

НАО КазНИТУ им. К.И. Сатпаева

УДК 004.94

На правах рукописи

**ЯКУНИН КИРИЛЛ ОЛЕГОВИЧ**

**Разработка моделей и методов сбора, анализа и классификации медиа-публикаций на базе методов обработки естественных языков**

6D070400 – Вычислительная техника и программное обеспечение

Диссертация на соискание степени  
доктора философии (PhD)

Научные консультанты  
доктор инженерных наук,  
профессор  
Р.И. Мухамедиев

доктор технических наук,  
доцент  
В.Б. Барахнин

Республика Казахстан  
Алматы, 2021

## СОДЕРЖАНИЕ

<b>НОРМАТИВНЫЕ ССЫЛКИ</b> .....	3
<b>ОПРЕДЕЛЕНИЯ</b> .....	4
<b>ВВЕДЕНИЕ</b> .....	5
<b>1 ЗАДАЧА МЕДИА-МОНИТОРИНГА И ОЦЕНКИ МЕДИА-ПРОСТРАНСТВА</b> .....	8
1.1 Задача медиа-мониторинга.....	8
1.2 Оценка влияния открытых информационных источников на социум	13
Выводы по 1-му разделу.....	22
<b>2 МЕТОДЫ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ В КОНТЕКСТЕ BIG DATA</b> .....	24
2.1 Проблема классификации текстовых документов и классические подходы к решению.....	24
2.2 Проблема векторизации документов, Topic Embeddings.....	26
2.3 Мера межкорпусного тематического дисбаланса.....	29
Выводы по 2-му разделу.....	32
<b>3 МЕТОД МУЛЬТИКРИТЕРИАЛЬНОЙ ОЦЕНКИ МЕДИА-ИСТОЧНИКОВ НА БАЗЕ БАЙЕСОВСКОЙ МОДЕЛИ АГРЕГАЦИИ ММА</b> .....	34
3.1 Байесовская модель агрегации гетерогенных данных.....	34
3.2 Метод мультикритериальной оценки медиа-источников ММА.....	37
Выводы по 3-му разделу.....	44
<b>4 РАЗРАБОТКА ИНФОРМАЦИОННОЙ СИСТЕМЫ МОНИТОРИНГА МЕДИА-ПРОСТРАНСТВА КАЗАХСТАНА НА БАЗЕ МЕТОДОВ NLP</b> .....	46
4.1 Описание программной архитектуры разработанной системы.....	46
4.2 Основной функционал разработанной системы.....	54
Выводы по 4-му разделу.....	58
<b>5 ДАННЫЕ, ЭКСПЕРИМЕНТЫ И ВАЛИДАЦИЯ РЕЗУЛЬТАТОВ РАЗРАБОТАННЫХ МЕТОДИК И СИСТЕМЫ</b>	61
5.1 Корпус новостных публикаций.....	61
5.2 Методика оценки социальной значимости.....	65
5.3 Результаты валидации модели и системы на базе размеченных экспертами данных.....	68
5.3.1 Прочие кейсы использования предложенных моделей и методов.....	75
Выводы по 5 разделу.....	77
<b>ЗАКЛЮЧЕНИЕ</b> .....	80
<b>СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ</b> .....	83
<b>ПРИЛОЖЕНИЕ А – Акт внедрения</b> .....	89
<b>ПРИЛОЖЕНИЕ Б – Листинг функций, реализующих метод ММА calc_mma.py</b> .....	94
<b>ПРИЛОЖЕНИЕ В – Пример отчета по мониторингу СМИ по теме "Образование"</b> .....	100

## **НОРМАТИВНЫЕ ССЫЛКИ**

В настоящей диссертации использованы ссылки на следующие стандарты:

ГОСТ 19.101-77. Описание программы.

СТ РК 34.014-2002. Информационные технологии.

ГОСТ 19.201-78. Единая система программной документации.

ГОСТ 28195-99. Оценка качества программных средств.

## ОПРЕДЕЛЕНИЯ

**Байесовская система вывода** – статистический метод, в котором свидетельство (наблюдение) используются, чтобы обновить или вывести вновь вероятность того, что гипотеза может быть верной.

**Модель многокритериального принятия решений (мультипликативная многофакторная модель)** – математическая модель принятия оптимального решения одновременно по нескольким критериям. Эти критерии могут отражать оценки различных качеств объекта (или процесса).

**Точность (англ. accuracy)** – относительное количество корректно классифицированных примеров.

**F1-Score** – мера оценки качества модели, учитывающая дисбалансность классов выборки, представляющая собой гармоническое среднее между точностью (precision) и охватом (recall/полнота)

**Экспертная система** – компьютерная система, способная частично заменить эксперта в решении проблемной ситуации.

**Back-end** – основная программно-аппаратная часть компьютерной системы.

**Front-end** – интерфейс взаимодействия между пользователем и программно-аппаратной частью.

**Full-stack разработчик** – универсальный разработчик, способный создавать как основную программно-аппаратную часть компьютерной системы, так и интерфейс взаимодействия.

**Javascript** – прототипно-ориентированный сценарный язык программирования, который обычно используется как встраиваемый язык для программного доступа к объектам приложений.

**Тематическая модель** – статистическая модель, позволяющая находить скрытые латентные структуры в корпусе текстов (топики/темы), представляющие собой две матрицы – матрицу распределения слов и фраз по топикам, и матрицу распределения документов корпуса по топикам.

**NLP** – Natural Language Processing, обработка естественных языков – область применения алгоритмов искусственного интеллекта, занимающаяся задачами анализа текстов на естественных языках, включая предобработку (нормализацию) текстов, векторизацию, классификацию текстов, задачи поиска, построения чат-ботов и т.п.

**Трансформеры** – класс современных моделей глубокого обучения (Deep learning), основанные на применении глубоких рекуррентных сетей (RNN, LSTM, GRU, bi-LSTM), в комбинации с так называемым attention-слоями (механизм внимания). Такие модели требуют больших вычислительных мощностей для эффективного обучения (сотни и тысячи независимых процессоров, например в составе GPU).

## ВВЕДЕНИЕ

**Актуальность темы.** Задачи медиа-мониторинга пользуются большой популярностью на мировом рынке, в первую очередь в виде продуктов по менеджменту репутации. Однако более высокоуровневые задачи, связанные с принятием решений на основе данных из медиа-пространства на государственном уровне, на данный момент практически не решаются в автоматическом режиме.

Несмотря на значительную проработанность темы классификации текстов, современные SoTA (state of the art – лучшие на текущий момент) модели требуют больших объемов разметки для классификации по заданному критерию, а также высокой производительности (сотни и тысячи процессоров для обучения). Следовательно, существует потребность в более эффективных с точки зрения требуемой экспертной разметки и вычислительной производительности моделях классификации и анализа текстовой информации.

**Цель работы.** Целью работы является разработка моделей и методов для автоматической многокритериальной оценки текстовой информации из медиа-источников и социальных сетей в рамках распределенной информационной системы.

### **Задачи исследования**

Для достижения цели в работе поставлены следующие задачи:

- выявить основные факторы, ограничивающие возможность построение эффективных (с точки зрения производительности и требуемой экспертной разметки) классификационных моделей для текстов;
- разработать метод эффективной векторизации текстов на базе неразмеченных данных (Big Data);
- разработать метод самообучения (self-learning) классификационной модели на основе объективных явных показателей, для распознавания скрытых параметров документов;
- разработать методику многокритериальной оценки документов и медиа-источников;
- разработать распределенную информационную систему для сбора, хранения, обработки и классификации текстовой информации из масс-медиа и социальных сетей и верифицировать точность результатов работы информационной системы.

**Объекты исследования.** Объектом исследования в диссертационной работе являются методы классификации текстовых документов.

**Предмет исследования.** Предметом исследования в диссертационной работе являются методы многокритериальной оценки текстовых документов на базе тематического моделирования.

**Методы исследования.** В диссертационной работе применяются следующие методы исследования: методы классификации, методы векторизации текстовых документов, методы многокритериального анализа для

принятия решения, технологии проектирования и разработки информационных систем.

#### **Научная новизна**

1 Предложен метод векторизации текстовых документов с помощью тематической модели BigARTM.

2 Предложен метод оценки тематического межкорпусного дисбаланса для самообучения классификационной модели.

3 Предложена методика многофакторной оценки социальной значимости публикации.

4 Предложена методика многокритериальной оценки масс-медиа ММА на базе байесовской системы агрегации, метода анализа иерархий (АНР) и тематического моделирования.

**Основные результаты исследования** заключаются в: разработке методики многокритериальной оценки текстовых документов и медиа-источников и сопутствующих методов для самообучения и обучения соответствующих моделей; разработке распределённой информационной системы для сбора, хранения, обработки и классификации данных масс-медиа и социальных сетей, являющейся программной реализацией предложенных методов; многофакторной методики оценки социальной значимости публикаций в СМИ и социальных сетях.

#### **Положения, выносимые на защиту**

1. Разработана методика мультифакторной оценки социальной значимости на базе многокритериальной методики оценки масс-медиа ММА с использованием тематической векторизации текстовых документов.

2. Разработана архитектура и программная реализация распределенной информационной системы для сбора, обработки, оценки и визуализации текстовых данных масс-медиа и социальных сетей Media Analytics, включающий предложенную методику оценки, а также ее оценку и верификацию.

**Связь темы с планами научно-исследовательских программ.** Представленные результаты получены при выполнении проекта ИИВТ КН МОН РК (источник финансирования Комитет науки МОН РК): программно-целевого финансирования (ПЦФ) КН МОН РК BR05236839 «Разработка информационных технологий и систем для стимулирования устойчивого развития личности как одна из основ развития цифрового Казахстана» в 2018-2020 годы;

#### **Апробация результатов исследования**

Основные положения и результаты исследования докладывались на: Procedia 9th International Young Scientist Conference (Scopus, 2020), International Conference on Digital Transformation and Global Society (Springer, 2019, 2020) и других конференциях. Основные результаты исследования опубликованы в журнале Symmetry, имеющем импакт-фактор 2.51 (Q1). Разработанная информационная система внедрена в МОН РК (Приложение А).

**Публикации.** По теме диссертации опубликовано 16 работ, из которых 6 статьей опубликованы журналах, входящих в базы Scopus и Thomson Reuters (2 статьи – Q1, 2 статьи – Q2, 2 статьи – Q3), одна из них опубликована в изданиях, рекомендованных Комитетом по контролю в сфере образования и науки МОН РК, 3 статьи в журналах имеющих CiteScore (2 статьи – CiteScore 17%, 1 статья – CiteScore 69%) в Scopus без присвоенного квартиля, 2 статьи в журналах, входящих в РИНЦ (импакт-фактор 0.482 и 1.385), 6 статей опубликованы в сборниках международных научно-практических конференций. Оформлено 4 авторских свидетельства на результаты работы.

**Структура и объем диссертации.** Диссертация состоит из введения, пяти разделов, заключения, списка использованных источников и приложений, содержит 88 страницы основного текста (16 рисунков, 9 таблиц). Список использованных источников содержит 92 наименований источников.

# 1 ЗАДАЧА МЕДИА-МОНИТОРИНГА И ОЦЕНКИ МЕДИА-ПРОСТРАНСТВА

## 1.1 Задача медиа-мониторинга

Средства массовой информации и информационные сообщества в социальных сетях не только отражают деятельность государственных органов, но также формируют информационный контекст, настроения и уровень значимости, приписываемые определенным государственным инициативам и общественным мероприятиям. Многосторонняя количественная (насколько это практически возможно) оценка деятельности СМИ важна для понимания их объективности, роли, направленности и, в конечном итоге, качества «четвертой власти» общества.

Согласно исследованию Edelman Trust Barometer 2019 года, проведенному в 27 странах, доверие к правительственной информации и каналам СМИ остается низким. Разрыв между информированной общественностью и большинством населения увеличивается (13 баллов в 2018 году, 16 баллов в 2019 г.) [1].

В случаях, когда аудитория не имеет существенных знаний или опыта в отношении событий и информационного контекста, она в основном зависит от информации, предоставляемой СМИ [2]. Согласно исследованиям [3, 4], СМИ используют различные методы и механизмы манипуляции, такие как формирование мнения или фокусирование внимания аудитории на определенных темах. Доступность в Интернете разнообразных новостей, часто содержащих противоположную информацию об одних и тех же событиях, является дополнительным фактором, влияющим на наше восприятие, которое может создать путаницу, вызванную личными субъективными мыслями, и выливаться различные фейковые публикации [5], теории заговора, освещаемые в личных аккаунтах социальных сетей и т.п.

В этой связи важно понимать, как средства массовой информации используют свое влияние, чтобы смягчить негативное влияние средств массовой информации и стимулировать положительное влияние на общественное развитие [6].

Исследователи сосредотачивают свои усилия на оценке медиаконтента в связи с его практической значимостью для информационных агентств, рекламных компаний и государственного сектора. Анализ медиаконтента позволяет нам прогнозировать вероятную популярность новостных статей [7], таргетировать рекомендациями отдельных групп пользователей, планировать и оценивать PR-стратегии для продвижения товаров или услуг [8, 9]. Госсектор получает инструмент для продвижения и информирования об инновациях, PR-планировании и выявлении запрещенного законом негативного контента. Отдельные пользователи могут быстро и эффективно фильтровать большие объемы информации на гораздо более высоком уровне семантических запросов, чем в случае с простыми системами полнотекстового поиска.



В [10] дано следующее определение мониторинга СМИ: «Мониторинг СМИ – это процесс постоянного наблюдения за потоком новостей СМИ с целью выявления, фиксации и анализа медиа-контента, содержащего ключевые слова определенной темы». Однако это определение связывает процесс мониторинга с методом (на основе ключевых слов). Таким образом, мы обобщаем определение мониторинга СМИ как процесса контент-анализа СМИ. Мониторинг СМИ можно использовать для сравнения больших корпусов текстов (Задача 1). Например, в [11] рассматривается задача сравнения турецкого и английского корпусов новостей, связанных с наукой. Кроме того, мониторинг СМИ включает в себя такие задачи, как анализ социального поведения, выявление общественного мнения [12, 13], анализ тенденций продаж в Интернете [14] и др. (Задача 2). Другая задача – сравнение предпочтений и характеристик социальных слоев населения (задача 3). Например, в [15] анализируется гендерное неравенство. Инструменты мониторинга СМИ – важная часть управления репутацией (задача 4). Наиболее часто выполняемые функции: поиск упоминаний бренда (с прямой маркировкой/тегом или без нее), релевантные хэштеги (брендовые и небрендовые), упоминания конкурентов, общие тенденции, применимые к той или иной отрасли. Главный вопрос, на который можно ответить с помощью таких инструментов: «Вы знаете, что говорят о вашей компании или бренде в СМИ?». Существует значительное количество продуктов, которые отслеживают СМИ и социальные сети, некоторые из которых перечислены в таблице 1 [16-18].

Таблица 1 – Примеры инструментов медиа-мониторинга

Название	Источники данных (пометки)	Ссылка
1	2	3
Agorapulse	Facebook, Twitter, Instagram и YouTube	agorapulse.com
Awario	Интернет (полный поиск), Twitter, Facebook, Instagram, Google+, YouTube, Reddit, новостные сайты и блог-платформы	awario.com
Brandwatch	Интернет (полный поиск), Facebook, Twitter, Instagram, YouTube, Google+, Pinterest, Sina Weibo, VK, QQ, новостные сайты и блог-платформы	www.mediatoolkit.com
BuzzSumo	Facebook, Twitter, Pinterest и Reddit	buzzsumo.com
Crimson Hexagon, Brandwatch.	Интернет (полный поиск), Facebook, Twitter, Instagram, YouTube, Google+, Pinterest, Sina Weibo, VK, QQ, новостные сайты и блог-платформы	brandwatch.com

Продолжение таблицы 1

1	2	3
Google Alerts	Интернет (полный поиск)	google.com/alerts
Hootsuite	Twitter	hootsuite.com
Keyhole	Twitter, Instagram, YouTube, Facebook, новостные сайты и блог-платформы	keyhole.co
Sprout Social	Интернет (полный поиск), Twitter, Instagram, Reddit, YouTube, Tumble	sproutsocial.com
Social Mention	Интернет (полный поиск), Twitter, Facebook, YouTube, Reddit, Google+	socialmention.com
SentiOne	Интернет (полный поиск), Facebook, Instagram, Twitter, LinkedIn, VK, новостные сайты и блог-платформы	sentione.com
Signal AI	Real-time unlimited information and insights for media monitoring, reputation management and market intelligence	signal-ai.com
Talkwalker	Интернет (полный поиск), Flickr, Foursquare, SoundCloud, Twitch, Pinterest, and others.	www.talkwalker.com
TweetDeck	Twitter	tweetdeck.twitter.com
Примечание – Составлено по источнику [19]		

В системах, перечисленных в таблице 1, в основном решаются задачи, связанные с управлением репутацией, наряду с ручными методами анализа, такими как запросы на основе ключевых слов с применением индикаторов TF-IDF [13, p. 184; 20].

Хотя применение запросов, по ключевым словам, обеспечивает определенный уровень интерпретируемости, оно все же накладывает некоторые ограничения на эти онлайн-системы и сервисы:

1. Как правило, они ограничены по семантике запроса, результаты запроса требуют дальнейшего ручного выбора [9].

2. Результаты выполнения запросов зависят от алгоритма поиска и текущего состояния системы/базы данных. Такие сервисы обычно не имеют возможности сохранять результаты запросов [20, p. 133].

3. Эти инструменты не решают вопросов оценки самих медиаисточников.

4. Эти инструменты ограничены в критериях оценки; обычно оцениваются только анализ настроек и некоторый индекс освещения в СМИ.

Для задач, более ориентированных на исследования (сравнение корпусов, относящихся к разным странам/культурным пространствам, анализ предпочтений слоев населения, анализ общественного мнения, актуальных

тенденций и сезонности и т.д.), автоматическая кластеризация текстов часто выполняется с применением латентного распределения Дирихле (LDA) [21, 22], а классификация документов выполняется с использованием моделей машинного обучения [14, р. 2; 18; 23]. Следует отметить, что подход к обучению с учителем возможен, если доступен размеченный набор данных значительного объема. Современные state-of-the-art модели классификации текста, такие как BERT и GPT, требуют не менее десятков тысяч обучающих примеров для точной настройки, при этом они чрезвычайно сложны (миллиарды параметров/весов) и требуют графических процессоров или иных суперкомпьютерных вычислительных комплексов для достаточной вычислительной производительности. В то же время для исследовательских целей необходим инструмент, который позволил бы решать более широкий круг исследовательских задач, включая сравнительный анализ СМИ. Основная цель такого инструмента – предоставить экспертам, исследователям, менеджерам и руководителям полный и мощный набор аналитических инструментов для получения актуальных релевантных отчетов, визуализаций и оценок публикаций в публичных средствах массовой информации в интересующей их сфере деятельности. Для решения этих задач в ситуации, когда получение больших объемов размеченных текстов затруднительно или слишком затратно, подход MCDM может применяться в сочетании с тематическим моделированием корпусов новостей.

Медиа-мониторинг представляет собой крупный рынок услуг различного характера, которыми активно пользуются все сегменты бизнеса от персональных бизнесов (личный бренд, блогеры, инфлюенсеры и т.п.) до МСБ, крупных корпораций и даже государственного сектора. Если попытаться классифицировать основные цели, с которыми бизнесы обращаются за услугами к провайдерам медиа-мониторинга, то можно выделить:

1. Менеджмент репутации – мониторинг и ретроспективная аналитика репутации тех или иных организаций, ведомств и персоналий в СМИ и социальных сетях. Оперативный поиск потенциально опасных публикаций (фейки, клевета, компромат).

2. Маркетинговые исследования. Сюда можно отнести анализ освещенности тех или иных маркетинговых и PR компаний в социальных сетях, тональности релевантных публикаций, сравнительная аналитика с показателями продаж. Также можно выделить проведение исследования прямых конкурентов – их имиджа в медиа-поле, освещенность и отношение клиентов к акциям и кампаниям конкурентов и т.п.

3. Оценка деятельности PR-отделов – например в виде KPI показателей, рассчитываемых из освещенности тех или иных направлений деятельности компаний, мероприятий и пр.

4. Оценка контрагентов – например при закупках, когда требуется всесторонняя оценка рейтинга доверия контрагенту, что не должно ограничиваться классическими инструментами скоринга контрагентов

(аналитика предыдущих закупок, наличие арестов счетов, наличие в черных списках и пр.).

5. Поддержка принятия решений – особенно актуально для крупных корпораций и государственных органов управления. Часто для принятия решений требуется информация по общественному мнению, касающемуся ряда высокоуровневых вопросов, а также автоматический поиск современных трендов и инфоповодов в определенной сфере.

При этом нужно отметить, что разработка систем медиа-мониторинга является очень высокотехнологичной сферой разработки ПО. Например, Cambridge Analytica потратили порядка 7 миллионов долларов США только на сбор данных, необходимых для создания политических портретов пользователей Facebook, не считая саму аналитику. Объем начального фондирования компании Signal AI, занимающейся вопросами бизнес-аналитики и мониторинга СМИ на основе ИИ, которая собирает, анализирует и предоставляет лидерам бизнеса информацию о цифровых, печатных и вещательных СМИ, новостях и нормативных данных, составил 49.7 миллионов долларов США. В 2019 63% акций одного из крупнейших российских провайдеров услуг медиа-мониторинга «Медиалогия» были куплены Банком ВТБ за, по разным оценкам, 500-700 миллионов рублей (8-11 миллионов долларов США).

Исследования, проводимые в области СМИ, охватывают только определенные области СМИ. В частности, проводится государственный мониторинг за соблюдением требований законодательства в области СМИ. Систематический мониторинг и оценка влияния СМИ в Казахстане осуществляется РГП «Центр анализ информации» с соблюдением требований казахстанского законодательства. Основные мероприятия по мониторингу и оценке эффективности СМИ, как правило, проводятся в контексте одной конкретной задачи, например, оценки эффективности информационной кампании, деятельности организации и других мероприятий, целью которых является продвижение инноваций и популяризация государственной политики.

Консалтинговые услуги по проведению медиа-мониторинга и контент анализа в основном нацелены на выявление общественного мнения относительно деятельности организаций, персон, брендов, мероприятия и маркетинговых кампаний, и предназначены в большей степени для продвижения брендов, товаров и услуг и т.п.

Согласно базе данных АО «Национальный центр государственной научно-технической экспертизы», в период с 1995 по 2016 год в Казахстане зарегистрировано 27 научно-исследовательских работ, косвенно относящихся к теме диссертационного исследования. В то же время проводимые исследования немногочисленны, охватывают лишь отдельные области изучения роли СМИ и лишь косвенно затрагивают проблемы комплексной оценки воздействия СМИ и влияния открытых тестовых источников информации на общество.

Автоматизированные средства медиа-мониторинга достаточно широко применяются на этапе сбора данных, однако автоматические анализ медиа-

данных применяется редко, либо только в ограниченном наборе кейсов (определение эмоциональной тональности, поиск именованных сущностей), имеющих спрос на рынке, и представляющие низкую техническую сложность. В области оценки медиавоздействия анализ данных в основном проводится вручную экспертами. Однако сейчас, ввиду экспоненциального роста количества данных, вопрос автоматизации решения таких задач приобретает все большую актуальность.

## **1.2 Оценка влияния открытых информационных источников на социум**

В работе предлагается рассматривать анализ СМИ и социальных сетей в контексте концептуальной схемы взаимного влияния социума, органов государственного управления и СМИ. Рисунок 1 иллюстрирует эту схему, включающую непосредственный объект исследования – информационное поле (СМИ и социальные сети), основного потенциального заказчика и пользователя алгоритмов – органы государственного управления и основной объект управления и получатель информационного воздействия – социум. При этом нужно отметить, что элементы данной схемы не являются изолированными, и в реальности как СМИ, так и во многом органы государственного управления, включены в состав того же общества, влияние на которое исследуется. Также нельзя рассматривать упрощенную модель, в которой гос. управление через информационные средства проводит свою политику влияния на общество (пропаганда). Модель, в которой общество придает огласке свои проблемы и чаяния, дабы повлиять на действия государственных регуляторов и исполнительных органов (концепция СМИ как четвертой власти [24]), также является отчасти верной, но не полной. Поэтому методологическая основа данной работы предполагает рассмотрение этих трех элементов концептуальной схемы в качестве относительно независимых акторов (субъектов), постоянно оказывающих влияния друг на друга. Несмотря на то, что в работе рассматривается основной use-case использование системы в качестве поддержки принятия решений для органов государственного управления, позволяющий получать информацию о состоянии социума посредством информации в СМИ и социальных сетях, данная постановка задачи возможна только исходя из концептуальной схемы, в которой все три элемента находятся в активной взаимосвязи, без жестко заданной иерархии или однонаправленных векторов управления.

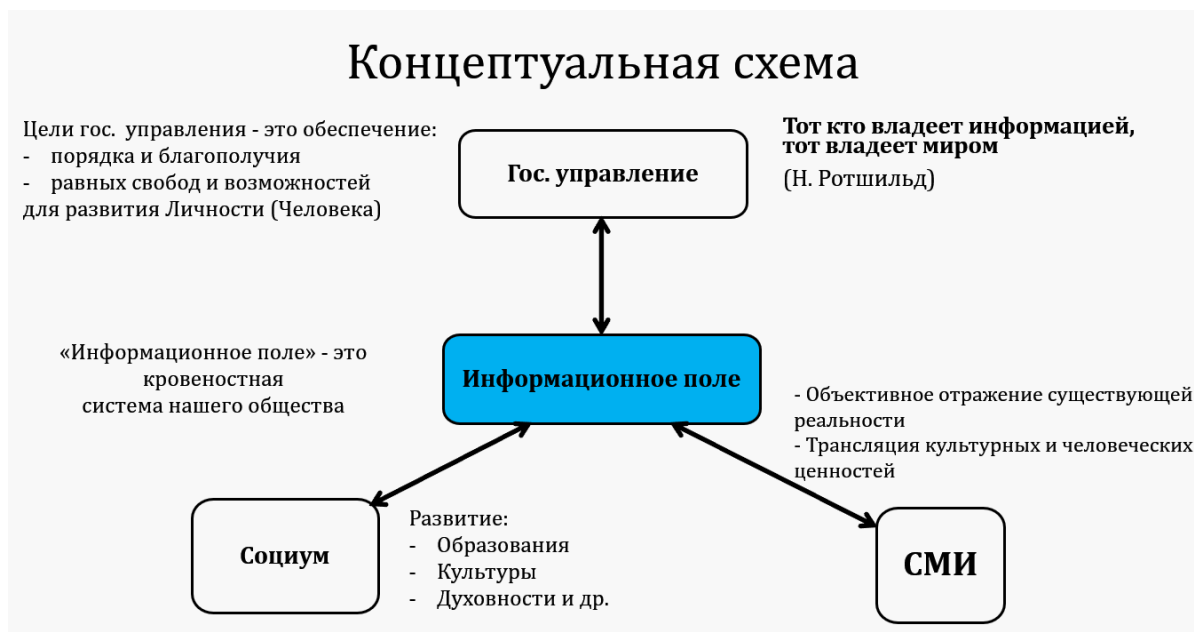


Рисунок 1 – Концептуальная схема взаимного влияния социума, органов гос. управления и СМИ

Следовательно, результаты работы могут быть потенциально могут быть использованы всеми тремя акторами концептуальной схемы. Рассмотрим основные сценарии использования системы:

1. Использование гос. органами. Может включать оперативную идентификацию опасных инфоповодов и информационных атак, высокоуровневую аналитику по трендам и тематическим тенденциям, построение рейтингов ведомств, организаций и персоналий, отслеживание освещенности государственных программ и т.п.

2. Использование обществом (физическими лицами). Результаты работы могут использоваться для интеллектуального информационно-разведывательного поиска [25] интересующей людей информации. Основным отличием от классических систем полнотекстового поиска (таких как Google, Yandex, и другие) является возможность задания более сложных семантических требований к искомым публикациям. Так, например, могут быть заданы запросы вида «получить тематические группы публикаций, в которых наблюдался экспоненциальный рост в последние 7 дней, обладающими высокими показателями социальной значимости и отношения пропаганды».

3. Использование СМИ. Как обсуждалось в разделе 1.1, современные системы медиа-мониторинга обычно не предоставляют инструментов для оценки медиа-источников в целом, а если такой функционал имеется, то он обычно ограничен анализом базовых статистических показателей (количество публикаций, просмотров, комментариев и пр.) и тональности. Следовательно, результаты работы могут быть использованы СМИ как для более глубокого анализа своего собственного тематического и параметрического портрета, в том числе на уровне отдельных авторов/редакторов/разделов, так и для подобного анализа своих конкурентов. Это может быть полезно как маркетинговый

инструмент для уточнения своего позиционирования на рынке, выявления сильных и слабых сторон, оптимизации контента и т.п.

В современном мире СМИ играют значительную роль в устойчивом развитии общества. Понимание психосоциальных механизмов, посредством которых СМИ влияет на человеческие мысли, мнения и действия имеет большое значение. Социально-когнитивная теория Бандуры [26] обеспечивает концептуальную основу для изучения и описания таких эффектов. Поведение человека часто объяснялось с помощью однонаправленной причинно-следственной связи. В рамках данных концепций поведение формируется и контролируется либо под влиянием окружающей среды, либо внутренним миром. Социально-когнитивная теория объясняет психосоциальное функционирование путем взаимного влияния всех факторов, как когнитивных, так и эмоциональных, а также поведенческих [27].

Когнитивно-социальная теория основана на теории социального научения (social learning theory). Согласно данной теории человек, усваивает модели поведения, наблюдая, как окружающие ведут себя, а затем имитируя их действия. Таким образом примеры, демонстрируемые в СМИ, становятся источником научения. Сначала человек наблюдает модель поведения в СМИ. Далее он запоминает модель поведения и начинает о ней думать («когнитивное проигрывание»). Затем человек оценивает увиденное с помощью когнитивных способностей, а затем совершает определенные действия. При этом должна присутствовать мотивация, подталкивающая человека к совершению действий

Другой концептуальной основой для анализа и оценки влияния СМИ на личность является теория культивирования Д. Гербера, разработанная в 1960-х годах. Данный подход, исследует то, как экстенсивное, многократное воздействие СМИ (в первую очередь телевидения) на протяжении продолжительного времени постепенно меняет представление человека о мире и социальной реальности.

Поскольку важность СМИ в формировании мировоззрения человека и их роль в формировании общественного сознания не вызывает сомнений, оценка воздействия СМИ на общество является одним из самых популярных направлений прикладных исследований. На сегодняшний день акцент в таких исследованиях сместился с оценки влияния телевидения на оценку влияния социальных сетей в связи с растущей популярностью последних.

Исследование, проведенное в 2016 году маркетинговой компанией MediaKix о том, сколько времени пользователи проводят в социальных сетях, показало, что электронные СМИ и социальные сети занимают второе место после телевидения. В то же время в 2015 году пользователи в США тратили больше времени на просмотр социальных приложений, чем на телевизор. Для большинства населения Казахстана (59,8%) основными источниками информации новостного характера являются электронные источники: социальные сети/блогеры (30,3%), а также новостные сайты (29,5%) [28]. Несмотря на большой интерес научного сообщества к оценке СМИ, в данной области представлены разные подходы. Нет единых общепринятых концепций

и методов. К тому же теоретические разработки влияния на общество многочисленны, не структурированы и отличаются друг от друга.

Методология воздействия СМИ на общество широко включает методы социологического исследования, анализа и ряд количественных показателей. Наиболее распространенные методы включают опросы общественного мнения, экспертные оценки, анализ дискурса, контент-анализ, кейсовое исследование, графематический и синтаксический анализ.

СМИ используют различные виды психологического воздействия, чтобы транслировать определенную позицию в отношении конкретной ситуации. По данным социологического опроса населения Казахстана [28, с. 1-10], большинство респондентов считают, что казахстанские информационные Интернет-сайты используются как инструмент для дискредитации определенных лиц (43,4%), формирования положительного имиджа (55,1%). При этом 52,7% населения считает, что казахстанские Интернет-источники СМИ используются для освещения событий и деятельности в определенной сфере в рамках интерпретации, выгодной для определенных групп лиц или организаций. Эти данные позволяют сделать вывод о том, что, по мнению казахстанцев, СМИ являются инструментом формирования и манипулирования общественным сознанием.

В рамках работы предложена методика для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации, включающая:

- групповые (сводные) и индивидуальные индексы по различным свойствам медиа-публикаций;
- система показателей влияния открытых текстовых информационных источников на социум;
- возможность получения многоуровневой системы оценок (по новостным порталам в целом, по отдельному новостному portalу, по отдельной публикации).

Методика предполагает определение влияния открытых информационных источников на социум на основе измеряемых показателей на базе информационной системы.

Основные результаты применения методики:

1. Определение уровня влияния медиа-текстов казахстанских СМИ по различным критериям.
2. Генерация аналитики, статистических данных на основе оценки.
3. Мониторинг медиа-потребления.

Предложенный метод оценки влияния открытых текстовых информационных источников представляет собой совокупность оценочных процедур.





Рисунок 2 – Пример декомпозиции показателей оценки влияния открытых текстовых информационных источников на социум

Как следует из рисунка 2, на самом верхнем (первом) уровне осуществляется общая (интегральная) оценка влияния. Декомпозиция интегрального индикатора влияния осуществляется по набору содержательных критериев, выделяемой в зависимости от задач (например тональность, потенциальная резонансность, социальная значимость и т.п.).

Процесс выбора информативных признаков (критериев) происходил многокритериально – учитывался мировой опыт других платформ медиа-мониторинга, доступность данных, возможность достаточно точной оценки критериев (например, такие критерии как манипулятивность, наличие сарказма, призыв к действию плохо поддаются классификации даже методами глубокого обучения) и полезности для решения определенных конкретных целей и задач.

В плане вовлеченности пользователей были определены информативные признаки на основе анализа критериев оценки, применяемых на практике с сфере оценки влияния СМИ. Изучены широко применяемые показатели оценки влияния СМИ такие как медиа-охват, индекс вовлеченности, индекс отказа, темп роста аудитории сообщества, соотношение лайков и дизлайков, тональность упоминаний публикации и комментариев.

Всесторонний анализ теоретических выкладок и практических разработок в области СМИ для общества показал, что сегодня средства массовой информации имеют широкий набор методов и приемов для аудитории и активно их используют. Эффективное владение такими техниками определяет роль СМИ как одного из важнейших факторов в формировании мировоззрения человека.

Несмотря на многочисленность показателей, позволяют оценить аспекты влияния СМИ, общую структуру анализа, как и сущность, характер и методы расчета количественных показателей, и определение информативных признаков зависит от целей исследования.

В методике предлагается независимо оценивать информативны критерии оценки статьи (тональность, социальная значимость и т.п.), и вклад статьи в ситуацию в медиа-поле, и следовательно степень влияния публикации на общество. Вклад статьи может определяться исходя из следующих соображений:

1. Заинтересованность читателей.
2. Реакция читателей.
3. Наличие вирусного эффекта распространения публикации.

Для оценки влияния конкретной публикации на социум с помощью автоматизированной информационной системы предлагается использовать следующие показатели:

– медиа-охват. Данный показатель позволяет оценить, сколько человек просмотрело публикацию и соответственно, в курсе опубликованной информации. Характеризует масштабность ее распространения. Обычно сюда включают всевозможные варианты просмотра – например встроенные в мессенджеры ссылки, RSS ленты, рассылки, просмотры по шейрам;

– количество шейров. Данный показатель позволяет оценить, сколько человек поделилось данной публикацией в социальных сетях. Высокий показатель показывает, что новость имеет резонанс и вызывает пользовательскую активность, желание поделиться данной информацией, а следовательно, является одним из индикатором вирусного распространения новости или иной публикации. Показатель может рассчитываться как сумма количества шейров публикации к общему количеству просмотров;

– количество комментариев. Данный показатель измеряет интерес новости у аудитории и может считаться опосредованным индикатором того насколько сильные эмоции или желание дискуссии вызывает публикация. Показатель может рассчитываться как сумма количества комментариев к одной публикации к общему количеству просмотров данной публикации;

– показатель тональности комментариев обычно рассчитывается с применением методов частотного лингвистического анализа, которые считают частоту вхождения слов-маркеров негативного, позитивного или нейтрального отношения отдельных читателей, либо с помощью моделей глубокого обучения, таких как рекуррентные нейронные сети и трансформеры. Часто имеет смысл не учитывать саму направленность тональности (позитив или негатив), а только ее величину (*magnitude*). Это связано с тем, что при оценке степени влияния новости, конкретные мнения могут быть не так важны, как общий эмоциональный фон в комментариях – либо относительно спокойный, нейтральный, либо в повышенный, причем распределения негатива и позитива не может говорить об общей степени влияния статьи. Также стоит заметить, что в комментариях в социальных сетях и медиаресурсах практически всегда превалирует негатив, а позитив часто связан с работой заказных медиаагентств, что в какой-то мере нивелирует информативность направленности тональности при анализе комментариев.

Определение тональности является одной из наиболее востребованных и наиболее изученных областей NLP. Существует огромное количество практических применений, и качественных датасетов, включая комментарии в социальных сетях, отзывы, обзоры, сообщения в мессенджерах и форумах и т.п. Можно выделить два основных подхода к решению задачи определения тональности – словарный, предполагающий составление списка

положительных и негативных слов и фраз, и подход машинного обучения с учителем, предполагающий обучение моделей разной степени сложности на размеченных наборах данных. Существенным препятствием в адекватном анализе тональности текста является наличие в тексте иронии и сарказма, поскольку при использовании в тексте элементов иронии и сарказма реальная тональность текста может быть прямо противоположна прямому смыслу сообщения. Так, по данным исследования Высшей школы экономики, некоторыми признаками наличия иронии/сарказма в тексте могут выступать усиленные восклицания – избыточное использование восклицательных или вопросительных знаков, наличие междометий, маркеры ирреалиса, а также их сочетание [29].

Можно выделить следующие критерии, позволяющие определить тональность:

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

Объективность информации, отсутствие предвзятости и нейтральное освещение разных точек зрения являются основными принципами журналистской этики [30]. Следовательно, классическое определение тональности не может в полной мере применяться к журналистским текстам, поскольку они по определению являются нейтральными. Следовательно, здесь и далее в данной работе под тональностью понимается альтернативное определение – не эмоциональность или оценочные суждения, а позитивное или негативное влияние описанных событий на общество и устойчивое развитие личности.

Другой важной задачей в оценке свойств медиа-публикаций является определение социальной значимости. Термин социальная значимость может иметь несколько определений, но в данной работе под социальной значимостью публикации понимается мера ее соответствия приоритетам государственного развития, интересам общества и отношение к проблеме развития личности. Таким образом, можно выделить несколько критериев для определения социальной значимости публикации:

- 1) соответствие приоритетам государственного развития;
- 2) отношение к развитию личности и социума;
- 3) резонансность, или, иными словами, общественный интерес;
- 4) соответствие запросам и волнующим проблемам населения.

При этом эти факторы следует рассматривать независимо, так, например социальная значимость тематики далеко не всегда напрямую связана с ее резонансностью, хотя резонансность должна учитываться при расчете интегральной оценки. Так, по данным «Яндекса» [31], в 2018 году пользователи из Казахстана наиболее часто делали запросы на следующие темы: 1. Чемпионат мира по футболу. 2. Зимняя Олимпиада в Пхёнчхане. 3. Бой

Нурмагомедова и Макгрегора. 4. Пожар в ТРЦ «Зимняя вишня». 5. Лига наций УЕФА. 6. Убийство Дениса Тена. 7. Повышение пенсионного возраста в России. 8. Лунное затмение 27 июля. 9. Вспышка менингита в Казахстане. 10. Президентские выборы в России. Среди топ 10 тем запросов к социально значимым для казахстанцев можно отнести лишь две темы: «Вспышка менингита в Казахстане» и «Убийство Дениса Тена».

В свете сказанного, при определении оценкам социальной значимости тематик можно пользоваться в том числе экспертными оценками. Так, Президент Казахстана Н.А. Назарбаев на заседании Совета безопасности РК 7 ноября 2018 года назвал наиболее острые вопросы, которые волнуют казахстанцев [32]: «Итоги социологического исследования показали, что на первом месте из шести проблем, которые выделили, стоит высокая стоимость коммунальных услуг». Среди других, особо волнующих казахстанцев вопросов Президент указал дорогое медицинское обслуживание и обучение, низкое качество образования.

Исследование, проведенное Центром политического анализа и стратегических исследований партии «Нур Отан» по итогам 2017 года, показало, что 10 наиболее острых для населения Казахстана проблем таковы [33]:

1. Рост цен на продукты питания, товары первой необходимости.
2. Низкие доходы, нехватка денег.
3. Высокие тарифы на коммунальные услуги.
4. Выплата кредита.
5. Низкое качество медицинского обслуживания.
6. Коррупция.
7. Цены на ГСМ.
8. Страх потерять работу.
9. Отсутствие собственного жилья.
10. Отсутствие работы.

Исходя из вышесказанного можно составить предварительный список социально-значимых тем: Цены на коммунальные услуги, Цены на продукты питания, Цены на ГСМ, Цены на жилье, Уровень доходов населения, Кредитование населения, Коррупция, Медицинское обслуживание и Образование.

Для определения так называемых подозрительных (потенциально-опасных) публикаций предлагается следующая схема, включающая несколько последовательных этапов классификации (рисунок 3).



Рисунок 3 – Этапы классификации текстов с целью выявления «подозрительных» публикаций

Примечание – Составлено по источнику [34]

Обобщенно процесс анализа текста СМИ выглядит следующим образом:

- 5) формируется перечень параметров, определяющих принадлежность текста к одному из классов;
- 6) оценивается сравнительная важность информативных параметров;
- 7) для каждого текста, вычисляется оценка параметров;
- 8) оценки параметров и их сравнительная важность агрегируется для получения интегральной оценки принадлежности текста к классам;
- 9) используя полученные оценки текстов, формируется обобщенные оценки СМИ.

Использование байесовского подхода позволяет формировать вероятность соответствующей гипотезы при неполной информации, например в случае отсутствия части текстов, или когда классификация по определенным информативным признакам затруднительна или невозможна.

На рисунке 4 представлена общая концептуальная схема оценки новостных текстов.

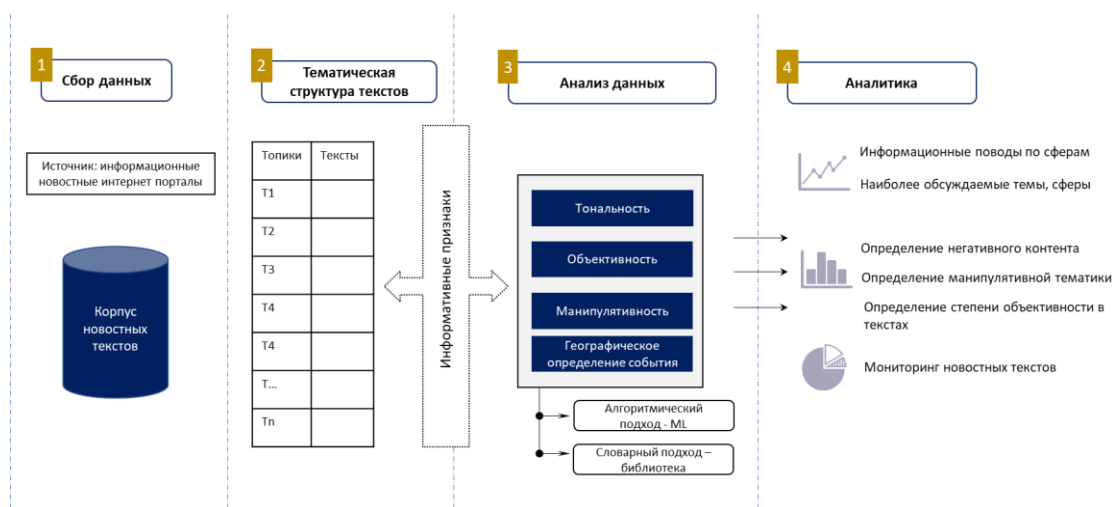


Рисунок 4 – Концептуальная схема оценки новостных текстов

На первом этапе осуществляется сбор новостных публикаций из открытых информационных интернет-источников.

На втором этапе осуществляется тематическое моделирование. Большинство тематических моделей в той или иной мере основаны на латентном размещении Дирихле (LDA, предложенной Дэвидом Блеем в 2003 году [35]). Это современный активно развивающийся вероятностный инструмент, который применяется в задачах анализа данных [36], в том числе для анализа текстов новостных публикаций [37], а также является одним из инструментов автоматизированного анализа корпуса новостных текстов и оценки их влияния на социум [38].

Тематическое моделирование активно применяется в различных гуманитарных исследованиях, включая социологические и политологические, а также в области принятия решений на государственном уровне. Результаты тематического моделирования могут быть полезны при определении повестки дня, информационных поводов и трендов, а также динамики изменения публикационной активности в отдельных темах во времени [37, р. 366].

На третьем этапе на основании результатов тематического анализа происходит построение, валидация и применения моделей (машинного обучения, либо словарных/на правилах).

На последнем этапе осуществляется визуализация результатов анализа данных в аналитическом инструменте, позволяющем определить информационные поводы по различным тематическим областям, наиболее обсуждаемые тематики, наиболее негативные тематики и источники (СМИ) и т.п.

### **Выводы по 1-му разделу**

Задачи медиа-мониторинга являются в высшей степени востребованными в самых разных областях деятельности от простого поиска информации до принятия решений на государственном уровне. При этом такие системы для использования в продуктовой среде являются высокотехнологичными, сложными с технической точки зрения программными решениями.

Главные недостатки и ограничения существующих систем можно обобщить следующим образом:

1. Ограниченность функционала классификации ввиду необходимости дорогостоящей объемной экспертной разметки, и использования вычислительно-затратных моделей глубокого обучения. Обычно функционал классификации ограничен только анализом тональности, решений с другими критериями классификации на рынке практически нет.

2. В большинстве систем медиа-мониторинга использование простых запросов, по ключевым словам, при поиске, без возможности задания комплексных семантических запросов.

3. Не решается вопрос оценки самих медиа-источников. В некоторых решениях есть возможность получения простой описательной статистики по базовым показателям (количество просмотров, лайков, средняя тональность и

пр.), однако многокритериальной семантической оценки СМИ в существующих продуктах не производится.

В работе предлагается концептуальная схема взаимодействия общества, СМИ и социальных сетей, и государства, на основе которой предлагается каскад вычислительных процедур, позволяющий проводить мультикритериальную оценку как отдельных публикаций, так и СМИ в целом. При этом предлагаемая процедура оценки не требует большого объема размеченных документов, поскольку в качестве базовой единицы анализа предлагается использовать топики (темы), полученные в ходе работы тематических моделей на базе латентного Дирихле размещения (LDA) на больших неразмеченных текстовых корпусах новостных публикаций.

Был исследован мировой опыт и проведен литературный анализ исследований, связанных как с качественной, так и с численной оценкой влияния СМИ на социум. Был сделан вывод, что анализ влияния СМИ в Казахстане может происходить и быть достаточно полным даже при концентрации только на данных из открытых электронных средств массовой информации, без учета телевидения, радио и других форматов вещания, поскольку анализ статистики показал, что на сегодняшний день в Казахстане влияния электронных текстовых СМИ равномерно для всех социальных страт.

Среди важнейших для классификации свойств новостных публикаций было выделено:

1. Тональность. Нужно отметить, что здесь ввиду особенностей задачи под тональностью понимается не эмоциональность текста или характер оценочных суждений, а оценка описываемых событий с точки зрения их влияния на развитие личности и социума.

2. Социальная значимость. Под социальной значимостью понимается отношение статьи к приоритетам государственного развития, отношение к волнующим общество вопросам и потенциальная резонансность публикации.

3. Потенциальная резонансность публикации – потенциал взрывной (вирусной) реакции общества на публикацию. Этот показатель достаточно легко оценивать постфактум (по показателям активности публикации), однако оценка этого показателя на момент публикации представляет как техническую сложность, так и большой научный интерес.

## 2 МЕТОДЫ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ В КОНТЕКСТЕ BIG DATA

### 2.1 Проблема классификации текстовых документов и классические подходы к решению

Обработка естественного языка (natural language processing, NLP) – бурно развивающаяся область исследований, результатами которых в виде технологий обработки речи и текстов активно пользуются многие направления индустрии.

Потребность в развитии данного направления связана с огромным количеством генерируемой в настоящее время информации. Тексты и иные материалы, распространяемые посредством Интернет, могут становиться как положительным, так и отрицательным фактором влияния на общество. Допустимо позволить себе некоторую метафору, сравнив информацию, существующую и питающую общественные процессы, с воздухом. Чем чище «воздух» в обществе, тем более защищено оно от «отравления» и «болезней», что, безусловно, является определяющим фактором устойчивого развития личностей, составляющих данное общество. Таким образом, на пути информационных потоков требуются фильтры, затрудняющие деструктивное воздействие. Как отмечается в [39], необходимы новые защитные меры, полученные в ходе междисциплинарных исследований, направленные на ограничение распространения ложной информации. Разумеется, решение этой задачи лежит не только в сфере информационных технологий, но информационные технологии являются важнейшим инструментом в обеспечения достоверности информации в медиaprостранстве. В рамках создания защитного механизма одной из важнейших задач является задача классификации текстов, в частности новостного характера. Хотя достигнуты несомненные успехи в различных областях NLP, тем не менее, ряд задач далек от решения. К числу таких задач относятся, в частности, задача поиска недостоверных текстов, задача точного определения тональности текста и т.п. Эти задачи пока не могут быть решены напрямую – путем создания большого корпуса текстов и обучением классификаторов. Причина состоит в том, что во многих случаях результат классификации зависит от контекста, объем которого может быть слишком велик или слишком разнороден для представления его в цифровом виде. Тем не менее, о содержании текста можно судить по ряду признаков, оценив которые можно, по крайней мере, сузить зону анализа, выделяя публикации, заслуживающие дополнительного рассмотрения.

Классическая задача классификации текстовых документов формулируется следующим образом [40]. Пусть  $D=\{d_1,\dots,d_{|d|}\}$  – это множество документов,  $C=\{c_1,\dots,c_{|c|}\}$  – множество заранее заданных классов. Также имеется некоторая неизвестная целевая функция  $\Phi:D\times C\rightarrow[0,1]$ , задаваемая формулой (1):

$$\Phi(d_j,c_i) = \begin{cases} 0, & \text{если } d_j \notin c_i \\ 1, & \text{если } d_j \in c_i \end{cases} \quad (1)$$



Задача текстовой классификации состоит в построении классификатора  $\Phi'$  максимально близкого к  $\Phi$ .

Если рассматривать историю развития моделей для решения задачи классификации текстов (например, анализ тональности), можно отметить тенденцию к постепенному усложнению применяемых моделей и методов. В начале 2010 года самые популярные модели были основаны на векторизации с помощью алгоритма мешок слов (bag of words) с применением tf-idf [41] с последующим применением простых моделей машинного обучения. Затем с появлением векторизации word2vec возникла проблема построения глубоких моделей для комбинирования векторных представлений слов в так называемые текстовые эмбединги (embeddings), а также применение рекурсивных глубоких сетей, таких как LSTM и RNN, для построения эмбедингов с учетом порядка слов [42]. Таким образом, количество параметров (весов) в моделях увеличилось с сотен и тысяч до миллионов.

Следующей вехой в развитии NLP моделей, включая классификацию текстов, стало применение подхода Transfer Learning (обычно переводится на русский как стратегия переноса обучения) используя парадигму semi-supervised learning (обучение частично с учителем). Transfer Learning предполагает, что модели, обученные на одном наборе данных, частично используются для получения знания о построении эффективных признаков (features), объединенных в векторное представление (embedding), и использования на других, ранее неизвестных наборах данных из другого (иногда непересекающегося) домена (иногда с переобучением либо дообучением отдельных слоев, либо используя модель как есть). При этом в случае NLP, этот набор данных обычно размечен автоматически. Например, в случае с Word2Vec происходит маскировка либо слов контекста (CBOW - continuous bag of words), либо наоборот слова, оставляя контекст (skip-gram). А в случае с BERT применяется две задачи для обучения – в первой маскируются случайные слова в предложении и модель должна их предсказать, а во второй на вход подается два предложения и модель должна определить следовало ли второе предложение в изначальном тексте непосредственно за первым. Именно поэтому такой подход называют semi-supervised learning – поскольку обучение происходит как в классическом supervised learning подходе (машинное обучение с учителем), однако разметка сформирована в полностью автоматическом режиме из самих данных, без какой-либо информации от экспертов или других внешних источников данных. Этот подход был впервые применен в сетях CNN для визуального анализа данных (например, ImageNet [43]). Первым успешным представителем этого подхода в области NLP является модель BERT [44] с более чем сотнями миллионов параметров.

С одной стороны, такой подход позволяет получить очень хорошие результаты по ряду задач NLP, включая классификацию текстов, вопросно-ответные системы, определение меры близости предложений и текстов, распознавание именованных сущностей и другие. С другой стороны, у этого подхода есть несколько ограничений:

– необходимость высокой вычислительной мощности как во время обучения, так и при использовании моделей (feed-forward). Например, современная state-of-the-art transformer модель GPT-3 от OpenAI содержит несколько сот миллиардов параметров и при обучении на одной видеокарте Tesla V100 (лучшая в своем классе на 2021 год) процесс обучения занял бы примерно 355 лет [45] (в реальности модель обучалась на целом кластере из сотен GPU);

– необходимость в большом количестве размеченных данных для каждой конкретной задачи;

– низкая интерпретируемость результатов, ввиду большой сложности моделей.

С развитием моделей и подходов эти проблемы постоянно усугублялись возрастающей сложностью моделей, что приводило к необходимости суперкомпьютерных систем для их обучения и использования и сотен тысяч или более размеченных примеров для успешного обучения. Например, стоимость предварительного обучения модели OpenAI GPT-3 без дополнительного обучения под конкретную задачу составляет более 6 миллионов долларов.

В этой работе предлагается подход, нацеленный на решение этих проблем и обход существующих ограничений, сводя к минимуму неизбежные потери качества для широкого круга задач, применяя тематические модели для построения эффективных векторных представлений.

## **2.2 Проблема векторизации документов, Topic Embeddings**

Во многих задачах NLP, включая классификацию текстов, формирование эффективных векторных представлений для текстов (text embedding) является ключевой задачей. После получения эффективного векторного представления сам процесс классификации можно считать тривиальным. Даже в продвинутых моделях, таких как BERT, после получения векторного представления к нему применяется либо линейная регрессия (во большинстве случаев), двух-трех-слойный перцептрон (нейронная сеть), либо простые алгоритмы на деревьях решений (обычно random forest либо gradient boosting). Следовательно, вопрос векторизации является ключевым, поскольку вопрос оптимизации и обучения вышеозначенных моделей можно считать решенным.

Если рассмотреть историю развития методов векторизации текстов, то можно заметить тенденцию к все большему использованию больших корпусов для получения более точных оценок.

Наиболее примитивный алгоритм векторизации – bag-of-words (мешок слов) представляет собой простой counter vector, каждая компонента которого показывает сколько раз в тексте встретилось то или иное слова из словаря. И уже на этом этапе есть определенная информация, которая должна быть получена из корпуса – сам перечень слов, а также желательно их частотные характеристики, поскольку в общем случае общий размер словаря может быть слишком велик (сотни тысяч-миллионы слов в зависимости от особенностей

языка), следовательно на основании частотных (и иных) характеристик может быть принято решение о том какие слова (и фразы) должны войти в словарь.

Следующей вехой можно выделить Tf-IDF [46] (term frequency-inverse document frequency) – метод, позволяющий оценить важность отдельных слов на основании частоты их появления в заданном корпусе. Этот метод позволяет значительно улучшить качество классификации bag-of-words моделей, поскольку позволяет уменьшить влияние фоновых слов, и увеличить влияние значимых слов.

Дальнейшим этапом развития вопроса векторизации текстов стал Word2Vec. Word2Vec фактически представляет собой простой Variational AutoEncoder, который обучается с некоторыми изменениями [47]. Как итог работы, полученные после обучения веса с одного из слоев рассматриваются в качестве векторных представлений соответствующих слов. Таким образом, Word2Vec использует большие корпуса текстов, чтобы получить максимально эффективные векторные представления слов.

Однако на этом этапе встало два ключевых вопроса: первый – каким образом возможно получение векторных представлений на уровне текстов на базе Word2Vec, и второй – проблема снятия омонимии и в целом вопрос учета контекста слова для уточнения его смысла (а значит и векторизации).

Здесь был сделан последний на текущий момент значимый прорыв – использование трансформеров (стеков рекуррентных моделей) вкупе с так называемыми слоями внимания (attention layers [48]) позволили создавать наиболее эффективные векторные представления как слов (с учетом контекста, в отличие от Word2Vec), так и текстов. К этому семейству моделей можно отнести EMO, BERT, GPT1-3, XLNet и другие.

Однако, как обсуждалось в разделе 2.1, у этих моделей есть ряд серьезных недостатков и ограничений, в основном связанных с их высокой сложностью (количеством параметров). Исходя из этого, в диссертационной работе предлагается метод векторизации, основанный на тематическом моделировании, нацеленный на решение этих проблем. Тематическое моделирование так же, как и другие модели, описанные выше, использует большие корпуса текстов для получения информации о скрытых внутренних структурах корпусов, что позволяет получать эффективные тематические текстовые векторные представления (topic embedding), причем обладающие меньшей размерностью, чем современные трансформеры, быстрее по производительности, и в ряде случаев показывающие сравнимые или лучшие результаты (качество классификации). Следовательно, предложенный подход следует, как и другие вышеописанные актуальные подходы к классификации, рассматривать в контексте Big Data и Transfer Learning, поскольку происходит извлечение знаний из больших неразмеченных корпусов текстов, с потенциалом использования в других (частично пересекающихся) доменах.

Рассмотрим более подробно что такое тематическая модель. Для построения тематической модели корпуса документов обычно используются: вероятностный латентно-семантический анализ (PLSA), ARTM (аддитивная

регуляризация тематических моделей) [49] и очень популярное латентное распределение Дирихле (LDA) [50, 51]. LDA можно выразить следующим равенством, представляющее собой сумму смешанных условных распределений по всем темам множества  $T$ :

$$p(w, m) = \sum_{t \in T} p(w | t, m) p(t | m) = \sum_{t \in T} p(w | t) p(t | m) = \sum_{t \in T} \varphi_{wt} \theta_{tm} \quad (2.2.1)$$

где  $p(w | t)$  – условное распределение слов в темах;

$p(t | m)$  – условное распределение тем в новостях. Переход от условного распределения  $p(w | t, m)$  к  $p(w | t)$  осуществляется благодаря гипотезе условной независимости, согласно которой появление слов в новости  $m$  по теме  $t$  зависит от темы, но не зависит от новости  $m$ , и она общая для всех новостей. Это соотношение верно, исходя из предположения, что нет необходимости сохранять порядок документов (новостей) в корпусе и порядок слов в новостях. Кроме того, метод LDA предполагает, что компоненты  $\varphi_{wt}$  и  $\theta_{tm}$  генерируются непрерывным многомерным вероятностным распределением Дирихле.

Цель алгоритма – поиск параметров  $\varphi_{wt}$  и  $\theta_{tm}$  путем максимизации функции правдоподобия с соответствующей регуляризацией:

$$\sum_{m \in M} \sum_{w \in m} n_{mw} \ln \sum_{t \in T} \varphi_{wt} \theta_{tm} + R(\varphi, \theta) \rightarrow \max \quad (2.2.2)$$

$$R(\varphi, \theta) = \sum_{i=1} \tau_i R_i(\varphi, \theta) \quad (2.2.3)$$

где  $n_{mw}$  – количество вхождений слова  $w$  в новость  $m$ ,  $R(\varphi, \theta)$ , определенное формулой (2.2.3) – логарифмический регуляризатор. Для определения оптимального количества тематических кластеров  $T$  часто используется метод максимизации значения когерентности, вычисляемого с помощью метрики UMass [52].

BigARTM предлагает набор регуляризаторов, реализованных на основе дивергенции Кульбака-Лейблера, в данном случае демонстрируя энтропийные различия между распределениями исходной матрицы  $p'(w | d)$  и модели  $p(w | d)$ :

1. Сглаживающий регуляризатор, основанный на предположении, что столбцы матрицы  $\varphi$  и  $\theta$  генерируются распределениями Дирихле с гиперпараметрами  $\beta_0 \beta_t$  и  $\alpha_0 \alpha_t$ , (идентично реализации модели скрытого размещения Дирихле LDA, где гиперпараметры могут быть только положительными):

$$R(\varphi, \theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \varphi_{wt} + \alpha_0 \sum_{d \in D} \sum_{w \in W} \alpha_{td} \ln \theta_{td} \rightarrow \max \quad (2.2.4)$$

Таким образом, мы можем выделить основные темы, определить словарный запас языка или вычислить общий вокабуляр в разрезе каждого документа.

2. Убывающий регуляризатор, регуляризатор обратного сглаживания направлен на выявление значимых предметных слов, так называемых

лексических ядер, а также предметных тем в каждом документе, обнуляя малые вероятности:

$$R(\varphi, \theta) = -\beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \varphi_{wt} - \alpha_0 \sum_{d \in D} \sum_{w \in W} \alpha_{td} \ln \theta_{td} \rightarrow \max \quad (2.2.5)$$

3. Декорректирующий регуляризатор, делает темы более разнообразными. Выбор тем позволяет модели избавиться от мелких, неинформативных, повторяющихся и зависимых тем:

$$R(\varphi, \theta) = -0.5 * \tau \sum_{t \in T} \sum_{s \in T/t} \text{cov}(\varphi_t \varphi_s) \rightarrow \max \quad (2.2.6)$$

$$\text{cov}(\varphi_t \varphi_s) = \sum_{w \in W} \varphi_{wt} \quad (2.2.7)$$

этот регуляризатор не зависит от матрицы  $\theta$ , здесь оценка различий в дискретных распределениях реализована как  $\varphi_{wt} = p(w|t)$ , где мерой является ковариация текущего распределения слов в темах  $\varphi_t$ , заданная формулой 2.2.7, относительно вычисленных распределений  $\varphi_s$ , где  $s \in T/t$ .

Описанные в работе результаты были получены путем применения тематической модели BigARTM с сглаживающим регуляризатором ( $\tau = 0,15$ ),  $\phi$ -декоррелятором ( $\tau = 0,5$ ) и «improve coherence  $\phi$ »-регуляризатором ( $\tau = 0,2$ ), количество тем - 200. Параметры подбирались эмпирически в процессе экспериментов.

Далее с помощью полученных в матрице  $\theta$  весов принадлежности документов к топикам строятся векторные представления каждого документа, которые далее используются для классификации. Процесс разметки может происходить как на уровне документов, так и на уровне слов (seed словарь, который затем проецируется на матрицу  $\varphi$ ), так и на уровне топиков – причем это может быть как экспертная разметка, так и разметка, полученная автоматическим способом из каких-либо внешних данных. Один из способов автоматического получения весов топиков описан в следующем разделе 2.3 – Мера межкорпусного тематического дисбаланса. Также результаты, приведенные в разделе 5 описывают примеры применения таких векторизаций для ряда задач оценки и классификации.

### 2.3 Мера межкорпусного тематического дисбаланса

Как упоминалось в разделе 2.2, для построения классификатора необходимо тем или иным способом получить оценку отдельных топиков по заданному критерию (например тональности или социальной значимости). Это оценка может быть как скалярной (например, оценка эксперта), так и векторной, например в случае, если оценки были получены из размеченного корпуса документов путем применения алгоритмов машинного обучения, например нейронных сетей. Поскольку разработка предложенных методов происходила в контексте Big Data, логичным является попытка использовать не только объем данных (например, для получения векторизаций), но и их

разнообразие (variety) данных. Так, возможно использование метаданных публикаций, кроме самого текста – например даты, источника, автора, тегов (категорий), показателей вовлеченности пользователей (количество лайков, шейров, репостов и просмотров), а также их комбинаций и производных, рассчитанных показателей (feature engineering). Либо, в наиболее простом виде в случае, если имеются несколько корпусов, отличающихся между собой доменом. Под доменом здесь может пониматься, например отличающийся тип текстов – новости, аналитические статьи, программные документы и т.п., кардинально разные источники (например СМИ разных стран) или наличие какого-то иного исходного признака, по которому новости были разделены по корпусам. Если явно такого разделение на корпуса нет, то возможно разделение на корпуса по упомянутым выше метаданным, либо рассчитанным показателям.

Затем предлагается использовать информацию о распределении документов по корпусам внутри каждого топика для оценки меры влияния информации о том, что документ принадлежит к данной теме на решение о принадлежности документа к заданному классу (например, к тональности или социальной значимости). Иными словами, предлагается использовать некий явный признак (например источник публикации) для оценки меры соответствия каждого топика некому скрытому признаку (например тональности). Таким образом, этот подход позволяет еще сильнее уменьшить количество разметки, т.к. разметки уже требуют не документы, не топики, а некие признаки публикаций, количество которых в ряде случаев может быть значительно меньше, чем количество топиков.

Примеры таких соответствий между метаданными (явными показателями) и классами (скрытой, неявной информацией) рассмотрены в разделах 5.3 и 5.3.1 этой работы.

Одной из основных проблем в реализации предложенного выше подхода является необходимость учета несбалансированности полученных корпусов по объему. Следовательно, простые инструменты описательной статистики не позволяют делать достаточно точных выводов, и необходим метод, позволяющий оценить межкорпусный дисбаланс по топикам (тематический межкорпусный дисбаланс) с учетом исходного дисбаланса объемов публикаций по корпусам. При этом должна учитываться нечеткая, непрерывная природа отношения документов к топикам, иными словами, ввиду не-бинарности отношения документов к топикам, простой подсчет количества и нормализация на объем является допустимым, но не идеальным вариантом, поскольку в таком случае мы теряем слишком много информации (значения весов принадлежности документов к топикам).

Таким образом, после получения списка топиков, необходимо провести разделение корпуса текстов на несколько корпусов (либо воспользоваться существующим разделением), с учётом заданного целевого критерия. При этом необходима консультация с экспертами в области в формате брейн-шторминга.

Необходимо дать оценку дисбаланса по корпусам по распределению новостей внутри каждого топика. Эта мера дисбаланса будет являться оценкой влияния принадлежности, данного топика к целевому показателю. После этого требуется провести валидацию полученных результатов, чтобы проверить насколько изначальная гипотеза, выработанная с экспертами, была точна. Если результаты будут неудовлетворительными, рекомендуется итеративно повторять сессии работы с экспертами с повторным разделением на под-корпусы и валидацией. Если данный подход не будет показывать требуемых результатов, рекомендуется усилить его, путем использования ручной разметки документов для автоматической донастройки параметров, либо классического машинного обучения модели с использованием на входе тематических векторизаций текстов и метаданных.

Предлагается следующая (2.3.1) формула для оценки межкорпусного тематического дисбаланса:

$$D_{t_i c_j} = \frac{\sum_k w_{d_k t_i c_j}}{\sum_k \sum_l w_{d_k t_l c_j}} / \sum_m \sum_k \sum_l w_{d_k t_l c_m} \quad (2.3.1)$$

где  $D_{t_i c_j}$  – это мера дисбаланса присутствия документов из корпуса  $c_j$  в топике  $t_i$ , а

$w_{d_k t_l c_m}$  – вес принадлежности документа  $d_k$  из корпуса  $c_m$  к топике  $t_l$ .

После этого возможно несколько сценариев использования полученным мер дисбалансов по топикам. Рассмотрим их:

1. Простое средневзвешенное среднее. Нужно отметить, что несмотря на то, что этот подход является тривиальным, не параметризуемым (то есть не настраиваемым и не обучаемым), результаты данного подхода на практике достаточно близки к результатам применения модели ММА основанной на байесовской агрегации, описанных в разделах 5.3 и 5.3.1. Следовательно, такой подход вполне может использоваться в случаях, когда скорость обработки данных является критичной, поскольку на современных процессорах, и тем более GPU такое, по сути дела, векторное перемножение векторов длиной в несколько сотен компонент можно считать атомарной операцией (то есть получить более оптимальное по производительности решение практически невозможно). Предлагается рассчитывать с помощью формулы (2.3.2):

$$r_{d_i c_k} = \sum_j w_{d_i t_j} * D_{t_j c_k} \quad (2.3.2)$$

где  $r_{d_i c_k}$  – оценка документа  $d_i$  по заданному классу.

2. Байесовский подход – этот подход рассматривает субъективную вероятность соответствия документа заданному критерию. Его преимущества описаны в [53, 54]. Его преимущества заключаются в том, что он позволяет обеспечить устойчивость работы в условиях, когда часть данных недоступна, либо заведомо неточна (например, содержит ошибки), а также в целом

обеспечивает больший уровень нелинейности, чем простое взвешенное среднее и имеет несколько параметров, позволяющих оптимизировать модель без отхода к итеративному изменению изначальных признаков для разделения на под-корпусы. Метод агрегации более подробно описывается в разделе 3.1 данной работы.

3. Semi-supervised подход (обучение частично с учителем). Можно предварительно обучить модель машинного обучения на результатах, полученных с помощью подходов 1, 2 или какого-либо другого подхода обучения без учителя или MCDM, а затем использовать под-корпус размеченных вручную документов для дообучения (fine-tuning) модели, тем самым улучшая ее производительность.

Таким образом, данная модель позволяет проводить обучение (получение оценок дисбаланса) на ограниченном подкорпусе и после этого экстраполировать полученные оценки на остальной корпус, для которого целевой показатель может быть неизвестен. Также такой подход требует минимальный объем разметки, при том, что в ряде задач качество работы такой модели сопоставима с классическими моделями на мешках слов и даже с глубокими предобученными моделям (transfer learning).

### **Выводы по 2-му разделу**

В разделе рассмотрена краткая история развития классических моделей машинного обучения для классификации текстовых данных. Были сделаны выводы о том, что современное развитие методов классификации текстов следует рассматривать в контексте Big Data, применения больших объемов, часто слабоструктурированных и не размеченных данных.

Так, даже наиболее примитивные модели основанные на bag-of-words векторизации (так называемый мешок слов) должны были использовать корпусную информацию для определения релевантного словаря, поскольку использование полного словаря не представляется целесообразным, как в виду вычислительной сложности, так и из математических соображений сходимости модели машинного обучения с сотнями тысяч входов.

Дальнейшее развитие идей bag-of-words в лице tf-idf метода также использует корпусную информацию для улучшения качества работы модели.

Следующей вехой можно выделить интеллектуальный алгоритмы векторизации слов Word2Vec и подобные (FastText, GloVe). Эти алгоритмы также позволяют использовать информацию из неразмеченных текстов для улучшения качества векторизаций, в дальнейшем подаваемых на вход моделей машинного обучения.

И последним на текущий момент прорывом стало появление так называемых трансформеров (transformer) – моделей глубокого обучения построенных на базе рекуррентных нейронных сетей с attention слоями, которые обеспечивают высочайшее качество распознавания, а также показывают отличные результаты в ряде других NLP задач. К трансформерам относят такие модели как ELMo, BERT, Open AI GPT-1-3, XLNet и другие.



Однако, у этих моделей есть ограничения, связанные с их высокой сложностью – низкая производительность, большой объем размеченных данных, требуемых для обучения, и низкая интерпретируемость.

Исходя из текущего состояния области, предлагается альтернативный способ векторизации также использующие большие данные в виде неразмеченных слабоструктурированных корпусов текстов – тематическая векторизация (topic embedding). Данный подход позволяет решить как вопросы производительности, так и вопрос интерпретируемости, при этом обеспечивая на ряде задач сопоставимое качество результатов. В данной работе в качестве реализации предложенного метода используется модель BigARTM, представляющая собой классическую тематическую модель на базе LDA с добавлением группы настраиваемых регуляризаторов.

Другой вывод, сделанный из исследования текущего состояния области классификации текстов, говорит о том, что не все особенности больших данных (Big Data) используются в современных подходах. Так, Volume (размер), Velocity (скорость обновления) и Veracity (достоверность) данных в какой-то мере используются в современных моделях. Так, BERT обучается на больших объемах данных, при этом может дообучаться в режиме реального времени, что позволяет адаптироваться к изменчивым условиям информационного контекста и изменчивой природе языка. Однако последнее свойство, которое как принято считать, определяет понятие Big Data – Variety (разнообразие) [55] фактически не используется в современных моделях классификации текстов. В этой связи, предлагается методика, позволяющая использовать метаданные о публикации такие как дата, источник, типа документа и пр. как источник данных и базовую единицу анализа для разметки. В разделе описывается метод определения межкорпусного тематического дисбаланса, которая позволяет реализовать эту идею, а также описаны рекомендации по практическому применению предложенных методов.

### 3 МЕТОД МУЛЬТИКРИТЕРИАЛЬНОЙ ОЦЕНКИ МЕДИА-ИСТОЧНИКОВ НА БАЗЕ БАЙЕСОВСКОЙ МОДЕЛИ АГРЕГАЦИИ ММА

#### 3.1 Байесовская модель агрегации гетерогенных данных

Оценка текста с точки зрения принадлежности к заданному набору классов заключается в вычислении некоторой агрегированной оценки, складывающейся из оценок параметров. В обобщенной модели это могут быть любые параметры и информативные признаки, а в предложенной конкретной методике, рассматриваемой в этой работе, предлагается векторные представления. Другими словами, на основании оценок параметров необходимо получить некоторое значение, свидетельствующее о принадлежности текста к классу или, наоборот, о том, что текст к заданному классу не относится.

С точки зрения многопараметрического агрегирования, в простейшем случае, можно использовать способ, основанный на суммировании оценок (аддитивная модель), возможно с весами, либо на перемножении (мультипликативная модель).

Мультипликативная многофакторная модель представлена формулой (3.1.1).

$$F = \prod_{i=1}^N f_i \quad (3.1.1)$$

где  $f_i$  = значение  $i$ -го параметра ( $i = \{1, \dots, N\}$ )

Многофакторная мультипликативная модель агрегирует все данные в виде произведения значений факторов. В случае если конкретный фактор не влияет на оценку – его значение устанавливается в 1, если оказывает резко негативное воздействие – в 0.

В результате, агрегированная оценка  $F$  может быть в диапазоне 0-1. Чем она ближе к 1, тем более больше наша уверенность в принадлежности текста к классу.

Несмотря на простоту, при использовании многопараметрической мультипликативной модели возникает проблема достоверности результатов в случае невозможности получения некоторых оценок некоторых параметров. Если значение параметра неизвестно, то каким должно быть значение параметра  $f_i$ ? Если мы принимаем его за 1, то тем самым полагаем, что данный параметр присутствует в полной мере, а если 0, то полностью отсутствует и фактически «запрещает» относить текст к рассматриваемому классу.

В случаях, когда необходимо принять решение на основе множества разнородных параметров и альтернатив, может также использоваться несколько методов для учета относительной важности параметров и агрегирования их значений в виде отдельной оценки или небольшого числа оценок. По своей природе такие задачи относятся к области многокритериального принятия решений (MCDM), широко используемой в системах поддержки принятия решений (DSS) [56-58], что также применимо в NLP [59]. MCDM использует

множество методов для принятия решения на основе разнородных критериев, включая [60]: взвешенную линейную комбинацию (WLC) и упорядоченное взвешенное усреднение (OWA) [61] (по сути дела являющиеся реализациями аддитивного метода с весами и мультипликативного метода, описанных выше), “Potentially All Pairwise Rankings of all possible Alternatives” (PAPRIKA, метод попарного сравнения всех потенциальных альтернатив) [62], ELECTRE [63], метод ранжирования предпочтений по сходству с идеальным решением (TOPSIS) [64], MAUT, “Preference Ranking Organization Method for Enrichment of Evaluations” (PROMETHEE, метод ранжирования предпочтений для улучшения оценок) [65], VIKOR, аналитической иерархический процесс (АИП) [66], нечеткая логика [67], байесовские сети [68, 69] и др.

В целом, из проведенного литературного обзора можно сделать вывод, что мультипликативную модель можно рассматривать как baseline подход, с которым могут сравниваться более продвинутые подходы. Как отмечено выше, мультипликативная система получения консолидированной оценки не дает возможность оценить качество полученных результатов, особенно в случае недостаточности данных или не полной уверенности в их достоверности. Частичным выходом в такой ситуации является применение системы логического вывода, основанного на субъективных вероятностях, предложенного Дж. Перлом [70], который традиционно рассматривается в контексте искусственного интеллекта и широко используется в экспертных системах [71].

В этом случае знания о предметной области представляются в виде правил:

«Если  $h$  является истиной, то  $e$  будет наблюдаться с некоторой вероятностью  $p$ ».

Например, можно рассматривать  $h$  как событие, означающее принадлежность к классу, а  $e$  – событие, заключающееся в поступлении определенного свидетельства, подтверждающего правильность оценки принадлежности текста к классу.

Другими словами, обозначим  $h$  как событие, которое означает что некая гипотеза (гипотеза о принадлежности текста к классу) верна, и  $e$  - событие, заключающееся в том, что поступило определенное доказательство (свидетельство, фактор), который может подтвердить указанную гипотезу.

В соответствии с формулой Байеса мы можем выразить вероятность события  $h$  при условии наступления события  $e$  в виде формулы (3.1.2).

$$p(h|e) = \frac{p(e|h) \times p(h)}{p(e|h) \times p(h) + p(e|\sim h) \times p(\sim h)} \quad (3.1.2)$$

где  $p(e|h)$  – условная вероятность наступления события  $e$  при справедливости  $h$ ;

$p(h)$  – априорная вероятность гипотезы  $h$ ;

$p(e | \sim h)$  – условная вероятность  $e$  при отсутствии  $h$ ;  
 $p(\sim h)$  – вероятность того, что событие  $h$  не верно, которое, в соответствии с определением полной вероятности, может быть вычислено по формуле (3.1.3).

$$p(\sim h) = 1 - p(h) \quad (3.1.3)$$

Таким образом, для вычисления условной вероятности  $p(h|e)$  достаточно знать вероятности  $p(e|h)$  и  $p(e | \sim h)$  и априорную вероятность  $p(h)$ . В наших терминах, условные вероятности могут быть интерпретированы как веса параметров, свидетельствующих за или против гипотезы о принадлежности текста к классу. Практически можно рассматривать и несколько событий  $\{h_1, h_2, \dots, h_n\} \subset H$ , заключающихся, например, в принадлежности текста к нескольким классам, а также против. Обычно факторов(свидетельств) множество  $e \subset E$ . Тогда алгоритм вычисления вероятности  $p(h|e)$  состоит из следующих этапов:

Для каждой гипотезы  $h \in H$  и всех свидетельств  $e \subset E$  необходимо определить априорные вероятности  $p(h)$  и условные вероятности  $p(e|h)$  и  $p(e | \sim h)$ . Последовательно вычислить вероятности  $p(h|e)$  и  $p(h | \sim e)$  для каждой гипотезы  $h$  и свидетельств  $e$ . При этом для каждого последующего свидетельства априорная вероятность гипотезы  $p(h)$  устанавливается равной найденной на предыдущем шаге условной вероятности  $p(h|e)$ , то есть  $p(h) := p(h|e)$  (здесь  $:=$  - оператор присваивания, в отличие от знака равенства  $=$ ). Более подробно, алгоритм расчета состоит из следующих шагов:

1. Пусть имеем некоторую априорную вероятность события  $p(h)$ .
2. Для данного свидетельства (фактора)  $e_i$  определяем  $p(e_i|h)$  и  $p(e_i | \sim h)$
3. Используя формулу 3.1.2, вычисляем  $p(h|e_i)$  и  $p(\sim h|e_i)$  в зависимости от исхода  $e_i$ , т. е. вычисляем апостериорную вероятность события  $h$  и  $\sim h$ .
4. Изменим текущую априорную вероятность события  $h$ , приравняв ее полученному значению апостериорной вероятности  $p(h) := p(h|e_i)$ .
5. Выберем новое свидетельство  $e_j$  для рассмотрения и перейдем к шагу 2.

В итоге для каждой гипотезы будут получены две оценки  $p(h|e)$  и  $p(h | \sim e)$ , которые можно сравнить между собой и выбрать те  $p(h|e)$  и  $p(h | \sim e)$ , которые будет обладать максимальной и минимальной величиной, соответственно.

Часто стоит вопрос об уверенности в наличии свидетельств или факторов, поскольку в ряде случаев мы можем только предполагать наличие некоторых параметров без полной уверенности в их реальном присутствии или отсутствии. Например, мы можем предполагать наличие некоторого свидетельства с некоторой долей уверенности, которую можно оценивать в диапазоне от 1 (фактор точно присутствует) до 0 (фактор точно отсутствует). В этом случае, предполагаемая оценка находится между  $\max=p(h|e)$  и  $\min=p(h|e)$  и возможный простой способ ее вычисления – линейная экстраполяция. Так, если

доля уверенности составляет  $p_c$ , то окончательное значение апостериорной вероятности можно рассчитать по формуле (3.1.4):

$$p(h|e, p_c) = (p(h|e) - p(h| \sim e)) \times p_c + p(h| \sim e) \quad (3.1.4)$$

Когда сведений о параметрах текста мало, модели, основанные на многофакторном анализе, выдают единственную величину, которой может быть недостаточно для оценки его принадлежности тому или иному классу. В то же время, байесовская система, учитывающая степень уверенности в наличии параметров, дает две оценки («за» и «против»), которые позволяют сделать более обоснованный выбор итогового результата.

Следует отметить, что обе модели в общем случае не являются самообучаемыми и требуют привлечения экспертов для получения оценок влияния факторов, либо наличия процедур для автоматического получения оценок (например метод, предложенные в разделе 2.3 работы).

Подход, предлагающий использование описанного метода байесовской агрегации в комбинации с аналитическим иерархическим процессом (АНР) и применением нечеткой логики для нормализации входных параметров был впервые описан в [54, p. 122283].

### **3.2 Метод мультикритериальной оценки медиа-источников ММА**

Основным методом, предлагаемым в работе, объединяющим остальные описаны методы и подходы в рамках одной модели является ММА (Multicriteria Mass-media Assessment) – метод мультикритериальной оценки медиа-источников. Данный метод был впервые описан в [19, p. 5].

Целью модели является, используя распределение вероятностей документов корпуса, агрегировать показатели соответствия статьи тематикам, тематик признакам (словарям) и классам для получения оценок соответствия средств массовой информации (СМИ) в трех модальностях: тематикам, признакам и классам.

Для оценки указанных соответствий используется вероятностный байесовский подход в рамках гипотезы, что вероятностные распределения статей, тематик, классов и признаков являются независимыми.

Ожидаемый конечный результат работы модели – оценка "принадлежности" СМИ к тематикам, признакам и классам в виде распределений вероятностей.

Используя множество тематик корпуса, полученных в ходе работы тематической модели, во-первых, получаем дискретное распределение вероятности статей по тематикам ( $p_2$ ) (фактически матрица  $\theta$ , описанная в разделе 2.2). Во-вторых, получаем распределение словарей по тематикам ( $p_1$ ), то есть, определяем, в какой степени словарь описывает конкретную тематику (фактически под словарем может пониматься матрица  $\varphi$ , либо словарь может быть получен другим методом). В-третьих, с помощью аналитического иерархического процесса (АНР) рассчитываем важность словарей для классов (отдельно для каждого класса) ( $p_3$ ). Затем, используя  $p_1$  и  $p_3$ , рассчитываем

условное распределение тематик по классам ( $p_4$ ). Зная распределение вероятностей тематики по классам ( $p_4$ ) и распределение вероятностей статьи по тематикам ( $p_2$ ) можно вычислить распределение статьи по классам ( $p_5$ ). В свою очередь, распределение статьи по признакам или словарям ( $p_6$ ) зависит от  $p_1$  и  $p_2$ . Исходные данные и получаемые матрицы условных вероятностей показаны на рисунке. Корпус документов описывается словарем (Corpus Dictionary). СМИ (MMS) являются источником  $m$  документов (рисунок 5).

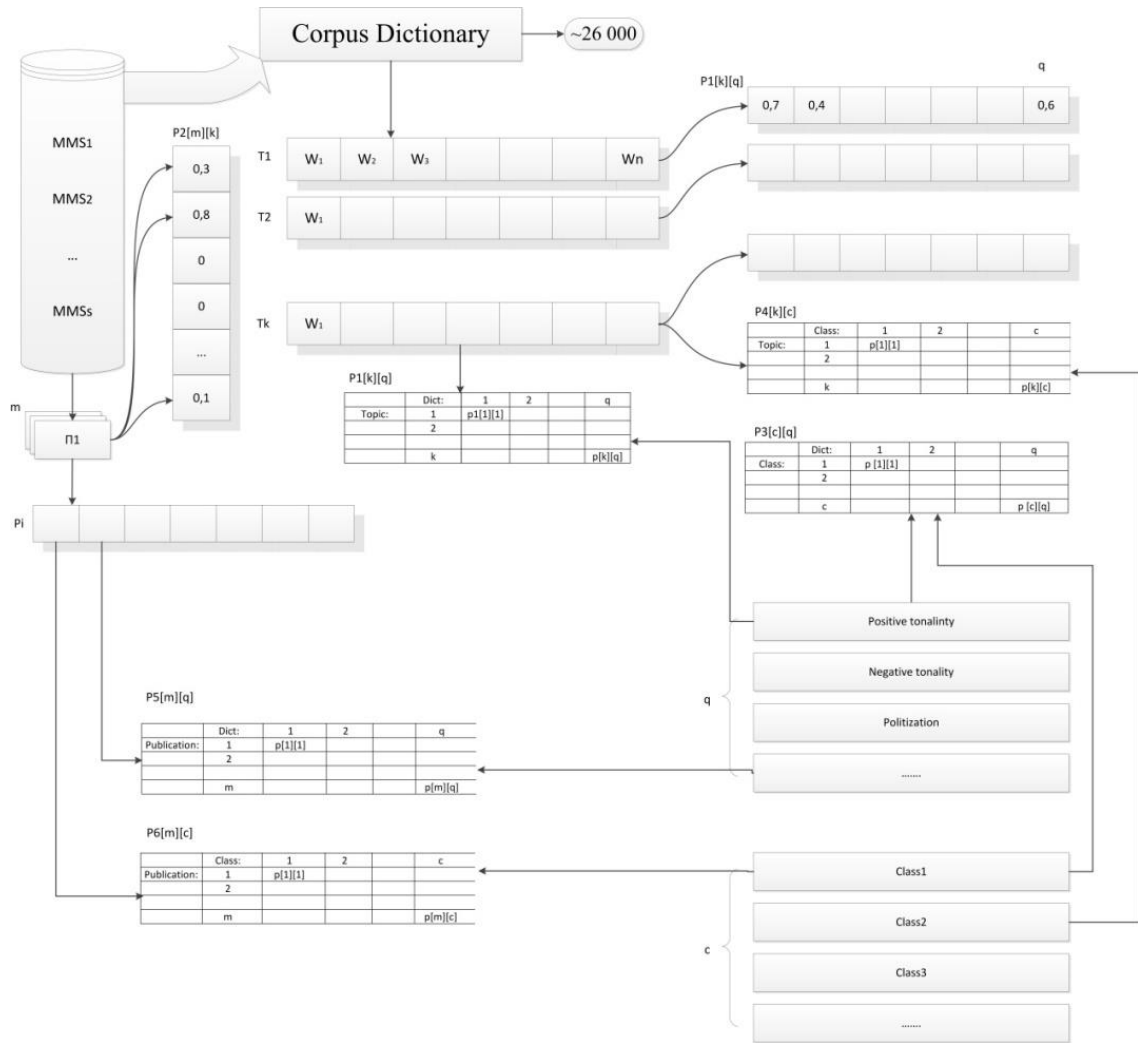


Рисунок 5 – Процессы определения условных вероятностей в методе ММА

Процесс принятия решений с помощью модели АНР, упомянутой выше, состоит из трех этапов [72]:

1. Оценка вектора весов факторов.
2. Расчет матриц мнений экспертов.
3. Ранжирование альтернатив.

В случае предложенного метода (ММА) первый этап используется для ранжирования факторов и применения разработанной байесовской модели агрегации.

Опишем процесс вычисления вектора весов факторов:

Проводится попарное сравнение всех факторов путем заполнения квадратной матрицы размера, соответствующего числу оцениваемых факторов, одним или несколькими экспертами. Для каждой пары факторов  $i$  и  $j$   $a_{ij}$  оценивается экспертом, который может выбрать одно из следующих значений, если  $j$  более важно, чем  $i$ :

$a_{ij} = 1$  -  $j$  и  $i$  одинаково важны;

$a_{ij} = 3$  -  $j$  немного важнее, чем  $i$ , эксперт может назвать это своим субъективным мнением;

$a_{ij} = 5$  -  $j$  важнее  $i$ , эксперт может привести практические доказательства этого;

$a_{ij} = 7$  -  $j$  сильно важнее  $i$ , у эксперта есть устоявшееся, проверенное на опыте мнение по этому поводу;

$a_{ij} = 9$  -  $j$  абсолютно важнее, чем  $i$ , эксперт может это строго доказать (например, строгой логической индукцией или математически).

Следует отметить, что  $a_{ij} * a_{ji} = 1$ , что означает, что если  $j$  важнее, чем  $i$ , то  $i$  пропорционально менее важен, чем  $j$ , а матрица  $A$   $a_{ij}$  симметрична относительно главной диагонали.

Затем вычисляется нормализованная матрица попарного сравнения, так что сумма каждого столбца равна 1:

$$a_{ij} = a_{ij} / \sum_k a_{kj} \quad (3.2.1)$$

В итоге вычисляется вектор весовых коэффициентов  $w$  путем усреднения записей каждой строки:

$$w_i = \sum_k a_{ik} + \sum_k \sum_l a_{kl} \quad (3.2.2)$$

Этот вектор представляет веса каждого из факторов на основе попарного сравнения.

Полученные значения весов используются в процессе нормализации факторов влияния.

В процессе анализа следует исходить из пресуппозиции, что к текстам, оказывающим значительное влияние на социальную информационную среду, следует относиться более внимательно, чем к текстам, связанным с частными или повседневными бытовыми проблемами, юмором и т.д. Предлагается сначала рассматривать популярные статьи, вызывающие бурную реакцию аудитории, а затем выделить из них группу новостей, имеющих социальную значимость, и, наконец, проанализировать последние более тщательно, оценив тональность содержания статьи.

Предлагаемая модель рассматривает оценку каждой публикации в трех модальностях:

1. Темы, полученные посредством тематического моделирования. Например, спорт, образование, экономика, несчастные случаи и т.п.

2. Свойства (критерии оценки) – можно использовать произвольный список свойств. Необходимо наличие процедуры для получения оценок влияния каждой темы на каждое из выбранных свойств. Примеры таких процедур описаны в разделе 5. Эти процедуры могут представлять собой экспертную разметку отдельных топиков, получение оценок из метаданных (метод межкорпусного дисбаланса, раздел 2.3 работы) или процесс машинного обучения или иного способа оптимизации параметров модели классификации. Примерами свойств являются тональность, социальная значимость, объективность, манипулятивность, пропаганда и т.п.

3. Классы. Идентификация статей, относящихся к каждому классу, в целом является конечной целью предлагаемого метода. Выбранные свойства должны иметь некоторую корреляцию с окончательными классами или влиять на них. В случае данной работы основной финальный оцениваемый класс – это негативная информация по социально значимым темам (потенциально опасные новости), однако в целом количество и состав класс может быть произвольным.

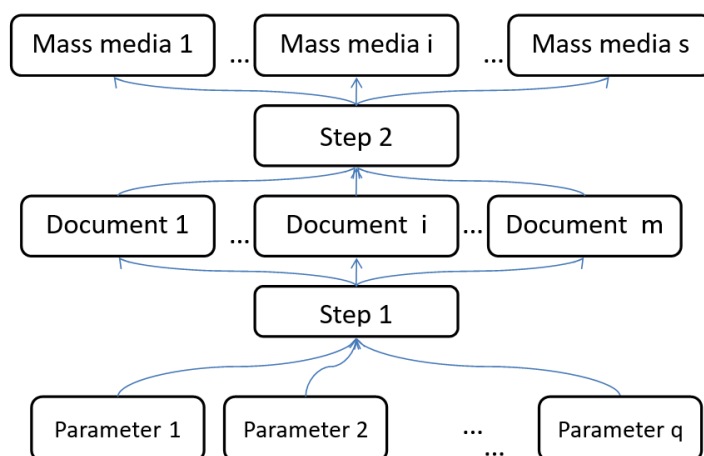


Рисунок 6 – Общая схема предлагаемой модели оценки

Процесс анализа медиатекстов можно кратко описать следующим образом (рисунок 6):

- 1) постановка задачи – выбор финального класса (классов) для оценки;
- 2) составление списка свойств, по которым можно определить, принадлежит ли текст к одному или нескольким из классов;
- 3) оценка сравнительной значимости объектов на основе АНР;
- 4) применение процедуры расчета оценок каждого из свойств для каждого текста на основе тематической модели корпуса;
- 5) агрегирование оценок свойств и их сравнительной значимости для принятия решения о принадлежности к классу (классам), к которым может быть отнесен текст (Этап 1);
- 6) оценка СМИ на основе полученных классификаций текстов (Этап 2).

Перечисленные этапы реализованы в виде метода мультимодальной оценки (ММА). Основная идея многокритериальной оценки и агрегирования субъективных и объективных свойств была первоначально изложена в [54,



p. 122280]. Для реализации описанного процесса в рамках работы были определены свойства текста, как описано в [73, 74], а также оценена их взвешенная значимость для задачи отнесения текста к указанным выше классам методом АНР. Применение тематической модели, созданной посредством кластерного анализа (машинное обучение без учителя), в сочетании с классами и атрибутами, определенными экспертами, примечательно в предлагаемом подходе. Таким образом, семантика распределения определяется пользователем (экспертом), а первоначальное моделирование темы зависит от корпуса документов. Применение байесовского подхода позволяет оценить вероятность гипотезы на основе неполной информации с частью текстовых корпусов. Другими словами, он позволяет получить оценочные оценки для источников для упомянутых модальностей путем обработки только части текстовых корпусов, хотя и с меньшей точностью.

Цель алгоритма – агрегировать веса соответствия статей темам, а затем соответствия тем параметрам/свойствам и классам. Это позволяет получать оценки соответствия СМИ в трех модальностях: темы, свойства и классы.

Обсудим два основных этапа работы алгоритма. На первом этапе генерируется тематическую модель и вычисляются значения условных вероятностей  $p_1, \dots, p_b$ . На втором этапе агрегируются полученные условные распределения вероятностей и вычисляются оценки свойств каждого средства массовой информации и их отнесение к определенным классам.

#### Этап 1. Расчет условных вероятностей

Исходные данные и полученные матрицы условных вероятностей показаны на рисунке 3.2.3, где источники СМИ (MMS) представляют собой набор текстовых источников.  $S$  источников (MMS) являются источниками  $m$  документов (публикации), которые получены с применением систем сбора данных. Получившиеся корпуса  $M$  делятся на тематические кластеры  $T$ . Эксперты формируют классы  $C$ , определяя свойства классов  $Q$ . Свойства описываются словарями и функциями (процедурами расчетов). Опишем процесс расчета модели M4A. В контексте байесовской агрегации следует упомянуть несколько вычислительных деталей. Если субъективная вероятность  $p(e|h)$  равна 0,5, она не повлияет на целевую гипотезу  $p(h|e)$  и может быть интерпретирована как неубедительная или нерелевантная. Следовательно, если  $p(e|h)$  больше 0,5, его влияние на целевую гипотезу будет положительным, а если оно ниже 0,5 - отрицательным. Эта особенность подразумевает, что нормализация на каждом шаге процесса вычисления должна корректироваться соответствующим образом. Например, если свойство/событие имеет положительное влияние 0,8, но применяется к определенному объекту с весом влияния 0,5, было бы неправильно просто умножать эти два числа, а затем перенормировать их, поскольку результат умножения будет равен 0,4, что является небольшим отрицательным воздействием, которое не соответствует фактическому влиянию функции/события, равному 0,8 (положительное). И даже если позже попытаться перенормировать результаты, нет гарантии, что середина (0,5) останется незатронутой. Также следует уточнить, что вес (0,5) в

предыдущем примере обычно соответствует степени, в которой документ соответствует теме, которой было присвоено определенное значение воздействия (0,8). Однако это может измениться от одного шага модели к другому, как будет отмечено ниже. Чтобы решить эту проблему, был введен особый процесс нормализации взвешенного воздействия, который выполняется по следующей формуле (3.2.3):

$$(p - 0.5) * w + 0.5 , \quad (3.2.3)$$

где  $p$  – влияние признака/события/темы;

$w$  – вес (или степень, в которой объект относится к данной функции/событию/теме).

После этого применяется байесовский процесс агрегирования, описанный выше. Последний шаг – нормализация каждой строки результирующей матрицы, которая также настраивается - все значения ниже 0,5 нормализуются отдельно до диапазона [0; 0,5], а все значения выше 0,5 нормализуются до [0,5; 1]. Эта нормализация позволяет предотвратить затухание значений (weight decay), чтобы поддерживать значения на адекватном уровне насыщения (saturation) в процессе множественного преобразованиями матриц ( $p1 \rightarrow p6$ ).

Теперь опишем процесс вычисления матриц  $p$  (рисунок 7):

1. Матрица  $p1$  описывает взаимосвязь между темами и критериями оценки. Ее можно получить разными способами, включая генерацию словарей, ручную разметку и (полу) автоматическую разметку с использованием многокорпусного подхода [75, 76] (также описано в разделе 2.3 работы).

2. Матрица  $p2$  описывает связь между документами и темами и получается посредством тематического моделирования (в тематических моделях на основе LDA она соответствует тета-матрице).

3. Матрица  $p3$  получается в результате выполнения первого этапа АНР модели – приведения матрицы попарного сравнения критериев оценки к вектору-столбцу важности каждого критерия.

4.  $p4$  описывает отношения между темами и целевыми классами. Матрица  $p4$  вычисляется как особая процедура умножения матрицы  $p1$  на  $p3$ . Под особой процедурой умножения матрицы здесь понимается обычное матричное умножение, в котором оператор скалярного умножения  $*$  заменен на особый процесс нормализации с учетом особенности умножения субъективных вероятностей, как описано выше.

5.  $p5$  описывает оценки отношений каждого документа к каждому целевому классу и рассчитывается с использованием особой процедуры перемножения матриц, описанной выше.

6.  $p6$  описывает оценки отношений каждого документа к каждому критерию оценки, а также рассчитывается с использованием особой процедуры перемножения матриц, описанной выше.

Основным результатом всего процесса является матрица  $p5$ , которая описывает отношение каждого документа к целевым классам. Матрица  $p6$

также может иметь некоторую практическую применимость в зависимости от того, насколько полезны критерии оценки независимо от рассматриваемого класса (классов).

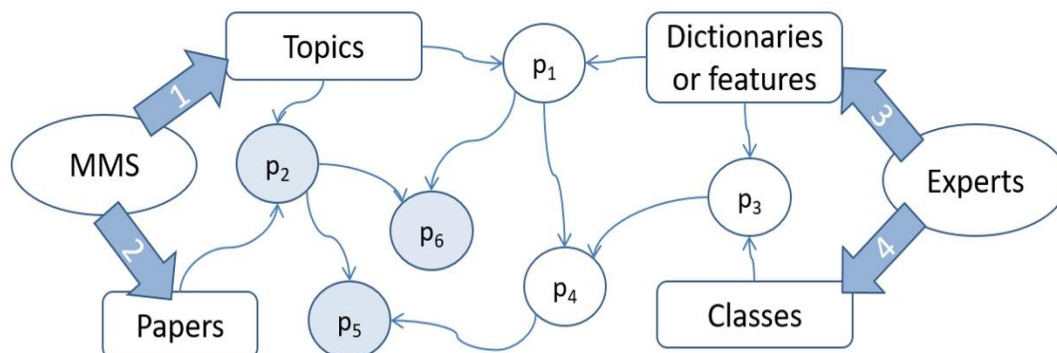


Рисунок 7 – Схема получения распределений субъективных вероятностей

После получения оценок по каждой статье СМИ оцениваются на втором этапе предлагаемой модели, описанной ниже.

#### Этап 2. Оценка СМИ

Второй этап модели – агрегирование оценок документов по каждому критерию и классу с целью получения общей оценки для каждого источника СМИ. Чтобы оценить общее влияние того или иного источника СМИ на общее информационное поле по данному критерию, необходимо рассмотреть два элемента: общий объем значений данных критериев среди документов, опубликованных в данном источнике, и среднее значение данных критериев. Если мы рассмотрим социальную значимость в качестве примера, мы могли бы назвать эти два элемента общим объемом социально значимых документов и концентрацией социально значимых документов. Это означает, что даже если концентрация социально значимых новостей высока, но в СМИ обычно публикуется лишь несколько новостных статей каждый месяц, их общее влияние невелико. Однако, если общий объем таких новостей высок в средствах массовой информации с высокой публикационной активностью, общее влияние также должно учитывать концентрацию таких новостей, которая может быть низкой. Чтобы объединить эти два элемента, была предложена следующая формула (3.2.4):

$$I_s = \frac{\frac{\sum e_d}{\max(\sum^s e_d)} + \frac{\sum e_d / N_s}{\max(\sum^s e_d / N_s)}}{2}, \quad (3.2.4)$$

где  $I_s$  – общее влияние источника  $s$ ,

$\sum e_d$  – сумма всех оценок документа,

$\max(\sum^s e_d)$  – максимальная сумма всех оценок документа среди всех других источников,

$N_s$  – количество документов внутри источника. Следовательно, эта формула (3.2.4) ограничена сверху значением 1 и ее значения неотрицательны. Это относительное значение, и его следует использовать в случаях, когда имеется

значительное количество источников, чтобы ранжировать их как по объему, так и по концентрации документов с учетом некоторых заданных критериев. Реализация данного метода (Приложение Б).

### **Выводы по 3-му разделу**

Таким образом, в разделе рассмотрены методы мультикритериальной оценки, в частности методы MCDM, проведен литературный обзор. Выявлены основные ограничения и недостатки существующих моделей. Предложено использовать разработанный и опубликованный ранее метод агрегации на базе байесовского определения субъективных вероятностей. Также предлагается использовать для сравнительной оценки важности показателей и информативных признаков метод анализа иерархий (АНР). Впервые подобный подход был описан и применен для решения задачи поддержки принятия решений в области установки генераторов возобновляемой энергии в рамках геоинформационной системы [54, p. 122277].

Формулу Байеса предлагается использовать последовательно для каждого известного/достоверного информативного признака. В соответствии с формулой Байеса мы можем выразить вероятность события  $h$  при условии наступления события  $e$  в виде формулы (3.1.2).

$$p(h|e) = \frac{p(e|h) \times p(h)}{p(e|h) \times p(h) + p(e|\sim h) \times p(\sim h)} \quad (3.1.2)$$

где  $p(e|h)$  – условная вероятность наступления события  $e$  при справедливости  $h$ ;

$p(h)$  – априорная вероятность гипотезы  $h$ ;

$p(e|\sim h)$  – условная вероятность  $e$  при отсутствии  $h$ ;

$p(\sim h)$  – вероятность того, что событие  $h$  не верно, которое, в соответствии с формулой полной вероятности, может быть вычислено по формуле (3.1.3).

Когда сведений о параметрах текста мало, модели, основанные на многофакторном анализе, выдают единственную величину, которой может быть недостаточно для оценки его принадлежности тому или иному классу. В то же время, байесовская система, учитывающая степень уверенности в наличии параметров, дает две оценки («за» и «против»), которые позволяют сделать более обоснованный выбор итогового результата.

Следует отметить, что обе модели в общем случае не являются самообучаемыми и требуют привлечения экспертов для получения оценок влияния факторов, либо наличия процедур для автоматического получения оценок (например метод, предложенные в разделе 2.3 работы).

На базе описанного метода байесовской агрегации, метода анализа иерархий и тематического моделирование предлагается метод мультикритериальной оценки масс-медиа ММА. Метод состоит из следующих стадий:

1. Постановка задачи – выбор финального класса (классов) для оценки.
2. Составление списка свойств, по которым можно определить, принадлежит ли текст к одному или нескольким из классов.
3. Оценка сравнительной значимости объектов на основе АНР.
4. Применение процедуры расчета оценок каждого из свойств для каждого текста на основе тематической модели корпуса.
5. Агрегирование оценок свойств и их сравнительной значимости для принятия решения о принадлежности к классу (классам), к которым может быть отнесен текст (Этап 1).

6. Оценка СМИ на основе полученных классификаций текстов (Этап 2).

Метод позволяет проводить мультикритериальную мультимодальную (темы, признаки, классы) оценку как отдельных публикаций, так и отдельных источников СМИ (или авторов/аккаунтов в социальных сетях).

Метод объединяет разработанные методы, описанные, в частности, в разделе 2 настоящей работы (тематическая векторизация, мера межкорпусного тематического дисбаланса), а также в работах [54, р. 122277; 77].

## **4 РАЗРАБОТКА ИНФОРМАЦИОННОЙ СИСТЕМЫ МОНИТОРИНГА МЕДИА-ПРОСТРАНСТВА КАЗАХСТАНА НА БАЗЕ МЕТОДОВ NLP**

### **4.1 Описание программной архитектуры разработанной системы**

Основная цель системы – предоставить экспертам, исследователям и руководителям полный набор аналитических инструментов для предоставления оперативных, актуальных отчетов, визуализаций и рейтингов публикаций в государственных СМИ в той или иной интересующей их области. Рассмотрим несколько сценариев использования:

– социологические и другие гуманитарные исследования - систему можно использовать для получения обобщенных данных по любой заданной теме/событию/объекту для анализа, а также количественных показателей для поддержки своих исследований;

– система может быть использована как KPI-система управления для PR-отдела. При отсутствии количественных показателей, подтверждающих выводы, оценить работу PR-отдела может быть затруднительно, в то время как система может предоставить инструменты для оценки KPI на нескольких уровнях, включая простую оценку объема освещения в СМИ, оценку настроений при освещении, а также позволяет проводить углубленный статистический анализ;

– извлечение наиболее социально значимых тем и направлений для различных демографических групп может стать мощным инструментом для местных администраций и государственных органов, поскольку разработанная система способна не только обеспечить простую фильтрацию и полнотекстовый поиск, но и извлечь скрытые латентные структуры корпуса с целью выявления тенденций и событий, которые могут остаться вне сферы внимания PR служб организаций;

– регулярная отчетность – система может быть использована для регулярного извлечения отчетов об актуальном состоянии по тому или иному аналитическому запросу с целью получения как обобщенной аналитики и визуализации, так и актуального списка наиболее резонансных опубликованных документов, или наиболее негативных документов и т.д.

Одна из трудноразрешимых проблем в области интеллектуального анализа данных - разработка универсального инструментария для анализа текста. Популярным направлением в разработке алгоритмов обработки корпусов текстовых документов является использование методов машинного обучения, позволяющих решать задачи NLP (Natural Language Processing).

Как было описано в [78], при проектировании системы учитывались требования к высокопроизводительным системам: модульность, возможность как горизонтального, так и вертикального масштабирования, условная независимость компонентов. Разработанная система представляет собой набор компонентов, каждый из которых сформирован и используется в виде контейнеров Docker. Основные уровни системы: уровень обработки (data processing/ETL level), хранение данных (data storage level), визуализация и

управление обработкой данных (visualization and control level/web-interface). На уровне обработки данных текстовые документы собираются (скраппинг) из медиапространства и социальных сетей и далее обрабатываются с помощью каскада различных предобработчиков и моделей, каждый из которых реализуется в системе как отдельная задача Airflow. Полученные результаты хранятся в реляционной базе данных (PostgreSQL); для увеличения производительности поиска данных (в базе порядка миллионов записей) используется инструмент Elasticsearch. Статистика, полученная в результате работы алгоритмов, визуализируется с помощью библиотеки Plotly. Администрирование и просмотр обработанных данных доступны через веб-интерфейс с использованием фреймворка Django. Общая схема взаимодействия компонентов в процессе обработки данных организована по принципу ETL (извлечение, преобразование, загрузка).

Как описано в разделах 2 и 3 данной работы, предложенный метод интеллектуального анализа данных позволяют обрабатывать большие текстовые документы (объемом 1 млн документов), используя особенности как тех или иных отдельных документов, входящих в корпус, так и общих закономерностей, характеризующих их совокупность. Поскольку задачи предполагают извлечение самых разнообразных характеристик из текстов (тональность, социальная значимость, отношение к теме, наличие именованной сущности и т.п.), что часто подразумевает использование сложных и далеко не всегда высокоскоростных алгоритмов, возникает необходимость хранить рассчитанные характеристики (вместе с самими документами) в гибридной высокоэффективной системе хранения разработанной системы.

В настоящее время область разработки программных систем для обработки текстовой информации является активно развивающейся отраслью информационных технологий. Обзор работ в этой области доступен, например, в [79-82]. Отметим, что в последнее десятилетие успешным направлением алгоритмов обработки корпуса текстовых документов является использование для этих целей методов машинного обучения [83-85].

Процесс обработки текстов на естественном языке сводится к следующим последовательным шагам:

- инициализация – формирование корпуса текстов и его предобработка для последующего анализа;
- семантический анализ – определение смысловых конструкций с учетом синонимии и связывание именованных сущностей (англ. Named Entity Linking, NEL); анализ научных текстов обычно ограничивается этим уровнем;
- прагматический анализ – определение жанровых и стилевых функций текстов; конструкции, определяющие деструктивное влияние новостных сообщений и т. п.
- синтез проведенных исследований – определение взаимосвязи между нижними уровнями и высшими, а также агрегирование результатов в форме, удобной для понимания и поиска.

Исходя из вышесказанного, сформулируем требования к функционалу системы, исходя из ее целевого назначения: скрапинг, хранение, потоковая аналитика и формирование аналитических отчетов с визуализацией.

1. Надёжное хранение корпусов текстов больших объёмов.

2. Быстрый параллельный доступ, фильтрация и агрегация данных с целью потоковой обработки: предобработка, построение тематических моделей, классификаторов, агрегация и выгрузка для отчётов в реальном времени и т.п.

3. Гибкость и возможность хранения неструктурированных и частично структурированных данных для поддержки возможности хранения и доступа к произвольным структурам данных с целью анализа и различных вычислительных экспериментов на основе современных методов анализа текста.

Структура программного комплекса позволяет решать масштабные задачи, заключающиеся в хранении корпусов из нескольких миллионов текстов и пакетной обработке тысяч документов в режиме онлайн.

Концептуальный дизайн включает требования, необходимые для обеспечения функционала системы. Созданный программный комплекс обладает следующими возможностями:

1. Обеспечение доступа к корпусам текстов.

2. Автоматизированная обработка корпуса текстов, хранящихся в БД.

3. Занесение полученных характеристик в хранилище.

4. Возможность гибкого планирования и мониторинга выполнения различных задач по обработке данных, а также оперативной обработке непредвиденных ситуаций.

5. Статистическая обработка полученных характеристик и их представление в удобном для исследователя виде.

6. Обновление и улучшение применяемых алгоритмов для анализа корпуса текстов.

Проектируемая система состоит из следующих подсистем:

1. Подсистема обработки данных.

2. Хранилища данных.

3. Подсистема доступа к данным (визуализация, отчеты и пр.).

Информационная система должна учитывать этапы анализа текста. В состав системы входят компоненты, перечисленные в описании постановки задачи. На этапе предварительной обработки текст подготавливается для дальнейшего анализа. Используемые методы предварительной обработки будут зависеть от алгоритма работы с данными, можно выделить следующие типы:

- основанные на bag of words; к этому виду можно также отнести метод TF-IDF;

- возвращающие в результате обработки каждой смысловой единицы корпуса (например, новости) его embedding (векторизацию), например, распределение по токенам/словам/фразам/предложениям; в этом случае возможно использование рекуррентных нейронных сетей (RNN);



– возвращающие в результате обработки каждой смысловой единицы корпуса один `text embedding`; для такой предобработки возможно использование стандартных методов классификации, современных `transformer` моделей, либо метод, предложенный в разделе 2.2 данной работы.

Семантический анализ может выполняться как на этапе предобработки текста, например, лемматизация слов, так и не выполняться вообще – выбранный инструментарий будет зависеть от методов машинного обучения и может изменяться с течением времени. Прагматический анализ в системе осуществляется с использованием совокупности алгоритмов машинного обучения и, в конечном итоге, составленных частотных словарей (в том числе опосредованно, используя тематическое моделирование). Синтез результатов обеспечивается путем агрегирования результатов в хранилище и вывод этих результатов в нужном виде.

На основе вышеописанных возможностей системы можно выделить требования для разрабатываемой системы:

– обеспечить работу подсистем в виде отдельных независимых компонентов, каждый из которых можно оперативно заменить при необходимости;

– организовать распараллеливание вычислений, в том числе на нескольких машинах;

– реализовать автоматизированную обработку корпуса текстов по запросу пользователя;

– вести мониторинг выполнения задач в реальном времени, в том числе оперативное информирование об исключениях;

– выводить данные по результатам анализа текстов в интерфейсе пользователя;

– обновлять применяемые в системе алгоритмы для улучшения качества анализа и расширения их области применения.

Все компоненты системы организованы в виде `Docker`-контейнеров. Все контейнеры имеют доступ к одной виртуальной сети, что обеспечивает возможность обмена данными с использованием стандартных сетевых протоколов (`TCP/IP`). Такая реализация обеспечивает работу подсистем в виде независимых компонентов, каждый из которых возможно заменить при необходимости, а также переместить на отдельную машину для реализации горизонтального масштабирования.

Взаимодействие компонентов между подсистемами осуществляется посредством системы хранения. Общая схема взаимодействия компонентов организована по принципу `ETL` (`extract, transform, load`). Когда от пользователя поступает запрос на получение данных, он перенаправляется в `ElasticSearch` для расчетов (если данные используются редко) или в `Redis` (если данные используются часто и уже были запрошены за определенный период). Кроме того, подсистема обработки использует `Airflow-scheduler`, который записывает в `Redis` информацию о распределении задач по вычислительным модулям; они, в свою очередь, отчитываются в `Redis` о статусе выполнения своих задач. В

процессе развития системы могут применяться компоненты согласно их целевому назначению. Визуализация структуры системы показана на рисунке 8.

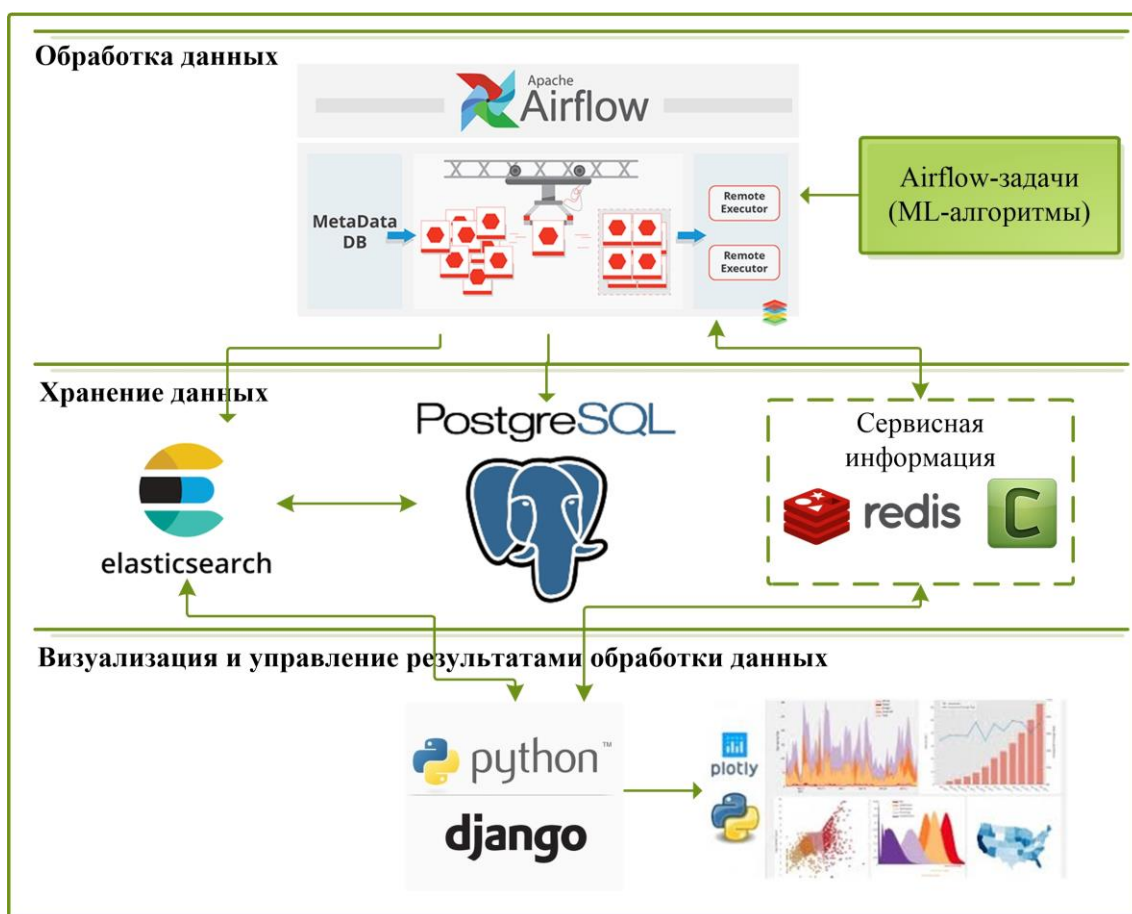


Рисунок 8 – Структура системы

Анализ текстовых корпусов (порядка миллионов новостных публикаций из электронных СМИ и социальных сетей Казахстана) осуществляется с помощью вычислительных модулей в рамках Airflow. Новые документы загружаются в подсистему обработки данных с помощью парсеров СМИ и социальных сетей. С заданной периодичностью выполняется генерация отчетов, требующих большого времени для вычисления; результаты размещаются в хранилище – такой подход уменьшает время ожидания результатов от подсистемы обработки данных. На основе собранных данных выполняется дополнительное обучение модели (1-2 раза в месяц), которое включает в себя пересчет набора ключевых характеристик текстового корпуса. В случае, если дообучение модели будет приводить к значительному уменьшению показателя точности (например, в задаче определения тональности текста), предусмотрено использование других ML-алгоритмов или их совокупности, а также откат к старым моделям.

Ролевая система включает в себя следующие роли:

1. Обычный пользователь – имеет доступ к базовой функциональности системы по доступу к данным: поиск, фильтрация, цифровые информационные панели (так называемые дашборды).

2. Расширенный пользователь – имеет доступ к настраиваемым отчётам, автоматическим оповещениям о «горячих темах», возможность проводить фильтрацию по именованным сущностям (например, человек, организация, регион) в статьях. Такое разделение пользователей обусловлено последующим использованием системы государственными органами.

3. Разработчик – имеет доступ к панели администратора Airflow и к репозиторию, в котором хранятся Airflow DAG. Может добавлять и менять свои задачи, запускать и отслеживать их выполнение.

4. Администратор – супер-пользователь, имеет полный набор прав по работе с системой.

Распределение доступной функциональности по ролям представлена на рисунке 9.

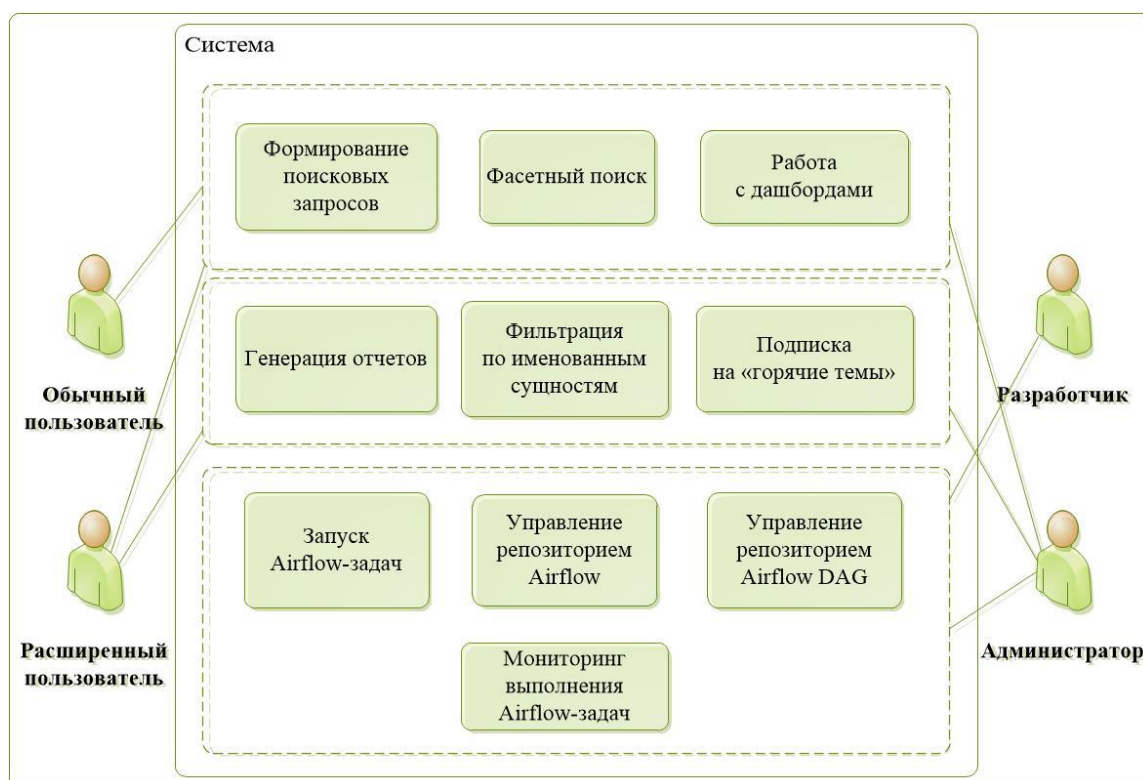


Рисунок 9 – Распределение доступа к системе по ролям

В ходе анализа для удовлетворения всех этих потребностей была выбрана программная платформа с открытым исходным кодом Apache Airflow [86]. Основные компоненты данной платформы:

1. Airflow-worker – основной компонент, выполняющий обработку данных. Может быть горизонтально масштабирован, в том числе на отдельные сервера/облачные VM. В текущем варианте архитектуры в образ контейнера Airflow-worker заранее встраиваются необходимые зависимости, однако принципиально процесс инъекции зависимостей может происходить различным образом, в том числе, путём динамического получения Docker-контейнеров из публичных, либо частных репозиториев.

2. Airflow-scheduler – компонент, отвечающий за назначение задач Airflow-workerам в порядке, и с учетом правил, ограничений и приоритетов, определённых Airflow DAG-ами. Airflow DAG – заданный программно нециклический направленный граф, описывающий порядок выполнения определённых задач, а также содержащий информацию о расписании, приоритетах, поведении в случае исключений и пр.

3. Airflow web server – веб-интерфейс, позволяющий отслеживать и контролировать ход выполнения задач.

Алгоритмы машинного обучения, а также скрапперы, различные виды предобработки и другие служебные задачи реализованы в системе как отдельные Airflow-задачи.

В системе предусмотрено три вида хранилищ:

1. PostgreSQL – выполняет роль персистентного хранилища для структурированных данных. Его использование обусловлено широкими возможностями данной реляционной БД (среди свободно распространяемых продуктов), взаимодействия с широким кругом инструментов. Основные типы данных, хранящиеся в этой базе:

- а) новости и метаданные;
- б) обработанные данные на уровне разных базовых единиц анализа (токен/слово/фраза/предложение/текст), в том числе векторизации, результаты лемматизации, очистки и пр.;
- в) результаты проведения тематического моделирования;
- г) результаты классификации новостей по различным признакам (тональность, политизированность, социальная значимость и пр.).

2. Elasticsearch – in-memory NoSQL хранилище, предназначенное для хранения неструктурированных или слабоструктурированных данных, а также быстрого поиска (в том числе полнотекстового), фильтрации и потокового доступа. В сравнении с другими NoSQL базами для хранения документов с произвольной структурой, такими как MongoDB и CouchDB, Elasticsearch выделяется расширенными инструментами для индексирования текста, позволяющими проводить полнотекстовый поиск по большим объёмам документов практически в реальном времени. Также в виду возможности построения продвинутых индексов для данных, возможно выполнение сложных агрегаций в самой базе, в том числе распределенно. Elasticsearch выполняет несколько функций:

- а) основное хранилище для доступа, поиска и фильтрации данных конечным пользователем;
- б) основное хранилище для ETL (Extract-Transform-Load) процессов обработки данных – в том числе запись любых промежуточных результатов в свободной форме;
- в) хранилище для кэширования определённых результатов вычислений, необходимых для построения дашбордов и отчётов в системе;

г) ElasticSearch дублирует данные, хранящиеся в PostgreSQL как персистентном хранилище, поскольку ElasticSearch является in-memory базой данных без гарантий относительно персистентности и целостности данных.

3. Redis – быстрое key-value хранилище, используемое для кэширования отдельных страниц и элементов, а также для кэширования сессий авторизации. В Redis хранятся служебные данные, а также кэш страниц и элементов, к которым происходит частый доступ.

Все три основных хранилища системы могут быть легко масштабированы на несколько отдельных компьютеров, поддерживается как горизонтальное масштабирование, так и репликация, при этом ElasticSearch и Redis показывают близкое к линейному увеличению производительности при горизонтальном масштабировании.

Для хранения служебных данных, таких как состояния выполнения задач, используется отдельный кластер PostgreSQL. Для запуска и отслеживания прогресса задач Apache Airflow используется связка Celery+Redis.

Интерфейс подсистемы отображения представляет из себя HTML+CSS+JS веб-сайт с доступом по протоколу HTTP. HTML+CSS+JS – выбор этого стека технологий для интерфейса оправдан тем фактом, что именно веб-интерфейсы являются наиболее распространённой и повсеместно поддерживаемой технологией построения UI с возможностью доступа с любых устройств, операционных систем из любой точки мира, при условии наличия веб-браузера и подключения к сети Интернет.

Веб-приложение реализовано на Python фреймворке Django, в качестве веб-сервера выступает Gunicorn, реверс-прокси Nginx. Веб-приложение имеет доступ как к персистентному хранилищу PostgreSQL, так и к ElasticSearch. В Django есть встроенный Cache Framework, который позволяет кэшировать страницы и элементы страниц в Redis. Например, если предполагается, что на страницу будут заходить часто, а считается она долго, то такую страницу целесообразно кэшировать в Redis, что позволит ускорить доступ к необходимым данным.

Фреймворк Django был выбран из следующих соображений:

1. Возможность быстрой Agile-разработки веб-интерфейса и модели хранения данных. Скорость разработки на фреймворке Django значительно выше, чем при использовании таких аналогов, как Spring (Java), Yii (PHP) и Node.js (JavaScript).

2. В виду того, что проект предполагает проведение анализа данных и построение моделей машинного обучения, в том числе NLP, язык Python является оптимальным выбором, так как большая часть State-of-the-art моделей и методов ML/AI и NLP разрабатывается сообществом именно на языке Python.

3. Django ORM лучше работает с БД PostgreSQL.

Веб-приложение реализует ряд страниц для фильтрации, поиска, доступа к различным дашбордам и отчётам. На текущий момент в большинстве модулей расчетов, связанных с отображением, используется фасетный поиск

(Faceted Search) из Elastic Search. Для отрисовки графиков и прочей визуализации используется Python-библиотека визуализации данных Plotly.

Пример того, какую информацию могут отображать графики:

1. Динамику по рассчитанным характеристикам (тональность, социальная значимость и пр.), тематикам, количеству просмотров/комментариев с фильтрацией по СМИ, тематикам, авторам, тегам + полнотекстовый поиск.

2. Распределение тематик, значений тональности и пр. в статике, с фильтрациями и поиском.

3. Выявление выбросов/аномалий для аналитических отчётов (самые «горячие» темы и пр.) [78, с. 8]

Таким образом, в разделе описана архитектура информационной системы, разработанной в ходе работы. Архитектура построена на современных Open Source решениях и соответствует требованиям масштабируемости, гибкости, надежности и производительности, в том числе за счет использования гибридного (SQL+noSQL) хранилища и ETL инструмента Apache Airflow, обеспечивающего возможность масштабирования вычислительных модулей независимо от распределения подсистемы хранения.

#### **4.2 Основной функционал разработанной системы**

Конкурирующими компаниями с похожим направлением развития являются Integrum.ru, Медиалогия, Brand Analytics. Рассмотрим основные конкурентные преимущества:

Использование машинного обучения без учителя:

– позволяет проводить поиск тем и инфоповодов без заранее заданных ключевых слов и запросов;

– позволяет выявить скрытые закономерности и тренды без вмешательства эксперта.

Малый объём экспертной разметки – на порядки меньший объём разметки больших корпусов, по сравнению с конкурентными системами:

– возможность тонкой настройки (fine tuning) моделей оценки за счёт дополнительной разметки публикации.

Возможность разметки по произвольным критериям, включая не только тональность, но и социальную значимость, резонансность, манипулятивность и т.д.:

– возможность формирования аналитических запросов как по ключевым словам, отношению к отдельным топиков, так и по значениями произвольных критериев.

Анализ информационного поля с точки зрения обработки временных рядов:

– возможность оценки поведения динамики топиков (инфоповод/вброс/информационная атака/фоновый топик/сезонный топик);

– возможность разработки прогнозирующих моделей;

– возможность учёта внешних, в том числе экономических, факторов.

Специализация под казахстанский сегмент:

- обработка казахского языка;
- тонкая настройка парсеров под широкий круг отечественных СМИ.

А также:

- решение вопросов безопасности передачи данных и задач аналитики для высоких уровней государственного управления зарубежным компаниям и облачным сервисам;

- возможность Ad-hoc разработки модулей обработки, визуализации и построения отчётов для отдельных государственных органов и департаментов.

Представленная система разработана с учётом современных подходов к информационной безопасности. В частности, можно выделить следующие основные элементы информационной безопасности системы:

1. Безопасность доступа к физическому серверу и ОС:

- а) доступ к физическому серверу осуществляется только из внутренней служебной сети (авторизация к Wi-Fi через логин пароль + регистрация MAC адреса), либо через защищённый SSL VPN;

- б) доступ происходит по протоколу SSH Protocol 2 по RSA ключу, доступ на нестандартном порту, пользователь root доступен только при физическом доступе;

- в) все пароли и данные для авторизации (credentials) хранятся в зашифрованном виде только на боевой машине через функционал docker secret.

2. Безопасность доступа к системе:

- а) трафик к web интерфейсу и API шифруется SSL;

- б) для доступа к web интерфейсу используется стандартная схема Session+Cookie (модуль Django Auth);

- в) для доступа к Rest API используются JWT токены с коротким временем истечения (expiration) (модуль Django Rest Framework).

3. Разграничение прав внутри системы:

Реализованы 4 основные роли – Администратор (суперпользователь), Эксперт, Роль для просмотра и загрузчик контента. Предусмотрена возможность расширения списка ролей и разработка иерархии ролей (модуль Django Permissions)

Предусмотрена возможность раздачи прав и разрешений на отдельные элементы данных, такие как: корпуса, тематические модели, топики и критерии (модуль Django Permissions).

Для доступа к данным, хранящихся в разработанной информационной системе в целях проведения исследований, аналитики, а также поддержки принятия решений, реализовано два основных вида инструментов:

1. Конфигурируемые дашборды.

2. Инструменты для аналитических запросов.

Конфигурируемые дашборды. Была разработана подсистема, позволяющая в реальном времени проводить конфигурацию отображений различных элементов визуализации и представления данных (виджетов). Преимущество такого подхода заключается с одной стороны в высокой гибкости интерфейса системы, с другой не требует дополнительной разработки

и обновления системы при изменениях. В частности, на текущий момент в системе настроен ряд дашбордов, используемых в тестовых, исследовательских и аналитических целях, в том числе для отслеживания показателей, связанных с деятельностью Министерства Науки и Образования Республики Казахстан. Перечень дашбордов на текущий момент:

1. Основной дашборд с обобщённой информацией по всему медиапространству Казахстана, включая тональность, основные темы, распределение позитива-негатива по СМИ и социальным сетям.
2. Дашборд по публикациям, связанным с образованием с аналогичным функционалом.
3. Дашборд с оперативной информации по ситуации с освещением COVID-19.
4. Рейтинги основных руководителей МОН РК по освещённости и тональности.
5. Дашборд социально-значимых тем и публикаций по данным опросов.

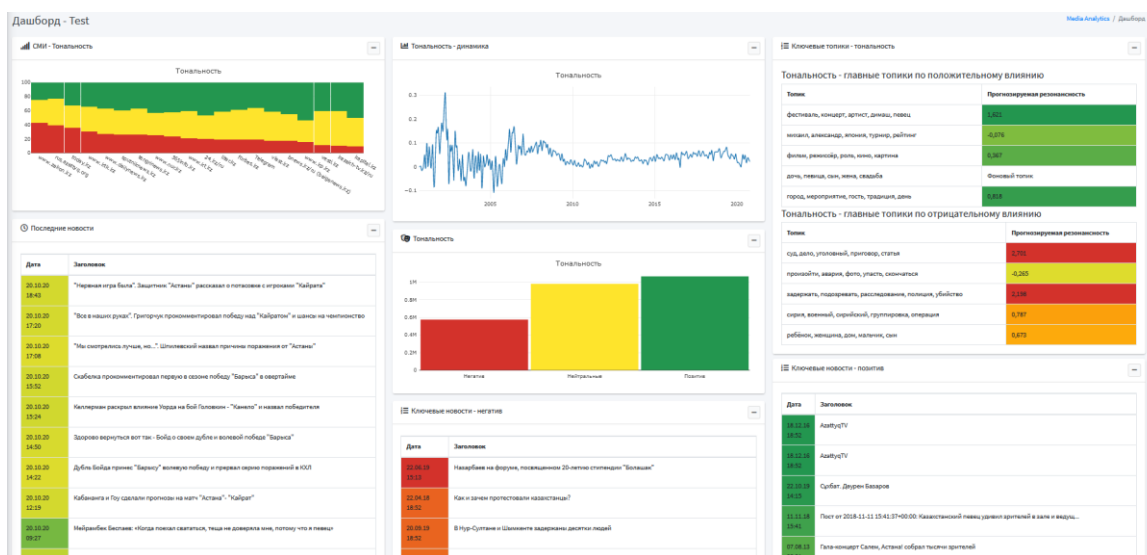


Рисунок 10 – Пример дашборда

Рисунок 10 показывает пример дашборда, реализованного в системе. Имеется возможность в течении нескольких минут настроить дашборд для мониторинга по заданным именованным сущностям, включая организации, персон, локации, темы и инфоповоды с требуемыми модулями визуализации (виджетами). В том числе в системе на текущий момент реализовано 10 видов виджетов:

1. Гистограмма объёма публикаций с низким, средним и высоким значением оценки по критерию (например, по тональности).
2. Динамика изменения критерия за выбранный период.
3. Список новостей с самым высоким показателем по критерию за период.
4. Список новостей с самым низким показателем по критерию за период.
5. Список последних новостей.



6. Список ключевых топиков по выбранному критерию. Например, список топиков, в которых концентрируется большая часть негатива за период.

7. Распределение освещённости и значений критериев по источникам.

8. Карта распределение освещённости выбранной темы/сущности по регионам и населённым пунктам Казахстана.

9. Карта распределение выбранного критерия по регионам и населённым пунктам Казахстана (например карта негативных публикаций).

10. Сравнение объектов мониторинга (например организация и персон) по освещённости и среднему значению выбранного критерия.

Инструмент аналитических запросов.

Несмотря на то, что дашборды позволяют настраивать мониторинг и визуализацию без дополнительных разработок, процесс настройки всё же занимает определённое время, в то время как работа аналитика зачастую требует более оперативного доступа к необходимым данным и проверки гипотез. Для таких целей в системе реализован ряд аналитических инструментов, позволяющих совершать такого рода запросы, в частности:

1. Полнотекстовый поиск с фильтрацией.

2. Аналитика по тематическим моделям.

3. Комплексный инструмент аналитический запросов.

Остановимся подробно на комплексном инструменте аналитических запросов.

Рисунок 11 – Интерфейс формирования аналитических запросов

Рисунок 11 показывает интерфейс формирования аналитических запросов. Рассмотрим какого рода фильтрации и запросы можно совершать с помощью данного инструмента:

1. Выбор конкретной тематической модели и критериев. При этом нужно отметить, что тематическая модель может быть построена с учётом ряда фильтров, в частности по дате, источникам, корпусам и темам/запросам.

2. Фильтрация по ключевым словам позволяет быстро фильтровать нужные публикации с помощью полнотекстового поиска, реализованного на поисковом движку Elasticsearch.

3. Фильтрация по группам топиков – множествам топиков, вручную отобранном экспертом, относящимся к близким тематикам.

4. Фильтрация по источникам – СМИ и соц. сети, с возможностью выбирать несколько источников сразу.

5. Фильтрация по значению критерия. Например, можно отфильтровать только публикации, имеющие значения критерия Тональность ниже 0.3 (негатив), или, например имеющий значения критерия Социальная значимость выше 0.8 (высокая социальная значимость).

6. Настройки отображения графиков, включая выбор гранулярности и включение/выключение сглаживания.

Таким образом, представленный инструмент позволяет проводить как исследовательскую, так и аналитическую работу, включая автоматическую выгрузку отчётов в формате PDF. Данный модуль является важнейшей частью разработанной системы, поскольку позволяет получать доступ к данным и результатам работы системы, в виде подходящем для дальнейших исследований.

Для организации процесса сбора и валидации данных необходимых для расчета отдельных информативных критериев в рамках разработанной информационной системы были разработаны парсеры для открытых информационных источников и общедоступных публикаций/постов в социальных сетях (YouTube, Telegram, Instagram, Facebook, Twitter, Вконтакте) и других источников. Составлен список ключевых источников, формирующих общественное мнение в социальных сетях и список ключевых слов для парсинга контента. А так же разработаны парсеры, позволяющие получать актуальную информацию по показателям резонансности публикаций (количество просмотров, комментариев, репостов и пр.). Отлажено регулярное обновление всех необходимых показателей внутри хранилищ системы, для обеспечения получения актуальных данных, по оценке информативных критериев.

### **Выводы по 4-му разделу**

В разделе описана архитектура и основной функционал информационной системы мониторинга медиапространства, на базе которой были реализованы описанные в диссертации предложенные модели и методы. Поскольку предложенные модели и методы непосредственно связаны с вычислительными технологиями и обработкой данных, такая система является абсолютно необходимым элементом для проверки, доработки и апробации предложенных моделей. Также данная система обладает высоким потенциалом коммерциализации, в частности была внедрена в Министерстве Образования и Науки Республики Казахстан в качестве инструмента для отдела мониторинга социальных сетей и СМИ. Информационная система позволила значительно

сократить объем работы, связанной с мониторингом – как за счет автоматического парсинга и сбора новостей из разных источников, так и за счет интеллектуальной системы фильтрации и оценки новостей по ряду критериев (мультикритериальная поддержка принятия решений, MCDM).

При разработке системы были учтены следующие требования:

- высокая производительность и масштабируемость, возможность параллельных вычислений;
- надежность хранения информации при достаточно высокой скорости доступа (включая полнотекстовый поиск и агрегацию);
- изолированность отдельных модулей системы для обеспечения безопасности, ролевая система доступа.

Исходя из этих требований предложена архитектура, состоящая из следующих основных модулей:

- подсистема обработки данных;
- хранилища данных;
- подсистема доступа к данным (визуализация, отчеты и пр.).

Подсистема обработки данных построена на базе Apache Airflow и состоит из следующих основных модулей:

1. Airflow-worker – основной компонент, выполняющий обработку данных. В текущем варианте архитектуры в образ контейнера Airflow-worker заранее встраиваются необходимые зависимости.

2. Airflow-scheduler – компонент, отвечающий за назначение задач Airflow-workerам в порядке, и с учетом правил, ограничений и приоритетов, определённых Airflow DAG-ами. Airflow DAG – заданный программно нециклический направленный граф, описывающий порядок выполнения определённых задач, а также содержащий информацию о расписании, приоритетах, поведении в случае исключений и пр.

3. Airflow web server – веб-интерфейс, позволяющий отслеживать и контролировать ход выполнения задач.

Хранилище состоит из модулей трех типов (технологий):

1. PostgreSQL – выполняет роль персистентного хранилища для структурированных данных. Основные типы данных, хранящиеся в этой базе:

- новости и метаданные;
- обработанные данные на уровне разных базовых единиц анализа (токен/слово/фраза/предложение/текст), в том числе векторизации, результаты лемматизации, очистки и пр.;
- результаты проведения тематического моделирования;
- результаты классификации новостей по различным признакам (тональность, политизированность, социальная значимость и пр.).

2. Elasticsearch – in-memory NoSQL хранилище, предназначенное для хранения неструктурированных или слабоструктурированных данных, а также быстрого поиска (в том числе полнотекстового), фильтрации и потокового доступа. Elasticsearch выполняет несколько функций:

– основное хранилище для доступа, поиска и фильтрации данных конечным пользователем;

– основное хранилище для ETL (Extract-Transform-Load) процессов обработки данных – в том числе запись любых промежуточных результатов в свободной форме;

– хранилище для кэширования определённых результатов вычислений, необходимых для построения дашбордов и отчётов в системе;

– Elasticsearch дублирует данные, хранящиеся в PostgreSQL как персистентном хранилище, поскольку Elasticsearch является in-memory базой данных без гарантий относительно персистентности и целостности данных;

– Redis – быстрое key-value хранилище, используемое для кэширования отдельных страниц и элементов, а также для кэширования сессий авторизации. В Redis хранятся служебные данные, а также кэш страниц и элементов, к которым происходит частый доступ.

Касательно функционала, в системе представлено два основных вида модулей для доступа:

1. Дашборды – настраиваемые статичные (с минимальными возможностями интерактивности) страницы для отображения информации по определенной тематике.

2. Аналитический инструмент, позволяющий задавать сложные семантические запросы по целому ряду параметров – более сложный, но более гибкий инструмент, по сравнению с дашбордами.

В данных инструментах реализовано 10 основных видов визуализации:

1. Гистограмма объёма публикаций с низким, средним и высоким значением оценки по критерию (например, по тональности).

2. Динамика изменения критерия за выбранный период.

3. Список новостей с самым высоким показателем по критерию за период.

4. Список новостей с самым низким показателем по критерию за период.

5. Список последних новостей.

6. Список ключевых топиков по выбранному критерию. Например, список топиков, в которых концентрируется большая часть негатива за период.

7. Распределение освещённости и значений критериев по источникам.

8. Карта распределение освещённости выбранной темы/сущности по регионам и населённым пунктам Казахстана.

9. Карта распределение выбранного критерия по регионам и населённым пунктам Казахстана (например карта негативных публикаций).

10. Сравнение объектов мониторинга (например организация и персон) по освещённости и среднему значению выбранного критерия.

## 5 ДАННЫЕ, ЭКСПЕРИМЕНТЫ И ВАЛИДАЦИЯ РЕЗУЛЬТАТОВ РАЗРАБОТАННЫХ МЕТОДИК И СИСТЕМЫ

### 5.1 Корпус новостных публикаций

Предлагаемые в работе методы предполагают наличие корпуса публикаций большого объема, поскольку можно сказать, что разработанные модели находятся в контексте понятия Big Data. Следовательно, одной из инженерных задач при выполнении данной работы является сбор датасета (корпуса), соответствующего требованиям объема, репрезентативности, точности и разнообразия данных.

Собранный датасет содержит новости из российских и казахстанских источников новостей с 2000 по 2020 год из 50 основных источников, включая социальные сети и новостные сайты. Он включает 4233990 документов из казахстанских источников и 2027963 документа из российских источников. Также в системе имеется отдельный датасет фрагментов текстов государственных программ развития, состоящий из примерно 4000 фрагментов. Датасет был опубликован в научном репозитории Data Mendeley [87]. Документы могут содержать следующие поля:

1. Заголовок.
2. Источник.
3. URL.
4. Текст публикации.
5. Дата публикации.
6. Автор.
7. Количество просмотров, лайков, шейров, комментариев.
8. Категория, теги.

Также следует отметить, что датасет не проверялся и не редактировался вручную. Следовательно, из-за технических трудностей и ограничений в процессе скрапинга, набор данных может содержать:

- куски кода HTML и JavaScript;
- неверные дата и время публикации из-за проблем с форматом;
- неверный URL;
- в редких случаях текстовое поле может содержать текст из разных публикаций.

Однако такие случаи нечасты и, согласно проверке на небольшом подмножестве данных, встречаются только менее чем в 5% публикаций.

Таблица 2 – Топ 20 казахстанских новостных источников по количеству публикаций

Источник	Количество публикаций
1	2
<a href="https://inbusiness.kz/ru">https://inbusiness.kz/ru</a>	741542
<a href="https://www.nur.kz/">https://www.nur.kz/</a>	514134
<a href="https://tengrinews.kz/">https://tengrinews.kz/</a>	427687

Продолжение таблицы 2

1	2
<a href="https://24.kz/ru/">https://24.kz/ru/</a>	270809
<a href="https://forbes.kz/">https://forbes.kz/</a>	252699
<a href="https://kapital.kz/">https://kapital.kz/</a>	239588
<a href="http://vesti.kz/">http://vesti.kz/</a>	213740
<a href="https://rus.azattyq.org/">https://rus.azattyq.org/</a>	145517
<a href="https://365info.kz/">https://365info.kz/</a>	135889
<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	131350
VK	130860
<a href="http://www.newsfactory.kz/">http://www.newsfactory.kz/</a>	125294
<a href="http://today.kz">http://today.kz</a>	123892
<a href="https://www.ktk.kz/">https://www.ktk.kz/</a>	119820
<a href="https://www.kt.kz/">https://www.kt.kz/</a>	107171
<a href="http://www.kp.kz/">http://www.kp.kz/</a>	89203
<a href="https://kazakh-tv.kz/ru">https://kazakh-tv.kz/ru</a>	78674
<a href="https://liter.kz/">https://liter.kz/</a>	73749
Telegram	71340
<a href="http://www.dailynews.kz/">http://www.dailynews.kz/</a>	70564

Таблица 3 – Топ 10 российских новостных источников по количеству публикаций

Источник	Количество публикаций
Lenta.ru	800970
Интерфакс	381659
РБК	275474
Лента	112139
РТ	111204
Ведомости	104059
Бизнес ФМ	66224
Радио Свобода	57059
Sputnik (Ru)	54264
Deutsche Welle	33922

В таблице 2 и 3 показаны основные источники казахстанских и российских новостей, представленные в корпусе. Ниже представлен полный перечень источников:

Казахстанские источники:

<https://inbusiness.kz/ru>, <https://www.nur.kz/>, <https://tengrinews.kz/>,  
<https://24.kz/ru/>, <https://forbes.kz/>, <https://kapital.kz/>, <http://vesti.kz/>,  
<https://sputniknews.kz/>, <https://rus.azattyq.org/>, <https://365info.kz/>, VK,  
<http://today.kz>, <http://www.newsfactory.kz/>, <https://www.ktk.kz/>, <http://www.kp.kz/>,

<https://www.kt.kz/>, <https://kazakh-tv.kz/ru>, Telegram, <https://liter.kz/>,  
<http://www.dailynews.kz/>, <https://bnews.kz/ru> (baigenews.kz),  
<https://www.zakon.kz/>, <https://vlast.kz/>, <http://www.spik.kz/>, <https://rezonans.kz/>,  
<http://kz.mir24.tv/>, <https://kaztrk.kz/ru>, egemen.kz, <https://aif-kaz.kz/>,  
<https://www.interfax.kz/>, Instagram, <https://menshealth.kz/>, <http://rk-news.com/>,  
<http://infonedra.kz/>, <https://kz.expert/>, YouTube, <https://www.kazpravda.kz/>,  
<https://www.sports.kz/>, <http://alashinform.kz/>, <https://baribar.kz/>,  
<https://korrespondent.net/>, VKontakte, <http://edunews.kz/>, Facebook,  
<https://adebiportal.kz/kz>, <http://turantv.kz/>

Российские источники:

Lenta.ru, Интерфакс, РБК, RT, Лента, Ведомости, Бизнес ФМ, Радио Свобода, Sputnik (Ru), Deutsche Welle (DW), Настоящее время

В корпус также включено около 4000 документов, которые представляют собой фрагменты документов государственной программы развития. Использовано 25 государственных программ развития, в том числе 18 программ для отдельных регионов и крупных городов, 2 долгосрочные государственные программы развития и 5 тематических программ (цифровизация, развитие села, программы образования, здравоохранения и социального обеспечения). Они были вручную разделены экспертами на 4000 независимых фрагментов. Причина такой предварительной обработки заключается в том, что правительственные программы, как правило, очень длинные и могут быть очень тематически разнообразными, что затрудняет использование всего документа в тематическом моделировании.

Сравнение представления этих правительственных документов в тематиках – один из подходов, используемых для оценки социальной значимости новостей в [19, р. 6]. Сравнение производилось посредством метода, предложенного в разделе 2.3 работы.

Метод сбора данных – это веб-парсинг (скрапинг) новостных и медиасайтов с открытым бесплатным доступом, а также парсинг социальных сетей либо по списку учетных записей социальных сетей (пользователей, групп, каналов и т.д.), либо по перечню поисковых запросов.

Алгоритмы парсинга были реализованы в виде операторов Apache Airflow. Apache Airflow – это инструмент ETL (извлечение-преобразование-загрузка, extract-transform-load), который позволяет программно планировать и управлять задачами, отслеживать их и обрабатывать ошибки и исключения. Apache Airflow используется как базовое ETL-решение для системы мониторинга СМИ разработанной в ходе работы информационной системы [88].

Алгоритмы парсинга были реализованы с использованием Python библиотеки Scrapy версии 1.7.3. Для веб-сайтов с динамическим контентом (например, веб-сайтов, основанных на React.JS, Angular и других современных фреймворках для разработки веб-интерфейсов) использовалась библиотека scrapy-splash вместе с официальным образом Docker scrapyhub/splash: 3.3.1. Scrapy-Splash был применен для имитации запуска кода JavaScript внутри

клиентского веб-браузера с целью получения содержимого веб-сайта (списка новостных публикаций и текстов новостных публикаций вместе с другими метаданными).

Был реализован конфигурируемый Scrapy Spider, который принимает начальный URL и список правил для получения определенных метаданных. Такой подход позволяет свести к минимуму объем необходимой разработки программного обеспечения для внедрения новых источников парсинга, поскольку для этого требуется только список правил парсинга и начальный URL. Конечно, можно создать универсальный синтаксический анализатор, который мог бы обрабатывать любой источник информации, однако такие универсальные синтаксические анализаторы, как правило, предоставляют гораздо более низкое качество метаданных, и результаты очистки такими универсальными решениями обычно неструктурированы или слабо структурированы. По этой причине предлагается подход, который требует набор правил парсинга для каждого веб-сайта, но имеет гораздо более высокое качество и точность метаданных, включая дату и время публикации, количество просмотров, автора, теги, комментарии, лайки, шейры и т.д.

Под правилами парсинга здесь понимается набор CSS-селекторов для каждого из доступных элементов метаданных новостных публикаций на данном веб-сайте. В принципе, также возможно использовать регулярные выражения вместо CSS-селекторов, однако CSS-селекторы выглядят более подходящим выбором, поскольку подавляющее большинство HTML-страниц построено в соответствии с определенной методологией CSS, что упрощает навигацию с помощью CSS-селекторов. Однако, исходя из полученного опыта, есть несколько редких случаев, когда требуется либо применение регулярных выражений, либо некоторые дополнительные условия и проверки в коде, реализующем Spider.

Пример правил лома для новостных СМИ Deutsche Welle (<https://www.dw.com/ru/>):

Дата и время публикации - `.col1 .group .smallList li`.

Текст публикации - `.intro, .longText`> p.

Название публикации - `#bodyContent h1`.

Еще одна техническая проблема с парсингом заключается в том, что новостные веб-сайты обычно применяют некоторые меры против автоматического парсинга и DDoS-атак (distributed denial of service – атаки, при которых на сайт совершается большое количество запросов, с целью помешать нормальному функционированию сайта), поэтому было реализовано использование случайных прокси-серверов, а также случайный выбор HTTP-заголовка User-Agent для снижения шансов идентификации и бана. Также Scrapy позволяет настроить период времени между запросами, параллелизм и другие параметры, настройка которых позволяет свести к минимуму вероятность быть заблокированным веб-сайтом.

Парсинг социальных сетей – в общем случае более сложная проблема, по сравнению с парсингом открытых новостных сайтов, поскольку социальные



сети имеют тенденцию вводить ряд технических ограничений для парсинга. В большинстве случаев парсинг возможен только через официальный API (доступ к которому может быть затруднен или может быть очень ограничен) или путем тщательной симуляции действий пользователя в браузере с помощью Selenium или аналогичного программного обеспечения. Второй вариант очень дорог в реализации, поэтому по возможности использовался первый подход.

## 5.2 Методика оценки социальной значимости

Социальная значимость является сложным критерием не только для автоматической классификации, но и для оценки экспертами, поскольку определение социальной значимости требует комплексного учёта многих внешних факторов, включая текущую социально-экономическую и политическую ситуацию, текущие тренды, а также глубокий контекстуальный анализ текста.

Таким образом, автоматическое определение социальной значимости требует многокритериального подхода и учёта различных внешних факторов. Для решения данной задачи был разработан многофакторный подход к определению социальной значимости текстов, включающий агрегацию информации из трёх основных источников информации:

- 1) корпус программ государственного развития по направлениям (промышленность, образование, экология и т.д.) и по регионам;
- 2) результаты социальных опросов, проведённых АО ИАЦ, целью которых было, в том числе, выявление наиболее социально значимых новостей и тем;
- 3) публично доступные объективные показатели вовлечённости читателей – количество просмотров, комментариев, репостов.

Для каждого из этих трёх источников информации был разработан свой подход по интеграции до уровня оценок отдельных публикаций в автоматическом режиме.

Учёт гос. программ проводился с помощью оценки меры межкорпусного тематического дисбаланса, значения которого затем использовались в качестве весов влияния тем/топиков на социальную значимость публикации. Формула (5.2.1) для оценки дисбаланса:

$$D_{t_i c_j} = \frac{\sum_k w_{d_k t_i c_j}}{\sum_k \sum_l w_{d_k t_l c_j}} / \sum_m \sum_k \sum_l w_{d_k t_l c_m} \quad (5.2.1)$$

где  $D_{t_i c_j}$  – это мера дисбаланса присутствия документов из корпуса  $c_j$  в топике  $t_i$ ;

$w_{d_k t_l c_m}$  – вес принадлежности документа  $d_k$  из корпуса  $c_m$  к топикам  $t_l$ .

Результаты социальных опросов были агрегированы по топикам, а также были рассчитаны замещённые значения оценок социальной значимости респондентами. Полученные оценки использовались в качестве весов влияния тем/топиков на социальную значимость публикации.

Для учёта данных о резонансности публикаций был создан корпус социально резонансных публикаций, в который вошли те публикации, количество просмотров, комментариев в которых больше, чем на одно стандартное отклонение больше, чем математическое ожидание статистического распределения количества просмотров в источнике данной публикации. В корпус нерезонансных публикаций попали все остальные документы. После этого аналогично с корпусом гос. программ, были рассчитаны оценки межкорпусного тематического дисбаланса.

Затем были рассчитаны оценки каждой публикации по каждому из трёх критериев (соответствие гос. программам, соответствие результатам опроса по соц. значимости и резонансность) с помощью простой взвешенной суммы:

$$v_{d_i c_k} = \sum_j w_{d_i t_j} * D_{t_j c_k} \quad (3)$$

где  $v_{d_i c_k}$  – оценка документа  $d_i$  по критерию;

$w_{d_i t_j}$  – вес принадлежности документа  $d_i$  к топикю  $t_j$ .

Таким образом, были получены 3 отдельные оценки социальной значимости, однако необходимо также провести их агрегацию, поскольку каждый из трёх критериев учитывает лишь отдельные из аспектов социальной значимости: так, например соответствие гос. программа является важным индикатором социальной значимости, не является ни достаточным, ни необходимым условием для классификации текста как социально значимого. То же можно сказать и об остальных критериях.

Для агрегации была применена модель мультикритериальной оценки масс-медиа (ММА), описанная в разделе 3.2 работы, использующая метод агрегации, основанный на теореме Байеса и рассматривающей социальную значимость документа как субъективную вероятность, зависящую от других вероятностных параметров – трёх вышеописанных критериев, а также от принадлежности документов к темам/топикам.

В итоге были сформированы дашборды по социальной значимости, а также проведена аналитическая работа. Результаты верификации полученных результатов были опубликованы в [19, р. 5].

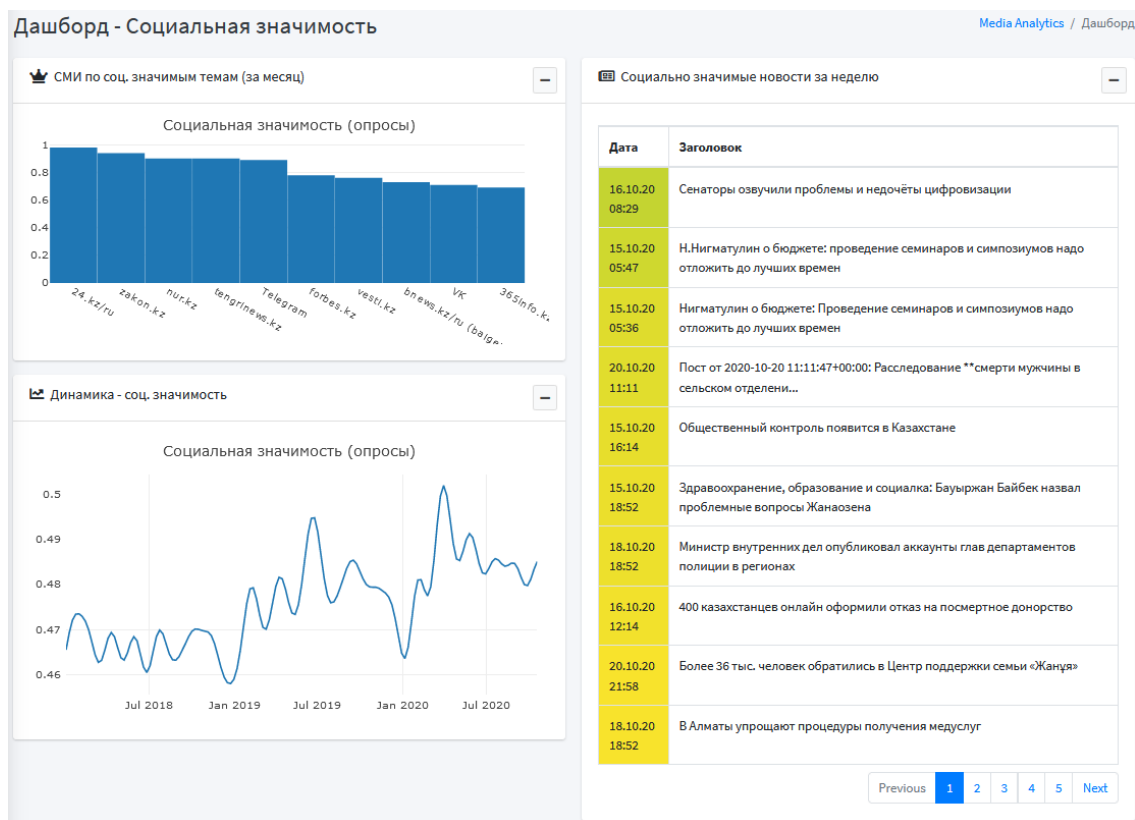


Рисунок 12 – Дашборд по социальной значимости

Гос программы + Опросы + Резонансность_m4a_class	Дата	Заголовок	Источник
0,954	2020-07-03T09:12:00+00:00	Более 130 тысяч рабочих мест уже создано в рамках Дорожной карты занятости в Казахстане	<a href="https://www.kt.kz/">https://www.kt.kz/</a>
0,954	2020-06-30T04:54:00+00:00	Министерству цифрового развития, инноваций и аэрокосмической промышленности надо актуализировать комплекс мероприятий госпрограммы "Цифровой Казахстан", отметил премьер	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>
0,954	2020-06-29T18:52:00+00:00	Ежегодная потребность в ИТ-специалистах в РК составляет более 30 тыс. человек	<a href="http://www.dailynews.kz/">http://www.dailynews.kz/</a>
0,953	2020-09-07T06:44:00+00:00	Как обеспечиваются чистой водой села Карагандинской области	<a href="https://www.zakon.kz/">https://www.zakon.kz/</a>
0,953	2020-08-21T05:47:00+00:00	Цифровые возможности в селах	<a href="https://tengrinews.kz/">https://tengrinews.kz/</a>
0,952	2020-07-22T13:56:00+00:00	Первый проект ГЧП в области здравоохранения - открытие медицинского центра	<a href="https://tengrinews.kz/">https://tengrinews.kz/</a>
0,952	2020-06-29T18:52:00+00:00	Аскар Мамин поручил реализовать масштабный проект по повышению цифровой грамотности населения	<a href="http://www.dailynews.kz/">http://www.dailynews.kz/</a>
0,952	2020-06-25T14:45:00+00:00	Мамин провел заседание комиссии по вопросам внедрения цифровизации	<a href="https://www.kt.kz/">https://www.kt.kz/</a>
0,952	2020-03-21T18:52:00+00:00	Как развивается цифровизация сферы строительства, рассказал Премьер	<a href="https://forbes.kz/">https://forbes.kz/</a>
0,951	2020-09-01T16:47:00+00:00	В РК создадут электронный реестр лицензий в сфере строительства	<a href="https://kapital.kz/">https://kapital.kz/</a>
0,951	2020-04-17T07:02:00+00:00	В акимате Акмолинской области обсудили вопросы реализации проекта «Ауыл – ел бесігі»	<a href="https://www.nur.kz/">https://www.nur.kz/</a>
0,950	2020-05-29T11:49:00+00:00	Как реализуется проект "Ауыл - ел бесігі" в Акмолинской области	<a href="https://tengrinews.kz/">https://tengrinews.kz/</a>

Рисунок 13 – Топ социально-значимых новостей за 2020 год

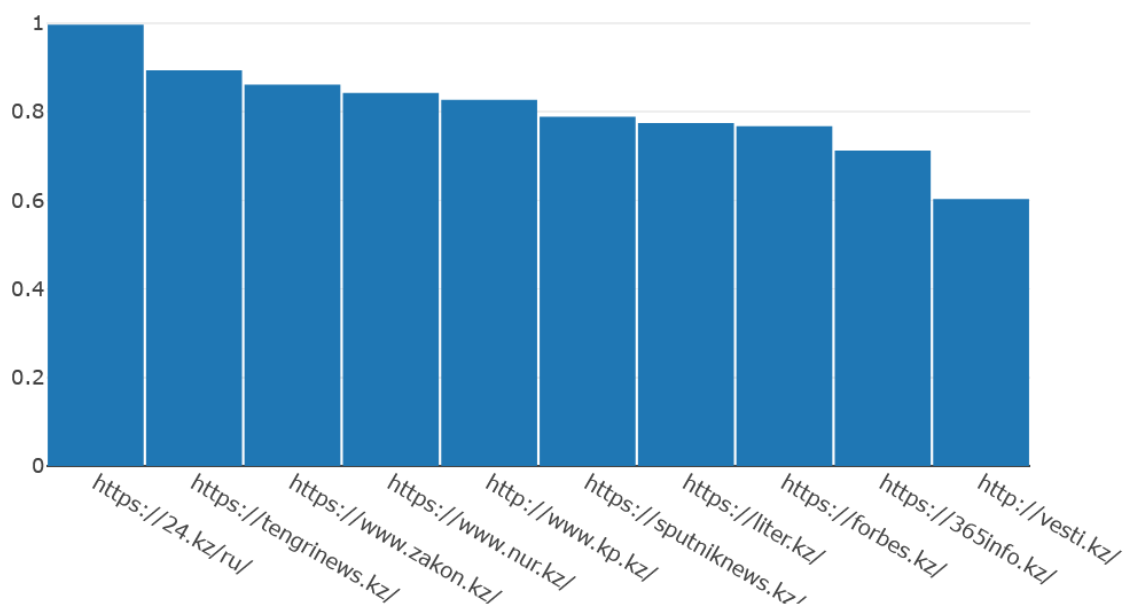


Рисунок 14 – Топ 10 новостных ресурсов по вкладу социально-значимой информации в медиа-поле

Рисунки 12, 13 и 14 демонстрируют часть полученных результатов, визуализированных в рамках разработанной информационной системы.

### 5.3 Результаты валидации модели и системы на базе размеченных экспертами данных

В данном разделе описаны результаты применения модели мультикритериальной оценки масс-медиа (ММА), предложенной в данной работе в разделе 3.2. Была поставлена задача выявления так называемых “потенциально опасных” новостей – тех новостей, которые могут вызвать острую реакцию в обществе. Таким новостям дается следующее определение – опасными новостями можно считать социально-значимые негативные новости, которые могут вызвать общественный резонанс. Представленные здесь результаты являются основным обоснованием применимости и актуальности предложенных в диссертации моделей и методов, выносимых на защиту. Представленные ниже результаты были впервые опубликованы в [19, р. 5].

Матрица P2 была получена посредством тематического моделирования BigARTM на корпусе из 804829 новостных публикаций из 40 казахстанских источников, опубликованных с 01.01.2018 по 31.12.2019. Корпус и данные, использованные для проверки, доступны в [89].

Тематическая модель BigARTM была применена с сглаживающим регуляризатором ( $\tau=0,15$ ), phi-декоррелятором ( $\tau=0,5$ ) и “improve coherence phi”-регуляризатором ( $\tau=0,2$ ), количество тем - 200. Параметры были выбраны в процессе экспериментов (grid-search оптимизация).

Для проведения вычислительного эксперимента был выбран набор из трех критериев оценки - популярность/резонансность, социальная значимость и негативная тональность. Предлагается агрегировать эти три признака в

финальный класс, описывающий потенциальную опасность новости, как это было описано выше. Таблица 4 АНР была заполнена экспертами, рассчитана и были получены следующие значения для матрицы P3:

- вес 0,58 для негативной тональности;
- вес 0,22 для резонансности/популярности;
- вес 0,19 для социальной значимости.

Таблица 4 – (АНР) таблица, заполненная экспертами

Критерий	Негативная тональность	Социальная значимость	Резонансность	Рассчитанные веса
Негативная тональность	1	2	5	0,58
Социальная значимость	1/2	1	1/2	0,19
Резонансность	1/5	2	1	0,22

Каждый из предложенных критериев был получен по отдельной методике, что демонстрирует гибкость предложенной модели:

1. Значения отрицательной тональности по каждой теме были получены путем ручной разметки двумя экспертами. Разметка была проверена на несоответствия, откорректирована и усреднена. Каждая тема была представлена 25 ключевыми словами/фразами, каждый эксперт дал оценку настроения этой темы по шкале от 0 до 10, где 5 – нейтрально, 0 – очень положительно, а 10 – очень отрицательно. Следует отметить, что оценка тональности в данном случае не отражает мнение автора о событии или сильную эмоциональную окраску, это скорее негативность или позитивность события, описанного в данной новостной публикации в контексте влияния на общественное и индивидуальное развитие.

2. Резонансность/популярность – данный критерий был автоматически размечен методом анализа межкорпусного дисбаланса, так как количество просмотров и другие показатели активности известны для определенной части корпуса. Каждая тема была размечена автоматически в зависимости от доли популярных/резонансных новостей, относящихся к теме.

3. Социальная значимость также оценивалась автоматически, но на основе экспертной оценки источников. В данном случае такая разметка была тривиальной – документы из корпуса официальных государственных программ развития, в том числе планы развития различных регионов, отраслей и направлений, считались социально значимыми, а документы из обычных источников в СМИ считались в общем случае социально незначимыми или нейтральными по отношению к социальной значимости. Это можно рассматривать как оценку тематической асимметрии (дисбаланса) между корпусом новостных публикаций и корпусом текстов государственных программ развития. В общем случае, данный подход позволяет назначать различные источники к разным группам/корпусам с помощью экспертной

разметки, или такое разделение может быть выполнено с использованием какого-либо другого явного свойства (метаданные).

Таким образом, проведенный эксперимент демонстрирует основные сценарии получения оценок тем по заданному критерию:

1. Ручная разметка на уровне тем.
2. Автоматическая или полуавтоматическая разметка на основе объективных параметров публикаций.
3. Высокоуровневая разметка источников, корпусов или некоторых других явных свойств публикаций с последующим присвоением меры межкорпусного дисбаланса в качестве тематических весов.

Рассмотрим топы новости по каждому из критериев и топовые новости общего целевого класса.

Таблица 5 – Топ новостей по негативной тональности

Тональность (-)	Дата	Заголовок
0,857	2019-11-01	Целый арсенал оружия изъяли у жителя Алматы
0,857	2019-05-06	Взрыв боеприпаса произошел в Арыси
0,853	2018-03-13	Движение «Демократический выбор Казахстана» признано экстремистским
0,851	2020-05-28	Еще одну активистку вызвали на допрос после запрета движения «Көше партиясы»
0,846	2020-01-05	Подозреваемый в покушении на убийство задержан в Темиртау

Таблица 6 – Топ новостей по оцененной резонансности

Резонансность	Дата	Заголовок
0,99	2019-09-12	Экс-капитан «Барыса» забросил 250-ю шайбу в КХЛ и помог своему клубу выиграть пятый матч подряд
0,958	2018-12-01	Разгромное поражение потерпели хоккеисты «Барыса»
0,944	2019-12-26	Данэлия Тулешова представила клип на песню "Don't cha"
0,923	2019-10-27	Димаш Кудайберген выступил на концерте Игоря Крутого в Нью-Йорке
0,917	2018-10-08	Шиповник: отвар при беременности

Таблица 7 – Топ новостей по социальной значимости

Социальная значимость	Дата	Заголовок
0,997	2019-04-09	210 млрд тенге было выделено на программу "Дорожная карта бизнеса-2020" за четыре года
0,974	2018-04-30	14 объектов водоснабжения сдали в ЮКО в 2017
0,957	2020-05-05	Более 1,2 млн. человек планируется трудоустроить в 2020 году
0,943	2020-04-03	690 тыс. казахстанцев получают рабочие места
0,938	2018-09-17	В Казахстане растет число малых компаний

Таблица 8 – Топ новостей по финальному классу (потенциально опасные новостные публикации)

Финальный класс	Дата	Заголовок
0,872	2018-03-13	Движение «Демократический выбор Казахстана» признано экстремистским
0,856	2018-03-29	Добавление в группы ДВК в Telegram, какие могут быть последствия для граждан
0,818	2019-10-25	Сотрудников Антикоррупционной службы ВКО обвиняют в применении пыток
0,818	2019-02-04	Задержан житель Астаны, намеревавшийся создать ячейку «Хизб-ут-Тахрир»
0,817	2019-11-01	Целый арсенал оружия изъяли у жителя Алматы

В таблицах 5, 6, 7 и 8 показаны 5 главных новостей по каждому критерию и по финальному классу соответственно. Следует отметить, что главные новости по резонансности/популярности в таблице 5 были отобраны вручную из 500 лучших новостей, поскольку топ новостей очень однотипен, а главные новости в основном состоят из публикаций, связанных со спортом. Главные новости по негативным настроениям состоят в основном из информации об убийствах, информации о расследованиях запланированных террористических актов, катастроф (таких как взрыв в Арыси в 2019 году) и вопросов, связанных с правами человека (митинги, свобода собраний, жестокость полиции и т.п.). Топовые новости по популярности в основном состоят из новостей спорта, новостей о знаменитостях, различных публикациях бытового характера. Главные социально значимые новости состоят в основном из государственных программ занятости, инфраструктурных программ и охватывают основные направления социальной поддержки со стороны государства. Важно отметить, что главные новости по финальному агрегированному классу в общем случае не пересекаются с топовыми новостями по трем критериям. Отсюда можно сделать вывод, что три свойства между собой независимы, что может указывать на потенциал для того, чтобы агрегированный класс был более

информативным. Главные новости финального класса состоят в основном из информации о популярном незаконном оппозиционном движении «ДВК» (Демократический выбор Казахстана, запрещенная в Казахстане организация), экстремистских религиозных движениях, планировании террористической активности и фактах жестокого обращения со стороны полиции. Действительно, эти новости популярны, связаны с социально значимыми темами (в основном, с вопросами национальной безопасности) и являются негативными, но при этом не наблюдается полная корреляций ни с одним из трех первоначальных критериев.

Таблицы 5-8 показывают, что топовые новости по рассчитанным значениям критериев действительно соответствуют заявленным критериям. Однако такая ручная проверка, естественно, не может применяться для полноценной валидации результатов модели. Следовательно, предлагается использовать методологию перекрестной проверки (cross-validation):

1. Подмножество из 10% новостных публикаций было случайным образом помечено как тестовая выборка.

2. Начиная с этапа тематического моделирования, расчет описанной модели проводился без учета новостей из тестовой выборки.

3. Затем, когда были получены все необходимые веса (матрицы P1, P2, P3 и P4), матрицы P5 и P6 были рассчитаны для всего корпуса, включая тестовый набор.

Затем, для каждого из критериев были сформированы случайные подвыборки из 1000 новостей из тестового набора, при этом учитывались новости из 10 верхних перцентилей и 10 нижних перцентилей по рассчитанному моделью значению соответствующего критерия. Причина такого подхода заключается в том, что эксперименты показали, что значительная часть новостей не может быть корректно размечена ни экспертом, ни предложенным методом, поскольку не все новости, например, являются социально значимыми или незначительными – большинство новостей являются неопределенными по данному критерию – в какой-то мере социально значимыми. Следует отметить, что 20 процентов корпуса по-прежнему представляют собой значительный объем новостей – около 160 000. Следовательно, цель проверки – убедиться, что точность (precision) модели высока, в то время как охват (recall) не считается критически важным, согласно предложенной методологии оценки.

Эти подвыборки были вручную размечены экспертами. Экспертам были предоставлены заголовок, дата публикации, URL-адрес и медиа-источник каждого документа. Разметка была проведена для негативной тональности, социальной значимости и для финального класса. Для резонансности такая разметка не потребовалась, поскольку истинные значения количества просмотров и другие показатели активности известны и были использованы для проверки.

По описанной методике были рассчитаны метрики качества предложенной модели.



Таблица 9 – Результаты валидации модели

Критерий	MAE	F1-Score	ROC AUC
Резонансность	0.36	0.49	0.56
Негативная тональность	0.14	0.93	0.93
Социальная значимость	0.27	0.65	0.69
Финальный класс – опасные новости	0.25	0.81	0.81

В таблице 9 отображены результаты проверки качества работы модели. Можно заметить, что самые низкие результаты были получены для резонансности/популярности, хотя объективные показатели активности были известны. Это может быть связано с тем, что популярность (количество просмотров) в большей степени связана с названием статьи, а не с темами, содержащимися в тексте, в то время как предлагаемая модель анализирует тексты на основе высокоуровневых тематических структур. Оценки социальной значимости также показывают результаты ниже среднего, в то время как отрицательная тональность, будучи более простым критерием для оценки, была предсказана с ROC AUC 0,93. Модель оценки по финальному классу также продемонстрировала значительную предсказательную способность – ROC AUC 0,81. В целом, можно сделать вывод, что модель демонстрирует значительную предсказательную способность, но из-за неопределенной субъективной природы рассматриваемых критериев необходим компромисс между точностью (precision) и охватом (recall). В данном случае точность достаточно высока: точность классификации резонансности для положительного класса составляет 0,94, для социальной значимости – 0,97, для отрицательной тональности – 0,94, а для финального класса была достигнута точность 0,81. В то время как охват колеблется от 0,13 (резонанс) до 0,94 (негативная тональность). Следует отметить, что такие результаты были получены на относительно простой модели (без каких-либо элементов рекуррентных нейронных сетей или других методов глубокого обучения), а расчет всех необходимых весов/параметров был либо автоматическим, либо требовал минимальной ручной разметки.

Чтобы сравнить качество классификации предложенной модели, к данным был применен современный подход Transfer Learning, в частности, модель глубокого обучения BERT. Предварительно обученная модель из библиотеки DeepPavlov (RuBERT, русский язык, с учетом регистра, 12 основных слоев, 768 – размерность векторизации, 12 attention head, порядка 180 миллионов параметров) использовалась для получения векторизаций отдельных предложений (sentence embeddings), которые затем усреднялись для получения векторных представлений текстов (text embeddings). Затем модель Gradient Boosting из пакета scikit-learn была обучена на 1000 размеченных текстах для каждого критерия. K-Fold кросс-валидация полученной модели показала, что результаты предложенной модели сопоставимы с современными моделями глубокого обучения в ситуациях с небольшим количеством размеченных объектов, что имеет место со многими проблемами

классификации, за исключением небольшого подмножества хорошо исследованных (например, проблема анализа тональности). Значения ROC AUC для моделей на основе BERT-векторизаций следующие: 0,65 для финального класса, 0,7 для социальной значимости, 0,72 для негативной тональности и 0,88 для резонансности, что сопоставимо с результатами предлагаемой модели MMA. Также следует отметить, что современные исследования показывают, что векторизации, полученные предобученными моделями BERT можно использовать для получения качества, близкого к state-of-the-art, по ряду задач, включая классификацию текста, даже без дообучения модели (fine tuning) [90]. По этой причине, дообучение модели BERT не рассматривалась в этой работе, особенно в условиях небольшого объема размеченных данных (рисунок 15).

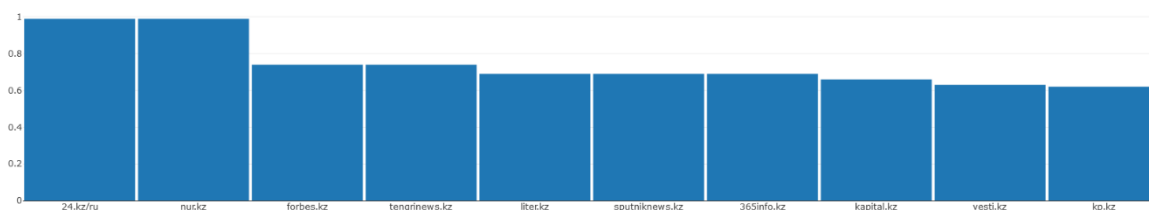


Рисунок 15 – Сравнение оценок главных медиа-источников по финальному классу (социально-значимый резонансный негатив)

Что касается тональности, можно применить другой тип анализа – например, рассмотреть распределение настроений между различными медиа-источниками. На рисунке 16 нейтральные новости соответствуют новостям с оценками тональности в диапазоне [0,4; 0,6].

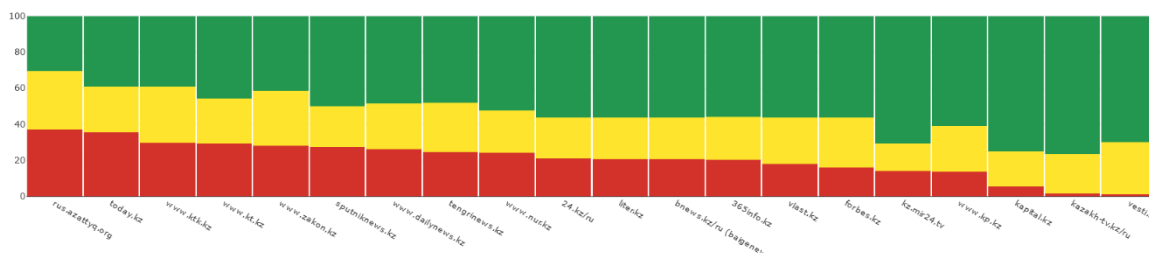


Рисунок 16 – Распределения тональности каждого источника, отсортированные по негативному воздействию

Примечание – Красный цвет – негатив, зеленый – позитив, желтые – нейтральные публикации

Таким образом, в разделе описаны основные результаты, полученные предложенными методами, демонстрирующие применимость методов в заданных условиях. Проведено сравнение с современными моделями глубокого обучения (BERT), которое показало, что результаты предложенной модели сопоставимы при значительно более низкой вычислительной сложности, более высокой интерпретируемости и небольшом объеме ручной разметки.

### 5.3.1 Прочие кейсы использования предложенных моделей и методов

Кроме того, с помощью предложенных моделей и методов был получен ряд других результатов, которые описаны в этом подразделе. Подраздел не нацелен на то, чтобы подкрепить обоснование применимости методом, а на то, чтобы продемонстрировать их универсальность.

Пропаганда. В [76, р. 5] описан пример применения предложенного подхода по оценке межкорпусного дисбаланса для выявления публикаций и с пропагандистским содержанием.

Здесь, в сотрудничестве с экспертами в области политологии и журналистики, пропаганде было дано следующее определение – пропаганда это нацеленное на изменение общественного и личного мнения информационное воздействие, предоставляющее информацию в определенном свете. С точки зрения журналистики, под пропагандой автоматически понимается информация, опубликованная аффилированными, зависимыми источниками, которая выдается за независимое нейтрально мнение. Как обсуждалось в разделе 1.2 работы, с точки зрения определения журналисткой этики пропаганда является не этичной, однако все еще относится к журналистике. Другие виды нацеленного информационного воздействия, как то: публичные открытые выступления, официальные заявления государственных органов и других организаций, публикация отчетов, справок и других документов не является пропагандой, в случае если не происходит попытки выдать подаваемую информацию за нейтральное, непредвзятое, объективное мнение.

Согласно этому определению, можно дать четкий критерий для определения пропагандистских источников – пропагандистские СМИ напрямую аффилированы и/или финансируются государственными и квази-государственными организациями, поскольку в таких условиях сохранять нейтральность в общем случае не представляется возможным. Частные, независимые, в свою очередь, СМИ должны финансироваться из собственных источников монетизации, добровольных открытых пожертвований и из средств частных спонсоров.

При этом нужно заметить, что из определения выше не следует, что все новости, опубликованные в СМИ, финансируемых напрямую государством, автоматически являются пропагандой; неверно также и обратное утверждение, касательно “независимых” СМИ. Однако, это определение позволяет сделать вывод, что в СМИ, финансируемых напрямую государством, содержание пропаганды в общем случае будет выше, чем в независимых СМИ. Следовательно, если предложенная гипотеза верна, то путем сравнительного анализа условно пропагандистских и условно независимых СМИ можно выявить определенные статистические закономерности, позволяющие сделать вывод о том какие темы преобладают в пропагандистских СМИ, а также построить классификатор для распознавания потенциально пропагандистских публикаций.

Для решения этой задачи был собран корпус российских СМИ. К пропагандистским СМИ были отнесены:

1. Russia Today.
2. Настоящее время.
3. Радио Свобода.
4. Deutsche Welle.
5. Спутник.

К условно объективным/независимым СМИ были отнесены:

1. Ведомости.
2. Interfax.
3. Lenta.ru.
4. Бизнес FM.
5. RBC.

Можно заметить, что при выборе пропагандистских СМИ не было использовано политических критериев, а только объективный критерий независимости финансирования. Так, в список попали прокремлевские (Russia Today), либеральные с американским финансированием (Настоящее Время, Радио Свобода), а также русскоязычные СМИ, нацеленные на российскую аудиторию, но действующие на территории других стран (Deutsche Welle). То же можно сказать и о объективных СМИ – при выборе списка мы руководствовались критерием разнообразия и репрезентативности.

В итоге, к полученным двум подкорпусам был применен метод межкорпусного тематического дисбаланса, описанный в работе, а также предложенный метод агрегации. Полученный классификатор был провалидирован на размеченном экспертом датасете, состоящем из 1000 документов. В зависимости от порога классификации, ROC AUC классификатора варьируется от 0.73 до 0.95.

Также в качестве побочного результата исследования были выявлены наиболее пропагандистские темы в российском медиапространстве – это темы, связанные советской историей (особенно период правления Сталина и Вторая Мировая Война), этническими вопросами и вопросами прав человека.

Анализ темы возобновляемой энергии. Другой кейс использования предложенных наработок был опубликован в [91]. В работе рассматривается применение тематических моделей и методов динамической обработки результатов их работы для численной и качественной оценки ситуации по освещенности вопросов, связанных с использованием возобновляемых источников энергии в СМИ и социальных сетях. При этом аналитика проводилась в сравнении между СМИ Казахстана и России.

В результате, были сделаны следующие основные выводы:

- проблемы атомной энергетики значительно более освещены в СМИ РФ, чем в СМИ Казахстана;
- вопросы, связанные с тарифами на электроэнергию, становятся всё менее обсуждаемыми в обществе;
- несмотря на то, что в казахстанских СМИ гораздо больше обсуждаются проблемы экологии, связанные с энергетикой, информация, касающаяся зелёной энергетики, освещена значительно меньше, чем в российских медиа.

Анализ трендов в области информационной безопасности. Работа, опубликованная в [92] демонстрирует применение предложенных методов к другой задаче оценки тематической структуры пространства – к анализу освещенности вопросов информационной безопасности. В данной работе был впервые опубликован метод каскадного применения тематических моделей.

Обоснование данного метода заключается в том, что на текущий момент классические LDA тематические модели даже с различными регуляризаторами не позволяют эффективно разбить большой корпус (сотни тысяч публикаций) на большое количество топиков (500 и более), поскольку модель построена на основе метода максимизации правдоподобия, а следовательно пытается аппроксимировать латентные семантические структуры максимально эффективно, и вклад каждого последующего топика в общее распределение постепенно уменьшается. Следовательно, менее объемные темы могут быть не выделены при изначально тематическом моделировании.

Тогда, предлагается провести ручной отбор тем, относящихся к искомой, с выставлением порога веса принадлежности документа к выбранным темам. После этого в случае, если результаты все еще не удовлетворительны, повторять итеративно этот процесс.

Таким образом, при поиске тем, связанных с информационной безопасностью, в первой тематической модели подходящих тем не было. Были отобраны темы, опосредованно связанные с информационной безопасностью (цифровизация, интернет, гаджеты и т.п.), после чего тематическое моделирование было проведено снова на отфильтрованном подмножестве публикаций. Такие итерации повторялись 4 раза, пока не была получена тематическая модель, отражающая реальные тренды и темы в области информационной безопасности, такие как:

- банковское мошенничество;
- скандалы с взломом крупных компаний с утечкой личных данных пользователей;
- достижения в области шифрования и технологий авторизации;
- познавательные статьи по практике информационной безопасности
- и другие.

Пример отчета, выгруженного из системы, по темам, связанным с образованием (Приложение В).

### **Выводы по 5-му разделу**

В разделе были рассмотрены собранные данные, полученные результаты экспериментов и выводы по валидации разработанных методик и систем.

В ходе работы был разработан конфигурируемый алгоритм скрапинга, позволяющий совершать высококачественный парсинг новостных и других сайтов. На момент публикации работы, собранный набор данных включает 4233990 документов из казахстанских источников и 2027963 документа из российских источников. Также в хранилище системы содержится порядка 4000 фрагментов государственных программ развития, которые применялись для

оценки соответствия отдельных новостей приоритетам государственного развития.

В ходе работы было разработано определение социальной значимости, а также методология ее оценки в рамках отдельных публикаций (текстов). Методология заключается в многофакторном подходе, с учетом как минимум трех источников информации:

- тексты государственных программ развития (для определения соответствия текста государственным приоритетам);
- результаты опроса населения (для учета мнения граждан);
- объективные данные вовлеченности пользователей.

На основе полученного классификатора социальной значимости затем решается более комплексная задача – мультикритериальная оценка опасности новостей, с учетом их тональности, социальной значимости и резонансности.

Задача решается предложенным в разделе 3.2 работы методом ММА основанном на применении байесовской агрегации субъективных вероятностей. Однако, нужно заметить, что даже при применении тривиального метода агрегации (например, взвешенного среднего), подход с использованием тематических векторизаций текстов все еще показывает высокий уровень эффективности.

После валидации на размеченных вручную наборах данных, а также сравнении полученных метрик качества классификации с применением современных моделей глубокого обучения, обученных на больших объемах данных, можно сделать вывод о том, что предложенная методика, по сравнению с алгоритмом BERT и аналогичными предлагает значительно более простое (тысячи параметров против сотен миллионов у BERT), производительное и интерпретируемое решение, которое в то же время требует гораздо меньшего объема ручной разметки.

Также нужно отметить, что ввиду нечеткой природы решаемых задач, построение классификаторов, которые показывают как высокую точность (precision), так и высокий охват (recall), не представляется возможным – это подтверждается в том числе и результатами экспертной разметки, в процессе которой разные эксперты соглашались касательно невозможности отнесение отдельных публикаций к заданному классу.

Также в разделе рассмотрены и другие результаты, полученные на базе разработанной в ходе исследования информационной системы с применением предложенных моделей и методов.

Эти сценарии использования можно условно разделить на две группы:

1. Построение классификаторов определенных признаков текстов (например пропаганда, резонансность, социальная значимость и т.п.).
2. Проведение анализа публикационной активности в определенной предметной области.

В случае второго сценарий, был кратко описан предложенный метод каскадного построения тематических моделей. Он позволил, в частности, построить более детальные тематические модели для численного и

качественного анализа публикаций, касающихся возобновляемых источников энергии, а также информационной безопасности.

## ЗАКЛЮЧЕНИЕ

Диссертация является научной квалификационной работой, в которой были рассмотрены модели и методы классификации текстов на базе тематических моделей, а также информационная система для мониторинга медиапространства и его оценки с помощью вышеозначенных моделей и методов. Основные научные результаты диссертации, практические **выводы** и рекомендации, полученные при выполнении исследований, заключаются в следующем:

1 Проведен анализ рынка систем, предоставляющих услуги медиа-мониторинга, а также анализ нормативно-правовой основы и технических особенностей. Выявлены слабые стороны существующих решений, сформированы рекомендации;

2 Исследован вопрос влияния открытых информационных источников на обществе, выявлены основные направления влияния, сформирован перечень информативных признаков, на основе которых можно оценить это влияние;

3 Исследованы существующие подходы классификации документов и векторизации текстов, выявлены проблемы и слабые стороны текущих решений, сформированы рекомендации;

4 Разработан подход векторизации текстов на основе тематической модели;

5 Разработан метод оценки межкорпусного тематического дисбаланса, позволяющий автоматически или полуавтоматически получать веса топиков по отношению к заданному признаку;

6 Разработан метод мультикритериальной оценки медиа-источников ММА на базе байесовской модели агрегации;

7 Разработана распределенная информационная система на базе Open Source решений, позволяющая производить сбор (скрапинг), хранение, обработку текстовой информации, а также построение тематических моделей и классификаторов с возможностью визуализации полученных результатов;

8 Собран корпус, состоящий из более чем 6 миллионов публикаций из казахстанских и российских источников, включая как тексты публикаций, так и метаданные;

9 Проведена валидация предложенных моделей и методов. Основная серия вычислительных экспериментов была связана с определением так называемых опасных новостей (социально значимые, резонансные негативные публикации). Было проведено сравнение с моделью глубокого обучения BERT, метрики качество предложенных моделей сопоставимы, при меньшей вычислительной сложности и меньших требованиях к объему ручной экспертной разметки.

Достоверность научных результатов, выводов и положений, сформулированных в диссертации, базируются на предложенных математических моделях и алгоритмах.



Научные результаты позволили разработать модели, методы и программные инструменты, позволяющие классифицировать текстовые документы по ряду признаков (критериев) с минимальным объемом ручной разметки, с применением так называемой высокоуровневой разметки топики, либо метаданных статей. Разработанные в рамках проводимого исследования информационная система является легко масштабируемой (как функционально, так и с точки зрения производительности) и позволяет хранить, обрабатывать, агрегировать и визуализировать большие объемы текстов данных.

*Рекомендации и исходные данные по конкретному использованию результатов*

Результаты научного исследования, в частности разработанная информационная система может быть использована для ряда целей:

1. Использование исследователями и учёными. Как показано в разделе 5.3.1 работы, существует большой потенциал для использования разработанной информационной системы для самых разных гуманитарных исследований.

2. Использование крупными компаниями и государственными органами для поддержки принятия решений.

3. Использование крупными компаниями для решения задачи менеджмента репутации.

4. Использование обычными пользователями для разведывательного поиска интересующих данных.

При этом нужно отметить универсальность системы – она может быть применена к самым разным корпусам текстовых данных, например к внутренним документам организаций, научным публикациям, личными перепискам и т.п.

*Оценка технико-экономической эффективности внедрения.*

Полученные в диссертационной работе результаты исследований могут принести существенный технико-экономический вклад в области медиа-мониторинга, как для потенциальных клиентов, так и в научно-исследовательских целях. При этом, ввиду эффективности предложенных моделей, при достаточном масштабе, использование разработанной информационной системы может быть экономически более выгодным, ввиду меньшей потребности в вычислительных мощностях.

Разработанная информационная система была внедрена в Министерстве Образования и Науки Республики Казахстан (Приложение А).

*Оценка научного уровня выполненной работы в сравнении с лучшими достижениями в данной области*

Оценка научного уровня выполненной работы в сравнении с лучшими достижениями в данной области проведена на основании анализа научно-практических литературных источников, посвященных тематике «Классификация текстовых данных» и «Медиа-мониторинг». Выбор индекса классификации и глубина поиска 10 лет соответствующие теме обеспечили надежность и достоверность поиска актуальных информационных материалов. В результате проведенного анализа определено, что научный уровень

выполненной диссертационной работы обладает достаточной новизной и в целом соответствует мировому техническому уровню и тенденциям развития технологий классификации текстовых данных.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 Edelman R. Edelman Trust Barometer // <https://www.edelman.com/research/2019-edelman-trust-barometer>. 25.04.2020.
- 2 Miller D. Promotional strategies and media power // In book: *The Media: an Introduction*. – London: Longman, 1998. – P. 65-80.
- 3 Bushman B., Whitaker J. Media Influence on Behavior // In book: *Encyclopedia of Human Behavior*. – London; Burlington, MA, Elsevier, 2012. – P. 571-575.
- 4 Don W., Zongchao C., Cylor S. Media Effects // In book: *International Encyclopedia Of The Soc. & Behavioral Sciences*. – Ed. 2nd. – London; Burlington, MA, Elsevier, 2015. – Vol. 3. – P. 29-34.
- 5 Ko H., Jong Y., Sangheon et al. Human-machine interaction: A case study on fake news detection using a backtracking based on a cognitive system // *Cogn. Syst. Res.* – 2018. – Vol. 55. – P. 77-81.
- 6 Bushman B., Whitaker J. Media Influence on Behavior // *Reference Module Neuroscience and Biobehavioral Psychology*. – 2017. – Vol. 1 – P. 571-575.
- 7 Mishra S., RizoIU M.A., Xie L. Feature driven and point process approaches for popularity prediction // In *Procced. 25th ACM internat. on conf. on information and knowledge management*. – NY.: ACM, 2016. – P. 1069-107.
- 8 Tatar A., Antoniadis P., Amorim M.D. et al. Ranking News Articles Based on Popularity Prediction // *Procced. internat. conf. on Advances in Soc. Networks Analysis and Min (2012 IEEE/ACM)*. – NY., 2012. – P. – 106-110.
- 9 Bandari R., Asur S., Huberman B.A. The Pulse of News in Social Media: Forecasting Popularity // <https://arxiv.org/pdf/1202.0332.pdf>. 21.08.2021.
- 10 Bauer M.W., Suerdem A. Developing science culture indicators through text mining and online media monitoring // *Procced. OECD Blue Sky Forum on Science and Innovation Indicators*. – Ghent, 2016. – P. 19-21.
- 11 Ataman D. Bianet: A parallel news corpus in turkish, kurdish and english // <https://arxiv.org/pdf/1805.05095v1.pdf>. 21.08.2021.
- 12 Willaert T., Van Eecke P., Beuls K. et al. Building Social Media Observatories for Monitoring Online Opinion Dynamics // *Social Media+ Society*. – 2020. – Vol. 6(2). – P. 1-12.
- 13 Neresini F., Lorenzet A. Can media monitoring be a proxy for public opinion about technoscientific controversies? The case of the Italian public debate on nuclear power // *Public Understanding of Science*. – 2016. – Vol. 25. – P. 171-185.
- 14 Thanasopon B., Sumret N., Buranapanitkij J. et al. Extraction and evaluation of popular online trends: A case of Pantip.com // *Procced. 2017 9th internat. conf. on Information Technology and Electrical Engineering (ICITEE)*. – Phuket, Thailand, 2017. – P. 1-5.
- 15 Macharia S. Global Media Monitoring Project (GMMP) // In book: *The International Encyclopedia of Gender, Media, and Communication*. – Hoboken, New Jersey, 2020. – P. 1-6.

- 16 Barysevich A. Top of the best social media monitoring tools // <https://www.socialmediatoday.com/news/20-of-the-best-social-media>. 25.02.2021.
- 17 Agilitypr. Media monitoring ultimate guide // <https://www.agilitypr.com/media-monitoring-ultimate-guide>. 25.02.2021.
- 18 Newberry C. Social media monitoring tools // <https://blog.hootsuite.com/social-media-monitoring-tools>. 25.02.2021.
- 19 Mukhamediev R.I., Yakunin K., Mussabayev R. et al. Classification of Negative Information on Socially Significant Topics in Mass Media // *Symmetry*. – 2020. – Vol. 12(12). – P. 1-23.
- 20 Barile F., Ricci F., Tkalcic M. et al. A News Recommender System for Media Monitoring // *Proc. Internat. conf. on Web Intelligence (IEEE/WIC/ACM)*. – Thessaloniki Greece, 2019. – P. 132-140.
- 21 Guo Y., Barnes S.J., Jia Q. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation // *Tourism Management*. – 2017. – Vol. 59. – P. 467-483.
- 22 Curiskis S., Drake B., Osborn T., Kennedy P. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit // *Information Processing & Management*. – 2020. – Vol. 57, №2. – P. 1-21.
- 23 Basnyat B., Anam A., Singh N. et al. Analyzing Social Media Texts and Images to Assess the Impact of Flash Floods in Cities // *Proc. 2017 IEEE internat. conf. on Smart Computing (SMARTCOMP)*. – Hong Kong, China, 2017. – P. 1-6.
- 24 Van Aelst P., Brants K., Van Praag P. et al. The fourth estate as superpower? // *Journalism Studies*. – 2008. – Vol. 9(4). – P. 494-511.
- 25 Chowdhury G.G. *Introduction to modern information retrieval*. – London, UK: Facet publishing, 2010. – 474 p.
- 26 Bandura A. *Social Learning Theory*. – Kent, USA, Prentice-Hall, 1977. – 51 p.
- 27 Bryant J., Oliver M.B. *Media Effects: Advances in Theory and Research*. – Ed. 3rd. – NY.; London: Routledge, 2009. – 657 p.
- 28 Результаты социологического опроса по оценке влияния открытых информационных источников (электронных СМИ) на социум. – Астана: АО «Информационно-аналитический центр» МОН РК, 2018. – 109 с.
- 29 Рыжакова Е.В. Автоматическое определение иронии и сарказма в тексте. – М.: Высшая школа экономики, 2014. – 100 с.
- 30 Ward St.J.A. Journalism ethics // In book: *The handbook of journalism studies*. – NY., 2009. – P. 295-309.
- 31 Яндекс назвал самые популярные запросы казахстанцев // <https://kapital.kz/lifestyle/74330/yandeks-nazval-samyepopulyarnye>. 20.09.2019.
- 32 Байманов Д. Волнующие казахстанцев вопросы назвал Президент Республики Казахстан // [https://www.inform.kz/ru/volnuyushchie-kazahstancsev-voprosy-nazval-prezident-rk\\_a3449476](https://www.inform.kz/ru/volnuyushchie-kazahstancsev-voprosy-nazval-prezident-rk_a3449476). 20.09.2019.
- 33 Жулмухаметова Ж. Тройку самых острых вопросов, волнующих казахстанцев, назвали социологи // <https://informburo.kz/novosti/troyku-samyh-ostryh-voprosov-volnuyushchih-kazahstancsev-nazvali-sociologi.html>. 20.09.2019.

- 34 Барахнин В.Б., Кучин Я.И., Мухамедиев Р.И. К вопросу о постановке задачи выявления фейковых новостей и алгоритмах их мониторинга // Информатика и прикладная математика: матер. 3-й междунар. науч. конф. посв. 80-лет. Р.Г. Бияшева и 70-лет. М.Б. Айдарханова. – Алматы, 2018. – С. 113-118.
- 35 Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation // Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 993-1022.
- 36 Воронцов К.В. Вероятностное тематическое моделирование // <http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf>. 14.08.2019.
- 37 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.
- 38 Атанаева М.К., Булдыбаев Т.К., Оспанова У.А. и др. Методика для определения информативных признаков новостных текстов и проверка их значимости // Научный аспект. – 2019. – Т. 3, №3. – С. 277-295.
- 39 Lazer D.M.J. et al. The science of fake news // Science. – 2018. – Vol. 359, №6380. – P. 1094-1096.
- 40 Батура Т.В. Методы автоматической классификации текстов Automatic text classification methods // Программные продукты и системы. – 2017. – Т. 23, №30. – С. 85-99.
- 41 Zhang Y., Jin R., Zhou Z. Understanding bag-of-words model: a statistical framework // Int. J. Mach. Learn. & Cyber. – 2010. – Vol. 1. – P. 43-52.
- 42 Wang Sh., Jing J. Learning natural language inference with LSTM // Proceed. of the 2016 conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – San Diego, 2016. – P. 1442-1451.
- 43 Deng J. et al. Imagenet: A large-scale hierarchical image database // IEEE conf. on computer vision and pattern recognition. – Miami, 2009 – P. 248-255.
- 44 Tenney I., Das D., Ellie P. Bert rediscovers the classical nlp pipeline // ACL, Florence, Italy – 2019. – Vol. 1 – P. 4593-4601.
- 45 Brown T.B., Mann B., Ryder N. et al. Language models are few-shot learners // Proceed. 34th conf. on Neural Information Processing Systems (NeurIPS 2020), Vancouver, 2020. – P. 2339-2352.
- 46 Ramos J. Using tf-idf to determine word relevance in document queries // In Proceed. of the first instructional conf. on machine learning. – Amsterdam, Netherlands, 2003. – P. 29-48.
- 47 Mikolov T., Grave E., Bojanowski P. et al. Advances in pre-training distributed word representations // ELRA, Miyazaki, Japan – 2017. – Vol. 1 – P. 52-55.
- 48 Vaswani A., Shazeer N., Parmar N. et al. Attention is all you need // Proceed. 31st conf. on Neural Information Processing Systems. – Long Beach, 2017. – P. 1-11.
- 49 Vorontsov K., Frei O., Apishev M. et al. BigARTM: Open Source Library for Regularized Multimodal Topic Modelling of Large Collections // Proceed

internat. conf. on Analysis of Images, Social Networks and Texts – M., 2015. – P. 370-381.

50 Blei D., Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research. – 2013. – Vol. 3. – P. 993-1022.

51 Jelodar H., Wang Y. et al. Latent Dirichlet allocation (LDA) and topic modelling: models, applications, a survey // Multimedia Tools and Applications. – 2018. – Vol. 78(5). – P. 1-43.

52 Mimno D., Wallach H., Talley E. et al. Optimizing Semantic Coherence in Topic Models // Proceed. conf. on Empirical Methods in Natural Language Processing (EMNLP 2011). – Edinburgh, 2011. – P. 262-272.

53 Barakhnin V.B. et al. Methods to identify the destructive information // Journal of Physics, IOP Publishing, Bristol, UK. – 2019. – Vol. 1405(1). – P. 1-9.

54 Mukhamediev R.I., Mustakayev R., Yakunin K. et al. Multi-Criteria Spatial Decision Making Support system for Renewable Energy Development in Kazakhstan // IEEE Access. – 2019. – Vol. 7. – P. 122275-122288.

55 Kitchin R., McArdle G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets // Big Data & Society. – 2016. – Vol. 3(1). – P. 1-10.

56 Scott J. A decision support system for supplier selection and order allocation in stochastic, multi-stakeholder and multi-criteria environments // International J. of Production Economics. – 2015. – Vol. 166. – P. 226-237.

57 Mardani A. Sustainable and renewable energy: An overview of the application of multiple criteria decision-making techniques and approaches // Sustainability. – 2015. – Vol. 7, №10. – P. 13947-13984.

58 Wanderer T., Stefan H. Creating a spatial multi-criteria decision support system for energy related integrated environmental impact assessment // Environmental Impact Assess. – 2015. – Vol. 52. – P. 2-8.

59 Hoceini Y., Mohamed C., Moncef A. Towards a new approach for disambiguation in NLP by multiple criterion decision-aid // The Prague Bull. of Mathematical Linguistics. – 2011. – Vol. 95. – P. 19-32.

60 Kumar A. A review of multi criteria decision making (MCDM) towards sustainable renewable energy development // Renewable and Sustainable Energy Rev. – 2017. – Vol. 69. – P. 596-609.

61 Yager R. On ordered weighted averaging aggregation operators in multi criteria decision making // IEEE Transactions on systems, Man, and Cybernetics. – 1988. – Vol. 18, №1. – P. 183-190.

62 Hansen P., Franz O. A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives // J. of Multi-Criteria Decis. Analysis. – 2008. – Vol. 15, №3-4. – P. 87-107.

63 Figueira J., Vincent M., Bernard R. ELECTRE methods // In book: Multiple criteria decision analysis: State of the art surveys. – New York, 2015. – P. 133-153.

64 Lai Y., Ting-Yun L., Ching-Lai H. Topsis for MODM // European J. of Operational Res. – 1994. – Vol. 76, №3. – P. 486-500.

- 65 Brans J. A Preference Ranking Organization Method: (The PROMETHEE Method for Multiple Criteria Decision-Making) // *Management Science*. – 1985. – Vol. 31, №6. – P. 647-656.
- 66 Saaty, T. Group decision making and the AHP // In book: *The Analytic Hierarchy Process* – NY.: Springer Verlag, 1989. – P. 59-67.
- 67 Charabi Y., Adel G. PV site suitability analysis using GIS-based spatial fuzzy multi-criteria evaluation // *Renewable Energy*. – 2011. – Vol. 36, №9. – P. 2554-2561.
- 68 Abaei M. Developing a novel risk-based methodology for multi-criteria decision making in marine renewable energy applications // *Renewable Energy*. – 2017. – Vol. 102. – P. 341-348.
- 69 Mukhamediev R.I., Mustakayev R., Yakunin K. et al. Multi-Criteria Spatial Decision Making Support System for Renewable Energy Development in Kazakhstan // *IEEE Access*. – 2019. – Vol. 7. – P. 122275-122288.
- 70 Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference // *Journal of Approximate Reasoning*. – 2014. – Vol. 55, №4. – P. 1072-1092.
- 71 Подвесовский А.Г., Ешин С.В. Применение байесовых сетей в задачах анализа и прогнозирования спроса // *Вестник БГТУ*. – 2011. – №1. – С. 244-254.
- 72 Saaty T.L. Group decision Making and the AHP // In book: *The Analytic Hierarchy Process*. – NY.: Springer Verlag, 1989. – P. 60-65.
- 73 Ospanova U., Atanayeva M., Akoyeva I. et al. Informative features of bias and reliability of electronic Mass Media // *Sociologia*. – 2019. – Vol. 2. – P. 259-270.
- 74 Atanayeva M., Buldybayev T., Ospanova U. et al. Methodology for determining informative features of news texts and checking their significance // *Scientific aspect*. – 2019. – Vol. 3, №3. – P. 277-296.
- 75 Mukhamediev R.I., Musabayev R.R., Buldybayev T. et al. Experiments to evaluate mass media based on the thematic model of the text corpus // *Cloud of Science*. – 2020. – Vol. 7, №1. – P. 87-104.
- 76 Yakunin K., Ionescu G.M., Murzakhmetov S. et al. Propaganda Identification Using Topic Modelling // *Proceed. 9th internat. Young Scientist conf. on Computational Science (YSC)*. – Heraklion, 2020. – P. 1-8.
- 77 Yakunin K., Mukhamediev R., Mussabayev R. et al. Mass Media Evaluation Using Topic Modelling // *Proceed. internat. conf. on Digital Transformation and Global Society*. – Cham: Springer, 2019. – P. 130-135.
- 78 Барахнин В.Б. и др. Проектирование структуры программной системы обработки корпусов текстовых документов // *Бизнес-информатика*. – 2019. – Т. 13, №4. – С. 60-72.
- 79 Pang B., Lee L., Vaithyanathan Sh. Thumbs up?: Sentiment Classification using Machine Learning Techniques // *EMNLP*. – Stroudsburg, PA, 2002. – P. 79-86.
- 80 Choi Y., Cardie Cl., Riloff E., Patwardhan S. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns // *Proceed. of the Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*. – Vancouver, 2005. – P. 355-362.

81 Manning C.D. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? // Proceed. of the 12th internat. conf. “Computational Linguistics and Intelligent Text Processing” (CICLing 2011). – New Delhi, 2012. – P. 171-189.

82 Mukhamedyev R. et al. Assessment of the dynamics of publication activity in the field of natural language processing and deep learning // Proceed. of the internat. conf. on Digital Transformation and Global Society. – SPb.; Cham: Springer, 2019. – P. 130-135.

83 Tarasov D.S. Deep Recurrent Neural Networks for Multiple Language Aspect-Based Sentiment Analysis // Proceed. of the internat. conf. “Dialogue-2015”. – M., 2015. – P. 65-74.

84 García-Moya L., Anaya-Sanchez H., Berlanga-Llavori R. Retrieving product features and opinions from customer reviews // IEEE Intelligent Systems. – 2013. – Vol. 28(3). – P. 19-27.

85 Mavljutov R.R., Ostapuk N.A. Using basic syntactic relations for sentiment analysis // Proceed. of the internat. conf. “Dialogue-2013”. – Bekasovo, 2013. – P. 101-110.

86 Apache Airflow Documentation // <https://airflow.apache.org/> 03.09.2019.

87 Kazakhstani news corpus for social significance identification with topic modelling results // <https://data.mendeley.com/datasets/hwj24p9gkh/1>. 21.08.2021.

88 Barakhnin V., Kozhemyakina O., Yakunin K. et al. The design of the structure 710 of the software system for processing text document corpus // Business-informatics. – 2019. – Vol. 13, №4. – P. 60-72.

89 Yakunin, K. This repo presents data illustrating results obtained by applying Multi Model Mass Media Assessment (M4A) to a corpora of news publication from Kazakhstan media // <https://github.com/KindYAK/M4A-Data>. 14.09.2020.

90 Peters M., Ruder S., Smith N. To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks // Proceed. of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). – Florence, 2019. – P. 7-14

91 Якунин К.О., Мусабаев Р.Р., Елис М.С. и др. Тема энергетики в новостных публикациях // Матер. всеросс. науч. конф. с междунар. уч. и 12-й науч. молодеж. школы «Возобновляемые источники энергии». – М., 2020. – С. 451-460.

92 Якунин К., Красовицкий А.М., Уалиева И.М. и др. Анализ новостных тематических трендов в сфере информационной безопасности // Матер. междунар. науч.-практ. конф. «Актуальные проблемы информационной безопасности в Казахстане». – Алматы, 2020. – С. 247-254.



# ПРИЛОЖЕНИЕ А

## Акт внедрения

Приложение 1  
к приказу Министра  
по инвестициям и развитию  
Республики Казахстан  
от 14 ноября 2018 года № 791  
форма

### **Акт внедрения результатов научно-исследовательских, научно-технических работ, (или) результатов научной и (или) научно-технической деятельности**

*1. Наименование научно-исследовательских, научно-технических работ и (или) результатов научной и (или) научно-технической деятельности:*

Проект ПЦФ BR05236839 «Разработка информационных технологий и систем для стимулирования устойчивого развития личности как одна из основ развития цифрового Казахстана».

*2. Краткая аннотация:*

В рамках реализации ПЦФ разработан программный комплекс системы оценки влияния открытых текстовых информационных источников на социум, предназначенный для обработки большого (более миллиона единиц) корпуса текстовых документов, включая, в частности, реализацию автоматизированной обработки корпуса текстов, возможность распараллеливания вычислений и, включающая подсистему обработки данных на основе программного решения Apache Airflow, несколько типов хранилищ, обеспечивающих быстрый доступ к компонентам системы, и подсистему для построения аналитических отчетов, которая генерируется в приложении Python Django с использованием полноценной библиотеки визуализации.

Для внедрения выбран разработанный функционал системы, который выполнен в рамках раздела календарного плана проекта №14 «Разработка аналитических инструментов в рамках информационной системы оценки влияния открытых текстовых информационных источников на социум».

Внедряемый функционал включает:

1. Дашборд освещения новостей сферы образования и науки в СМИ и социальных сетях с графиками тональности (негативные/нейтральные/позитивные), с динамикой тональности за весь период (охвачены статьи с 2006 года), с динамикой тональности за последние 35 дней, с графиком тональности (в разрезе негатива, нейтральности и позитива) распределенная по СМИ и социальным сетям за последние 15 дней, с последними актуальными новостями, с главными позитивными и негативными новостями за последние 15 дней, с главными топиками (темы обсуждения) по положительному влиянию и по отрицательному влиянию.

2. Дашборд упоминаемости в новостях руководящих должных лиц МОН РК по таким показателям как тональность и освещенность.

3. Автоматическое формирование отчета «Мониторинг восприятия вопросов системы образования и науки на интернет порталах СМИ» за

определенный период. Отчет включает следующие параметры: 1) распределение позитивных, нейтральных и негативных публикаций за период; 2) динамика средних показателей различных критериев за период, в частности: тональности; освещённости заданных тем/гос. программ/событий; социальной значимости; 3) распределение негативных, позитивных и нейтральных сообщений по СМИ, отсортированные по содержанию негатива; 4) список главных новостей по позитиву/негативу, либо другим критериям за период.

Разработанные аналитические инструменты, такие как алгоритмы тематического моделирования, численной оценки в публикациях тональности (негативные, нейтральные, позитивные), социальной значимости, определения геолокации в пределах Республики Казахстан имеют свою уникальность, что подтверждается публикациями в рецензируемых научных журналах и материалах конференций, входящих в базы данных Scopus и web of Science:

1) Kirill Yakunin etc. Mass Media Evaluation Using Topic Modelling //International Conference on Digital Transformation and Global Society. – Springer, Cham, 2019. – С.130-135. (in Communications in Computer and Information Science, Scopus Q3, 31%, CiteScore = 0.7).

2) Yakunin K., Ionescu G.M., Murzakhmetov S., Mussabayev R., Filatova O., Mukhamediev R. Propaganda Identification Using Topic Modelling // Procedia 9th International Young Scientist Conference on Computational Science (YSC 2020).

3) Kirill Yakunin etc. Classification of negative publication in mass media using topic modeling// Journal of Physics: Conf. Series (Scopus (17%), CiteScore =0.7).

4) Ospanova U., Baimakhanbetov M., Buldybayev T., Akoyeva I., Nurumov K., Atanayeva M. The effect of data triangulation on performance of K-nearest neighbors and Naïve Bayes ML algorithms // In: Sukhomlin V., Zubareva E. (eds.) Proc. of the 4th Int. sc. conf. "Convergent Cognitive Information Technologies". – Convergent 2019. Communications in Computer and Information Science. Springer, Cham. – 2020.

5) Yakunin K., Mukhamediev R., Mussabayev R., Buldybayev T., Kuchin Ya., Murzakhmetov S., Yunussov R., Ospanova U. Mass Media Evaluation Using Topic Modelling // Digital Transformation and Global Society. – 2020 (Q3, SJR 0.17).

*3. Эффект от внедрения (экономический, социальный, экологический), подчеркнуть область эффекта):*

Экономический эффект от внедрения заключается в сокращении времени на обработку, анализ и формирование аналитического отчета публикаций СМИ и в социальных сетях в области образования и науки. В настоящее время отчет формируется в среднем за 2 минуты, благодаря технологиям Big Data, что позволяет сократить трудозатраты. Если вручную 2 сотрудника тратят в среднем минимум 16 часов на поиск информации,

обработку, анализ и формирование отчета, то система это сделает в среднем за 2 минуты и сформирует по заданным параметрам аналитический отчет.

Социальный эффект от внедрения заключается в быстром реагировании МОН РК на происходящие события, так как система каждый час собирает публикации в СМИ и социальных сетях, анализирует и выдает социально-значимую информацию на дашборд.

*4. Место и время внедрения:*

Министерство образования и науки Республики Казахстан для деятельности в области мониторинга информированности в области образования и науки.

Время внедрения: август-октябрь 2020 года

*5. Форма внедрения:*

Предоставление доступа к дашборду аналитической системы и регулярная выгрузка отчетов по мониторингу восприятия вопросов системы образования и науки.

Подписи:

1. Представитель/представители заявителя (налогоплательщик), внедривший результаты научно-исследовательских, научно-технических работ и (или) результаты научной и (или) научно-технической деятельности

Вице-министр образования и науки РК,  
**Бигари Рустем Айдарбекулы**



(подпись)

2. Представитель/представители организации исполнителя научно-исследовательских, научно-технических работ и (или) научной и (или) научно-технической деятельности (внедренных)

Заведующий лабораторией  
«Анализ и моделирования информационных  
Процессов» ИИВТ,  
**Мусабаев Рустам Рафикович**



(подпись)

Генеральный директор  
РГП на ПХВ ИИВТ КН МОН РК,  
**Калимолдаев Максат Нурадилович**

Министерство образования и науки Республики Казахстан  
Комитет науки  
РГП «ИНСТИТУТ ИНФОРМАЦИОННЫХ И ВЫЧИСЛИТЕЛЬНЫХ  
ТЕХНОЛОГИЙ»  
(ИИВТ)

МРНТИ 20.53.00  
УДК 004.94  
№ гос.регистрации 0118РК01201  
Инв. №

УТВЕРЖДАЮ  
Генеральный директор  
РГП на ПХВ ИИВТ КН МОН РК  
академик НАН РК, д.ф.-м.н., профессор  
 М.Н. Калимолдаев  
« 14 » 10 2020 г.  


ОТЧЕТ  
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ  
по теме  
РАЗРАБОТКА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И СИСТЕМ ДЛЯ  
СТИМУЛИРОВАНИЯ УСТОЙЧИВОГО РАЗВИТИЯ ЛИЧНОСТИ КАК ОДНА ИЗ  
ОСНОВ РАЗВИТИЯ ЦИФРОВОГО КАЗАХСТАНА  
(заключительный)  
Шифр программы О.0861


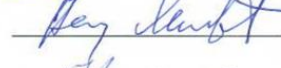
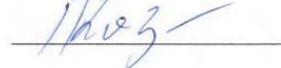



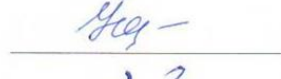











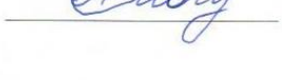
Научный руководитель НИР, к.т.н.



Р.Р. Мусабаев

Алматы, 2020

## СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель темы, канд. техн. наук		Р.Р. Мусабаев (Введение, Разделы 1 – 9, Заключение)
Гл. науч. сотр., PhD		Н. Младенович (Раздел 9)
Гл. науч. сотр., д-р техн. наук		Г.З. Казиев (Раздел 9)
Вед. науч. сотр., PhD		Р.И. Мухамедиев (Разделы 1, 4, 5, 6, 9)
Вед. науч. сотр., д-р техн. наук		В.Б. Барахнин (Разделы 1, 4, 5, 6, 9)
Вед. науч. сотр., д-р техн. наук		А.А. Хорошилов (Раздел 4)
Вед. науч. сотр., канд. физ.-мат. наук		И.М. Уалиева (Разделы 4, 8, 9)
Вед. науч. сотр., PhD		А.М. Красовицкий (Разделы 1, 4, 8, 9)
Ст. науч. сотр., Канд. филолог. наук		О.Ю. Кожемякина (Раздел 1, 4, 5, 6, 9)
Науч. сотр.		В.М. Ибраева (Разделы 5, 6, 7, 9 Заключение)
Науч. сотр.		А.А. Ашимов (Раздел 9)
Мл. науч. сотр.		Ж. Мейрамбеккызы (Разделы 1, 2, 4, 5, 6, 9)
Мл. науч. сотр.		Н. Тасболатулы (Разделы 4)
Математик- программист		О.Б. Козбагаров (Раздел 9)
Математик- программист		А. Жангабылова (Раздел 9)
Инженер- программист		И.Р. Ахметов (Разделы 4, 6, 8, 9)
Инженер- программист		К. Якунин (Разделы 1, 4-9)
Инженер- программист		Я. Кучин (Разделы 1, 4, 8)
Инженер- программист		А. Сымагулов (Разделы 1, 4, 8)

## ПРИЛОЖЕНИЕ Б

### Листинг функций, реализующих метод ММА calc\_mma.py

```
def calc_mma(**kwargs):
    from nlpmonitor.settings import ES_CLIENT, ES_INDEX_TOPIC_MODELLING
    from elasticsearch_dsl import Search
    import numpy as np
    from .util import calc_p1, calc_p2, calc_p4, calc_p5, calc_p6,
    create_delete_index, bulk_factory

    import logging
    es_logger = logging.getLogger('elasticsearch')
    es_logger.setLevel(logging.ERROR)

    topic_modelling_name = kwargs['topic_modelling_name']
    criterion_ids = kwargs['criterion_ids']
    criterion_weights = kwargs['criterion_weights']
    class_ids = kwargs['class_ids']
    perform_actualize = kwargs['perform_actualize']

    tm = Search(using=ES_CLIENT, index=ES_INDEX_TOPIC_MODELLING).filter('term',
**{'name': topic_modelling_name}) \
        .source(['number_of_topics']).execute()[0]
    topics_number = tm.number_of_topics
    p1_matrix = calc_p1(topic_modelling_name=topic_modelling_name,
                        criterion_ids=criterion_ids,
                        topics_number=topics_number)
    p2_matrix, document_es_guide =
    calc_p2(topic_modelling_name=topic_modelling_name,
            topics_number=topics_number)
    p3_matrix = np.array(list(a for a in zip(*criterion_weights))).reshape(-1,
len(criterion_weights))
    p4_matrix = calc_p4(p1=p1_matrix, p3=p3_matrix) # prob x weight
    p5_matrix = calc_p5(p2=p2_matrix, p4=p4_matrix) # weight x prob
    p6_matrix = calc_p6(p1=p1_matrix, p2=p2_matrix) # weight x prob

    index_kwargs = {
        'perform_actualize': perform_actualize,
        'topic_modelling_name': topic_modelling_name,
        'scored_documents': p6_matrix,
        'is_criterion': True,
        'crit_or_class_ids': criterion_ids,
        'document_es_guide': document_es_guide
    }

    create_delete_index(**index_kwargs)
    bulk_factory(**index_kwargs)

    index_kwargs['crit_or_class_ids'] = class_ids
    index_kwargs['is_criterion'] = False
    index_kwargs['scored_documents'] = p5_matrix
    create_delete_index(**index_kwargs)
    bulk_factory(**index_kwargs)

    return 'Done'
```

#### util.py

```
def calc_p1(topic_modelling_name, criterion_ids, topics_number):
    from sklearn.preprocessing import MinMaxScaler
```

```

from datetime import datetime
from evaluation.models import TopicIDEval
from collections import defaultdict, OrderedDict
import numpy as np

scaler = MinMaxScaler(feature_range=(0, 1))
print('!!! p1 matrix calculating started', datetime.now())
p1_matrix = None
for crit_id in criterion_ids:
    column = defaultdict(list, {i: [0] for i in range(topics_number)})
    evals = TopicIDEval.objects.filter(topics_eval__criterion_id=crit_id,
topic_id__topic_modelling_name=topic_modelling_name)
    if not evals:
        continue

    for eval_ in evals:
        value = eval_.topics_eval.value
        topic_id = int(eval_.topic_id.topic_id.split('_')[1])
        column[topic_id].append(value)

    for key, value in column.items():
        column[key] = np.mean(value)

    ordered_column = OrderedDict(sorted(column.items()))
    if crit_id == 1:
        for key, value in ordered_column.items():
            ordered_column[key] = -value

    column =
scaler.fit_transform(np.array(list(ordered_column.values())).reshape(-1, 1))
    if p1_matrix is None:
        p1_matrix = column
        continue
    p1_matrix = np.hstack((p1_matrix, column))
print('!!! p1 matrix calculated', p1_matrix.shape, datetime.now())
return p1_matrix

def calc_p2(topic_modelling_name, topics_number):
    from elasticsearch_dsl import Search
    from nlpmonitor.settings import ES_CLIENT, ES_INDEX_TOPIC_MODELLING,
ES_INDEX_TOPIC_DOCUMENT
    from datetime import datetime
    import numpy as np
    from collections import defaultdict

    print('!!! p2 matrix calculating started', datetime.now())

    theta = Search(using=ES_CLIENT,
index=f'{ES_INDEX_TOPIC_DOCUMENT}_{topic_modelling_name}') \
        .source(['document_es_id', 'datetime', 'document_source',
'topic_weight', 'topic_id'])

    theta_dict = defaultdict(list)
    document_es_ids = dict()
    total = theta.count()
    for i, t in enumerate(theta.scan()):
        if i % 10_000_000 == 0:
            print(f'!!! {i}/{total} thetas passed in dict creating')
            theta_dict[t.document_es_id].append([t.topic_id, t.topic_weight])
        if t.document_es_id not in document_es_ids.keys():
            document_es_ids[t.document_es_id] = {'datetime': getattr(t,

```

```

"datetime", None),
'document_source': getattr(t,
'document_source', None)}

total = len(document_es_ids)
p2_matrix = np.zeros((total, topics_number))
for i, document_id in enumerate(document_es_ids):
    if i % 100_000 == 0:
        print(f'!!! {i}/{total} documents passed in p2 matrix creating')
    column = np.zeros(topics_number)
    for topic_doc in theta_dict[document_id]:
        id_in_column = int(topic_doc[0].split('_')[1])
        column[id_in_column] = topic_doc[1]
    p2_matrix[i] = column
print('!!! p2 matrix calculated', p2_matrix.shape, datetime.now())
return p2_matrix, document_es_ids

def calc_p4(p1, p3):
    from datetime import datetime
    print('!!! p4 matrix calculating started', datetime.now())
    """Вероятность совпадения тематики и класса: p4[k][c]"""
    p4 = custom_dot(matrix_1=p3.T, matrix_2=p1.T, agg_type='bayes')
    print('!!! p4 matrix calculated', p4.T.shape, datetime.now())
    return p4.T

def calc_p5(p4, p2):
    from datetime import datetime
    print('!!! p5 matrix calculating started', datetime.now())
    """Распределение вероятностей статей по классам: p5 [m][c]"""
    p5 = custom_dot(matrix_1=p2, matrix_2=p4, agg_type='bayes')
    print('!!! p5 matrix calculated', p5.shape, datetime.now())
    return p5

def calc_p6(p1, p2):
    from datetime import datetime
    print('!!! p6 matrix calculating started', datetime.now())
    """Распределение статей по признакам"""
    p6 = custom_dot(matrix_1=p2, matrix_2=p1, agg_type='bayes')
    print('!!! p6 matrix calculated', p6.shape, datetime.now())
    return p6

def create_delete_index(**kwargs):
    from nlpmonitor.settings import ES_CLIENT, ES_INDEX_DOCUMENT_EVAL
    from mainapp.documents import DocumentEval
    from util.util import shards_mapping
    from elasticsearch_dsl import Index

    crit_or_class_ids = kwargs['crit_or_class_ids']
    is_criterion = kwargs['is_criterion']
    perform_actualize = kwargs['perform_actualize']
    topic_modelling_name = kwargs['topic_modelling_name']
    scored_documents = kwargs['scored_documents']

    for crit_id in crit_or_class_ids:
        if not perform_actualize:
            es_index =
Index(f"{ES_INDEX_DOCUMENT_EVAL}_{topic_modelling_name}_{crit_id}{'_m4a' if
is_criterion else '_m4a_class'}", using=ES_CLIENT)
            es_index.delete(ignore=404)

```



```

        if not
ES_CLIENT.indices.exists(f"{ES_INDEX_DOCUMENT_EVAL}_{topic_modelling_name}_{crit
_id}{'_m4a' if is_criterion else '_m4a_class'}"):
    settings = DocumentEval.Index.settings
    settings['number_of_shards'] =
shards_mapping(scored_documents.shape[0])

ES_CLIENT.indices.create(index=f"{ES_INDEX_DOCUMENT_EVAL}_{topic_modelling_name}
_{crit_id}{'_m4a' if is_criterion else '_m4a_class'}", body={
    "settings": settings,
    "mappings": DocumentEval.Index.mappings
    }
    )

def bulk_factory(**kwargs):
    from elasticsearch.helpers import parallel_bulk
    from nlpmonitor.settings import ES_CLIENT, ES_INDEX_DOCUMENT_EVAL
    from datetime import datetime

    crit_or_class_ids = kwargs['crit_or_class_ids']
    scored_documents = kwargs['scored_documents']
    is_criterion = kwargs['is_criterion']
    topic_modelling_name = kwargs['topic_modelling_name']
    perform_actualize = kwargs['perform_actualize']
    document_es_guide = kwargs['document_es_guide']
    print(f"!!! start elastic sending for {"criteria" if is_criterion else
"classes"}", datetime.now())
    for i, ids in enumerate(crit_or_class_ids):
        total_created = 0
        failed = 0
        success = 0
        for ok, result in parallel_bulk(ES_CLIENT,
document_eval_generator(class_crit=scored_documents,
document_guide=document_es_guide,
enum_id=i),
index=f"{ES_INDEX_DOCUMENT_EVAL}_{topic_modelling_name}_{ids}{'_m4a' if
is_criterion else '_m4a_class'}",
                                chunk_size=10000 if not
perform_actualize else 500, raise_on_error=True,
                                thread_count=4):
            if (failed + success) % 100_000 == 0:
                print(f"!!!{failed+success}/{len(scored_documents)} processed",
datetime.now())
            if failed > 5:
                raise Exception("Too many failed ES!!!")
            if not ok:
                failed += 1
            else:
                success += 1
                total_created += 1

def document_eval_generator(class_crit, document_guide, enum_id):
    from mainapp.documents import DocumentEval

    for i, document_es_id in enumerate(document_guide.keys()):
        bayes_value = class_crit[i, enum_id]
        doc = DocumentEval(value=bayes_value,
                            document_es_id=document_es_id,

```

```

document_datetime=document_guide[document_es_id]['datetime'],
document_source=document_guide[document_es_id]['document_source'])

    yield doc.to_dict()

def bayes(values):
    """
    :param values:
    :return:
    """
    hypothesis = 0.5
    for val in values:
        hypothesis = val * hypothesis / (val * hypothesis + (1 - val) * (1 - hypothesis))
    return hypothesis

def custom_dot(matrix_1, matrix_2, agg_type='mean'):
    import numpy as np
    """
    1.берем строку m1 берем столбец m2
    2.попарное умножение со "стагиванием"
    стягивание это - оценка значений столбца и строки (столбец это вероятность,
    строка это вес)
    логика стягивания - threshold = 0.5, (P-0.5) * w + 0.5
    3.логика агрегации вероятностей ??? среднее
    4.шкалирование по матрице ,если меньше 0.5 одна своя шкала, если 0.5 то
    другая своя шкала
    """
    new_matrix_rows = matrix_1.shape[0]
    new_matrix_cols = matrix_2.shape[1]
    new_matrix = np.zeros(shape=(new_matrix_rows, new_matrix_cols))
    for index, weights in enumerate(matrix_1):
        for col in range(new_matrix_cols):
            probs = matrix_2[:, col]
            assert len(probs) == len(weights)
            values = [(p - 0.5) * w + 0.5 + 2 ** -20 for p, w in zip(probs,
weights)] # 2 ** -20 bcs of prob 0 issue
            if agg_type == 'mean':
                new_matrix_element = np.mean(values)
            elif agg_type == 'bayes':
                new_matrix_element = bayes(values)
            else:
                new_matrix_element = sum(values)

            new_matrix[index, col] = new_matrix_element

    if agg_type == 'bayes':
        return new_matrix

min_low, max_low, min_up, max_up = 1, 0.5, 1, 0.5

for row in new_matrix:
    for col_elem in row:
        if col_elem < 0.5:
            if col_elem < min_low:
                min_low = col_elem
            if col_elem > max_low:
                max_low = col_elem
        else:

```

```

        if col_elem < min_low:
            min_low = col_elem
        if col_elem > max_low:
            max_low = col_elem

    for index, row in enumerate(new_matrix):
        for col, col_elem in enumerate(row):
            if col_elem <= 0.5:
                new_matrix[index, col] = 0.5 * (col_elem - min_low) / (max_low -
min_low)
            else:
                new_matrix[index, col] = 0.5 * (col_elem - min_up) / (max_up -
min_up) + 0.5

    return new_matrix

```

## ПРИЛОЖЕНИЕ В

### Пример отчета по мониторингу СМИ по теме "Образование"

#### Мониторинг восприятия вопросов системы образования на интернет-порталах СМИ

1. Для каждого топика исследуются периодичность/сезонность изменения его популярности (если такая периодичность есть). Предполагаем, например, что событие по теме X вызывает резонанс R, и такой подъём интереса длится D дней.

2. Топик – совокупность слов, характеризующих определённое множество статей.

3. Прогнозируемая резонансность – средний пиковый резонанс такого периода подъёма интереса к данной тематике. Для 1 и 2 - значения больше нуля означают, что показатель выше среднего по корпусу, значения меньше нуля — что ниже среднего по корпусу (сигма нормального распределения).

4. Прогнозируемая продолжительность – средняя продолжительность одного цикла исходя из исторических данных.

5. Тренд - текущее направление изменения популярности (растёт или спадает сейчас интерес к этой теме).

6. Вес - относительный объём топика.

1. Период мониторинга: январь 2018 г. – декабрь 2019 г.

2. Источники мониторинга: 34 информационных интернет-портала СМИ (tengrinews, nur, 365info и др.).

3. Мониторинг тем:

- ЕНТ;
- стипендии;
- курсы повышения квалификации;
- обновленное содержание образования;
- учебник.

# 1. Освещение ЕНТ



Рисунок В.1 – Динамика восприятия ЕНТ

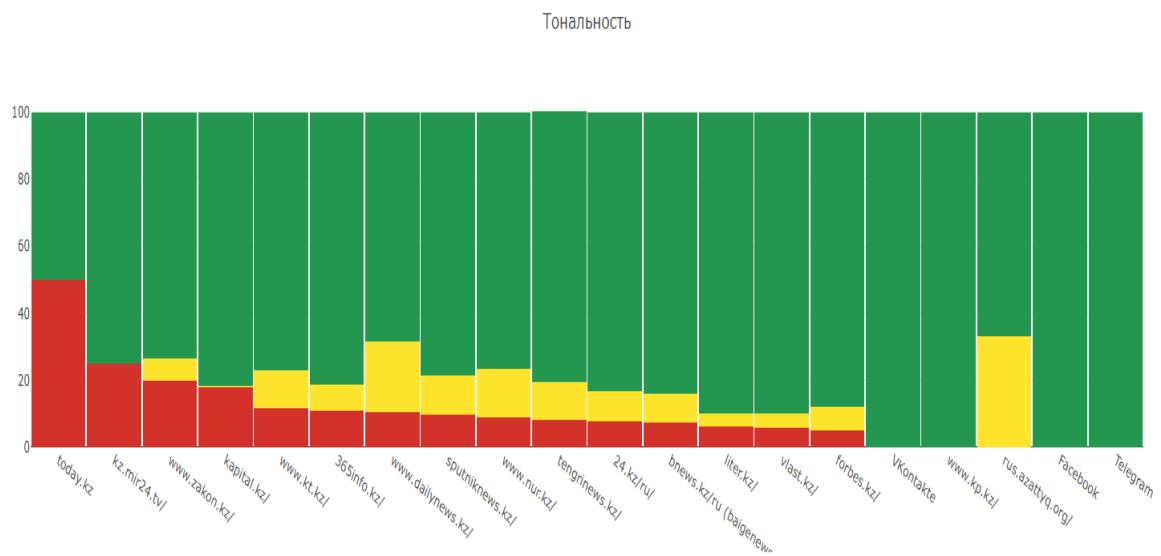


Рисунок В.2 – Восприятие ЕНТ в разрезе источников

Таблица В.1 – Топ негативных новостей

Тональность	Дата	Заголовок	Источник	Гиперссылка
-0,846	2019-10-12T06:06:00+00:00	Бывший замглавы Национального центра тестирования осужден на 7 лет	<a href="https://vlast.kz/">https://vlast.kz/</a>	Ссылка
-0,818	2019-07-03T06:01:00+00:00	Как охраняют тестовые задания ЕНТ	<a href="https://tengrinews.kz/">https://tengrinews.kz/</a>	Ссылка
-0,767	2019-05-16T18:52:00+00:00	В СКО наградили полицейских, спасших школьников из горящего микроавтобуса	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
-0,711	2019-06-25T17:50:00+00:00	Новый руководитель Национального центра тестирования назвал имена отважных женщин, которые под взрывами помогли роженице и организовывали эвакуацию	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
-0,707	2019-05-24T14:31:00+00:00	Прежний и нынешний руководители центра, ответственного за проведение ЕНТ, оказались замешаны в коррупционных делах	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
-0,701	2019-07-03T19:18:00+00:00	Выпускница из СКО мечтает стать следователем	<a href="https://bnews.kz/ru(baigenews.kz)">https://bnews.kz/ru(baigenews.kz)</a>	Ссылка
-0,663	2019-12-10T06:14:00+00:00	Экс-замглавы Национального центра тестирования осужден на 7 лет за взятку	<a href="https://bnews.kz/ru(baigenews.kz)">https://bnews.kz/ru(baigenews.kz)</a>	Ссылка
-0,639	2019-05-16T18:52:00+00:00	В СКО наградили полицейских, спасших школьников из горящего микроавтобуса	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
-0,634	2019-06-24T09:37:00+00:00	ЕНТ-2019: усилен контроль санэпиднадзора	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка
-0,632	2019-07-24T07:50:00+00:00	600 тыс. тенге за помощь в ЕНТ: замдиректоров школ задержали в Атырауской области	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка

Таблица В.2 – Топ позитивных новостей

Тональность	Дата	Заголовок	Источник	Гиперссылка
0,836	2019-07-17T18:52:00+00:00	В Нур-Султане запускается учебное заведение нового формата	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
0,806	2019-07-17T18:52:00+00:00	В Нур-Султане запускается учебное заведение нового формата	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
0,79	2019-07-10T18:52:00+00:00	В Нур-Султане запускается учебное заведение нового формата - школа программирования alem	<a href="https://forbes.kz/">https://forbes.kz/</a>	Ссылка
0,713	2019-07-22T10:13:00+00:00	Выделять на образование по 2 млрд долларов в год предлагают в Казахстане	<a href="https://bnews.kz/ru">https://bnews.kz/ru</a> (baigenews.kz)	Ссылка
0,709	2019-02-01T14:24:00+00:00	Новый закон о статусе учителей будет распространяться и на студентов	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка
0,654	2019-06-18T11:36:00+00:00	294 столичных школьника наградили знаком "Алтын белгі"	<a href="https://bnews.kz/ru">https://bnews.kz/ru</a> (baigenews.kz)	Ссылка
0,64	2019-02-04T18:52:00+00:00	МОН не исключает введения повторной сдачи ЕНТ	<a href="https://vlast.kz/">https://vlast.kz/</a>	Ссылка
0,626	2019-10-25T11:25:00+00:00	Экс-главу центра тестирования обвиняют в получении 140 миллионов тенге	<a href="https://tengrinews.kz/">https://tengrinews.kz/</a>	Ссылка
0,623	2019-08-29T21:40:00+00:00	В Алматы обладателей грантов не зачислят в вуз из-за ошибки	<a href="https://tengrinews.kz/">https://tengrinews.kz/</a>	Ссылка
0,609	2019-08-29T21:32:00+00:00	Ошибка обнаружена в документах обладателей грантов	<a href="https://bnews.kz/ru">https://bnews.kz/ru</a> (baigenews.kz)	Ссылка

### Тональность - главные топики по положительному влиянию

Таблица В.3 – Данный анализ показывает, насколько тематическая область будет популярна

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Тестирование, выпускник, ент, национальный, балл	2,448	146,0 дней (Score: 1,708)	0,417	1
Грант, ЕНТ, балл, экзамен, тестирование	1,706	98,0 дней (Score: 0,480)	3,154	0,662
Педагог, учитель, закон, статус, заработный	Фоновый топик	Фоновый топик		0,155
Конкурс, участник, победитель, участие, команда	0,813	66,5 дней (Score: -0,326)	-1,169	0,099
Военный, оборона, сила, жас, выпускник	-0,113	180,0 дней (Score: 2,579)	-0,336	0,085
Учитель, педагог, школа, праздник, страна	0,881	71,5 дней (Score: -0,198)	1,007	0,085
Образование, министерство, организация, вопрос, учебный	1,808	60,0 дней (Score: -0,493)	9,732	0,056
Школа, ребёнок, выпускной, родитель, образование	-0,009	111,0 дней (Score: 0,813)	0,609	0,056
Ребёнок, саин, Казахстан, аружан, страна	-0,417	118,0 дней (Score: 0,992)	-0,558	0,056
Колледж, специалист, подготовка, кадр, специальность	0,82	72,0 дней (Score: -0,186)	9,812	0,042

Таблица В.4 – Тональность - главные топики по отрицательному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Суд, директор, дело, взятка, тенге	1,368	69,0 дней (Score: -0,262)	5,59	1
Арысь, город, оружие, область, житель	1,247	156,0 дней (Score: 1,964)	3,568	0,667
Полиция, несовершеннолетний, родитель, административный, сообщить	1,326	60,0 дней (Score: -0,493)	6,93	0,5
Вуз, студент, коррупция, университет, работа	Фоновый топик	Фоновый топик		0,417
Подросток, полиция, произойти, задержать, драка	1,941	60,0 дней (Score: -0,493)	3,999	0,25
Школа, анна, ученик, уборка, кластер	-0,364	172,0 дней (Score: 2,374)	1,195	0,25
Ребёнок, школа, больница, произойти, девочка	2,438	54,0 дней (Score: -0,646)	3,657	0,167
Ребёнок, родитель, категория, случай, информация	-0,432	83,5 дней (Score: 0,109)	3,054	0,083
Китай, китайский, страна, власть, Казахстан	-0,613	93,5 дней (Score: 0,365)	-1,388	0,083



## 2. Стипендии

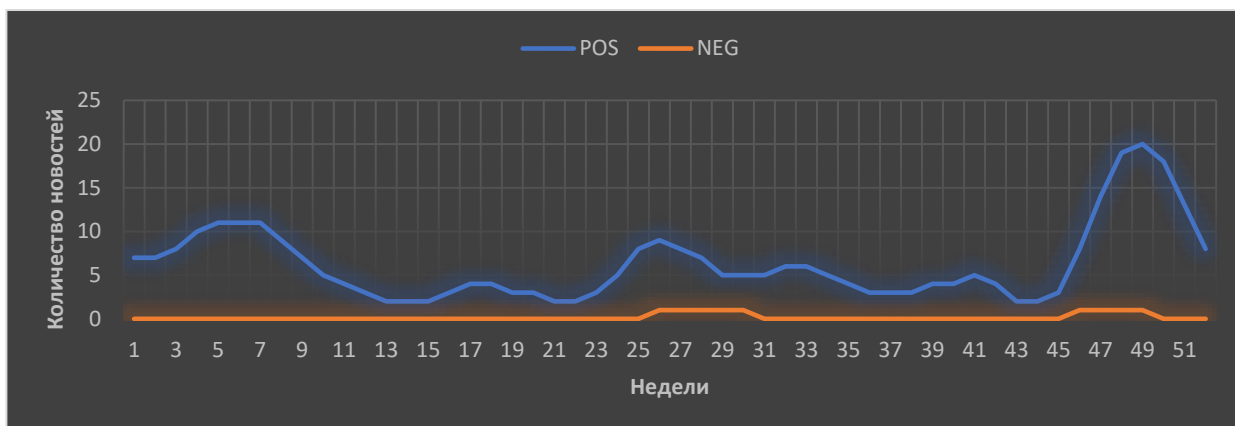


Рисунок В.3 – Динамика восприятия вопросов о стипендиях

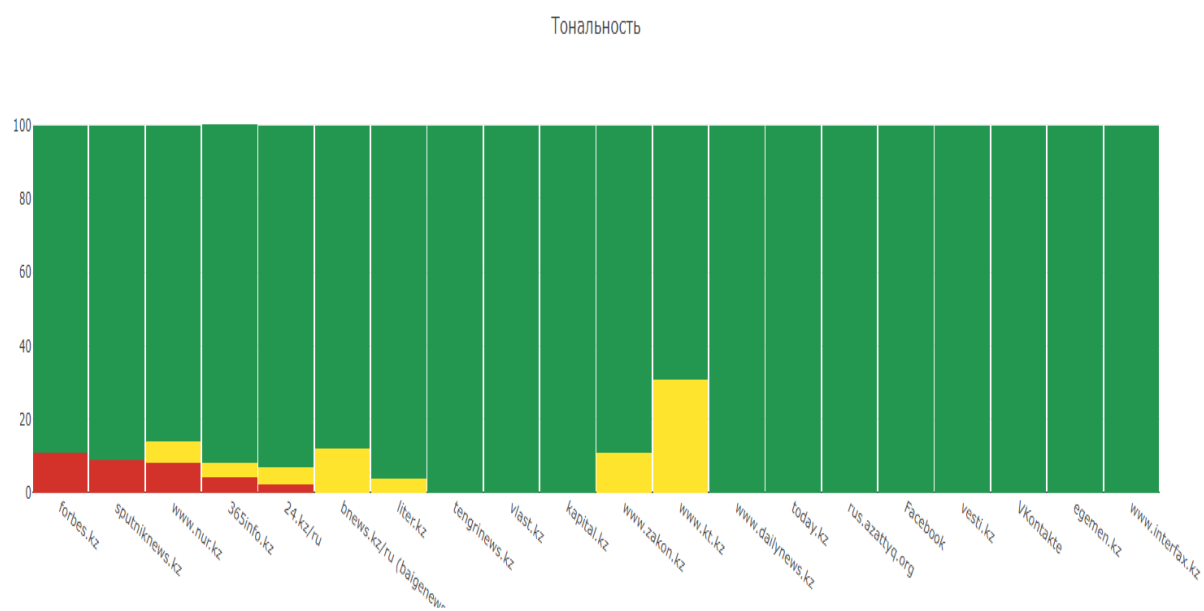


Рисунок В.4 – Восприятие вопросов о стипендиях в разрезе источников

Таблица В.5 – Топ негативных новостей

Тональность	Дата	Заголовок	Источник	Гиперссылка
-0,737	2019-06-06T05:36:00+00:00	Вице-министр Эльмира Суханбердиева арестована на 2 месяца	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка
-0,723	2019-07-05T18:52:00+00:00	40 тыс. долларов стоило госбюджету обучение Эльмиры Суханбердиевой в США	<a href="https://www.nur.kz/">https://www.nur.kz/</a>	Ссылка
-0,722	2019-06-06T11:12:00+00:00	В отношении вице-министра избрана мера пресечения в виде содержания под стражей сроком на два месяца - по 2 августа	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
-0,71	2019-06-05T18:52:00+00:00	Ведётся расследование в отношении вице-министра образования Эльмиры Суханбердиевой	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
-0,552	2019-11-30T04:44:00+00:00	Как девушка из маленького поселка стала стипендиатом «Болашак»	<a href="https://www.nur.kz/">https://www.nur.kz/</a>	Ссылка
-0,479	2019-12-18T12:22:00+00:00	Студента порезали в центре Алматы	<a href="https://365info.kz/">https://365info.kz/</a>	Ссылка
-0,455	2019-07-05T18:52:00+00:00	40 тыс. долларов стоило госбюджету обучение Эльмиры Суханбердиевой в США	<a href="https://www.nur.kz/">https://www.nur.kz/</a>	Ссылка
-0,417	2019-11-22T03:11:00+00:00	Президенты Казахстана и Швейцарии намерены обсудить инвестиционное сотрудничество двух стран в различных сферах	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
-0,387	2019-11-25T05:08:00+00:00	Бывшего чиновника от образования приговорили к реальному сроку лишения свободы за взятку в размере 5 миллионов тенге	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
-0,342	2019-10-11T18:52:00+00:00	МОН о вузах «печатающих» дипломы: Поступают 5 студентов – выпускают 150	<a href="https://www.nur.kz/">https://www.nur.kz/</a>	Ссылка

Таблица В.6 – Топ позитивных новостей

Тональность	Дата	Заголовок	Источник	URL
0,839	2019-02-04T18:52:00+00:00	С 2016 года будет возобновлено обучение специалистов по программе МВА в рамках стипендии "Болашак"	<a href="https://vlast.kz/">https://vlast.kz/</a>	Ссылка
0,828	2019-07-21T18:52:00+00:00	Повысить зарплату учителям за счёт средств Нацфонда предложил Рахим Ошакбаев	<a href="https://forbes.kz/">https://forbes.kz/</a>	Ссылка
0,763	2019-10-11T18:52:00+00:00	Токаев присудил премии ряду ученых	<a href="https://www.nur.kz/">https://www.nur.kz/</a>	Ссылка
0,758	2019-12-10T14:12:00+00:00	Президент РК присудил Госпремии в области науки и техники имени аль-Фараби	<a href="https://www.kt.kz/">https://www.kt.kz/</a>	Ссылка
0,728	2019-12-10T13:44:00+00:00	Самал Еслямовой присудили стипендию Первого Президента РК	<a href="https://bnews.kz/ru">https://bnews.kz/ru</a> (baigenews.kz)	Ссылка
0,723	2019-12-23T10:31:00+00:00	Токаеву показали выведенные с помощью генной инженерии новые сорта пшеницы	<a href="https://bnews.kz/ru">https://bnews.kz/ru</a> (baigenews.kz)	Ссылка
0,722	2019-09-27T09:24:00+00:00	Включать волонтерский опыт в трудовой стаж предлагают в Казахстане	<a href="https://tengrinews.kz/">https://tengrinews.kz/</a>	Ссылка
0,715	2019-12-23T10:20:00+00:00	Токаев обратился к молодым агротехникам	<a href="https://365info.kz/">https://365info.kz/</a>	Ссылка
0,713	2019-07-22T10:13:00+00:00	Выделять на образование по 2 млрд долларов в год предлагают в Казахстане	<a href="https://bnews.kz/ru">https://bnews.kz/ru</a> (baigenews.kz)	Ссылка
0,705	2019-01-24T12:04:00+00:00	Объявление 2019-го Годом молодежи даст толчок для всестороннего развития общества - эксперт	<a href="https://bnews.kz/ru">https://bnews.kz/ru</a> (baigenews.kz)	Ссылка

Таблица В.7 – Тональность - главные топики по положительному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Университет, студент, вуз, программа, обучение	1,315	55,3 дней(Score: - 0,612)	3,441	1
Тенге, тысяча, социальный, миллион, семья	1,569	63,5 дней(Score: - 0,403)	6,724	0,5
Проект, развитие, технология, страна, образование	0,159	62,5 дней(Score: - 0,429)	11,808	0,458
Наука, национальный, академия, научный, университет	-0,106	58,0 дней(Score: - 0,544)	10,389	0,292
Колледж, специалист, подготовка, кадр, специальность	0,82	72,0 дней(Score: - 0,186)	9,812	0,25
Центр, проект, обучение, служба, сотрудник	0,498	170,0 дней(Score: 2,323)	0,589	0,25
Педагог, учитель, закон, статус, заработный	Фоновый топик	Фоновый топик		0,208
Проект, развитие, молодёжь, страна, наш	0,124	96,0 дней(Score: 0,429)	3,255	0,167
Спорт, спортивный, клуб, ребёнок, тренер	0,604	112,0 дней(Score: 0,838)	3,135	0,167
Театр, искусство, еврей, димаш, советский	-0,53	53,0 дней(Score: - 0,672)	3,55	0,167

Таблица В.8 – Тональность - главные топики по отрицательному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Суд, директор, дело, взятка, тенге	1,368	69,0 дней (Score:- 0,262 )	5,59	1
Вуз, студент, коррупция, университет, работа	Фоновый топик	Фоновый топик		0,375
Китай, китайский, страна, власть, казахстан	-0,613	93,5 дней (Score:0,365 )	-1,388	0,25
Подросток, полиция, произойти, задержать, драка	1,941	60,0 дней (Score:- 0,493 )	3,999	0,125

### 3. Учебник

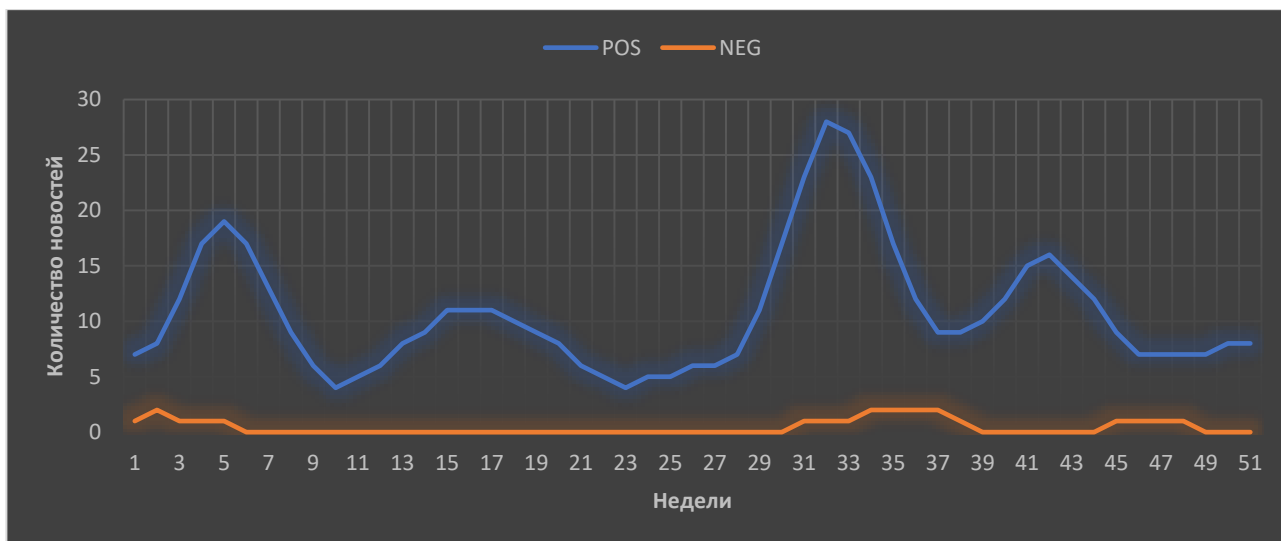


Рисунок В.5 – Динамика восприятия вопросов о учебниках

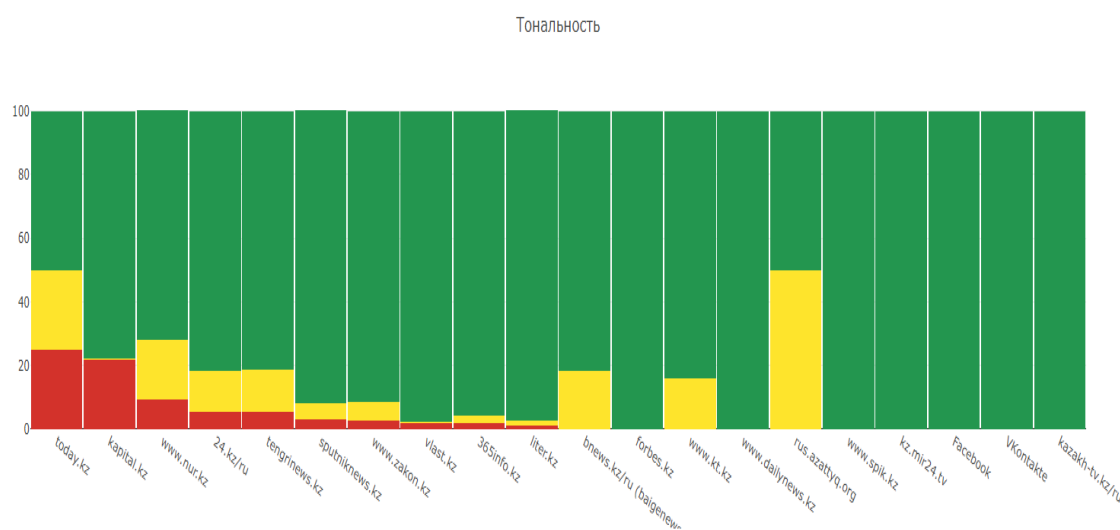


Рисунок В.6 – Восприятие вопросов о учебниках в разрезе источников

Таблица В.9 – Топ негативных новостей

Тональность	Дата	Заголовок	Источник	URL
1	2	3	4	5
-0,65	2019-09-10T11:49:00+00:00	Упал под тяжестью рюкзака: спецылисты взвесили портфели школьников Нур-Султана	<a href="https://tengrinews.kz/">https://tengrinews.kz/</a>	Ссылка
-0,449	2019-05-08T18:52:00+00:00	Нужна работа над ошибками	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
-0,445	2019-01-16T09:52:00+00:00	Что такое кластер в истории	<a href="https://www.nur.kz/">https://www.nur.kz/</a>	Ссылка

Продолжение таблицы В.9

1	2	3	4	5
-0,424	2019-09-12T16:01:00+00:00	Димаш Кудайберген попал в китайский школьный учебник	<a href="https://tengrinews.kz/">https://tengrinews.kz/</a>	Ссылка
-0,387	2019-09-10T11:44:00+00:00	Вес рюкзаков школьников превышает нормативы в 2,5 раза - специалист	<a href="https://www.zakon.kz/">https://www.zakon.kz/</a>	Ссылка
-0,31	2019-08-22T05:41:00+00:00	Ни в школе, ни в областном управлении образования Мангистауской области ситуацию пока не комментируют	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
-0,297	2019-11-25T16:06:00+00:00	Клопы завелись в актауской школе	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка
-0,286	2019-10-09T12:39:00+00:00	Вес рюкзаков школьников превышает нормативы в 2,5 раза в Нур-Султане	<a href="http://today.kz">http://today.kz</a>	Ссылка
-0,252	2019-09-10T12:32:00+00:00	Ученики начальных классов носят рюкзаки вдвое тяжелее нормы	<a href="https://365info.kz/">https://365info.kz/</a>	Ссылка
-0,218	2019-01-18T21:05:00+00:00	Учебная программа Бразилии отстает от мировой на 15 лет	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка

Таблица В.10 – Топ позитивных новостей

Тональность	Дата	Заголовок	Источник	URL
1	2	3	4	5
0,899	2019-09-01T06:36:00+00:00	Президент пожелал учащимся новых открытий, а педагогам - успехов в работе	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
0,875	2019-02-04T18:52:00+00:00	Мажилис одобрил проект закона, направленный на защиту детей от негативной информации	<a href="https://vlast.kz/">https://vlast.kz/</a>	Ссылка
0,775	2019-08-13T09:11:00+00:00	С 1 сентября бумажные учебники в школах можно будет заменить планшетами	<a href="https://www.kt.kz/">https://www.kt.kz/</a>	Ссылка

Продолжение таблицы В.10

1	2	3	4	5
0,741	2019-02-01T15:07:03+00:00	20 ВУЗов модернизируют в Казахстане	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка
0,736	2019-08-30T12:55:00+00:00	Накануне начала учебного года Sputnik Казахстан рассказывает, какой гаджет лучше выбрать для школьника	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
0,728	2019-04-23T18:52:00+00:00	На доступном языке	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
0,726	2019-04-17T07:26:00+00:00	На скольких языках говорят в Казахстане?	<a href="https://365info.kz/">https://365info.kz/</a>	Ссылка
0,717	2019-01-26T09:17:00+00:00	«Тіл біліміне кіріспе» («Введение в языкознание»)   Qazbooks	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка
0,707	2019-04-23T18:52:00+00:00	На доступном языке	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
0,706	2019-03-31T18:53:00+00:00	«Дети богачей смотрят на учителя, как на домработницу» — директор школы	<a href="https://365info.kz/">https://365info.kz/</a>	Ссылка

Таблица В.11 – Тональность - главные топики по положительному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Министр, образование, наука, учебник, асхат	1,932	164,0 дней(Score: 2,169)	5,71	1
Проект, развитие, технология, страна, образование	0,159	62,5 дней(Score: -0,429)	11,808	0,583
Приложение, ребёнок, знать, школа, мобильный	-0,518	98,0 дней(Score: 0,480)	0,453	0,306
Школа, ремонт, робототехника, район, кабинет	-0,026	65,5 дней(Score: -0,352)	0,058	0,278
Язык, казахский, английский, русский, обучение	0,884	125,0 дней(Score: 1,171)	4,564	0,222
Проект, развитие, молодёжь, страна, наш	0,124	96,0 дней(Score: 0,429)	3,255	0,194
Электронный, интернет, система, школа, доступ	0,093	78,0 дней(Score: -0,032)	2,139	0,194
Ребёнок, школьный, питание, лагерь, школьник	1,34	153,0 дней(Score: 1,888)	1,434	0,167
Образование, министерство, организация, вопрос, учебный	1,808	60,0 дней(Score: -0,493)	9,732	0,139
Музей, Сатпаев, История, Выставка, каньш	-0,233	139,0 дней(Score: 1,529)	2,941	0,139

Таблица В.12 – Тональность - главные топики по отрицательному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Школа, анна, ученик, уборка, кластер	-0,364	172,0 дней (Score:2,374)	1,195	1
Ребёнок, мальчик, школа, полиция, дело	-0,571	35,0 дней (Score:-1,133)	1,85	0,667
Китай, китайский, страна, власть, казахстан	-0,613	93,5 дней (Score:0,365)	-1,388	0,667
Конфликт, школа, штат, индия, медиация	-0,456	64,0 дней (Score:-0,390)	1,695	0,333
Ребёнок, родитель, категория, случай, информация	-0,432	83,5 дней (Score:0,109)	3,054	0,333
Ребёнок, видео, воспитатель, женщина, школа	Фоновый топик	Фоновый топик		0,333

#### 4. Обновленное содержания образования

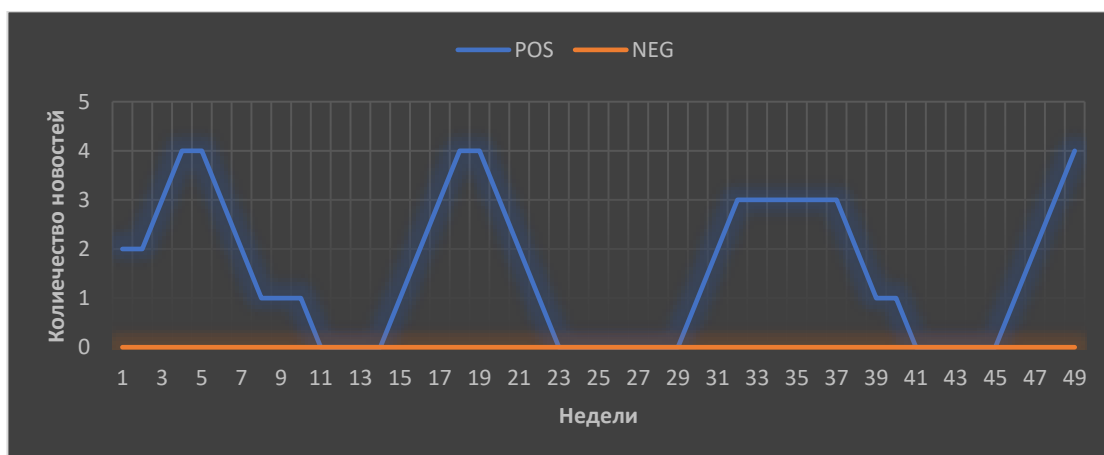


Рисунок В.6 – Динамика восприятия вопросов об обновленном содержании

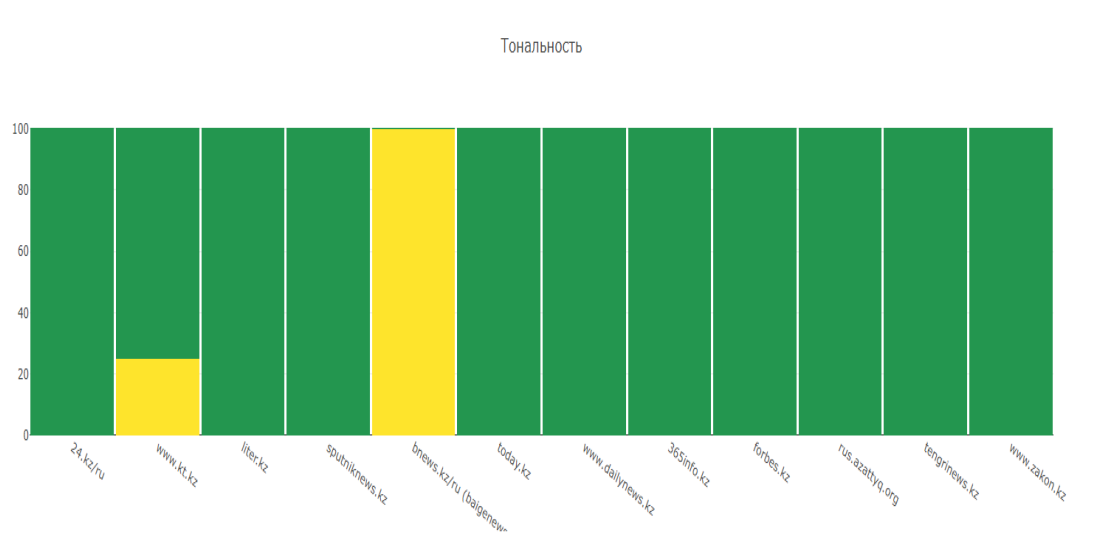


Рисунок В.7 – Восприятие вопросов об обновленном содержании в разрезе источников



Таблица В.13 – Топ позитивных новостей

Тональность	Дата	Заголовок	Источник	URL
0,714	2019-05-05T18:52:00+00:00	В ближайшие три года в РК будет построено 190 школ	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
0,709	2019-12-18T18:52:00+00:00	Расходы на образование и науку в Казахстане к 2025 году вырастут до 7% от ВВП	<a href="http://www.dailynews.kz/">http://www.dailynews.kz/</a>	Ссылка
0,697	2019-01-18T03:07:00+00:00	Какие новшества предусмотрены в законе «О статусе педагога»	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка
0,667	2019-02-04T18:52:00+00:00	Ускорить создание передовой системы образования в Казахстане поручил президент	<a href="https://vlast.kz/">https://vlast.kz/</a>	Ссылка
0,631	2019-01-11T09:45:00+00:00	Система оценки знаний школьников через СОР и СОЧ будет усовершенствована	<a href="http://today.kz">http://today.kz</a>	Ссылка
0,61	2019-02-04T18:52:00+00:00	380 тысяч первоклассников пойдут в школу в этом году – Сагадиев	<a href="https://vlast.kz/">https://vlast.kz/</a>	Ссылка
0,595	2019-08-27T18:52:00+00:00	Сколько будут получать учителя в Жамбылской области	<a href="https://forbes.kz/">https://forbes.kz/</a>	Ссылка
0,577	2019-12-19T10:05:00+00:00	В Казахстане увеличат расходы на образование и науку	<a href="https://www.zakon.kz/">https://www.zakon.kz/</a>	Ссылка
0,572	2019-05-04T07:38:00+00:00	Более 750 учебных заведений РК работают по программе НИШ	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка
0,521	2019-10-10T18:52:00+00:00	Концепция «слышащего государства»	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка



Таблица В.15 – Топ негативных новостей

Тональность	Дата	Заголовок	Источник	URL
-0,237	2019-10-19T18:52:00+00:00	Учителя, которая рассказала о поборах в школе на канале Навального, суд обязал выплатить компенсацию	<a href="https://rus.azattyq.org/">https://rus.azattyq.org/</a>	Ссылка
-0,171	2019-10-19T18:52:00+00:00	Учителя, которая рассказала о поборах в школе на канале Навального, суд обязал выплатить компенсацию	<a href="https://rus.azattyq.org/">https://rus.azattyq.org/</a>	Ссылка
-0,142	2019-11-11T18:52:00+00:00	Воспитательницу детского сада в Нур-Султане уволили за грубое обращение с детьми	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
-0,088	2019-11-11T18:52:00+00:00	Воспитательницу детского сада в Нур-Султане уволили за грубое обращение с детьми	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
-0,086	2019-10-29T13:17:00+00:00	Партия Nur Otan усилит работу по поддержке учителей	<a href="https://www.zakon.kz/">https://www.zakon.kz/</a>	Ссылка
-0,003	2019-11-08T18:52:00+00:00	Более 2 000 новых видеочкамер установят в детсадах Нур-Султана	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка

Таблица В.16 – Топ позитивных новостей

Тональность	Дата	Заголовок	Источник	URL
1	2	3	4	5
0,714	2019-05-05T18:52:00+00:00	В ближайшие три года в РК будет построено 190 школ	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
0,654	2019-12-12T11:24:00+00:00	Что доступно казахстанцам в рамках развития цифровой грамотности	<a href="https://www.zakon.kz/">https://www.zakon.kz/</a>	Ссылка
0,612	2019-12-20T07:52:00+00:00	В клиниках УМС внедрено 79 инновационных технологий	<a href="https://tengrinews.kz/">https://tengrinews.kz/</a>	Ссылка
0,601	2019-12-12T12:48:00+00:00	Дом учителя открыли в Восточном Казахстане	<a href="https://www.zakon.kz/">https://www.zakon.kz/</a>	Ссылка

Продолжение таблицы В.16

1	2	3	4	5
0,56	2019-02-25T04:38:00+00:00	Многоступенчатая система обучения - залог успеха израильских врачей	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка
0,559	2019-04-19T13:56:00+00:00	Космонавт Аимбетов участвует в IT-форуме в Павлодаре	<a href="https://bnews.kz/ru/(baigenews.kz)">https://bnews.kz/ru (baigenews.kz)</a>	Ссылка
0,546	2019-12-17T06:38:00+00:00	Курсы для учителей коррекционных классов проходят в Кокшетау	<a href="https://bnews.kz/ru/(baigenews.kz)">https://bnews.kz/ru (baigenews.kz)</a>	Ссылка
0,546	2019-08-21T07:35:00+00:00	Учить предпринимательству школьников и студентов будут в СКО	<a href="https://bnews.kz/ru/(baigenews.kz)">https://bnews.kz/ru (baigenews.kz)</a>	Ссылка
0,521	2019-10-10T18:52:00+00:00	Концепция «слышащего государства»	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
0,516	2019-07-29T18:52:00+00:00	Язык до Лондона доведет	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка

Таблица В.17 – Тональность - главные топики по положительному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Центр, проект, обучение, служба, сотрудник	0,498	170,0 дней(Score: 2,323)	2,272	1
Язык, казахский, английский, русский, обучение	0,884	125,0 дней(Score: 1,171)	0,741	0,667
Форма, школьный, школа, дочь, образование	-0,177	135,0 дней(Score: 1,427)	-1,675	0,5
Компания, производство, пластиковый, пластик, мусор	0,407	143,0 дней(Score: 1,632)	-0,961	0,333
Педагог, учитель, закон, статус, заработный	Фоновый топик	Фоновый топик		0,333
Проект, развитие, технология, страна, образование	0,159	62,5 дней(Score: -0,429)	-5,171	0,167
Образование, министерство, организация, вопрос, учебный	1,808	60,0 дней(Score: -0,493)	-2,313	0,167
Медицинский, университет, фото, медицина, казnm	-0,34	103,0 дней(Score: 0,608)	3,235	0,167
Система, оценка, учитель, работа, новый	0,659	44,0 дней(Score: -0,902)	0,898	0,167
Учитель, педагог, школа, праздник, страна	0,881	71,5 дней(Score: -0,198)	1,326	0,167

## 6. Болашак

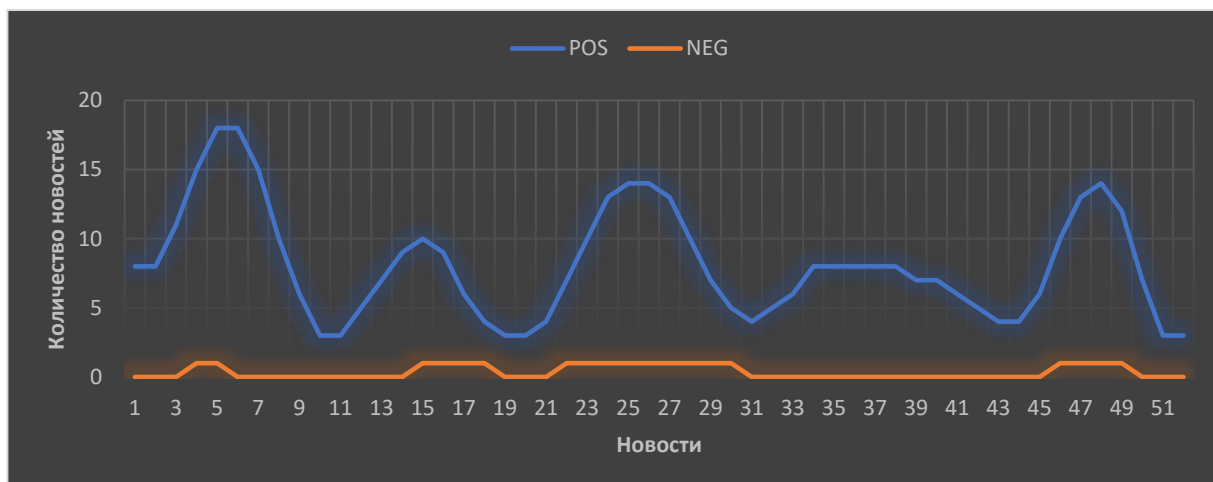


Рисунок В.9 – Динамика восприятия вопросов по теме Болашак

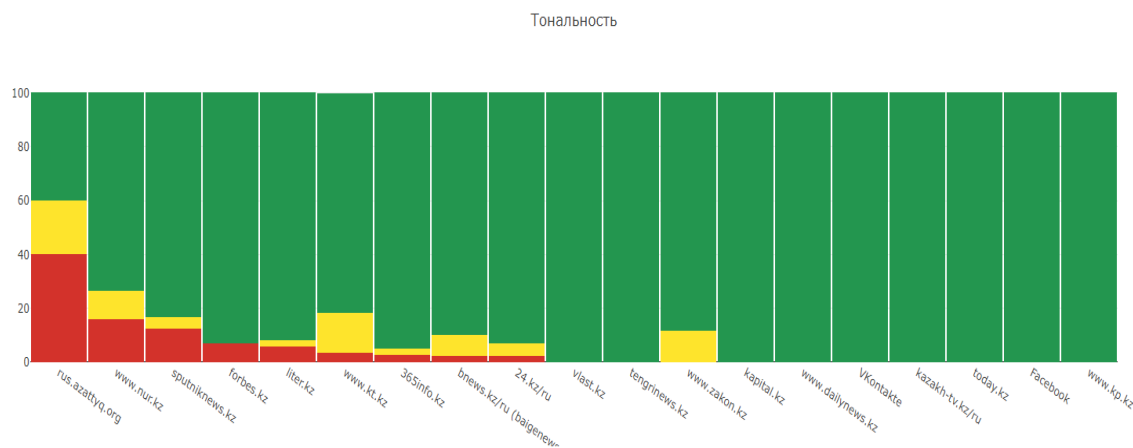


Рисунок В.10 – Динамика восприятия вопросов по теме Болашак

Таблица Б.18 – Топ позитивных новостей

Тональность	Дата	Заголовок	Источник	URL
1	2	3	4	5
0,839	2019-02-04T18:52:00+00:00	С 2016 года будет возобновлено обучение специалистов по программе МВА в рамках стипендии "Болашак"	<a href="https://vlast.kz/">https://vlast.kz/</a>	Ссылка
0,687	2019-03-07T11:50:00+00:00	Наши женщины олицетворяют зеркало души, характер и ментальность народа, – Нурсултан Назарбаев	<a href="https://365info.kz/">https://365info.kz/</a>	Ссылка

Продолжение таблицы В.18

1	2	3	4	5
0,685	2019-11-26T18:52:00+00:00	Кабинет поддержки инклюзии открыли в одной из школ Нур-Султана	<a href="http://www.dailynews.kz/">http://www.dailynews.kz/</a>	Ссылка
0,679	2019-06-24T18:52:00+00:00	В Мюнхене прошел Малый курултай казахов	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
0,673	2019-06-27T18:00:00+00:00	ВЕЧЕР БОЛАШАК: ЧТО Я СДЕЛАЛ ДЛЯ СТРАНЫ?	Vkontakte	Ссылка
0,656	2019-01-18T06:02:00+00:00	Финалистка «Startup Bolashak» реализует проект в отдаленных населенных пунктах	<a href="https://kazakh-tv.kz/ru">https://kazakh-tv.kz/ru</a>	Ссылка
0,625	2019-11-25T13:14:00+00:00	Как правильно сдавать мусор и как узнать качество воздуха: полезные казахстанские экоприложения	<a href="https://bnews.kz/ru">https://bnews.kz/ru</a> (baigenews.kz)	Ссылка
0,61	2019-04-19T18:52:00+00:00	Болат Султанкулов: Все самое лучшее дается трудом, а чудо реально при правильном отношении к жизни	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
0,604	2019-11-27T09:42:00+00:00	Кабинет поддержки инклюзии открыли в столичной школе	<a href="https://www.nur.kz/">https://www.nur.kz/</a>	Ссылка
0,601	2019-04-19T18:52:00+00:00	Болат Султанкулов: Все самое лучшее дается трудом, а чудо реально при правильном отношении к жизни	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка

Таблица В.19 – Топ негативных новостей

Тональность	Дата	Заголовок	Источник	URL
1	2	3	4	5
-0,737	2019-06-06T05:36:00+00:00	Вице-министр Эльмира Суханбер диева арестована на 2 месяца	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка
-0,723	2019-07-05T18:52:00+00:00	40 тыс. долларов стоило госбюджету обучение Эльмиры Суханбердиевой в США	<a href="https://www.nur.kz/">https://www.nur.kz</a> /	Ссылка

Продолжение таблицы В.19

1	2	3	4	5
-0,722	2019-06-06T11:12:00+00:00	В отношении вице-министра избрана мера пресечения в виде содержания под стражей сроком на два месяца - по 2 августа	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
-0,71	2019-06-05T18:52:00+00:00	Ведётся расследование в отношении вице-министра образования Эльмиры Суханбердиевой	<a href="https://litter.kz/">https://litter.kz/</a>	Ссылка
-0,581	2019-04-23T18:52:00+00:00	Карагандинские студенты признались в даче взяток преподавателям	<a href="https://litter.kz/">https://litter.kz/</a>	Ссылка
-0,455	2019-07-05T18:52:00+00:00	40 тыс. долларов стоило госбюджету обучение Эльмиры Суханбердиевой в США	<a href="https://www.nur.kz/">https://www.nur.kz/</a>	Ссылка
-0,417	2019-11-22T03:11:00+00:00	Президенты Казахстана и Швейцарии намерены обсудить инвестиционное сотрудничество двух стран в различных сферах	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
-0,387	2019-11-25T05:08:00+00:00	Бывшего чиновника от образования приговорили к реальному сроку лишения свободы за взятку в размере 5 миллионов тенге	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
-0,379	2019-04-23T18:52:00+00:00	Карагандинские студенты признались в даче взяток преподавателям	<a href="https://litter.kz/">https://litter.kz/</a>	Ссылка
-0,225	2019-07-20T04:32:00+00:00	В Казахстане остановилась трансплантация из-за ареста известного врача	<a href="https://365info.kz/">https://365info.kz/</a>	Ссылка

Таблица В.20 – Тональность - главные топики по положительному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Ребёнок, семья, работа, партия, социальный	1,262	82,5 дней(Score: 0,083)	0,336	1
Университет, студент, вуз, программа, обучение	1,315	55,3 дней(Score: -0,612)	3,441	0,778
Проект, развитие, технология, страна, образование	0,159	62,5 дней(Score: -0,429)	11,808	0,5
Государственный, университет, образование, имя, должность	0,188	55,0 дней(Score: -0,621)	1,203	0,444
Школа, ремонт, робототехника, район, кабинет	-0,026	65,5 дней(Score: -0,352)	0,058	0,278
Проект, развитие, молодёжь, страна, наш	0,124	96,0 дней(Score: 0,429)	3,255	0,278
Учитель, педагог, школа, праздник, страна	0,881	71,5 дней(Score: -0,198)	1,007	0,278
Этап, кандидат, отбор, резерв, кадровый	Фоновый топик	Фоновый топик		0,278
Детский, сад, ребёнок, дошкольный, учреждение	1,056	111,0 дней(Score: 0,813)	2,317	0,278
Наука, национальный, академия, научный, университет	-0,106	58,0 дней(Score: -0,544)	10,389	0,222

Таблица В.21 – Тональность - главные топики по отрицательному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Суд, директор, дело, взятка, тенге	1,368	69,0 дней (Score:-0,262 )	5,59	1
Вуз, студент, коррупция, университет, работа	Фоновый топик	Фоновый топик		0,5
Руслан, концерт, музыка, заведение, ночное	0,087	122,0 дней (Score:1,094 )	2,149	0,125
Школа, здание, арысь, город, дамир	-0,3	98,5 дней (Score:0,493 )	3,737	0,125
Китай, китайский, страна, власть, Казахстан	-0,613	93,5 дней (Score:0,365 )	-1,388	0,125



## 7. Финансовый центр

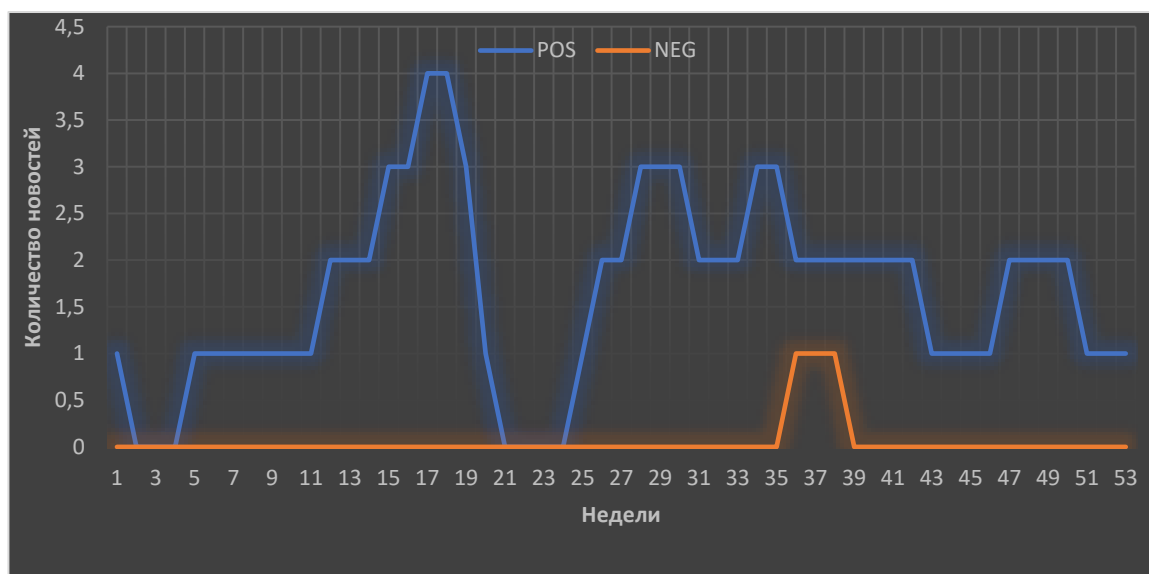


Рисунок В.11 – Динамика восприятия вопросов Финансового центра

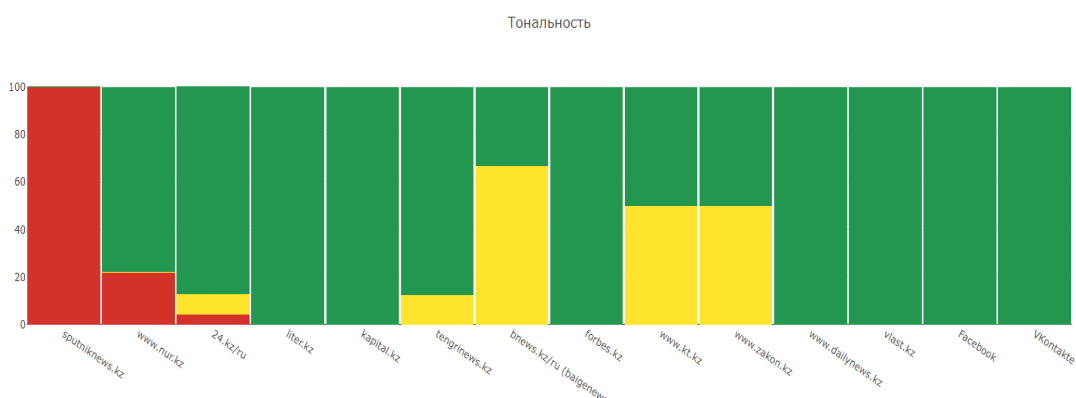


Рисунок В.12 – Восприятие вопросов Финансового центра в разрезе источников

Таблица В.22 – Топ позитивных новостей

Тональность	Дата	Заголовок	Источник	URL
1	2	3	4	5
0,939	2019-12-04T11:47:00+00:00	Мажилис одобрил ряд законопроектов	<a href="https://www.kt.kz/">https://www.kt.kz/</a>	Ссылка
0,689	2019-08-12T18:52:00+00:00	Бюро МФЦА займется обучением внутренних аудиторов	<a href="https://forbes.kz/">https://forbes.kz/</a>	Ссылка
0,644	2019-02-04T18:52:00+00:00	Механизм кредитования и строительства общежитий в Казахстане запустят в августе	<a href="https://vlast.kz/">https://vlast.kz/</a>	Ссылка

Продолжение таблицы В.22

1	2	3	4	5
0,634	2019-07-03T03:52:00+00:00	Строительство общежитий становится привлекательной сферой	<a href="https://kapital.kz/">https://kapital.kz/</a>	Ссылка
0,633	2019-08-28T03:36:00+00:00	Как будет развиваться Нур-Султан, рассказали в "Астанагенплане"	<a href="https://tengrinews.kz/">https://tengrinews.kz/</a>	Ссылка
0,628	2019-05-22T18:52:00+00:00	Гульшара Абдыкаликова выступила на форуме по инновационной экономике	<a href="http://www.dailynews.kz/">http://www.dailynews.kz/</a>	Ссылка
0,626	2019-10-15T18:00:00+00:00	ДАН СТАРТ СТРОИТЕЛЬСТВУ ШКОЛЫ, ОРИЕНТИРОВАННОЙ НА КОСМИЧЕСКИЕ ТЕХНОЛОГИИ	Facebook	Ссылка
0,622	2019-09-05T18:52:00+00:00	Ставку субсидирования по строительству студенческих общежитий подняли до 9%	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
0,621	2019-07-16T18:52:00+00:00	Где студенту жить хорошо?	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка
0,62	2019-07-16T18:52:00+00:00	Где студенту жить хорошо?	<a href="https://liter.kz/">https://liter.kz/</a>	Ссылка

Таблица В.23 – Топ негативных новостей

Тональность	Дата	Заголовок	Источник	URL
1	2	3	4	5
-0,512	2019-09-10T18:52:00+00:00	1,7 млн тенге обязали вернуть экс-студента за обучение по гранту в Павлодарской области	<a href="https://www.nur.kz/">https://www.nur.kz/</a>	Ссылка
-0,387	2019-11-25T05:08:00+00:00	Бывшего чиновника от образования приговорили к реальному сроку лишения свободы за взятку в размере 5 миллионов тенге	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка

Продолжение таблицы 23

1	2	3	4	5
-0,382	2019-09-10T18:52:00+00:00	1,7 млн тенге обязали вернуть экс-студента за обучение по гранту в Павлодарской области	<a href="https://www.nur.kz/">https://www.nur.kz/</a>	Ссылка
-0,193	2019-05-24T04:07:00+00:00	19 млрд тенге накопили казахстанцы на образовательном депозите AQYL	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка
-0,033	2019-11-19T11:28:00+00:00	Судья с высшим Бри танским юридическим образованием без сожалений уехал из столицы в регион	<a href="https://www.zakon.kz/">https://www.zakon.kz/</a>	Ссылка
-0,017	2019-02-16T10:50:00+00:00	Более одного миллиарда тенге задолжали государству выпускники за отказ работать в сёлах	<a href="https://bnews.kz/ru">https://bnews.kz/ru</a> (baigenews.kz)	Ссылка

Таблица В.24 – Тональность - главные топики по положительному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Общежитие, место, строительство, студент, объект	0,876	71,3 дней(Score: -0,203)	1,89	1
Тенге, тысяча, социальный, миллион, семья	1,569	63,5 дней(Score: -0,403)	8,463	0,171
Проект, развитие, технология, страна, образование	0,159	62,5 дней(Score: -0,429)	13,211	0,029
Проект, развитие, молодёжь, страна, наш	0,124	96,0 дней(Score: 0,429)	5,997	0,029
Электронный, интернет, система, школа, доступ	0,093	78,0 дней(Score: -0,032)	11,717	0,029
Грант, учебный, образовательный, заведение, предпринимательство	-0,348	65,5 дней(Score: -0,352)	1,405	0,029
Педагог, учитель, закон, статус, заработный	Фоновый топик	Фоновый топик		0,029

Таблица В.25 – Тональность - главные топики по отрицательному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Суд, директор, дело, взятка, тг	1,368	69,0 дней (Score:-0,262)	-3,786	1
Китай, китайский, страна, власть, Казахстан	-0,613	93,5 дней (Score:0,365 )	2,021	0,333

## 8. Центр дополнительного образования

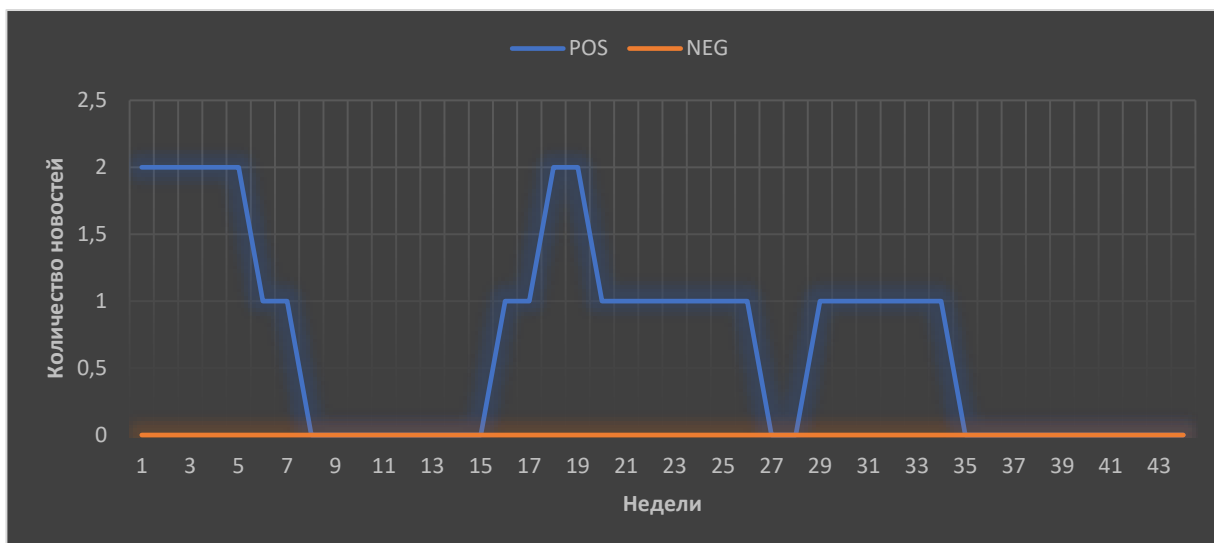


Рисунок В.13 – Динамика восприятия вопросов центра дополнительного образования

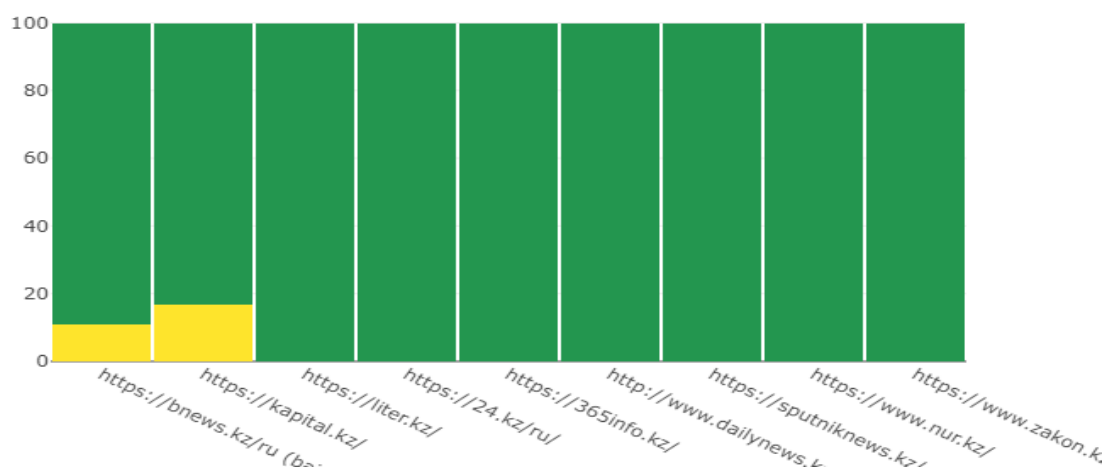


Рисунок В.14 – Восприятие вопросов центра дополнительного образования в динамике

Таблица В.26 – Топ позитивных новостей

Тональность	Дата	Заголовок	Источник	URL
1	2	3	4	5
0,664	2019-11-22T02:52:00+00:00	Вакансии для переводчиков и учителей на Rabotanur.kz	<a href="https://www.nur.kz/">https://www.nur.kz/</a>	Ссылка
0,613	2019-05-16T18:52:00+00:00	Касым-Жомарт Токаев посетил ряд социальных объектов в Нур-Султане	<a href="http://www.dailynews.kz/">http://www.dailynews.kz/</a>	Ссылка

Продолжение таблицы В.26

1	2	3	4	5
0,581	2019-01-29T09:50:00+00:00	Детский IT-ресурсный центр открыли в Кызылординской области	<a href="https://bnews.kz/ru">https://bnews.kz/ru</a> (baigenews.kz)	Ссылка
0,546	2019-09-01T18:52:00+00:00	Новые школы открылись в Атырау и Усть-Каменогорске	<a href="https://litter.kz/">https://litter.kz/</a>	Ссылка
0,518	2019-01-28T18:52:00+00:00	Детский IT-ресурсный центр открыли в Кызылординской области	<a href="https://bnews.kz/ru">https://bnews.kz/ru</a> (baigenews.kz)	Ссылка
0,496	2019-02-20T09:09:00+00:00	В 2019 году в Астане будут открыты 12 новых школ	<a href="https://kapital.kz/">https://kapital.kz/</a>	Ссылка
0,472	2019-05-17T16:44:00+00:00	Глава государства посетил школу-гимназию №68 в Нур-Султане	<a href="https://365info.kz/">https://365info.kz/</a>	Ссылка
0,469	2019-08-30T17:21:00+00:00	7 дней в ауле: как поменялся взгляд городских школьников	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка
0,437	2019-09-01T18:52:00+00:00	Новые школы открылись в Атырау и Усть-Каменогорске	<a href="https://litter.kz/">https://litter.kz/</a>	Ссылка
0,434	2019-09-01T18:52:00+00:00	Новые школы открылись в Атырау и Усть-Каменогорске	<a href="https://litter.kz/">https://litter.kz/</a>	Ссылка

Таблица В.27 – Тональность - главные топики по положительному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Школа, ремонт, робототехника, район, кабинет	-0,026	65,5 дней(Score: -0,352)	0,61	1
Проект, развитие, технология, страна, образование	0,159	62,5 дней(Score: -0,429)	-0,555	0,25
Ребёнок, школьный, питание, лагерь, школьник	1,34	153,0 дней(Score: 1,888)	-0,64	0,25
Учитель, педагог, школа, праздник, страна	0,881	71,5 дней(Score: -0,198)	2,612	0,25
Ребёнок, саин, Казахстан, аружан, страна	-0,417	118,0 дней(Score: 0,992)	0,519	0,25

## 9. Национальный центр тестирования

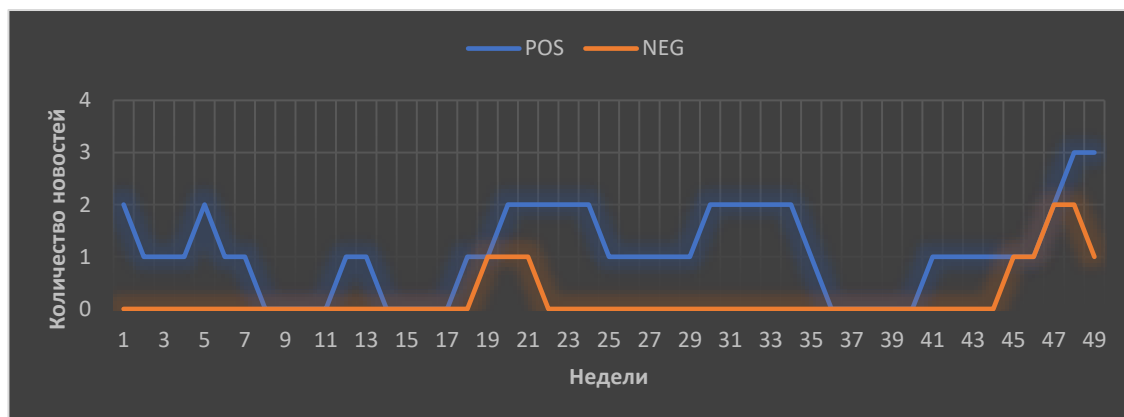


Рисунок В.15 – Динамика восприятия вопросов НЦТ

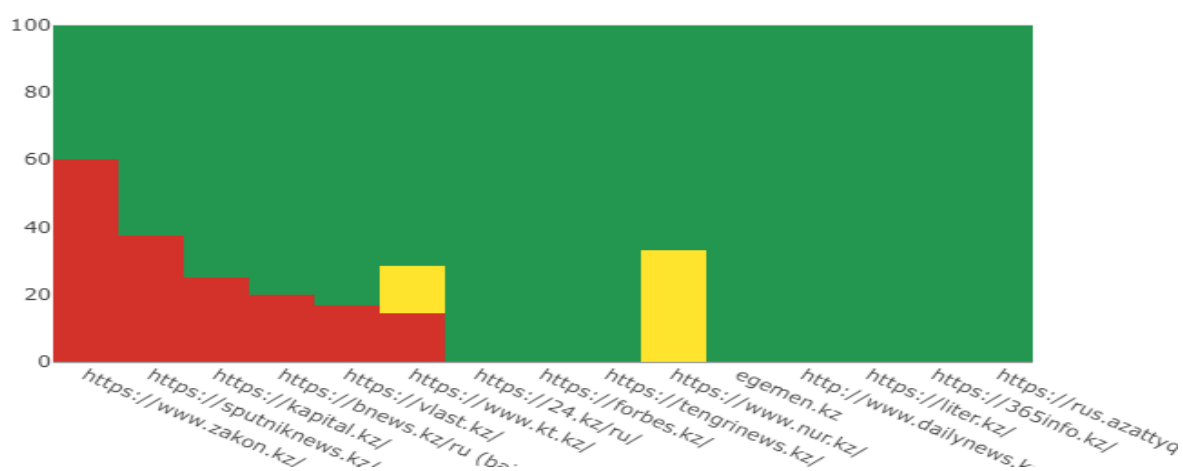


Рисунок В.16 – Восприятие вопросов НЦТ в разрезе источников

Таблица В.28 – Топ негативных новостей

Тональность	Дата	Заголовок	Источник	URL
1	2	3	4	5
-0,846	2019-10-12T06:06:00+00:00	Бывший замглавы Национального центра тестирования осужден на 7 лет	https://vlast.kz/	Ссылка
-0,83	2019-11-25T05:02:00+00:00	Вынесен приговор экс-директору "Национального центра тестирования"	https://www.zakon.kz/	Ссылка
-0,815	2019-11-25T05:17:00+00:00	Экс-директор Национального центра тестирования осужден на 3,5 года	https://kapital.kz/	Ссылка
-0,707	2019-05-24T14:31:00+00:00	Прежний и нынешний руководители центра, ответственного за проведение ЕНТ, оказались замешаны в коррупционных делах	https://sputniknews.kz/	Ссылка

Продолжение таблицы В.28

1	2	3	4	5
-0,663	2019-12-10T06:14:00+00:00	Экс-замглавы Национального центра тестирования осужден на 7 лет за взятку	<a href="https://bnews.kz/ru">(baigenews.kz)"/&gt;https://bnews.kz/ru (baigenews.kz)</a>	Ссылка
-0,421	2019-05-24T08:11:00+00:00	Задержан гендиректор Национального центра тестирования МОН РК	<a href="https://www.zakon.kz/">https://www.zakon.kz/</a>	Ссылка
-0,393	2019-12-10T06:52:00+00:00	Бывшего замуководителя центра признали виновным в получении взятки, сообщили в антикоррупционном ведомстве	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
-0,387	2019-11-25T05:08:00+00:00	Бывшего чиновника от образования приговорили к реальному сроку лишения свободы за взятку в размере 5 миллионов тенге	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
-0,371	2019-05-24T17:01:00+00:00	МОН: Задержание руководства Национального центра тестирования не повлияет на проведение ЕНТ	<a href="https://www.ktz.kz/">https://www.ktz.kz/</a>	Ссылка

Таблица В.29 – Топ позитивных новостей

Тональность	Дата	Заголовок	Источник	URL
1	2	3	4	5
0,57	2019-02-19T18:52:00+00:00	Прием в вузы Узбекистана хотят сделать платным	<a href="http://www.dailynews.kz/">http://www.dailynews.kz/</a>	Ссылка
0,479	2019-10-31T18:52:00+00:00	В Казахстане изменились правила приема в магистратуру и докторантуру	egemen.kz	Ссылка
0,462	2019-06-21T07:35:00+00:00	Дидар Смагулов возглавил Национальный центр тестирования	<a href="https://24.kz/ru/">https://24.kz/ru/</a>	Ссылка
0,409	2019-12-04T09:38:00+00:00	Ничего личного: онлайн-проверка профессиональной компетенции	<a href="https://365info.kz/">https://365info.kz/</a>	Ссылка

Продолжение таблицы В.29

1	2	3	4	5
0,343	2019-02-04T18:52:00+00:00	Участвовать в ЕНТ 2014 года изъявило желание 92 077 выпускников &mdash ; Национальный центр тестирования	<a href="https://vlast.kz/">https://vlast.kz/</a>	Ссылка
0,327	2019-06-20T18:52:00+00:00	Дидар Смагулов возглавил Национальный центр тестирования	<a href="https://forbes.kz/">https://forbes.kz/</a>	Ссылка
0,303	2019-01-10T18:52:00+00:00	Об изменении правил приёма в магистратуру и докторантуру сообщили в МОН РК	<a href="https://forbes.kz/">https://forbes.kz/</a>	Ссылка
0,302	2019-04-06T18:52:00+00:00	Систему распределения грантов в вузы предложили проверить казахстанцам	<a href="https://www.nur.kz/">https://www.nur.kz/</a>	Ссылка
0,298	2019-08-19T18:52:00+00:00	В Шымкенте абитуриенты обратились к министру после «путаницы» с грантами	<a href="https://rus.azattyq.org/">https://rus.azattyq.org/</a>	Ссылка

Таблица В.30 – Тональность - главные топики по положительному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Грант, ента, балл, экзамен, тестирование	1,706	98,0 дней(Score: 0,480)	1,647	1
Тестирование, выпускник, ента, национальный, балл	2,448	146,0 дней(Score: 1,708)	0,345	0,6
Казахстан, студент, гонконг, министерство, казахстанский	Фоновый топик	Фоновый топик		0,2
Конкурс, участник, победитель, участие, команда	0,813	66,5 дней(Score: - 0,326)	1,195	0,2
Педагог, учитель, закон, статус, заработный	Фоновый топик	Фоновый топик		0,1

Таблица В.31 – Тональность - главные топики по отрицательному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Суд, директор, дело, взятка, тенге	1,368	69,0 дней (Score:- 0,262)	5,616	1



## 10. Национальная академия образования

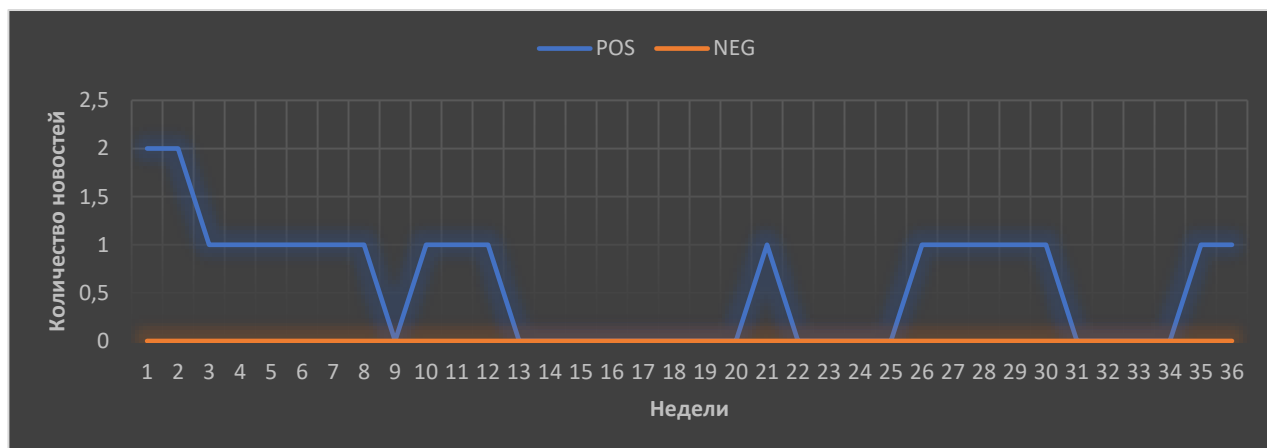


Рисунок В.17 – Динамика восприятия вопросов НАО

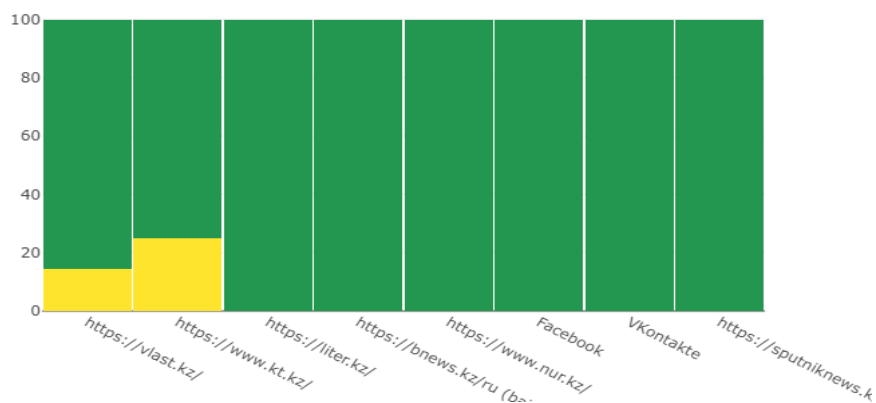


Рисунок В.18 – Восприятие вопросов НЦТ в разрезе источников

Таблица В.32 – Топ позитивных новостей

Тональность	Дата	Заголовок	Источник	URL
1	2	3	4	5
0,719	2019-10-10T18:00:00+00:00	Сегодня в г. Нур-Султан прошла VI международная научно-практическая конференция «инновации в образовании: поиск и решения»	Facebook	Ссылка
0,637	2019-10-10T18:00:00+00:00	Бүгін Нұр-Сұлтан қаласында «Білім берудегі инновациялар: ізденіс және шешімдер» тақырыбында VI халықаралық конференция өтті. сегодня в г. нур-султан прошла VI международная научно-практическая конференция «инновации в образовании: поиск и решения»	VKontakte	Ссылка

Продолжение таблицы В.32

1	2	3	4	5
0,501	2019-04-16T18:52:00+00:00	ЦИК РК в преддверии президентских выборов начинает обучение членов избиркомов	<a href="https://litter.kz/">https://litter.kz/</a>	Ссылка
0,499	2019-02-04T18:52:00+00:00	Содержание и форма ЕНТ изменятся в 2019 году	<a href="https://vlast.kz/">https://vlast.kz/</a>	Ссылка
0,49	2019-08-24T14:25:00+00:00	По образцу НИШ: Каким будет новый учебный год для акмолинских школьников	<a href="https://bnews.kz/ru">https://bnews.kz/ru</a> (baigenews.kz)	Ссылка
0,485	2019-03-15T06:59:00+00:00	Около 23% учителей ЕМЦ полностью или частично преподают предметы на английском языке	<a href="https://www.kt.kz/">https://www.kt.kz/</a>	Ссылка
0,447	2019-04-16T18:52:00+00:00	ЦИК РК в преддверии президентских выборов начинает обучение членов избиркомов	<a href="https://litter.kz/">https://litter.kz/</a>	Ссылка
0,36	2019-08-09T18:52:00+00:00	«Принести в подоле» – не стыдно!	<a href="https://litter.kz/">https://litter.kz/</a>	Ссылка
0,349	2019-08-07T01:53:00+00:00	Фонд ООН в области народонаселения (ЮНФПА) совместно с министерством образования начали внедрение курса нравственно-полового образования в школьную программу	<a href="https://sputniknews.kz/">https://sputniknews.kz/</a>	Ссылка
0,333	2019-03-14T12:43:00+00:00	МОН утвердило типовые учебные планы со снижением нагрузки для школьников	<a href="https://www.kt.kz/">https://www.kt.kz/</a>	Ссылка

Таблица В.33 – Тональность - главные топики по положительному влиянию

Топик	Прогнозируемая резонансность	Прогнозируемая продолжительность	Тренд	Вес
Проект, развитие, технология, страна, образование	0,159	62,5 дней(Score: -0,429)	6,055	1
Город, алматы, акимат, документ, президент	0,678	85,0 дней(Score: 0,147)	4,74	0,667
Проект, развитие, молодёжь, страна, наш	0,124	96,0 дней(Score: 0,429)	19,054	0,333
Ребёнок, проблема, человек, уровень, образование	-0,007	99,5 дней(Score: 0,518)	-2,457	0,333
Система, оценка, учитель, работа, новый	0,659	44,0 дней(Score: -0,902)	1,303	0,333