

ТЕХНИЧЕСКОЕ ЗАДАНИЕ / ПРОЕКТ ДЛЯ СОГЛАСОВАНИЯ

Разработка высокопроизводительной мультязычной системы выявления и идентификации объектов информационного поля с определением контекстной тональности по отношению к ним

Заказчик	ТОО «IMAS GROUP»
Предмет	Разработка независимых компонентов выявления объектов и направленной тональности с общим API
Формат	НИР → прототип двух этапов → независимые испытания → подготовка пилота
Статус документа	Проект; сроки, пороги качества, финансирование, права и приёмка уточняются по договору
Основание	Публикация технологической задачи Astana Hub; прилагаемое ТЗ для отбора исполнителя

Назначение: открытая технологическая задача на два связанных AI-модуля. Документ не содержит описания действующей инфраструктуры, клиентских выгрузок или внутренних алгоритмов iMAS.

1. Вводные разделы

№	Элемент шаблона	Содержание и требования
1.1.1	Наименование	Разработка высокопроизводительной мультязычной системы выявления и идентификации объектов информационного поля с определением контекстной тональности по отношению к ним
1.1.2	Краткая характеристика области применения	Медиамониторинг, репутационная и конкурентная аналитика публикаций СМИ, социальных сетей и других открытых источников. Автоматическое извлечение упомянутых организаций, органов власти, персон, брендов, продуктов и услуг с последующим определением тональности именно по отношению к каждому объекту.
1.2.1	Документ-основание	Публикуемая технологическая задача ТОО «IMAS GROUP» на площадке Astana Hub; настоящее ТЗ — проект для поиска исполнителя и согласования PoC.
1.2.2	Дата утверждения ТЗ	Не утверждена. Дата и версия утверждения определяются заказчиком после внутреннего согласования.
1.2.3	Орган утверждения	ТОО «IMAS GROUP», уполномоченный представитель заказчика; процедура утверждения устанавливается отдельно.
1.3.1	Наименование объекта	Независимый мультязычный программный модуль с двумя связанными функциями: (I) обнаружение, типизация и разрешение упоминаний объектов; (II) определение контекстной тональности в отношении выявленных объектов.
1.3.2	Заказчик / разработчик	Заказчик: ТОО «IMAS GROUP». Исполнитель определяется по результатам отбора технических предложений; контактные и договорные реквизиты — по отдельному соглашению.
1.3.3	Сроки	Продолжительность, график разработки, приёмочные точки и сроки пилота согласуются с исполнителем; дата начала работ не предопределяется данным ТЗ.

№	Элемент шаблона	Содержание и требования
1.3.4	Финансирование	Бюджет, условия вознаграждения, источник финансирования, порядок расчётов и права на результаты определяются отдельным договором.
1.4.1	Перечень терминов и сокращений	NER — автоматическое обнаружение и классификация упоминаний сущностей; Entity Linking — сопоставление упоминания с идентификатором объекта; Coreference — связь повторных указаний на один объект; Target-Based Sentiment — тональность по отношению к указанному объекту; ABSA — аспектная тональность; PoC — проверка технологической реализуемости; API — программный интерфейс.
1.5.1	Объект	Поток публикаций, в которых встречаются одно или несколько упоминаний объектов; одни и те же объекты могут обозначаться разными именами, аббревиатурами, местоимениями и вариантами написания.
1.5.2	Цель	Создать и экспериментально проверить автономный двухэтапный модуль: сначала обнаруживать и идентифицировать объекты независимо от наличия оценки, затем определять адресованную каждому объекту тональность с проверяемым текстовым основанием. Обеспечить пригодность результатов для поиска, фильтров и аналитики.
1.5.3	Методы	Сравнить правила/словари, NER и Entity Linking, компактные трансформерные модели, извлечение отношений, классификацию целевых оценок, каскадные схемы и выборочное применение LLM. Метод выбирается по качеству, скорости и себестоимости, а не по названию модели.
1.5.4	Новизна	Сквозная многоязычная обработка с независимым выявлением всех значимых объектов, сопоставлением повторных упоминаний и последующей привязкой позитивных/негативных/нейтральных оценок к их адресатам; языки Центральной Азии, неоднозначные названия и отсутствие оценки.
1.5.5	Эффективность	Повышение точности объектной и репутационной аналитики, уменьшение ложного переноса негативной оценки между объектами, сокращение ручной разметки и формирование переиспользуемых сущностей для быстрого поиска и аналитики. Эффект подтверждается отдельными и сквозными метриками, затратами и скоростью.

2. Назначение, цели, объект

№	Элемент шаблона	Содержание и требования
2.1.1	Назначение системы	Принимать текст публикации без обязательного входного списка объектов; находить все релевантные упоминания, их границы и типы, связывать повторные указания, при наличии оснований определять устойчивый идентификатор. Затем выявлять и оценивать направленное отношение к каждому объекту, включая отсутствие высказывания об объекте.
2.1.2	Показатели достижения цели	1) Precision/Recall/F1 обнаружения объектов и правильности их типов; 2) качество разрешения упоминаний и корректных отказов от идентификации; 3) F1 привязки оценочных выражений к объекту; 4) macro-F1 тональности по размеченным объектам и сквозное качество; 5) P50/P95/P99, публикаций/сек и полная стоимость обработки 1 млн публикаций.
2.2	Назначение разработки	Разработать два автономно вызываемых и совместно работающих AI-компонента, рабочий PoC с API, воспроизводимыми тестами, отчётами по качеству и стоимости и техническим планом промышленной интеграции в медиaplatformу.

№	Элемент шаблона	Содержание и требования
2.3	Характеристика предмета	Короткие посты, новости, заголовки, смешение языков, неоднозначные имена, сокращения, прямые/косвенные указания на объекты, цитаты, сравнения и отрицания. В одной публикации возможны противоположные оценки нескольких объектов или отсутствие оценок.
2.4	Существующее состояние	Базовая тональность публикации и обнаружение отдельных именованных сущностей не равнозначны направленной оценке. При наличии нескольких объектов общая полярность текста не показывает, кому именно адресован позитив или негатив; одно имя может обозначать разные организации или продукты.
2.5	Научная цель / гипотеза	Гипотеза: каскад из самостоятельного многоязычного обнаружения/разрешения упоминаний и контекстной модели связи «объект — оценочное высказывание — полярность» способен улучшить точность адресной аналитики при меньшей себестоимости, чем постоянный вызов ресурсоёмкой LLM для каждого объекта/материала. Проверяется на независимом многоязычном корпусе.

3. Требования к результату / системе

№	Элемент шаблона	Содержание и требования
3.1.1	Структура и функционирование	Этап I: нормализация текста, обнаружение упоминаний, типизация, объединение повторных отсылок и разрешение сущностей по доступному справочнику либо локальному ключу. Этап II: выявление оценки, адресата, полярности, при необходимости аспекта и подтверждающего фрагмента. Независимые API обоих этапов и общий конвейер; версионированные результаты.
3.1.2	Показатели назначения	Потоковая, пакетная и единичная обработка; возможность получить только объекты без запуска тональности; поддержка нескольких объектов и разных оценок в одном тексте; поиск и фильтры по типу, объекту и его тональности. Строгие целевые значения устанавливаются после базовых испытаний и согласования корпусов.
3.1.3	Надёжность	Идемпотентная повторная подача, сохранение согласованных результатов, возможность возобновить обработку после сбоя, управляемые тайм-ауты и повторные попытки. Недоступность второго этапа не должна блокировать получение объектов от первого.
3.1.4	Безопасность	Аутентификация API, минимально необходимые полномочия, защищённое хранение секретов, маскирование чувствительных данных в журналах, управление доступом к обучающим и контрольным корпусам.
3.1.5	Эргономика и техническая эстетика	Полноценный интерфейс не обязателен; нужна демонстрационная форма или тестовый клиент с подсветкой границ объектов и подтверждающих оценочных фрагментов, корректным показом языка, адресата, класса и неопределённости.
3.1.6	Эксплуатация и обслуживание	Документированные установка, настройка, обновление моделей и словарей, перезапуск, проверка качества, резервное копирование артефактов, наблюдаемость и откат версии. Воспроизводимый запуск без обязательного обращения к внешним сервисам.
3.1.7	Защита от несанкционированного доступа	Изоляция тестовых окружений, управление ролями/API-ключами, шифрование трафика, отсутствие закрытых данных и ключей в исходном коде и логах, протоколирование административных действий.
3.1.8	Сохранность при авариях	После перезапуска должны сохраняться подтверждённые задания и

№	Элемент шаблона	Содержание и требования
		результаты; повторные вызовы не должны размножать сущности и оценки. Демонстрируется восстановление после временной недоступности модели и отказа вычислительного узла.
3.1.9	Защита от внешних воздействий	Устойчивость к сетевым задержкам, пиковым очередям, чрезмерно длинным и некорректным текстам, тайм-аутам модели и недоступности необязательных зависимостей. Для физической инфраструктуры — стандартные меры эксплуатации вычислительных узлов.
3.1.10	Патентная чистота	Исполнитель подтверждает правомерность использования моделей, весов, библиотек, корпусов и компонентов; раскрывает лицензии и ограничения коммерческой эксплуатации, включая передачу и локальное развёртывание.
3.1.11	Стандартизация / унификация	Документированные JSON/JSONL или аналогичные контракты, стабильный item_id, границы/позиции каждого упоминания, типы объектов и идентификаторы (при наличии), адресная тональность и текстовые доказательства, уверенность, версии модели, поддержка экспортных форматов.
3.1.12	Дополнительные требования	Демонстрация отдельной работы обоих модулей и полного конвейера; инструкции по тестовым примерам, ознакомление специалистов заказчика с воспроизведением результатов. Разработка OCR/ASR и полноценных клиентских интерфейсов не входит в обязательный объём.
3.2.1	Функции и критерии отказов	I. Получение текста/title/lang/item_id; поиск всех релевантных сущностей с типами, символьными границами, группировкой повторных указаний и уверенной/неопределённой идентификацией. II. Для каждого объекта — адресованная оценка, полярность, необязательный аспект, подтверждающая цитата/диапазон и confidence. Не приписывать объекту общий негатив текста. Отдельно различать нейтральную оценку, отсутствие оценки и неопределённость. Ошибки: пропуск объекта, неверное объединение одноимённых сущностей, ошибочная привязка оценки, ложный негатив, несогласованная повторная обработка.
3.3.1	Информационное обеспечение	Открытые, синтетические и отдельно согласованные обезличенные материалы; разметка первого уровня: границы, типы, упоминания и идентичность; второго: связь высказывания с объектом, полярность, аспект/доказательство, случаи отсутствия оценки. Разделение обучения, настройки и независимого теста; контроль дублей между выборками.
3.3.2	Лингвистическое обеспечение	Обязательная оценка на казахском, кыргызском, узбекском, таджикском, туркменском, русском и английском; кириллица, латиница и встречающиеся варианты арабской графики, транслитерация и смешанные тексты. По редким языкам отдельно документировать ограничение разметки и достигнутое качество, а не переносить оценки с русского.
3.3.3	Программное обеспечение	Допускаются самостоятельные либо комбинированные библиотеки и модели NER, Entity Linking, Coreference и Target Sentiment. Передать запускные конфигурации, контейнеры/пакеты, исходный код или согласованный развёртываемый компонент, тесты и перечень сторонних зависимостей и лицензий.
3.3.4	Техническое обеспечение	Измерить достаточные CPU/RAM и GPU-варианты, скорость и задержку отдельных этапов и общей цепочки, поведение при параллельной обработке. Развёртывание на собственной инфраструктуре без обязательной платной внешней LLM; внешние API допустимы для сравнительного эксперимента.
3.3.5	Организационное обеспечение	Определить ответственных, порядок совместной разметки и разрешения спорных случаев, версионирование моделей и корпусов, протокол

№	Элемент шаблона	Содержание и требования
		проверки результатов, обмен техническими замечаниями и режим обращения с предоставленными материалами.
3.4	Требования к программному изделию	Независимые методы extract_entities, analyze_target_sentiment и общий process_document либо функциональные аналоги; стабильная схема ответа по каждому объекту; поддержка пакетной обработки и API; пригодность результатов для быстрых фильтров/отчётов; локальная установка и воспроизводимая проверка.
3.5	Информационная безопасность	Без доступа к production, клиентским базам и внутренним алгоритмам iMAS в рамках открытой публикации. Правомерные данные для тестов, разграничение доступа, отсутствие логирования исходных персональных данных без необходимости, документированные правила хранения и удаления тестовых наборов.
3.6	Промышленная безопасность / ESG	Промышленная и физическая безопасность неприменимы к программной NLP-задаче. Для оценки ресурсной эффективности учитывать энергопотребление CPU/GPU, долю тяжёлых LLM-вызовов и затраты на 1 млн публикаций.

4. Работы, сроки, приёмка

№	Элемент шаблона	Содержание и требования
4.1.1	Стадии, исполнители, документы	Этап 1 — согласование таксономии объектов, классов оценки, схемы разметки и baseline; этап 2 — отдельный работающий модуль обнаружения/разрешения сущностей с API и тестом; этап 3 — модуль адресной тональности, связывание двух этапов и прототип; этап 4 — независимые испытания, нагрузка, отчётность, документация и план пилотной интеграции. Сроки и контрольные документы по этапам — в договоре.
4.2.1	НИР → ОКР → ОПЭ / внедрение	НИР: исследование методов, разметка и проверяемая гипотеза; ОКР: реализация обоих модулей и общей цепочки; ОПЭ: контрольный прогон, интеграционный PoC, оценка стоимости и потребных ресурсов. Промышленное внедрение и длительное обслуживание — отдельный этап по договору.
4.3.1	Контроль и приёмка	Приёмка на заранее согласованной независимой выборке без обучения на контрольных данных. Отдельно проверяются: обнаружение всех объектов даже без оценок; различение омонимов; привязка разных тональностей разным объектам в одном тексте; цитаты и отрицания; отсутствие оценки; API, повторная обработка, сбои и воспроизводимость. Акт и состав комиссии — по договору.
4.4.1	Технико-экономические показатели	I: entity-level precision/recall/F1 по границам и типам, accuracy разрешения и корректные неопределённые случаи. II: precision/recall/F1 адресата оценки, macro-F1 классов тональности на эталонных объектах. End-to-end: точность полного результата «объект — адресат — класс» и ложный негатив по неочисляемому объекту. Раздельно по языкам. Нагрузки 100 тыс., 1 млн, 5 млн документов/сутки (испытательные сценарии, не текущий поток iMAS), P50/P95/P99, стоимость 1 млн документов и ресурсы. Пороговые значения утверждаются после baseline.
4.5.1	Порядок оплаты	Вознаграждение, цена, порядок этапных расчётов, сроки оплаты, права на исходный код/веса/разметку и условия дальнейшего внедрения согласуются отдельно. Публикация ТЗ сама по себе не является предложением оферты.

5. Подготовка объекта и документирование

№	Элемент шаблона	Содержание и требования
5.1	Подготовка к вводу в действие	Развернуть отдельное тестовое окружение без подключения к производственным системам; подготовить разрешённые многоязычные корпуса, справочные идентификаторы объектов при наличии, baseline и закрытую контрольную выборку. Для пилота отдельно согласовать формат интеграции без раскрытия внутренних схем.
5.2.1	Комплект документации	Архитектура и схема данных, описание алгоритмов и версий моделей, протокол двухуровневой разметки, спецификация независимых и объединённого API, отчёты по каждому этапу и всей цепочке, исходники/артефакты и лицензии, инструкции развертывания/обновления, нагрузочный протокол, сценарии отказов и руководство по тестовому воспроизведению. Степень оформления по ГОСТ — по согласованию.
5.3.1	Источники разработки	Настоящее ТЗ, публичное описание Astana Hub, правомерно используемые исследования и корпуса по NER/Entity Linking/Target Sentiment/ABSA, техническая документация предлагаемых инструментов. Производственные данные и алгоритмы iMAS для открытого поиска решения не предоставляются.

6. Квалификационные требования

№	Элемент шаблона	Содержание и требования
6.1.1	Материально-техническая база	Возможность разработать и воспроизвести работающий прототип на собственных или согласованных вычислительных ресурсах, выполнить раздельную и сквозную оценку качества и нагрузки, предоставить демонстрацию API. Конкретная марка GPU/модель не навязывается.
6.2	Разрешительные документы	Предварительно специальные разрешительные документы для разработки данного ПО не установлены; исполнитель отвечает за законное использование данных и компонентов. Документы, требуемые для договорных отношений, уточняются при отборе.
6.3	Гарантийные обязательства	На согласованный период после передачи PoC — устранение воспроизводимых критических дефектов, препятствующих воспроизведению приёмочных тестов. Объём гарантий, условия поддержки и сроки определяются договором; качество не обещается без независимых измерений.