Письменный отзыв официального рецензента на диссертационную работу Якунина Кирилла Олеговича на тему «Разработка моделей и методов сбора, анализа и классификации медиа-публикаций на базе методов обработки естественных языков», представленную на соискание степени «доктор философии» (Ph.D) по специальности 6D070400 – «Вычислительная техника и программное обеспечение»

№ п/п	Критерии	Соответствие критериям (необходимо отметить один из вариантов ответа)	Обоснование позиции официального рецензента
1.	Тема диссертации (на дату утверждения) соответствует направлениям развития науки и/или государственным программам	1.1 Соответствие приоритетным направлениям развития науки или государственным программам:	
		1) Диссертация выполнена в рамках проекта или целевой программы, финансируемого (ой) из государственного бюджета (указать название и номер проекта или программы) 2) Диссертация выполнена в рамках другой государственной программы (указать название программы) 3)Диссертация соответствует приоритетному направлению развития науки, утвержденному Высшей научно-технической комиссией при Правительстве Республики Казахстан (указать направление)	Диссертация работа выполнялась в рамках проекта 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 (Институт информационных и вычислительных технологии КН МОН РК).
2.	Важность для науки	Работа вносит/не вносит существенный вклад в науку, а ее важность хорошо раскрыта/не раскрыта	Вынесенные на защиту положения содержат явную научную новизну и были тщательно верифицированы в ходе экспериментов.
3.	Принцип самостоятельности	Уровень самостоятельности: 1) Высокий; 2) Средний; 3) Низкий; 4) Самостоятельности нет	Личный вклад соискателя диссертации состоит в формулировке и обосновании темы исследования, постановке задач, проведении теоретических и экспериментальных исследований, разработке методов мультикритериальной оценки медиа-источников ММА на основе Байесовской модели агрегации гетерогенных данных, разработке информационной системы мониторинга медиа-пространства Казахстана на базе методов NLP с

			описанием программной архитектуры и основным функционалом, разработке выводов и рекомендаций, что позволяет оценить уровень самостоятельности в проведений исследований, как высокий.
4.	Принцип внутреннего единства	4.1 Обоснование актуальности диссертации: 1) Обоснование; 2) Частично обоснована; 3) Не обоснована.	Задача эффективной векторизации текстовых данных с целью последующей мультикритериальной классификации – одна из важнейших задач в области обработки естественных языков. Современные подходы к векторизации текстов обладают рядом недостатков – они либо слишком тривиальны (bag-of-words) либо наоборот требуют сложнейших моделей глубокого обучения и, следовательно, являются слабо интерпретируемыми (BERT, GPT). Любой новый подход к решению данного комплекса задач вызывает интерес и определяет актуальность настоящего диссертационного исследования. В частности, предлагаемые автором модели и методы, реализованные в рамках распределенной информационной системы, позволяют решать задачу мультикритериального анализа информационных трендов в медиа-пространстве с небольшим объёмом необходимой ручной разметки.
		4.2 Содержание диссертации отражает тему диссертации: 1) Отражает; 2) Частично отражает; 3) Не отражает	Содержание диссертации отражает тему диссертации.
		4.3 Цель и задачи соответствуют теме диссертации: 1) соответствуют; 2) частично соответствуют; 3) не соответствуют	Цели и задачи, поставленные в работе, соответствуют теме диссертации и конкретизируют направление исследований.

		4.4 Все разделы и положения диссертации логически взаимосвязаны: 1) полностью взаимосвязаны; 2) взаимосвязь частичная; 3) взаимосвязь отсутствует	Полученные в диссертационном исследовании научные результаты обладают внутренним единством, так как все они являются следствием решения одной единой задачи. Оценка внутреннего единства основана на логической связи последовательности элементов разработанного метода, выраженной в составе исследований и изложенной в диссертационной работе. Степень этого единства очень высока, так как каждый полученный результат является следствием предыдущего результата.
		4.5 Предложенные автором новые решения (принципы, методы) аргументированы и оценены по сравнению с известными решениями: 1) критический анализ есть; 2) анализ частичный; 3) анализ представляет собой не собственные мнения, а цитаты других авторов	Диссертационная работа является целостным научным исследованием, содержащим системный анализ состояния вопросов в исследуемой области, проработку актуальных направлений и обоснование достигнутых научных результатов. Соискателем раскрыты актуальность, конкретизированы проблемы, связанные с исследуемой темой диссертации. Предложен метод оценки межкорпусного тематического дисбаланса для решения найденных в ходе литературного обзора ограничений, в особенности необходимости больших объёмов ручной экспертной разметки и низкой интерпретируемости современных моделей. Разработан мультикритериальный метод оценки медиа-источников ММА. Метод основан на синтезе байесовской модели агрегации критериев, также описанной в данном разделе, на методе анализа иерархий (AHP), позволяющем проводить относительную оценку важности критериев на целевой показатель, а также на применение тех или иных тематических моделей для поиска информационных трендов в корпусе текстов.

5.	Принцип научной новизны	5.1 Научные результаты и положения являются новыми? 1) полностью новые; 2) частично новые (новыми являются 25-75%); 3) не новые (новыми являются менее 25%)	Основные научные результаты и положения включают следующие пункты: 1 Исследован вопрос влияния открытых информационных источников на обществе, выявлены основные направления влияния, сформирован перечень информативных признаков, на основе которых можно оценить это влияние. 2 Исследованы существующие подходы классификации документов и векторизации текстов, выявлены проблемы и слабые стороны текущих решений, сформированы рекомендации. 3 Разработан подход векторизации текстов на основе тематической модели. 4 Разработан метод оценки межкорпусного тематического дисбаланса, позволяющий автоматически или полуавтоматически получать веса топиков по отношению к заданному признаку. 5 Разработан метод мультикритериальной оценки медиа-источников ММА на базе байесовской модели агрегации. 6 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. 7 Собран корпус, состоящий из более чем 6 миллионов публикаций из казахстанских и российских источников, включая как тексты публикаций, так и метаданные.
		5.2 Выводы диссертации являются новыми? 1) полностью новые; 2) частично новые (новыми являются 25-75%); 3) не новые (новыми являются менее 25%)	
		5.3 Технические, технологические, экономические или управленческие решения являются новыми и обоснованными: 1) полностью новые; 2) частично новые (новыми являются 25-75%); 3) не новые (новыми являются менее 25%)	
6.	Обоснованность основных выводов	Все основные выводы основаны/не основаны на весомых с научной точки зрения доказательствах либо достаточно хорошо обоснованы (для qualitative	Полученные и представленные в диссертационной работе научные результаты опираются на строгое математическое изложение

		research и направлений подготовки по искусству и гуманитарным наукам)	основных положений и применяемых методов и подтверждаются результатами проведенного моделирования. Достоверность результатов обеспечивалась использованием современных средств и методик проведения исследований. Это дает основание считать полученные результаты достаточно обоснованными и достоверными.
7.	Основные положения, выносимые на защиту	Необходимо ответить на следующие вопросы по каждому положению в отдельности: 7.1 Доказано ли положение? 1) доказано; 2) скорее доказано; 3) скорее не доказано; 4) не доказано 7.2 Является ли тривиальным? 1) да; 2) нет 7.3 Является ли новым? 1) да; 2) нет 7.4 Уровень для применения: 1) узкий; 2) средний; 3) широкий 7.5 Доказано ли в статье? 1) да; 2) нет	7.1 Основные положения, выносимые на защиту полностью доказаны вычислительными экспериментами для валидации предложенных моделей и методов. На базе этой системы были проведены вычислительные эксперименты для валидации разработанных моделей и методов. 7.2 Основные положения, выносимые на защиту, не являются тривиальными, так как содержат решения, отличающиеся научной новизной и практической значимостью. 7.3 Результаты, полученные автором и сформулированные в диссертации, являются новыми научными знаниями в области обработки естественных языков. Предложен метод векторизации текстовых документов с помощью тематической модели BigARTM; данный метод векторизации позволяет использовать данные, полученные из больших неразмеченных текстовых корпусов для получения эффективных векторных представлений, сопоставимых для ряда задач с векторизациями, полученным с помощью моделей глубокого обучения. Предложен метод оценки тематического межкорпусного дисбаланса для самообучения классификационной модели; данный метод

			представляет собой новый подход в области semi-supervised learning, который был верифицирован при решении ряда задачи классификации текстов. Предложена методика многофакторной оценки социальной значимости публикации; данная методика предлагает способ для решения тяжело формализуемой задачи оценки социальной значимости с учётом различных факторов. Методика была апробирована на размеченном корпусе достаточного объёма. Предложена методика многокритериальной оценки масс-медиа ММА на базе байесовской системы агрегации, метода анализа иерархий (AHP) и тематического моделирования; данная методика является новым подходом к решению задачи оценки масс-медиа документов (публикаций) и источников. 7.4 Положения выносимые на защиту имеют широкий уровень применения. 7.5 Основные положения достигнутых результатов опубликованы в открытой печати, обсуждены в ряде международных и научно-практическихконференций, а также опубликованы в рейтинговых журналах (Q1, Q2, Q3).
8.	Принцип достоверности Доверенность источников и предоставляемой информации	8.1 Выбор методологии – обоснован или методология достаточно подробно описана 1) да; 2) нет	8.1 Выбор методологии исследований проведен с учетом особенностей методов обработки естественных языков. 8.2 В работе использованы новые подходы к решению комплекса задач. В частности, предлагаемые автором модели и методы, реализованные в рамках распределенной информационной системы, позволяют решать задачу мультикритериального анализа информационных трендов в
		8.2 Результаты диссертационной работы получены с использованием современных методов научных исследований и методик обработки и интерпретации данных с применением компьютерных технологий: 1) да; 2) нет	

		8.3 Теоретические выводы, модели, выявленные взаимосвязи и закономерности доказаны и подтверждены экспериментальным исследованием (для направлений подготовки по педагогическим наукам результаты доказаны на основе педагогического эксперимента): 1) да; 2) нет	медиапространстве. Разработана распределённая интеллектуальная информационная системы Media Analytics, позволяющей проводить сбор, хранение и обработку текстовых данных. 8.3 Соискателем в полной мере обоснованы теоретические выводы, методы, разработанные в ходе исследовательской работы. В рамках работы над диссертационным исследованием была разработана распределённая интеллектуальная информационная системы Media Analytics, позволяющей проводить сбор, хранение и обработку текстовых данных. В частности, на базе этой системы реализован метод MMA (мультикритериальный метод оценки медиа-источников), а также проведены вычислительные эксперименты для валидации предложенных моделей и методов. На базе этой системы были проведены вычислительные эксперименты для валидации разработанных моделей и методов. Автором проведена всесторонняя валидация разработанных моделей и методов, было проведено сравнение эффективности предложенных методов с современными моделями классификации. 8.4 Важные утверждения, приводимые в диссертационной работе, подтверждаются ссылками на актуальную и достоверную научную литературу. 8.5 Список источников, использованных в диссертационной работе, состоит из 92 наименований, которых достаточно для литературного обзора по данной теме.
		8.4 Важные утверждения подтверждены/частично подтверждены/не подтверждены ссылками на актуальную и достоверную научную литературу	
		8.5 Использованные источники литературы достаточны/не достаточны для литературного обзора	
9.	Принцип практической ценности	9.1 Диссертация имеет теоретическое значение: 1) да; 2) нет	9.1 В научно-исследовательской работе внимание уделено исследованию, которое является актуальным

			и имеет теоретическую и прикладную значимость. Научные результаты применимы для многокритериальной оценки медиапространства и автоматического поиска информационных трендов.
		9.2 Диссертация имеет практическое значение и существует высокая вероятность применения полученных результатов на практике: 1) да; 2) нет	9.2 Диссертация имеет практическое значение и существует высокая вероятность реального использования распределённой интеллектуальной информационной системы Media Analytics, позволяющей проводить сбор, хранение и обработку текстовых данных.
		9.3 Предложения для практики являются новыми? 1) полностью новые; 2) частично новые (новыми являются 25-75%); 3) не новые (новыми являются менее 25%)	9.3 Разработанная система решает стандартные для систем медиа-мониторинга задачи, включая сбор текстовых данных, а также поиск, фильтрацию, визуализацию данных. Соискатель предлагает в работе другие варианты практического применения: применение автоматического тематического моделирования, оценку межкорпусного дисбаланса, а также применение предложенных в работе моделей классификации (включая метод ММА). Таким образом, предложения для практики являются в основном новыми.
10.	Качество написания и оформления	Качество академического письма: 1) высокое; 2) среднее; 3) ниже среднего; 4) низкое.	Диссертация написана грамотным научно-техническим, доступным для чтения, языком. В диссертационной работе имеются незначительные стилистические ошибки, не снижающие качества диссертационной работы.

Заключение. Проведена качественная научно-исследовательская работа, по результатам которой получены актуальные, новые результаты, имеющие теоретическую и прикладную значимость, разработана и предложена методика мультифакторной оценки социальной значимости на базе многокритериальной методики оценки масс-медиа ММА с использованием тематической векторизации текстовых документов. Разработана архитектура и программная реализация информационной системы для сбора, обработки и оценки масс-медиа текстов и социальных сетей Media Analytics,

включающий предложенную методику оценки, а также ее оценку и верификацию. На основании вышеизложенного рекомендую присудить Якунину Кириллу Олеговичу степень доктора философии (Ph.D) по специальности 6D070400 – «Вычислительная техника и программное обеспечение».

Официальный рецензент, к.т.н., ассистент-профессор, декан факультета
Компьютерных технологий и кибербезопасности
АО «Международный Университет Информационных Технологий **Сейлова Н.А.**

Подпись указанного лица удостоверяю
Сейловой
11.12.2021г