляет собой ту степень, в которой корпус содержит весь полный диапазон вариативности лингвистических характеристик,

присущих исследуемой области. Лингвистическая репрезентативность определяет пригодность выборки корпуса для ответа на конкретные вопросы исследования о конкретных языковых особенностях. Важно отметить, что языковая репрезентативность по своей сути связана с изучаемым языковым уровнем; один и тот же корпус может быть представителем общей грамматической структуры, но не лексических распределений.

Основной методологической проблемой при оценке репрезентативности корпуса является оценка языковых и ситуационных характеристик цензуса, с которым сравнивают корпус. Поскольку собрать всю совокупность данных, отражающих определенный феномен, как правило, не представляется возможным.

Дж. Грей и соавторы в своей работе, посвященной методам оценки представительности корпуса обобщили подходя к такой оценке, принятые в научно-исследовательской деятельности [83] в соответствии с представленным в таблице 30, где отображены параметры оценки представительности (1 и 2), исходя из возможности (А) или невозможности (В) определения характеристик цензуса, а также сценарии оценки представительности корпуса в этих случаях.

Таблица 30 – Подходы к оценке репрезентативности корпуса

	А. Возможность определения характеристик цензуса присутствует	**В. Возможность определения характеристик цензуса отсутствует**
1. Представительность в отношении исследуемой области	**А1.** Проведение детального анализа ситуационных характеристик каждого текста в корпусе. Сравнение таких характеристик с характеристиками исследуемой области. Другим вариантом является экстраполяция обобщений по корпусу только на определенную часть популяции.	**В1.** Сравнение встречаемости и / или пропорции ситуационных характеристик, представленных в полной популяции, с серией экспериментальных под-корпусов, которые были собраны при помощи различных методов формирования корпусов.
2. Лингвистическая представительность	**А2.** Разделение сформированного корпуса на более мелкие подмножества, отобранные методом случайной выборки. Сравнение дисперсии лингвистических характеристик в малых выборках с лингвистическими характеристиками всего корпуса.	**В2.** Сравнение распределения ряда лингвистических особенностей цензуса, с серией экспериментальных под-корпусов, которые были собраны при помощи различных методов формирования корпусов.

Под *сбалансированностью* корпуса понимается обеспечение пропорционального присутствия в корпусе текстов, отражающих категории исследуемой области. В сбалансированных корпусах в корпус включаются тексты, представляющие разные модусы дискурса, разнообразные по жанрам, стилям, тематике, классам и т.п.

В целом сбалансированность корпуса является качеством, достигаемым искусственным путем. Сбалансированность корпуса может достигаться за счет

равноценного распределения каждой из категорий (одинаковое количество текстов, представляющих определенную категорию) или за счет пропорционального распределения [84]. В последнем случае каждой категории текстов присваивается определенная весовая доля, в соответствии с которой и определяется количество текстов в корпусе. Категория (признак, класс), по которой оценивается сбалансированность корпуса, зависит от целей исследования. Кроме того, в зависимости от них же, при обеспечении сбалансированности корпуса можно учитывать не одну, а несколько категорий.

Таким образом, сбалансированные корпуса имеют фиксированный объем, и процедура пополнения таких корпусов происходит согласно строгой и четко выработанной процедуре отбора. Также принцип сбалансированности может быть соблюден в разрезе источников (определенное *пропорциональное соотношение публикаций из разных источников*).

11.2 Формирование корпуса

В ходе проведенного комплекса работ по формированию дополнительного сбалансированного, представительного, размеченного и проклассифицированного корпуса публицистических текстов информационных новостных порталов было сформировано 3 корпуса:

1) основной корпус, включающий 1 982 433 статьи из 22 источников с мета-разметкой по 14 категориям;

2) малый корпус текстов, включающий 245 352 статьи из 5 источников с мета-разметкой по 7 категориям;

3) тестовый корпус, включающий 35 643 статьи из 5 источников, с разметкой по 39 категориям и классификацией по части из признаков.

Методология формирования основного корпуса. Из общего списка доменов, состоящего из 2613 СМИ с указанием домена, типа СМИ (печатные, телевидение, радио, новостные агентства, интернет-СМИ), и масштаба (региональные, центральные), на основе специфики методов машинного обучения при анализе текстовой информации [85], был произведен отбор источников. При анализе и отборе источников учитывались требования по:

- исключению региональных СМИ;
- исключению узкоспециализированных источников;
- обеспечению включения в состав корпуса публикаций на русском языке;
- включению в корпус информационных порталов, специализирующихся на предоставлении ежедневных новостей о событиях в стране;

- включению в корпус публикаций информационных порталов СМИ, различных по уровню популярности среди населения.

Исходя из этих требований, из выборки были исключены региональные СМИ, узкоспециализированные СМИ, СМИ публикующие статьи на других языках, СМИ активность которых приостановлена.

В результате был сформирован основной корпус из 1 982 433 статей, основные характеристики которого представлены в таблице 31.

Таблица 31 - Основные характеристики дополнительного большого корпуса

№ п/п	Источник	Период	Количество публикаций
1	https://365info.kz	05.2014-02.2019	57 326
2	https://rus.azattyk.org	08.2003-02.2019	52 583
3	https://bnews.kz	04.2012-02.2019	7 631
4	https://www.caravan.kz	07.2001-02.2019	246 125
5	http://dailynews.kz	02.2012-02.2019	25 409
6	https://esquire.kz	05.2013-02.2019	1 510
7	http://www.exclusive.kz	10.2006-02.2019	7 184
8	https://kz.expert	01.2017-12.2018	671
9	https://forbes.kz	12.2011-02.2019	159 755
10	https://www.inform.kz	01.2010-02.2019	341 232
11	https://kapital.kz	09.2007-02.2019	73 468
12	https://www.kazpravda.kz	07.2013-02.2019	432
13	http://www.kp.kz	12.2011-02.2019	12 422
14	https://kursiv.kz	10.2015-02.2019	4 253
15	https://www.nur.kz	08.2018-02.2019	18 003
16	https://ru.sputniknews.kz	05.2016-01.2019	23 406
17	https://tengrinews.kz	01.2009-02.2019	198 797
18	https://time.kz	05.2016-02.2019	46 557
19	http://today.kz	12.2011-05.2017	102 456
20	http://vesti.kz	10.2008-08.2018	27 740
21	https://vlast.kz	04.2012-02.2019	21 051
22	https://www.zakon.kz	Информация отсутствует	554 422
	Всего		**1 982 433**

По результатам работ над формированием дополнительного большого корпуса были выделены следующие категории мета-разметки:

1. Автор публикации – имя автора;
2. Категория/ тема публикации – тематический раздел сайта, к которому относится публикация;
3. Дата выхода публикации;
4. Дата скачивания публикации;

5. Идентификационный номер публикации - автоматически присваиваемый статье идентификационный номер для ее поиска и исключения двойного учета или «неучтения» статьи в процессе анализа;
6. Ссылки - наличие гиперссылок на др. статьи или странице в тексте самой публикации;
7. Количество комментариев;
8. Количество просмотров (на дату скачивания);
9. Заголовок;
10. Подзаголовок;
11. Теги (теги и хеш-теги присвоенные публикации);
12. Текст публикации;
13. Текст публикации в формате html;
14. URL адрес публикации.

В нашем случае репрезентативность в отношении исследуемой области большого корпуса не подлежит сомнению, так как выборка дополнительного корпуса включила исключительно из текстов, представляющих исследуемую область (всю популяцию текстов открытых информационных, источников, оказывающих влияние на социум Казахстана). Поскольку основной целью проекта является исследование влияния открытых информационных текстовых источников на социум, данная задача была решена полностью: дополнительный корпус представляет собой собрание текстов исключительно из источников данного типа, а также охватывает практически весь период существования таких источников, и практически весь спектр источников, оказывающих влияние на социум.

В рамках исследования, исходя из того, что собранный корпус отражает всю совокупность казахстанских открытых текстовых источников СМИ с информацией новостного характера, оказывающих влияние на социум, требование к сбалансированности всего корпуса необходимо не воспринимать как критически важное, поскольку самое главное условие - отражение объективной реальности, полностью соблюдено (рисунок 12).

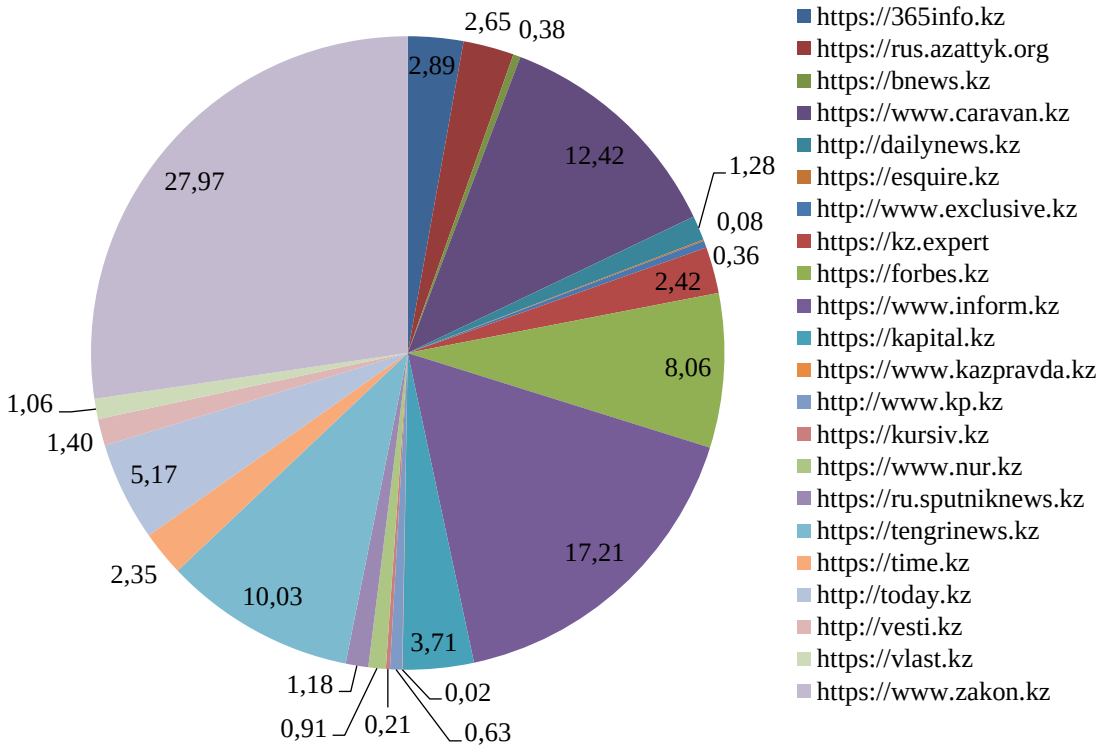

Рисунок 12 - Сбалансированность состава основного дополнительного корпуса по источникам, %

Необходимо также уточнить, что для целей машинного обучения и тестирования алгоритмов проектируемой в рамках проекта информационной системы, из общего корпуса осуществлена выборка, где принцип сбалансированности в отношении исследуемых признаков соблюден надлежащим образом.

Необходимо отметить, что во многих случаях данные по вышеуказанным категориям мета-разметки отсутствуют. Это обусловлено, прежде всего, существующей архитектурой данных. К примеру, лишь небольшая часть сайтов спроектирована таким образом, что можно собрать данные по количеству просмотров и количеству комментариев к публикациям, размещенным на них. На многих сайтах возможность комментирования публикаций отсутствует. Это значительно усложняет анализ данных, поскольку с точки зрения оценки влияния на социум, количество комментариев и просмотров публикаций является, почти единственным прямым индикатором заинтересованности населения.

11.3 Формирование малого и тестового корпусов и категории их разметки

Как было установлено в ходе предыдущих этапов исследования, для оценки влияния медиа-контента на социум основными показателями являются количество просмотров, количество комментариев и количество шейров/репостов новостной публикации. Для целей исследования, заключающихся в оценке влияния новостных СМИ на социум, необходимо обеспечить наличие этих показателей. По этой причине из списка источников,

вошедших в основной корпус, были отобраны те, которые содержат минимальный набор данных, позволяющих произвести оценку интереса казахстанского общества к теме, поднимаемой в публикации: количество комментариев и просмотров. Исходя из вышесказанного, был составлен малый корпус, в состав которого вошли только те источники, по которым имеются обе эти категории метаданных.

Таким образом, малый дополнительный корпус включает 245 352 публикации из 5 источников: tengrinews.kz, www.kp.kz, www.kazpravda.kz, vesti.kz, vlast.kz (таблица 32).

Таблица 32 – Состав малого дополнительного корпуса

№ п/п	Источник	Период	Количество публикаций
1	https://tengrinews.kz	01.2009-02.2019	197 983
2	http://www.kp.kz	12.2011-02.2019	12 422
3	https://www.kazpravda.kz	07.2013-02.2019	432
4	http://vesti.kz	10.2008-08.2018	27 740
5	https://vlast.kz	04.2012-02.2019	6 775
	Всего		**245 352**

Категории разметки малого дополнительного корпуса включают:

1. Идентификационный номер публикации;
2. URL-адрес публикации;
3. Дата выхода публикации;
4. Дата скачивания публикации;
5. Текст публикации;
6. Количество просмотров публикации;
7. Количество комментариев к публикации.

Для тестирования работы алгоритмов анализа тестирования эффективности применения словарей для дальнейшей классификации текстов и дальнейших работ по проекту из малого дополнительного корпуса был составлен тестовый корпус, который меньше по объему.

Составление тестового корпуса обусловлено необходимостью экспериментального подтверждения действенности алгоритмов, методов и походов, выработанных для целей проекта, а также тем, что при машинной обработке данных и вычислительных процессах тестирование на всем корпусе нецелесообразно ввиду больших потерь времени и возникновения ошибок.

Для того, чтобы соблюсти все методологические условия, позволяющие экстраполировать результаты, полученные на тестовом корпусе на всю совокупность данных, входящих в состав малого корпуса, отбор публикаций в тестовый корпус производился методом случайной выборки, при котором каждая публикация имела одинаковую вероятность попадания в состав тестового корпуса выборки. Общая статистика тестового корпуса представлена в таблице 33.

Таблица 33 - Общая статистика тестового корпуса

№ п/п	Источник	Количество публикаций
1	https://tengrinews.kz	29 128
2	http://www.kp.kz	1 873
3	https://www.kazpravda.kz	59
4	http://vesti.kz	3 693
5	https://vlast.kz	890
	Всего	**35 643**

Категории разметки тестового корпуса включают:

1. Идентификационный номер публикации;
2. Источник публикации;
3. URL-адрес публикации;
4. Дата выхода публикации;
5. Дата скачивания публикации;
6. Текст публикации;
7. Количество просмотров публикации;
8. Количество комментариев к публикации;
9. Количество слов
10. Количество знаков
11. Количество предложений
12. Количество существительных;
13. Количество глаголов;
14. Количество прилагательных;
15. Количество местоимений;
16. Количество наречий;
17. Количество позитивных слов;
18. Количество негативных слов;
19. Количество политизированных слов;
20. Количество манипулятивных слов;

21. Количество слов из словаря, составленного по признаку «призыв к действию»;

22. Количество слов из словаря, составленного по признаку «предвзятость»;

23-39. Количество слов из словарей по признаку «сфера» (17 сфер).

11.4 Экспериментальные корпуса для распознавания признака «тональность»

Для экспериментального подтверждения метода сентимент-классификации были собраны три малых экспериментальных корпуса.

Первый корпус представляет собой изначальный корпус и включает 5211 публикаций, отобранных методом систематической кластерной выборки из 5 казахстанских новостных порталов: tengrinews.kz, kz.expert, camonitor.kz, azattyq.org, zakon.kz. Сентимент-разметка (количественная оценка) корпуса производилась с применением Google Cloud Natural Language API [86].

Второй корпус является корпусом новостных текстов открытого доступа, размещенным на сайте kaggle.com [87] и включает 8260 новостных публикаций, собранных из разных социальных сетей и размеченных по трем категориям тональности.

Третий корпус, включающий 32292 текста, был составлен в рамках реализации проекта Linis Crowd Лаборатории Интернет исследований ВШЭ (г. Санкт-Петербург, Россия), где сентимент разметка производилась с применением алгоритма, предложенного Кольцовой и Кольцовым [10]. Сентимент классификация малых экспериментальных корпусов представлена ниже в таблице 34.

Таблица 34 – Сентимент классификация малых экспериментальных корпусов

Корпус	Категория классификации	Кол-во
1) 5211 публикаций	Позитивный	1436
	Нейтральный	2068
	Негативный	1707
2) 8260 публикаций	Позитивный	2794
	Нейтральный	4032
	Негативный	1434
3) 32292 публикаций	Позитивный	3180
	Нейтральный	16856
	Негативный	12256

Помимо этого, в соответствии с методологией, предложенной Ченло и Лосада [34], была проведена разметка всех трех корпусов по следующим группам информативных признаков:

- *Словарные признаки* (Vocabulary features). Данная группа включает N-граммы: униграмы (1) и биграммы (2), количество уникальных слов (3). Оставшиеся 4 предиктора

группы были определены с помощью программного обеспечения Statistica[10] и представляют собой концепты (оцифрованная репрезентация текстовой информации путем определения важных параметров с наибольшей вариативностью между словами и документами) , полученные в результате процедуры текст-майнинга – 4 концепта (4-7).

- *Лексические признаки* (Lexicon based features). Данные признаки были получены с применением подхода «мешок слов» (bag-of-words approach) и включают: количество точек (8), запятых (9), восклицательных знаков (10), вопросительных знаков (11), многоточий (12), кавычек (13), скобок (14), тире (15) и дефисов (16).

- *Признаки длины* (Length features). Предикторы данной группы включают количество предложений (17) и знаков (18). Количество знаков было рассчитано с учетом пробелов.

- *Частеречная разметка* (PoS tagging). Группа включает такие признаки как: количество прилагательных (19), предлогов (20) , наречий (22), вспомогательных слов (22), соединительных союзов (23), определений (24), междометий (25), существительных (26), числительных (27), частиц (28), местоимений (29), имен собственных (30), знаков пунктуации (31), подчинительных союзов (32), глаголов (33). Для частеречной разметки использовались библиотеки открытого доступа UDPipe [43] для статистического ПО R.

- *Тематические признаки* (Topic area features). Изучение научной литературы посвященной сентимент анализу позволила сделать предположение, что тематическая сфера, к которой относится новостная публикация, также может выступать предиктором тональности. На основе результатов лингвистического анализа 2000 текстов, проведенным привлеченными экспертами в области русской филологии, были составлены словари тематических сфер. Сферы были определены по результатам анализа тематических тегов новостных сайтов и консультаций с 7 экспертами в области журналистики, СМИ, массовых коммуникаций. Словари были составлены по следующим сферам: административные и территориальные вопросы (123 слова), экономика и финансы (238 слов), внешняя политика (95 слов), гражданское общество (188 слов), здоровье и здравоохранение (141 слово), индустриально-инфраструктурное развитие и технологии (115 слов), оборона и безопасность (77 слов), образование и наука (160 слов), правительство и парламент (97 слов), преступления и происшествия (326 слов), сельское хозяйство (61 слово), социальное обеспечение (28 слов), спорт, культура и развлечения (243 слова), труд (25 слов), энергетика (70 слов), экология и туризм (40 слов), чрезвычайные происшествия (53 сова). В общем, эта

[10] http://statsoft.ru

группа включает 17 предикторов (34-50), рассчитываемых как количество слов из соответствующего словаря в публикации.

- *Содержательные информативные признаки* (Content specific features). Данная группа включает в себя такие признаки как политизированность (51), манипулятивность (52), призыв к действию (53), предвзятость (54), позитивные (55) и негативные слова (56). Количественное выражение этих признаков основано на встречаемости в публикациях слов из соответствующих словарей: политизированность (30 слов), манипулятивность (652 слова), призыв к действию (89 слов), предвзятость (32 слова), позитивные (979 слов) и негативные слова (2112 слов).

Размеченные экспериментальные корпуса представлены в приложении на диске.

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

Целью работы является изучение и анализ фактических характеристик сформированного корпуса новостных текстов.

В ходе проведенного комплекса работ по анализу фактических характеристик сформированного корпуса были проанализированы состав, структура, представительность, сбалансированность, корректность и полнота корпуса.

Также были представлены особенности 3 корпусов: основного, малого и тестового:

1) Основной корпус. Описаны структура, состав, категории мета-разметки. Проведен анализ представительности, сбалансированности, а также корректности корпуса по закону Ципфа.

2) Малый корпус. Проанализирован состав, структура, сбалансированность в разрезе источников, уровни мета-разметки (7 категорий), и характеристики корпуса по статистике комментариев и просмотров.

3) Тестовый корпус. Проанализированы фактические характеристики корпуса: состав, структура, сбалансированность. Категории разметки корпуса включили, мета-разметку, морфологическую разметку, лексическую разметку по сформированным лингвистическим словарям. В их состав входят 39 категорий:

1. идентификационный номер публикации;
2. источник публикации;
3. URL-адрес публикации;
4. дата выхода публикации;
5. дата скачивания публикации;
6. текст публикации;
7. количество просмотров публикации;
8. количество комментариев к публикации;
9. количество слов;
10. количество знаков;
11. количество предложений;
12. количество существительных;
13. количество глаголов;
14. количество прилагательных;
15. количество местоимений;

16. количество наречий;

17. количество позитивных слов;

18. количество негативных слов;

19. количество политизированных слов;

20. количество манипулятивных слов;

21. количество слов из словаря, составленного по признаку «призыв к действию»;

22. количество слов из словаря, составленного по признаку «предвзятость»;

23-39. количество слов из словарей по признаку «сфера» (17 сфер).

На основе анализа этих данных описаны фактические характеристики корпуса по морфологической и лексической разметке, проведена классификация корпуса по признаку «тональность» и «объективность». Лексическая разметка выполнялась путем подсчета количества слов, отнесенных лингвистами к словарю по определенному признаку. В дальнейшем эти данные будут использоваться для классификации текстов словарно-лексическим подходом.

В общем составе тестового корпуса при его классификации по признаку «тональность» количество нейтральных статей составило 4,4%, позитивных – 56,46%, негативных – 39,13%. По признаку «объективность» корпус был проклассифицирован с использованием квинтилей, при этом распределение данных по этому классу по корпусу сбалансировано: классы представлены практически в равных пропорциях. Так, доля публикаций с очень высокой степенью объективности составила – 19,6%, с высокой степенью объективности - 20,31%, со средней степенью объективности – 19,10%, с низкой степенью объективности – 20,28%, с очень низкой степенью объективности -20,71%.

Аналогичный анализ был проведен в разрезе источников. При этом проанализированы фактические характеристики по объему публикаций, статистике просмотров и комментариев, морфологической разметке, лексической разметке, классификации по признакам «тональность» и «объективность».

В целом распределение данных в разрезе источников показало, что в составе публикаций из порталов vlast.kz и kazpravda.kz превалируют публикации с высокой и очень высокой степенью объективности, в источниках kp.kz и vesti.kz с низкой и очень низкой степенью объективности, а распределение данных по корпусу tengrinews.kz по признаку «объективность» между классами сбалансировано.

Как было описано в разделе 11 настоящего отчета для экспериментального подтверждения метода сентимент-классификации были собраны три малых экспериментальных корпуса (таблица 34).

В размеченном Корпусе 1 позитивные статьи составили 27,6%, нейтральные 39,7%, негативные 32,7%. В Корпусе 2 распределение статей по классам составило 33,8%, 48,8%, 17,4% для категорий позитивные, нейтральные и негативные соответственно. В Корпусе 3 нейтральный класс тоже является мажоритарным – 52,2%, негативный класс составил 38%, а позитивный миноритарный класс – всего 9,8%. Наибольшая несбалансированность наблюдается в Корпусе 3, относительная сбалансированность данных наблюдается в Корпусе 1.

Для описания фактических характеристик экспериментальных корпусов по оставшимся категориям разметки (53 предиктора, за исключением 4 концептов, полученных в результате текст майнинга), ниже в таблице 35 приведены данные о количестве маркеров отдельных информационных признаков в Корпусах 1, 2 и 3 соответственно.

Таблица 35 - Фактические характеристики экспериментальных корпусов

№	Информативные признаки	Количество		
		Корпус 1	**Корпус 2**	**Корпус 3**
Словарные признаки				
1.	униграмы	2107333	3960989	5301762
2.	биграммы	287521	819525	1257213
3.	уникальные слова	2020003	3711361	4952135
Лексические признаки				
4.	количество точек	126162	232672	243054
5.	количество запятых	182830	292003	488840
6.	количество восклицательных знаков	887	1752	7478
7.	количество вопросительных знаков	3481	4382	8421
8.	количество многоточий	298	905	3188
9.	количество кавычек	31393	74415	521
10.	количество скобок	7945	32518	72306
11.	количество тире	4918	14917	7905
12.	количество дефисов	37910	87810	123001
Признаки длины				
13.	количество предложений	130828	239711	262141
14.	количество знаков с пробелами	15454781	30009901	40019244
Частеречная разметка				
15.	количество прилагательных	238261	494421	658806
16.	количество предлогов	266924	487543	602896
17.	количество наречий	92285	142037	236148
18.	количество вспомогательных слов	18374	30917	41485
19.	количество соединительных союзов	74806	129569	217341
20.	количество определений	54966	80860	134723
21.	количество междометий	102	260	644
22.	количество существительных	789553	1636811	2167859
23.	количество числительных	68081	199502	145654
24.	количество частиц	48473	64832	146773
25.	количество местоимений	92189	115287	196125
26.	количество имен собственных	17502	35616	23896
27.	количество знаков пунктуации	392286	745173	1023077
28.	количество подчинительных союзов	42184	54134	92334
29.	количество символов	4375	23030	9970
30.	количество глаголов	268904	428934	624720

Тематические признаки				
31.	административные и территориальные вопросы	29349	55442	52541
32.	экономика и финансы	41325	185962	63793
33.	внешняя политика	22592	33435	58910
34.	гражданское общество	28863	45116	57618
35.	здоровье и здравоохранение	7019	8153	19288
36.	индустриально-инфраструктурное развитие и технологии	19472	64975	31210
37.	оборона и безопасность	6606	7865	69712
38.	образование и наука	20864	29720	27084
39.	правительство и парламент	46976	84866	52246
40.	преступления и происшествия	42722	47015	74262
41.	сельское хозяйство	2312	5166	3213
42.	социальное обеспечение	7186	19428	11776
43.	спорт, культура и развлечения	25469	30971	28625
44.	труд	3638	8903	5448
45.	энергетика	6173	22748	11614
46.	экология и туризм	3734	6777	6806
47.	чрезвычайные происшествия	3760	3185	6169
Содержательные информативные признаки				
48.	политизированность	6461	6259	11465
49.	манипуляция	21276	26696	68333
50.	призыв к действию	13764	28827	25722
51.	предвзятость	1014	1638	1414
52.	количество позитивных слов	89465	204687	188423
53.	количество негативных слов	56492	69773	269533

ЗАКЛЮЧЕНИЕ

В ходе проведения исследования в 2019 году в соответствии с календарным планом получены следующие результаты:

1. Проведен анализ эффективности разработанной в 2018 году методики для оценки влияния открытых информационных источников на социум на основе публикуемой текстовой информации.

2. На основе результатов анализа усовершенствована методика оценки влияния открытых текстовых информационных источников на социум на основе публикуемой информации.

3. Разработана методика расчета информативных признаков тональности, степени объективности публикации. Предложен подход распознавания манипулятивности новостных текстов.

4. Сформированы 27 лингвистических словарей по информативным признакам тональности, политизированности, манипулятивности, призыва к действию, предвзятости, именованных сущностей, тематических областей и др.

5. Экспериментально подтверждены словари и правила распознавания признака «тональность» с применением двух разных подходов, и определения признака «объективность» текста.

6. Сформированы 4 корпуса новостных текстов: (1) основной корпус текстов официальных СМИ, журнальных публикаций, новостных порталов, включающий 1 982 433 статьи из 22 источников; (2) малый дополнительный корпус текстов, включающий 245 352 статьи из 5 источников, где имеются данные по количеству просмотров и комментариев; (3) тестовый корпус, включающий 35643 статьи из 5 источников, с разметкой и классификацией; (4) экспериментальные корпуса (i) с сентимент-разметкой с применением Google Cloud Natural Language API, (ii) размеченный корпус открытого доступа по трем категориям тональности, размещенным на сайте kaggle.com, (iii) корпус, размеченный в рамках проекта Linis Crowd Лаборатории интернет Исследований ВШЭ (г. Санкт-Петербург, Россия).

7. Разработана методика определения степени объективности новостной текстовой информации.

8. Определены статистические и текстовые закономерности, присутствующие в размеченном корпусе новостных текстов.

9. Описаны фактические характеристики корпусов новостных текстов (состав, структура, представительность, сбалансированность, корректность, полнота корпуса, категории разметки).

Все поставленные задачи выполнены в полном объеме и в соответствии с календарным планом.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. О национальном корпусе русского языка. [Электронный ресурс] URL: http://www.ruscorpora.ru/corpora-structure.html (дата обращения: 30.09.2018)/

2. Reis J.C.S., Correia A., Murai F., Veloso A., Benevenuto F., & Cambria E. Supervised Learning for Fake News Detection // IEEE Intelligent Systems. – 2019. – Vol. 34(2). – P. 76–81. doi: 10.1109/mis.2019.2899143

3. Bushman B.J., Whitaker J.L. Media Influence on Behavior // Encyclopedia of Human Behavior. – 2012. – P. 571-575.

4. Geertzen, J. Inter-Rater Agreement with multiple raters and variables / [Электронный ресурс] URL: https://nlp-ml.io/jg/software/ira/ (дата обращения: 18.05.2019)

5. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // Journal of Machine Learning Research, 2003.Vol. 3. P. 993–1022.

6. К. В. Воронцов. Вероятностное тематическое моделирование / [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (дата обращения: 14.08.2019)

7. Korenčić D., Ristov S., Najder J.E. Document-based topic coherence measures for news media text // Expert Systems with Applications. – 2018. – Vol, 114. – P. 357-373.

8. Атанаева М.К., Булдыбаев Т.К., Оспанова У.А., Акоева И.Г., Нурумов К.С., Баймаханбетов М.А. Методика для определения информативных признаков новостных текстов и проверка их значимости // Научный аспект. – 2019. – №3.

9. Hansen L. K., Arvidsson, A., Nielsen, F. A., Colleoni, E., & Etter, M. Good Friends, Bad News - Affect and Virality in Twitter // Communications in Computer and Information Science Future Information Technology. – 2011. – Vol. 34–43. doi: 10.1007/978-3-642-22309-9_5

10. Koltsova O.Y., Alexeeva S.V., Koltsov S.N. An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”. – Moscow, 2016. –P. 277–287.

11. Visa S., Ramsay B., Ralescu, A.L. Van Der Knaap E. Confusion matrix-based feature selection. MAICS. – 2011. – Vol. 710. – P. 120–127.

12. Prasath S., Arafat A.A.H., Lasassmeh O., & Hassanat A. Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier / A Review. – 2017.

13. Chomboon K., Chujai P., Teerarassammee P., Kerdprasop K., & Kerdprasop N. An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm. 2015. P. 280-285. 10.12792/iciae2015.051.

14. Weiss G.M. Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? // Proceedings of the 2007 International Conference on Data Mining, CSREA Press, 2007. – P. 35-41

15. McCarthy K. Does Cost-Sensitive Learning Beat Sampling for Classifying Rare Classes? // Proceedings of the First International Workshop on Utility-Based Data Mining (at KDD-05). – ACM Press, 2005. – P. 69-75.

16. Nikulin V. N. On the Evaluation of the Homogeneous Ensembles with CV-passports. // PAKDD 2013, Springer LNCS 7867, J. Li et al. (Eds.) – 2013. – P. 109–120.

17. Chawla N.V., Hall L.O., Bowyer K.W., and Kegelmeyer W.P., "SMOTE: Synthetic Minority Oversampling Technique" // Journal of Artificial Intelligence Research. – 2002. – Vol. 16. – P. 321-357.

18. Chawla N.V., Lazarevic A., Hall L.O., and Bowyer K.W. SMOTEBoost: Improving Prediction of the Minority Class in Boosting // Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases. – 2003. – P. 107-119.

19. Han H., Wang W., and Mao B., "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning" // Proceedings of the International Conference on Intelligent Computing (ICIC '05), Lecture Notes in Computer Science. – 2005. – Vol. 3644. – P. 878–887.

20. Zhang H.X., Li M.F. RWO-Sampling: a random walk over-sampling approach to imbalanced data classification // Information Fusion. – 2014. – Vol. 20. – P. 99–116.

21. Guo H., Viktor H.L. Learning from Imbalanced Data Sets with Boosting and Data Generation: the DataBoost-IM Approach // SIGKDD Explorations: Special issue on Learning from Imbalanced Datasets. – 2004. – Vol. 6(1). – P. 30 - 39.

22. He H., Garcia E. Learning from Imbalanced Data // IEEE Transactions on Knowledge and Data Engineering. – 2009. – Vol. 21(9). – P. 1263-1284. doi:10.1109/tkde.2008.239

23. Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques // Proceedings of the Conference on Empirical Methods in Natural Language Processing. – Philadelphia, 2002. – P. 79-86.

24. Pang B. & Lee L. A Sentimental Education: Sentiment Analysis Using Subjective Summarization Based on Minimum Cuts // Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics. – Barcelona, 2004. – P. 271-278.

25. Pang B., Lee, L. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales // Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics. – Ann Arbor, Michigan, 2005. – P. 115-124.

26. McDonald R., Hannan K., Neylon T., Wells M., Reynar J. Structured Models for Fine-to-Coarse Sentiment Analysis // Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. – Prague, 2007. – P. 432-439.

27. Ikeda D., Takamura H., Ratinov L., Okumura M. Learning to Shift the Polarity of Words for Sentiment Classification. // Transactions of the Japanese Society for Artificial Intelligence. – 2008 or 2010. – Vol. 25(1). – P. 50-57.

28. Maas A., Daly R., Pham P., Huang D., Ng, A., & Potts C. Learning Word Vectors for Sentiment Analysis // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. – Portland, 2011. – P. 142-150.

29. Xia R., Xu F., Yu J., Qi Y., & Cambria, E. Polarity Shift Detection, Elimination and Ensemble: A Three-Stage Model for Documents-Level Sentiment Analysis // Information Processing and Management. – Vol. 52, 2015 or 2016. – P. 36-45.

30. Araque O., Corcuera-Platas I., Snachez-Rada F., & Iglesias C. Enhancing Deep Learning Sentiment Analysis with Ensemble Techniques in Social Applications // Expert System with Applications. – Vol. 77, 2017. – P. 236-246.

31. Chen T., Xu R., He Y., & Wang X. Improving Sentiment Analysis Via Sentence Type Classification Using BiLSTM-CRF and CNN // Expert Systems with Applications. – Vol. 72, 2016. – P. 221-230.

32. Jianqiang Z., Xiaolin G., & Xuejun Z. Deep Convolution Neural Networks for Twitter Sentiment Analysis // IEEE Access. – Vol. 6, 2018. – P. 23253-23260.

33. Zhang L., Wang S., & Liu B. Deep Learning for Sentiment Analysis: A Survey // Wiley Interdisciplinary Review: Data Mining and Knowledge Discovery. – 2018. – DOI: 10.1002/widm.1253.

34. Chenlo M.L., Losada, E.D. An Empirical Study of Sentence Features for Subjectivity and Polarity Classification // Information Sciences. – 2014. – Vol. 280. – P. 175-288.

35. Turney P. D., & Littman M. L. Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus (Report No. ERB-1094) NRC 44929. – 2002.

36. Denecke K. Using SentiWordNet for Multilingual Sentiment Analysis // Proceedings of the 2008 IEEE 24th International Conference on Data Engineering Workshop. – Cancun, 2008. – P. 507-512.

37. Ding X., Liu B., & Yu P. A Holistic Lexicon-Based Approach to Opinion Mining // Proceedings of the 2008 International Conference on Web Search and Data Mining. – Palo-Alto, 2008. – P. 231-240.

38. Taboada M., Brooke J., Tofiloski M., Voll K., & Stede M. Lexicon Based Methods for Sentiment Analysis // Computational Linguistics. – Vol. 37(2), 2011. – P. 267-307.

39. Tan S., Wang Y., & Cheng X. Combining Learn-Based and Lexicon-Based Techniques for Sentiment Detection Without Using Labeled Examples // Proceedings of the 31st Annual International ACM SIGIR Conference of Research and Development in Information Retrieval. – Singapore, 2008. – P. 743-744.

40. Mudinas A., Zhang D., & Levene M. Combining Lexicon and Learning Based Approaches for Concept-Level Sentiment Analysis // Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining. – New-York, 2012. – P. 1-8.

41. Hutto C. & Gilbert E. VADER: a parsimonious rule based model for sentiment analysis of social media text // In Proceedings of the Eight International Conference on Weblogs and Social Media. – Califaornia, 2014. – P. 216-225.

42. Nguyen T.H., Shirai K. Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. – 2015. – P. 1354–1364.

43. Wijffels J. Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the “UDPipe” // “NLP” Toolkit. [Электронный ресурс] URL: https://cran.r-project.org/web/packages/udpipe/index.html (дата обращения: 15.09.2019)

44. Elmurngi E., Gherbi A. Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques // Proceedings of DATA ANALYTICS 2017: The Sixth International Conference on Data Analytics. – 2017. – P. 65-72.

45. Patel D., Kapoor A., Sonawane S. Fake Review Detection using Opinion Mining // International Research Journal of Engineering and Technology (IRJET). – 2018. – Vol. 05(12). – P. 818-822

46. Samonte M.J. Polarity Analysis of Editorial Articles towards Fake News Detection // Proceeding ICIEB '18. –2018. – P. 10-112.

47. Brummette J., DiStaso M., Vafeiadis M., & Messner M. Read All About It: The Politicization of “Fake News” on Twitter // Journalism & Mass Communication Quarterly. – 2018. – Vol. 95(2). – P. 497–517. doi:10.1177/1077699018769906

48. Egelhofer J. L., & Lecheler S. Fake news as a two-dimensional phenomenon: a framework and research agenda // Annals of the International Communication Association. – 2019. – P. 1–20.doi:10.1080/23808985.2019.1602782

49. Nielsen R.K., Graves L. “News you don’t believe”: Audience perspectives on fake news / Reuters Institute for the Study of Journalism. [Электронный ресурс] URL: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2017-10/Nielsen&Graves_factsheet_1710v3_FINAL_download.pdf (дата обращения: 25.08.2019).

50. Тенева, Е. В. EMOTIONALISATION STRATEGY AS A MEANS OF MANIPULATION IN THE BRITISH MASS MEDIA DISCOURSE // Russian Linguistic Bulletin. – 2018. – Vol. 3(15). – P. 19-22. https://doi.org/10.18454/RULB.2018.15.3.4

51. Kramer A. D. I., Guillory J. E., & Hancock J. T. Experimental evidence of massive-scale emotional contagion through social networks // Proceedings of the National Academy of Sciences. – 2014. – Vol. 111(24). – P. 8788–8790.doi:10.1073/pnas.1320040111

52. Igartua J.J., Moral F, & Fernández I. Cognitive, Attitudinal, and Emotional Effects of News Frame and Group Cues, on Processing News About Immigration // Journal of Media Psychology Theories Methods and Applications. –2011. – Vol. 23. – P. 174-185. 10.1027/1864-1105/a000050

53. Otieno C., Spada H., & Renkl A. Effects of News Frames on Perceived Risk, Emotions, and Learning // PLoS ONE. – 2013. – Vol. 8(11). e79696.doi:10.1371/journal.pone.0079696

54. Papacharissi Z. Toward new journalism(s): Affective news, hybridity, and liminal spaces // Journalism Studies. – 2014. – Vol. 16(1). – P. 27–40. doi:10.1080/1461670X.2014.890328

55. Dafonte-Gómez, A. Audience as Medium: Motivations and Emotions in News Sharing // International Journal of Communication. – 2018. – Vol. 12. – P. 2133–2152.

56. Recasens M., Danescu-Niculescu-Mizil C., Jurafsky D. Linguistic Models for Analyzing and Detecting Biased Language. [Электронный ресурс] URL: https://nlp.stanford.edu/pubs/neutrality.pdf (дата обращения: 21.08.2019).

57. Bing L., Minqing H., and Junsheng C. Opinion Observer: analyzing and comparing opinions on the Web // Proceedings of WWW 2005. – 2005. –P. 342–351.

58. Turney P.D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // Proceedings of ACL. – 2002. – P. 417–424.

59. Balahur A., Steinberger R. Rethinking Sentiment Analysis in the News: from Theory to Practice and back // Proceedings of the '1st Workshop on Opinion Mining and Sentiment Analysis' 2009. – 2009. –. P. 1-12.

60. Zhang J., Kawai Y., Nakajima S., Matsumoto Y., & Tanaka K. Sentiment Bias Detection in Support of News Credibility Judgment // Proceedings of the 44th Hawaii International Conference on System Sciences. – 2011. doi:10.1109/hicss.2011.369

61. Analyzing Sentiment (2019). [Электронный ресурс] URL: https://cloud.google.com/natural-language/docs/analyzing-sentiment. (дата обращения: 03.09.2019).

62. Sentiment Analysis in Russian (2018). [Электронный ресурс] URL: https://www.kaggle.com/c/sentiment-analysis-in-russian/data. (дата обращения: 03.09.2019).

63. Колмогорова А. В., Калинин А. А., Талдыкина Ю. А. Языковые маркеры манипуляции в поляризованном политическом дискурсе: опыт параметризации // Политическая лингвистика. – 2016. – 4(58). – С. 194-199.

64. Парасуцкая М.И. Манипуляция и «Манипулятивный дискурс» в лингвистике: принципы исследования // Юрислингвистика. – 2011. – №11. – С. 493-501

65. Dijk. T. A. van. Discourse and manipulation // Discourse and Society. – 2006. – № 17. (2). P. 359—383.

66. Maillat D. Constraining context selection: On the pragmatic inevitability of manipulation // Journal of Pragmatics. – 2013. – № 59. – P. 190—199.

67. Доценко Е. Л. Психология манипуляции: феномены, механизмы и защита. – М.: Изд-во МГУ, 1997.

68. Герасимов В. М., Деркач А. А., Косопкин А. С., Нефедова Т. И. Психология парламентаризма. – М. : Статус-Кво, 1999.

69. Быкова О. Н. Языковое манипулирование: материалы к энциклопедическому словарю «Культура русской речи» // Теоретические и прикладные аспекты речевого общения : вестн. Рос. риторич. ассоциации. – Красноярск, 1999. – Вып. 1 (8). – С. 91—103.

70. Беляева И.В. Феномен речевой манипуляции: лингвоюридические аспекты // Автореферат дис. …доктора филол. наук. – 2009.

71. Гаврилова А.С. Словарь синонимов и антонимов современного русского языка. 50000 слов. – Москва, 2014

72. Ko H., Hong J.Y., Kim S., Mesicek L., Na I.S. Human-machine interaction: A case study on fake news detection using a backtracking based on a cognitive system // Cognitive Systems Research. – 2019. – Vol. 55. – P. 77-81. https://doi.org/10.1016/j.cogsys.2018.12.018.

73. Jang S.M., Geng T., Li J.-Y. Q., Xia R., Huang C.-T., Kim H., and Tang J. A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis // Computers in Human Behavior. – 2018. doi: 10.1016 /j.chb.2018.02.032

74. Баранов А.Н. Корпусная лингвистика // Введение в прикладную лингвистику. – М., 2003. – С. 118.

75. Biber D. “Representativeness in Corpus Design.” // Literary and Linguistic Computing – 1993. – Vol. 8.4. – P. 241–257.

76. Gries S.Th. Exploring variability within and between corpora: some methodological considerations // Corpora. – 2006. – Vol. 1(2). – P. 109-151.

77. Leech G. New resources, or just better old ones? The Holy Grail of representativeness // Corpus linguistics and the web. – 2007. – P. 133-149.

78. Váradi T. The linguistic relevance of corpus linguistics // Proceedings of the Corpus Linguistics 2001 Conference. UCREL Technical Papers. – 2001. – Vol. 13. – P. 587-593.

79. Gries S.Th. Dispersions and adjusted frequencies in corpora // International Journal of Corpus Linguistics. – 2008. – Vol. 13(4). – P. 403-437.

80. Sinclair J. Corpus, concordance, collocation. – Oxford: Oxford University Press, 1991.

81. Hanks P. The corpus revolution in lexicography // International Journal of Lexicography. – 2012. – Vol. 25(4). – P. 398-436.

82. McEnery T., Xiao R., & Tono Y. Corpus-based language studies: An advanced resource book. – New York: Routledge, 2006.

83. Gray B., Egbert J., Biber D. Exploring Methods for Evaluating Corpus Representativeness. [Электронный ресурс] URL: https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper277.pdf (дата обращения: 27.08.2019).

84. Corpus Linguistics. Method, Theory and Practice [Электронный ресурс] URL: http://corpora.lancs.ac.uk/clmtp/1-data.php (дата обращения: 29.08.2019).

85. Shang J., Liu J., Jiang M., Ren X., Voss C.R. and Han J. Automated Phrase Mining from Massive Text Corpora // IEEE Transactions on Knowledge and Data Engineering. – 2018. – Vol. 30, №. 10. – P. 1825-1837. doi: 10.1109/TKDE.2018.2812203.

86. Analyzing Sentiment (2019). [Электронный ресурс] URL: https://cloud.google.com/natural-language/docs/analyzing-sentiment. (дата обращения: 03.09.2019).

87. Sentiment Analysis in Russian (2018). [Электронный ресурс] URL: https://www.kaggle.com/c/sentiment-analysis-in-russian/data (дата обращения: 05.09.2019).

ПРИЛОЖЕНИЯ

1. Сформированные словари по 27 информативным признакам.
2. Экспериментальные размеченные корпуса (1) с использованием Google Cloud Natural Language API, (2) корпус, размещенный на ресурсе Kaggle (https://www.kaggle.com/) и (3) корпус, полученный в рамках проекта Linis-crowd.
3. Алгоритм машинного обучения распознавания тональности новостных текстов на языке R и Python.
4. Алгоритм определения полярности тональности и степени объективности лексико-словарным подходом.
5. Алгоритм тематического моделирования LDA.