РЕЦЕНЗИЯ

на дисертационен труд за придобиване на образователна и научна степен „доктор" в област на висше образование 4. „Природни науки, математика и информатика", Професионално направление: 4.6 „Информатика и компютърни науки", Научна специалност: 01.01.12 „Информатика"(Изкуствен интелект)

Автор на дисертационния труд: *Тодор Васков Цонков*

Тема: *„Изследване на мнения и чувства от текстове от социалните мрежи"*

Научен ръководител: *проф. д-р Иван Койчев*

Рецензент: *доц. д-р Светла Бойчева - ИИКТ-БАН*

Тази рецензия е написана и представена на основание на заповед РД38-99/17.02.2016 г. на ректора на СУ „Св. Климент Охридски", както и на решението на научното жури по процедурата (Протокол 1 от 19.02.2016). Тя е изготвена въз основа на ЗРАСРБ, Правилника за прилагане на ЗРАСРБ, Правилника за условията и реда за придобиване на научни степени и за заемане на академични длъжности във Факултета по математика и информатика на СУ „Св. Климент Охридски" и указания за изготвяне на рецензии и становища от членове на научни журита и за документите в електронен вид, подавани от кандидата по процедури за академични длъжности и научни степени на Факултета по математика и информатика на СУ "Св. Климент Охридски".

1. Съдържателен анализ на научните и научно-приложните постижения в дисертационния труд. Характеризиране на основните постижения.

Дисертационният труд се състои от 95 страници. Оформен е в пет глави, увод, заключение, авторска справка и списък на цитираните литературни източници. Съдържа 17 фигури и 24 таблици. Използваната литература е от 83-2 източника на английски език и от 3 интернет сайта, като 28 от печатните издания са публикувани през последните 5 години, което показва познаване на съвременното състояние на областта. Използваната литература е цитирана по подходящ начин в текста на дисертацията.

Актуалност на проблема

Съвременните комуникации навлизат все по-дълбоко в ежедневието ни и изместват традиционното общуване, за което са характерни емоционалните аспекти. Форумите и социалните мрежи генерират огромен обем информация, който е непосилно да се следи от човек. Ежедневието ни е изцяло диктувано от мнението на останалите, както в професионален, така и в личен план. В глобален план представлява интерес и от гледна точка на сигурността. Автоматичният анализ на мненията и чувствата в текстове би спомогнал за подпомагане вземането на решения в различните аспекти на нашия живот, ето защо това е

много актуална тема. От друга страна това е една много трудна и предизвикателна задача, която изисква задълбочени изследвания и разработването на модели и методи за анализ на мнения и чувства в съдържание в социалните мрежи.

Познаване на състоянието на проблема

В дисертационния труд е направен обстоен задълбочен анализ на изследвания проблем и различни негови аспекти. Подробно са описани съществуващите методи за решаване на проблема като са разгледани и подобни разработки. Всичко това показва задълбоченото познаване на проблема. Изложени са предимствата и недостатъците на различните методи. На базата на този анализ са предложени алгоритми за решаване на проблема. Използваните литературни източници също показват, че кандидатът познава естеството на проблема.

Подход и решение на проблема

Настоящият дисертационен труд има за цел да изследва двойното и обратно значение (ирония, сарказъм, сатира, хипербола, литота, лъжи и невярна информация) и разработване на подход за неговото автоматично откриване в текстове на български и английски език.

За постигане на целта на дисертационния труд, кандидатът е решил следните задачи:

- Обзор на съществуващите подходи за разпознаване на на ирония, сарказъм, хипербола и неверни твърдения в текст;
- Разработване на подход за автоматично откриване на двойно значение в текстове на английски език и текстове на български език;
- Изграждане на прототип на система за автоматично откриване на двойно значение в текстове;
- Оценяване на ефективността на предложения подход.

Основни приноси

Постигнатите резултати при изпълнение на задачите на дисертационния труд съответстват на поставента цел в дисертационния труд. Разработени са подходи за автоматично откриване на двойно значение в текстовете на английски език и текстове на български език. Реализирано е приложение на предложените подходи и са описани експерименти с някои от най-разпростанените социални мрежи. Получените резултати от експериментите подкрепят валидността на предложените подходи.

Анализ на научните и научно-приложните постижения вдисертационния труд

В **увода (първа глава)** са поставени целта и задачите на дисертацията. Дадени са някои основни дефиниции и е направена обосновка на актуалността на разглежданата тема.

Във **втора глава** е направено проучване и анализ на съвременното състояние проблема. По специално внимание е отделено на най-разпространените инструменти за машинно самообучение.

В **трета глава** е описан нов подход базиран на правила за автоматично откриване на двойно значение (ирония, сарказъм, сатира) в текстовете на английски език и текстове на български език. Дефинирани са две групи евристични правила – независими от езика и специфични за езика. Описани са основните характеристики на мнения. Направени са експерименти с три класификатора: Naïve Bayes, K-NN и Support Vector Machine. Описани са експерименти за проверка на предсказващата точност на предложения подход за текстове от социални мрежи на български и на английски език.

В **четвърта глава** са описани са нови подходи базирани на правила за:

- автоматично засичане на неверни твърдения от социалните мрежи Facebook и Twitter в английски текстове;
- автоматично определяне на мнения, които съдържат надценяване и подценяване в себе си върху текстове на български език от социалните мрежи Facebook и Twitter;
- автоматичното засичане на лъжи в текстови съобщения.

В **пета глава** са описани са нови подходи базирани на правила за:

- извличане на мнения от текст;
- автоматично засичане на слухове;
- подобрение на списъка от думи и изрази носещи определено отношение (положително или отрицателно) към даден обект.

В **шеста глава** е описано разработването на система за извличане на потребителски текстове и автоматично извличане на текст от социалните мрежи. Описани са подробно всички етапи – анализ на изискванията, дизайн, потребителски интерфейс, реализация на Java, тестване.

В края на дисертационния труд са представени **заключението** и **перспективите за бъдещо развитие**.

Приноси в дисертацията

Научни приноси

- Анализирани са двойното и обратно значение (ирония, сарказъм, сатира, хипербола, литота, лъжи и невярна информация) в текстове на български и английски език. Разработени са езиково зависими и независими правила за тяхното разпознаване. Дефинирани са основните характеристики за представянето на мнения и чувства.
- Анализирани са основните подходи за автоматично откриване на двойно значение в текстове, както и на мнения и чувства. Разработени са нови алгоритми базирани на правила за:
    - автоматично откриване на двойно значение (ирония, сарказъм, сатира) в текстовете на английски език и текстове на български език в социалните мрежи;
    - автоматично засичане на неверни твърдения от текстове на английски език в социалните мрежи;
    - автоматично определяне на мнения, които съдържат надценяване и подценяване в себе си в текстове на български език от социалните мрежи;
    - автоматичното засичане на лъжи в текстови съобщения.

- автоматично извличане на мнения от текст;
- автоматично засичане на слухове;
- подобрение на списъка от думи и изрази носещи определено отношение (положително или отрицателно) към даден обект.

Научно-приложни приноси

- Реализирана е система заопределяне на мнения и чувства в текстове от социалнит мрежи, базирана на разработените алгоритми.

Достоверност на получените резултати

За достоверността на дисертацията говорят петте публикации, изнесените доклади по темата на дисертацията и реализацията на система за извличане на мнения от социалните мрежи и тяхната оценка.

Лично посетих уеб приложението на разработената система и направих тестове. Имам положителни впечатления от работата на системата.

2. Общо описание на публикациите, които отразяват дисертацията – монографии, статии, свидетелства и патенти, класифицирани по тематика или друг признак и редуцирани поради съвпадение или препокриване

Според правилника на ФМИ за образователната и научна степен "доктор" се изискват поне 2 публикации в рецензирани издания, поне едно от които да е списание. По дисертацията има 5 публикации, като 1 е в научно списание издание на Пловдивски университет, 1 в сборник с доклади на международна конференция и 3 на национални конференции. Две публикации са на български език, а останалите три публикации са на английски език. Публикациите отразяват основните научни резултати, постигнати в дисертацията. Четири от публикациите са в съавторство с научния ръководител, а една от публикациите е в съавторство с друг наскоро защитил докторант със същия научен ръководител. Не са представени самостоятелни публикации за рецензия. Освен това част от резултатите са представени и в доклади на международни семинари в чужбина.

3. Отражение на резултатите от дисертацията в трудове на други автори. Числови показатели – цитати без автоцитатите), импакт-фактор и др.

Забелязано е 1 цитиране на публикацията „Automatic Detection of Double Meaning in Texts from the Social Networks" в наскоро защитен труд на тема „Computational Analysis on a Corpus of Political Satire Articles: A Theoretical and Experimental Study" на студент от Università Ca' Foscari Venezia. Това още веднъж показва, че темата е актуална и се провеждат сходни изследвания в тази област от студенти от други университети по света.

4. При колективни публикации да се отрази приносът на кандидата

Според мен приносът на кандидатът при колективните публикации е ясен и съществен. Той личи от засегната тематика. На всички публикации докторантът е първи автор.

5. Критични бележки и препоръки на рецензента

Бих направила някои технически бележки, които не намаляват стойността на представения дисертационен труд от гледна точка на постигнатите резултатите от кандидата, но биха способствали за подобряването на изложението:

- Текстът на дисертационния труд се нуждае от корекции, както като форматиране, така и от стилистична гледна точка. Има известни пунктуационни и правописни грешки, чието отстраняване само би спомогнало за подобряване на общото впечатление от представените научни резултати.
    - В предоставеното печатно копие липсват номера на страниците, както на дисертационния труд, така и на автореферата, което затруднява четенето и следването на препратките.
    - Фигури:
        - Липсва фигура под номер 9 – прескача се от фиг. 8 направо на 10;
        - Някои фигури са с лошо качество в печатното издание (например, фиг. 5, 8, 10).
    - Таблици:
        - Таблици с номера 20 и 26 са описани в списъка на таблиците, но липсват в текста;
        - В Таблица 1 има разностилови мерни единици – на английски и български език;
        - Таблиците са разностилно форматирани;
        - Използването на разделител за единиците при големите числа би спомогнало за по-бързата и лесна интерпретация. Само на няколко места е използван такъв разделител (Например числото 2,000 на стр. 46, трети параграф). При такова представяне би било по-удачно да се представят числовите данни подравнени вдясно.
    - Формули:
        - Някои формули са указани като фигури (фигура 6);
        - Липсва номерация на формулите като цяло;
        - Добре би било да се използват стандартните математически означения във формулите – като например знака за умножение вместо *;
        - Има сгрешени горни и долни индекси в някои формули.
    - Псевдокод (формално описание):
        - Въпреки че няма единен стандарт за представяне на псевдокод – може да се използват някои от по-често разпространените конвенции, като използването на C-подобни оператори например, които биха спомогнали за подобряване на четабилността. Добре би било да се използва друг шрифт, за да се отличават от текста на изложението. Част от представените алгоритми биха могли да се онагледят и с блоксхеми или друг вид диаграми, които ще улеснят възприятието.
- Добре би било дефинициите на основните понятия да се изнесат извън основния текст на изложението и да се номерират, за да може по-лесно да се

правят справки и препратки към тях. Едно по-формално дефиниране на понятията би предало по-голяма стегнатост на изложението. Необходимо е да се прецизира използваната терминологията в дисертационния труд.

- Примерите се сливат с цялостното изложение – използването на друг шрифт или курсив би подобрило четабилността. Въпреки, че на места са използвани кавички за по-малките фрази и изрази, не стои така въпросът с цели параграфи с примерен текст (като например на стр. 17, втория параграф от началото на секция 2.1), където примерното мнение е непосредствено следвано от неговия анализ и липсват кавички. Това е частично направено в автореферата.
- Целите на дисертационния труд могат да се прецизират като се обособи основна цел и задачи за решаването й. В заключението в последния параграф на секция 7.1 е описана целта на дисертацията, която е изпълнена, но липсва дефиницията й в този вид в глава 1. Добре би било да се дефинират и работни хипотези и изследователски въпрос.
- Използвана литература:
    - няма единен стил на форматиране;
    - на места има пропуски на номера на страници, издател и/или година на публикуване;
    - за някои източници е посочена само фамилия на автор без инициал на първото име;
    - за интернет източниците лиспва дата на последно посещение;
    - Липсва пълен линк до използваните ресурси на фиг. 1 и таблица 1;
    - Не е добре параграфите да започват с цитат на източник;
    - Списъкът е сортиран като за някои автори е използвана фамилията, а за други собственото име;
    - Има пропуски и неточности при цитиранията на източници - някои източници са цитирани в текста на дисертацията, но не са посочени в библиографията, има сгрешени години при цитирания на източници и многозначности;
    - Част от източниците се повтарят по два пъти в списъка с използвана литература;
    - Липсва цитиране в текста на дисертацията на някои източници.
- Интересно би било да се даде информация не само за точността на предложените алгоритми, но и за тяхната покриваемост и F-score, за да може да се оцени по комплексно ефективността на тези методи.
- За представените „регулярни изрази“ и шаблони би било добре да се използва някои стандартен формат –например Perl Regex.
- Не са унифицирани термините и наименованията –използвано е както Twitter, Facebook, Google+, така и Туитър, Фейсбук и Гугъл+. Погрешно за името на езика за обобщено програмиране Scala е използвана транслитерация – Скала. За класификаторите са използвани както иметнат Наивен Бейсов класификатор, К-най-близки съседи, така и оригиналните им имена Naive Bayes и K-NN.
- Добре е да се използва името на разработената система MentionGraph в описанието й. Това име е видно само от приложените екрани на системата. Трябва експлицитно да се зададе и линк към системата.
- Необходимо е да се прецизира формулировката на приносите, като се обобщят.
- Авторска справка:

- Заглавието на публикация 2 е на български език, а по принцип публикацията е на английски език;
- В библиографската справка липсва информация за броя страници на публикациите;
- В текста на дисертацията има цитирания на описаните публикации, но само една от тях е включена в списъка с използвана литература.

- Като препоръка към кандидата може да се отправи съвета да публикува резултатите от научните си изследвания в реномирани международни научни списания, както и да участва в международни конференции, тясно специализирани в тематиката на научните изследвания.

6. Качества на автореферата, включително доколко правилно отразява приносите на дисертацията

Авторефератът отразява основните резултати, постигнати в дисертацията. Обемът на представения автореферат позволява още малко да се разшири изложението, така че да могат да се включат още резултати от дисертационния труд, които биха представлявали интерес за научната общност.

7. Заключение

Представеният за рецензиране дисертационен труд отговаря на изискванията на Закона за РАСРБ и на съответните Правилници на МОНМ, СУ и ФМИ.

Предвид горното и поради научните приноси на кандидата в дисертационния труд, давам **положителна оценка и предлагам на уважаемото жури да присъди на *Тодор Васков Цонков* образователната и научна степен "доктор"** в област на висше образование, 4.0. Природни науки, математика и информатика, професионално направление 4.6. Информатика и компютърни науки, научна специалност: 01.01.12. Информатика.

София, 10 май 2016 г.

Рецензент:
/доц. д-р Светла Бойчева/