ОТЗЫВ

на диссертацию работу Мухсиной Куралай Женисбековны по теме: «Разработка системы анализа многоязычной текстовой информации на основе машинного обучения», представленную на соискание степени доктора философии (PhD) по специальности «6D070400 – Вычислительная техника и программное обеспечение»

1. Актуальность темы исследования и ее связь с общенаучными и общегосударственными программами

На сегодняшний день наблюдается колоссальный рост количества информации, создаваемой людьми и машинами на естественном языке. Ввиду такого стремительного увеличения объемов текстов, исследования в области компьютерной лингвистики и искусственного интеллекта становятся все более актуальными. Именно данным научным областям и посвящена диссертационная работа Мухсиной К. Ж., связанная с автоматизацией семантического анализа слабоструктурированной и неструктурированной текстовой информации на казахском, русском и английском языках.

Тема диссертационной работы соответствует приоритетным направлениям и программам развития науки Республики Казахстан, связанным интеллектуальными информационными технологиями и машинным обучением, а также с совершенствованием употребления казахского языка в области информатизации и коммуникации. Работа осуществлялась в рамках проекта по грантовым исследованиям МОН РК «AP05131073 Методы и модели поиска и анализа криминально значимой информации в неструктурированных и слабоструктурированных текстовых массивах» (2018-2020 гг.).

2. Научные результаты и их обоснованность

Основная идея работы — разработка и внедрение новой информационной технологии, позволяющей осуществлять автоматический семантический анализ текстов казахского, русского и английского языков. Реализуемый семантический анализ направлен на извлечение из неструктурированных и слабоструктурированных текстов информации в виде структурированных триплетов фактов.

Научные результаты Мухсиной К. Ж. по теме диссертации получены на основе корректной постановки задач исследования и их последовательным решениям. В диссертации приведены следующие полученные результаты:

- Разработана модель извлечения фактов из слабоструктурированных и неструктурированных текстовых массивов, которая адаптирована для казахского, русского и английского языков; обоснован выбор математического аппарата алгебры конечных предикатов для моделирования семантики предложений естественного языка;
- Модифицирован метод автоматической морфологической и семантической разметки текстовых корпусов казахского, русского и английского языков;
- Разработан метод определения семантической близости текстовых документов на казахском языке к узкоспециализированной предметной области; показана практическая реализация метода для определения принадлежности документов к криминально окрашенным текстам;
- Разработана методика экспертной оценки качества работы системы анализа семантической близости текстов;

Создан программный комплекс, позволяющий определить наличие криминального оттенка и осуществляющего семантическую разметку текстов казахского, русского и английского языков.

3. Степень обоснованности и достоверности каждого научного результата (научного положения), выводов и заключения соискателя, сформулированных в диссертации

Обоснованность и достоверность научных положений, выводов и рекомендаций основывается на корректном применении общей теории систем, системном анализе, а также на совместном использовании уже известных моделей: Скрытой Марковской модели (HMM) и векторной модели (VSM) с классическими алгоритмами машинного обучения. Обоснованность и достоверность результатов исследования подкрепляется внедрением разработанных рекомендаций, что подтверждается имеющимся актом внедрения.

Основные результаты диссертационной работы неоднократно докладывались и обсуждались на семинарах кафедры информатики Казахского Национального университета им. аль-Фараби, Института информационных и вычислительных технологий МОН РК и на международных научных конференциях. Большая их часть опубликована в профильных научных изданиях, в том числе, имеющих высокий рейтинг. Решение каждой задачи опирается на полученные результаты предыдущих этапов исследования, что обуславливает их взаимосвязь и взаимозависимость, а также внутреннее единство полученных результатов.

4. Степень новизны каждого научного результата (положения), выводов и заключения, сформулированных в диссертации

Результаты, полученные в диссертации Мухсиной К. Ж., являются новыми и дополняют известные. А именно:

– разработана логико-лингвистическая модель семантического анализа, позволяющая явным образом идентифицировать факты в текстах казахского, русского и английского языков и представлять их структурированно в виде RDF-триплетов, формируя семантически размеченные обучающие корпуса;

– модифицирован метод автоматической морфологической и семантической разметки текстовых корпусов казахского языка, отличительной особенностью которого является одновременное использование модели HMM (Hidden Markov Model) и правил, представленных регулярными выражениями;

– усовершенствован метод определения семантической близости текстовых документов казахского языка к узкоспециализированной предметной области;

– разработана методика экспертной оценки качества работы системы анализа семантической близости текстов, базирующаяся на вычислении среднего значения косинусного сходства.

Также Мухсиной К. Ж. разработан программный комплекс, позволяющий определить наличие некоторого криминального содержания в текстах казахского, русского и английского языков и осуществить их семантическую разметку.

5. Практическая и теоретическая значимость научных результатов, направленных на решение актуальной проблемы, теоретической и прикладной задачи.

Диссертационная работа является квалификационным научным трудом, содержащим научно обоснованные результаты, решение которых направленно на улучшение качества автоматической обработки текстов казахского языка и повышение эффективности работы систем мульти язычного автоматического анализа.

Значимость полученных в диссертационном исследовании Мухсиной К. Ж. практических результатов заключается во внедрении программного комплекса, позволяющего определить наличие криминального оттенка в текстах казахского, русского и английского языков.

Прикладная ценность результатов работы заключается в возможности государственным органам получать информацию о наличии противоправных текстов в социальных сетях.

6. Соблюдение в диссертации принципа самостоятельности

Представленная диссертационная работа является самостоятельным исследованием, имеющим научную и практическую значимость. Результаты работы подтверждены актом внедрения и апробированы публикациями в журналах, рекомендованных ККСОН МОН РК, и в международных журналах, входящим в базу Scopus и Web of Science. Полученные результаты также докладывались на международных конференциях.

Основные научные результаты докторской диссертации опубликованы в 17 научных трудах, в том числе: 2 статьи (одна из которых с процентилем – 76) и 5 конференций опубликованы в изданиях, индексируемых в базах данных Web of science и Scopus; 4 – в научных изданиях, рекомендуемых ККСОН МОН РК; 6 – в материалах международных научных конференций.

На разработанную программу получено свидетельство авторского права. Индекс Хирша Мухсиной К. Ж. равен 3.

7. Соответствие аннотации содержанию диссертации

Содержание аннотации соответствует содержанию диссертации, отражает все четыре главы, заключение, научные положения, научную новизну и практическую значимость.

8. Замечания, предложения по диссертации

Последовательность основных этапов выполнения и структура диссертации соответствуют логике научного исследования, полностью отвечают ее цели и задачам. При этом соблюдено внутренне единство результатов диссертации. Однако, к работе имеются отдельные замечания:

1. Недостаточно обосновано использование весового коэффициента PPMI в матрицах термин-документ разработанного метода определения семантической близости текстов к узкой специализированной области знаний (подраздел 3.3).
2. Следовало более подробно описать ограничения предложенной логико-лингвистической модели извлечения фактов из многоязычной текстовой информации. Например, ко всем ли естественным языкам она может быть применима и как именно.
3. Не совсем понятна файловая структура разработанного параллельного корпуса, приведенная на рисунке 4.2 диссертации.

Тем не менее, работа соответствует требованиям, указанные замечания не снижают актуальность и качество выполненных исследований и полученных результатов.

9. Заключение о возможности присуждения соискателю степени доктора философии (PhD) по специальности 6D070400 – «Вычислительная техника и программное обеспечение»

Диссертационная работа Мухсиной К. Ж. выполнена на высоком научном уровне и представляет собой завершённую научно-квалификационную работу. Полученные автором результаты являются новыми, обоснованными и достоверными.

На основании вышеизложенного считаю, что диссертационная работа Мухсиной Куралай Женисбековны на тему «Разработка системы анализа многоязычной текстовой информации на основе машинного обучения», представленная на соискание степени доктора философии (PhD) по специальности «6D070400 – Вычислительная техника и программное обеспечение», соответствует всем требованиям «Правил присуждения ученых степеней» ККСОН МОН РК, предъявляемым к работам такого рода, как по содержанию, так и по объему.

Соискатель Мухсина Куралай Женисбековна заслуживает присуждения степени доктора философии (PhD) по специальности «6D070400 – Вычислительная техника и программное обеспечение».

Официальный рецензент:
PhD, ст.преподаватель
кафедры «Информационных систем»,
КазНУ имени аль-Фараби Омаров Б.С.

ЗАВЕРЯЮ
Начальник управления подготовки
научных кадров КазНУ им. аль-Фараби
Р.Е. Кудайбергенова
«_____» _____ 20__

«ӘЛ-ФАРАБИ АТЫНДАҒЫ ҚАЗАҚ ҰЛТТЫҚ УНИВЕРСИТЕТІ» КОММЕРЦИЯЛЫҚ ЕМЕС АКЦИОНЕРЛІК ҚОҒАМЫ
НЕКОММЕРЧЕСКОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО «КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ ИМЕНИ АЛЬ-ФАРАБИ»
NON-PROFIT JOINT-STOCK COMPANY «AL-FARABI KAZAKH NATIONAL UNIVERSITY»
ӘЛ-ФАРАБИ УНИВЕРСИТЕТІ
БИН 990140001154