УТВЕРЖДАЮ

Ректор Российско-Таджикского
(Славянского) университета
д.э.н., профессор

_____ Файзулло М.К.

«_22_» _____05_____ 2024 г.

ОТЗЫВ

ведущей организации, Российско-Таджикского (Славянского) университета, на диссертационную работу **Косимова Абдунаби Абдурауфовича** на тему **«Статистические закономерности распознавания однородности текстов с помощью γ-классификатора»**, представленную на соискание ученой степени доктора технических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» в разовый диссертационный совет 6D.KOA-049 при Таджикском техническом университете имени академика М.С. Осими

1. Актуальность темы диссертации

Диссертационная работа Косимова Абдунаби Абдурауфовича «Статистические закономерности распознавания однородности текстов с помощью γ-классификатора» посвящено комплексному анализу важной научно-теоретической проблемы, имеющей очевидную прикладную значимость, проблемы стилеметрической идентификации субъекта идиолектной деятельности и рассмотрению возможности классификации и/или кластеризации текстов по набору заданных стилеметрических признаков. Работа представляет собой многоплановое и междисциплинарное исследование на стыке языкознания (теория речевой деятельности, лингвистика текста, корпусная лингвистика), психологии (теория деятельности, восприятие человека) и прикладной информатики (прежде всего, модели и методы машинного обучения).

Результаты диссертации, в равной мере применимые для разных буквенно-алфавитных языков, будут полезны для решения разнообразных практических задач, в частности, для выявления плагиата, распознавания авторства анонимных текстов, определения подозреваемого по составу преступления и др.

Сказанное говорит в пользу актуальности избранной темы диссертации также и потому, что исследования по данной проблеме в Таджикистане начали

разворачиваться лишь в последние несколько лет.

В работе для распознавания однородных текстов используются математические модели принятия решений, среди которых особо успешными являются нейронные сети, машина опорных векторов, метод ближайшего (по расстоянию) соседа и недавно разработанный в Институте математики имени А. Джураева НАНТ γ-классификатор.

2. Структура и содержание работы

Диссертация состоит из введения, шести глав, заключения и списка литературы из 407 наименования. Основная часть диссертации изложена на 271 страницах. Диссертация содержит 107 таблиц и 9 рисунков.

Введение включает обоснование актуальности работы, цели и задачи исследования, объект и предмет исследования, научную новизну и практическую значимость диссертации, а также сведения о публикациях и апробации работы.

На первом этапе автором рассмотрена научная литература по теме диссертации, ставится проблема распознавания однородностей, вводится γ-классификатор.

В §1.1 сообщается обзор литературы по автоматическому распознаванию однородности текста. Применение методов математического моделирования к идентификации однородности текстов опирается в своей основе на модель текста, то есть количественное описание объекта исследования. В настоящее время по подсчетам J. Rudman используется около 1000 групп характеристик в качестве текстовых моделей, среди которых – морфѣлогические, лексические, идиосинкразические, синтаксические, структурные, контентно-специфические и другие характеристики. В дополнение к сказанному уместно отметить, что в монографии А.А. Шелупанова, А.С. Романова и Р.В. Мещерякова представлен обширный обзор работ по распознаванию однородности текста на основе разнообразных ЦП текстов и применяемых методов классификации.

В §1.2 описана постановка задач, решение которых формирует полное представление об эффективности применения γ-классификатора для распознавания однородности произведения.

В §1.3 вводятся терминология и понятие, которые используются при описании математической модели текста.

В §1.4 дается описание используемого в работе метода принятия решения с помощью так называемого γ-классификатора.

γ-классификатор – это математическая триада, состоящая из ЦП текста, формулы расстояний между текстами и алгоритма обучения по прецедентам.

На втором этапе исследована однородность произведений классиков

таджикско-персидской литературы, современных поэтов и современных прозаиков с помощью γ-классификатора и метода ближайшего соседа. Путем применения метрического классификатора и метода ближайшего (по расстоянию) соседа удалось идентифицировать авторов убывающих по размерам последовательности текстовых фрагментов от величины в 7000 слов (40000 символов) вплоть до 20 слов (100 символов).

На третьем этапе рассмотрена однородность тематики текста, языка, группа языков, оригинал и его перевода, стиля произведений и шифров научных работ. Очевидно, что решение такой задачи имеет чрезвычайно важное практическое значение.

В §3.1 автором интересует способность классификатора настраиваться на определение авторства и тематики произведений. В качестве рабочей гипотезы в первом случае будет приниматься утверждение об однородности произведений одного автора и неоднородности произведений различных авторов; во втором случае – однородность произведений по одной тематике и неоднородность по различным тематикам. На примере небольшой коллекции *C* произведений художественной литературы советского периода изучается совместное влияние ЦП, метрического пространства и классификатора текстов на принятие решения об «однородности» и «неоднородности» произведений. С помощью γ-классификатора на предмет возможной «однородности» изучаются пары основных произведений М.А. Шолохова, Н. Островского, Б. Полевой, К. Симонова, А. Фадеева, Д. Фурманова, А.С. Серафимовича и Ф.Д. Крюкова, представляемые девятью различными ЦП.

В §3.2 на примере модельной коллекции текстов устанавливается применимость γ-классификатора для автоматического распознавания языка произведения на основе частотности алфавитных букв.

В §3.3 на примере модельной коллекции текстов на русском и таджикском языках и их переводов на таджикский и русский языки с помощью γ-классификатора и ЦП, характеризующих в текстах распределения частотности буквенных униграмм, исследуется статистическая «однородность» оригинальных и переводных произведений.

В §3.4 определяется применимость γ-классификатора для автоматического распознавания шифра специальности на основе распределения частотности униграмм. Были взяты научные труды, авторефераты разных ученых, написанные на русском языке. Авторефераты были взяты в следующих научных областях: история, педагогика, политология, филология и экономика.

В §3.5 на основе применения γ-классификатора к обработке 68 произведений 7 литературных школ устанавливаются оценки эффективности

распознавания авторства и стилей в рамках таджикско-персидской литературы.

На четвертом этапе исследована однородность реальных и сгенерированных текстов.

В §4.1 устанавливается применимость γ-классификатора для автоматического распознавания языка произведения на основе частотности общих 26 кириллических алфавитных букв на примере корпуса, состоящего из 70 текстов на 20 языках с использованием кириллической графики.

В §4.2 устанавливается применимость γ-классификатора для автоматического распознавания языка произведения на основе частотности общих 26 латинских алфавитных букв на примере корпуса, состоящего из 70 текстов на 20 языках (по 8 произведений на 5 языках: английском, венгерском, латинском, литовском и голландском, и по 2 произведения на других 15 языках) с использованием латинской графики.

В параграфе 4.3 устанавливается применимость γ-классификатора для автоматического распознавания автора произведения на основе частотности 26 кириллических алфавитных букв на примере корпуса, состоящего из 70 поэтических текстов 20 таджикско-персидских авторов (по 8 произведений 5 авторов: А. Суруш, А. Фирдоуси, К. Худжанди, Л. Шерали и Дж. Руми, и по 2 произведения от других 15 авторов) с использованием кириллической графики. На примере корпуса рассмотрены случаи с 5, 10, 20 предполагаемыми авторами, а также 10, 20, 40 текстов, выявляются особенности применения γ-классификатора при распознавании автора текста.

На пятом этапе исследовано влияние порядка ЦП текста на распознавание однородности произведения.

В §5.1 на примере модельной коллекции, количественные описания произведений которой основываются на различных вариантах упорядочения буквенных *N*-грамм (*N*=1,2,3) с пробелами, выявляются особенности применения γ-классификатора при распознавании автора текста. Из огромного количества всевозможных вариантов упорядоченного расположения элементов текста было рассмотрено только четыре: два из них – связаны с алфавитным порядком, и два других – с учётом частотности элементов. Именно в этих двух случаях, прямого и обратного порядков упорядочения элементов, расстояния между любыми парами произведений оказывались равными, вследствие чего равными оказывались коэффициенты π эффективности γ-классификатора, а также и полуинтервалы оптимальных значений γ. В §§5.2.-5.4. исследуются другие допустимые варианты.

На шестом этапе описан объектно-ориентированный компьютерный программный комплекс, созданный для идентификации однородности неизвестного текста на практике.

Заключение содержит основные выводы, которые подтверждают успешное решение поставленных автором задач.

3. Научная новизна основных положений и результатов работы заключается в следующем:

- исследована информативность нетрадиционных признаков на предмет количественного описания таджикских текстов;

- установлена статистическая эффективность π математической модели опознавания авторов произведений таджикской классической поэзии ($\pi = 1.00$) на основе триграмм, современной поэзии ($\pi = 0.98$) с помощью униграмм и современной прозы ($\pi = 0.96$) на основе распределения длин предложений (в словах);

- установлена 100% статистическая эффективность путем применения метрического γ-классификатора и метода ближайшего (по расстоянию) соседа идентифицировать авторов произведений – убывающих по размерам последовательности текстовых фрагментов от величины в 7000 слов (40000 символов) вплоть до 20 слов (100 символов);

- для целей существенного сокращения объёма вычислительных процедур установлена возможность эффективного использования не всех, а только высокочастотных элементов ЦП текстов;

- установлена статистическая эффективность применения γ-классификатора и исследована пригодность ЦП на основе распределения частотности различных алфавитных элементов текста для распознавания других признаков однородности, таких как тематики текста, язык, группа языков, оригинал и его перевод, стиль произведений и шифры научных работ;

- исследованы статистические закономерности опознавания авторов и языков произведений на корпусах художественных литературных произведений с помощью γ-классификатора;

- γ-классификатор и метод ближайшего соседа были протестированы на случайных выборках текстов, распознаются с достаточно высокой точностью признаки однородности произведений различных модельных коллекций и корпусов;

- установлена эффективность применения γ-классификатора для атрибуции искусственно сгенерированных поэм «Шахнаме» А. Фирдоуси по обучению рекуррентных нейронных сетей LSTM (Long short-term memory);

- исследовано влияние порядка ЦП текста на распознавание однородности произведения с помощью γ-классификатора;

- впервые в Таджикистане создан объектно-ориентированный компьютерный программный комплекс распознавания (идентификации) однородности текста на основе различных ЦП текста и γ-классификатора среди сколь угодно большого числа текстов.

4. Практическая значимость работы и реализация её результатов

Комплекс программ под названием **«THR»** (text homogeneity recognition) применён в следующих организациях:

1. Академия Министерства внутренних дел Республики Таджикистан.
2. Государственный комитет национальной безопасности Республики Таджикистан.
3. Институт языка и литературы имени Рудаки НАНТ.
4. Институт математики имени А.Джураева НАНТ.
5. ТТУ имени академика М.С. Осими.

Построенный с широким использованием математических моделей и высокого уровня программирования комплекс, в частности, предназначен для развития таджикского языка с использованием возможностей информационных технологий.

Данный комплекс программы является важным как с точки зрения компьютерной лингвистики, так и с точки зрения литературоведения, и направлен на оказание практической помощи исследователям в области языка, литературы, математики и информационных технологий. Среди них призвано определить и распознать стиль каждого автора, особенности отдельных произведений разных авторов, частоту встречаемости букв, слогов, слов, словосочетаний, состав слов в отдельных произведениях, создание различных математических моделей.

5. Рекомендации по использованию результатов и выводов работы

Спроектированный комплекс рекомендуется для применения в автоматизации процесса обработки текстовой информации в государственной административной деятельности, для установления авторства анонимных текстов в сфере криминалистики, для обнаружения плагиата в курсовых и дипломных проектах в области образования, а также для использования в изучении разнообразных научных проблем, связанных с вопросами распознавания авторства печатных текстов.

6. Публикации и апробация работы

По теме диссертации опубликовано 73 статей, из них – 34 (14 без соавторов) наименований в изданиях, рекомендованных ВАК Республики Таджикистан.

7. Оценка содержания диссертации

Диссертация написана четким и ясным языком с большим количеством графического материала, поясняющего и иллюстрирующего соответствующие результаты научных положений и технических решений. По содержанию работы можно сделать следующие замечания:

1. Имеется целый ряд ошибок технического характера (стр. 15, 22, 85, 89, 101, 105, 131, 233).

2. Неудачное обозначение некоторых формул (глава 1, параграф 1.3, формула (1.7)).

3. Многократно повторяются одни и те же факты (например, что в таджикском языке 35 букв), а также имена некоторых авторов.

4. В диссертационной работе разработан целый ряд программ, относящиеся к разным комплексам, однако в диссертации говориться об одном комплексе.

8. Заключение

Отмеченные недостатки не снижают научной и практической ценности диссертационной работы. Результаты диссертационной работы отражены в периодических изданиях, рекомендованных ВАК при президенте Республики Таджикистан, и доложены на международных научно-практических конференциях и семинарах международного и местного уровня. Автореферат диссертации правильно и полно отражает ее содержание.

Анализ диссертационной работы в целом позволяет сделать следующие выводы:

1. Содержание диссертации соответствует паспорту специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

2. Представленная диссертационная работа Косимова А.А. является самостоятельной, законченной научной квалификационной работой, обладающей признаками актуальности, новизны, внутреннего единства, научной и практической значимости. В ней решена научно-практическая задача – создание программного комплекса **«THR» (text homogeneity recognition)**, предназначенного для распознавания однородности незнакомого текста.

Диссертация полностью соответствует требованиям, предъявляемым к научно-квалификационным работам на соискание ученой степени доктора наук, соответствующее требованиям Положения о присуждении учёных степеней, утвержденного Постановлением Правительства Республики

Таджикистан от 30 июня 2021 № 267 года (в редакции пост. Правительства РТ от 26.06.2023 г. № 295), а ее автор, Косимов Абдунаби Абдурауфович, заслуживает присуждения ученой степени доктора технических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Косимов А.А. выступил с докладом по материалам диссертации на расширенном научном семинаре естественнонаучного факультета РТСУ 17 мая 2024 года.

Отзыв подготовили доктор физико-математических наук, профессор Курбаншоев С.З. и доктор физико-математических наук, доцент Кабилов М.М. Отзыв был заслушан, обсуждён и утверждён на расширенном заседании научного семинара естественнонаучного факультета Российско-Таджикского (Славянского) университета 17 мая 2024 г., протокол №10.

Председатель заседания,
заведующая кафедрой информатики и ИТ,
кандидат экономических наук, доцент **Лешукович А.И.**

Эксперты:
доктор физико-математических наук **Курбаншоев С.З.**

доктор физико-математических наук **Кабилов М.М.**

Секретарь **Ахмедова З.М.**

Сведения о ведущей организации:
Российско-Таджикский (славянский) университет
734025, Республика Таджикистан,
г. Душанбе, ул. Мирзо Турсунзаде, 30
Тел. +992 372 21-35-50,
e-mail: p.rektora@mail.ru
Сайт: http://www.rtsu.tj

Подписи Лешукович А.И., Курбаншоева С.З., Кабилова М.М., Ахмедовой З.М. заверяю.

Начальник управления кадров РТСУ **Рахимов А.А.**