СОФИЙСКИ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФИЛОСОФСКИ ФАКУЛТЕТ

Проектът „Социологията в дигиталните медии“ се състои в разработване на софтуер и инструментариум за контент-анализ на големи бази данни, и по-специално за контент-анализ на присъствието на българската социология в българските медии.

Съответствие на резултатите с поставените в проекта цели

Проектът има няколко основни цели:

- Създаване на алгоритъм за извличане на тестови материали от интернет и тестването му чрез създаването на база данни от такива
- Разработване на алгоритъм за количествено изследване и респективно,
- Осъществяване на количествено изследване под формата на анализ на съдържанието (контент-анализ) с помощта на този алгоритъм и също така, валидиране и апробиране на алгоритъма
- Популяризиране на резултатите от количественото изследване чрез публикации/ участие в конференции

В следствие на работата по проекта не само беше създаден и апробиран алгоритъм за извличане на големи текстови данни, отделно от това бяха и разработени насоки за работа с такъв тип извличане, които ще помогнат не само в научно-изследователската дейност и на други екипи, но също и в обучителния процес в Софийския университет.

На базата на извлечения текстов масив бяха осъществени две изследвания:

- Пилотно изследване, базирано на извадка от извлечените материали
- Същинско изследване върху целия текстов масив, което имаше две основни цели: да отговори на въпроса доколко социологията е популярна/често споменавана в големите български дигитални медии и от друга страна, на базата на реализирането изследване да се създаде алгоритъм (поредица от стъпки) за количествен контент-анализ, който да служи за работа в други проблемни области.

Изпълнение на дейностите по проекта съгласно работната програма

Работата по двете изследвания беше предвидена и се извърши онлайн. По обективни причини, работата по проекта стартира по-късно, а това наложи хронологично разместване на някои от дейностите в работната програма, с цел пестене на време.

Тематичен проект:

„СОЦИОЛОГИЯТА В ДИГИТАЛНИТЕ МЕДИИ “

Ръководител:
Гл. ас. д-р Антоанета Гетова

Договор № 80-10-164/ 24.04.2020

Период на изпълнение на проекта: април – ноември 2020

Обща стойност:
5 644 лв.

Друго, което наложи разместване на дейностите във времето, беше работата по текстовия масив, което наложи и допълнително валидиране на данните по два начина: първият се осъществи чрез повторно извличане на извадка от масива със скрипт на Python (дейност, извършена от външния технически изпълнител) и допълнителна ръчна проверка на данните от източници, при които извличането на материали беше затруднено (дейност, която се извърши от ръководителя на проекта).

Независимо че тези дейности допринесоха за нанасяне на корекции и респективно апробиране на основния алгоритъм за извличане на големи текстови данни, това наложи допълнително преструктуриране на поредицата от извършвани дейности, като някои бяха изтеглени по-напред в пилотното изследване (например качествения анализ на данни). Това се наложи отново с цел реорганизация на времето за работа по проекта.

Проектът включва следните дейности:

- Създаване на списък с критерии за подбор на медии, от които да бъдат извлечени материалите, обект на анализа. Тази дейност беше изключително важна, тъй като на базата на списъка с критерии бяха подбрани и източниците на извличане на материалите. С помощта на тези критерии бяха подбрани 10 популярни български новинарски сайта, от които бяха извлечени публикации съдържащи думата „социология" или някоя от производните й (социологически, социолог) във всичките им форми(ед.ч и мн.ч., членувани и нечленувани и т.н.). Подробното описание на този списък с медийни сайтове е част от приложения анализ на данни от основното изследване.
- След това беше извлечена база от данни с публикации от 3 от тези 10 сайта, на база на които е направено пилотното изследване. Респективно, извличането е направено на базата на предварително създадения за целта алгоритъм. След това е направена проверка и изчистване на базата от данни и нанасяне на корекции в самия алгоритъм.

Пилотното изследване включва следните основни дейности:

- Реализиране на количествен контент-анализ на медийните материали на тези три сайта.
- Реализиране на качествен контент-анализ на същите медийни публикации
- На база на получените резултати от предварителното (пилотното) проучване са извлечени хипотези, които са проверени в хода на основното изследване.
- Междувременно е извлечена и базата данни за основното изследване (т.е. целият текстов масив). Базата е проверена и изчистена от грешки ръчно, чрез допълнителна статистическа проверка и чрез втори алгоритъм на базата на скрипт в Python.
- След проверките са нанесени съответните корекции в алгоритъма, който е разписан под формата на основни насоки (инструкция) за извличане на големи текстови данни.
- Коригирания и изчистен от грешки пълен текстов масив на публикациите, съдържащи социология и/или производните й за периода януари 2016-юни 2020 г. е използван за база на основното изследване (количествен контент-анализ на публикации). Извлечени са над 15000 публикации, съдържащи поне една от тези думи, които са анализирани чрез набора от индикатори, извлечен по време на пилотното изследване и допълнен с методи за автоматизирано извличане на теми/ключови думи, предложен от външния консултант . На база на получените резултати е апробиран и коригиран алгоритъма за количествен контент-анализ, предложен предварително от външния консултант.
- Получените резултати от основното изследване се предвижда да бъдат представени на предстоящата конференция на Международната социологическа асоциация, а в последствие обобщени и в научна публикация на български език.

Обосновка на извършените разходи по проекта

Бюджетът е предвиден за следните две пера: възнаграждения на външни консултанти и технически изпълнители и разходи за участие в конференцията на Международната социологическа асоциация.

Разходите за хонорар на външния технически изпълнител са предвидени по отношение на следните дейности: разработване на алгоритъма за извличане на статии (което е една от основните заложни цели на настоящия проект) и самата работа по извличане на статии и създаване на база от данни с текстови материали за контент-анализ.

В процеса на валидиране на базата с данни от изследването са се наложили множество допълнителни дейности по изчистването на грешки и проверката на базата с текстови данни още на ниво пилотно

изследване. Тези допълнителни дейности се са извършени поради неизвестни и непредвидени бъгове в сайтовете с публикации, които са възпрепятствали автоматизирането на извличането, налага се с цел проверка част от данните да бъдат извличани ръчно, в която техническа дейност се включва ръководителят на проекта. Тази дейност е компенсирана чрез прехвърляне на разходи от второто перо на бюджета (предвидените разходи за заплащане на такса участие в конференцията), които остават неизползваеми тъй като поради пандемичната обстановка се променят условията за участие и заплащане на таксите за конференцията. Прехвърлянето на средствата е описано във финансовия отчет.

Останалите разходи за хонорар за външен консултант, са предвидени за работата на външен експерт който участва в основното изследване, реализирано по проекта в две основни направления: разработване на алгоритъма за количествен контент-анализ на големи текстови данни и респективно избора и тестването на методи, които да бъдат включени в този алгоритъм. Работата по количественото изследване на външния консултант също беше завършена качествено и в срок.

Обобщение на постигнатите научни резултати от проекта

В следствие на работата по проекта са постигнати няколко важни научни резултата.

Първият беше реализацията на самото изследване, което търси отговор на въпроса доколко конкретна научна сфера (в случая социологията) изобщо намира място в публикациите на популярни дигитални медии, което само по себе си представлява и значим обществен проблем. По-подробно резултатите са описани в приложения анализ на основното изследване, тук следва само да се отбележи че академичната социология не се радва на твърде висока популярност в големите дигитални български медии.

Важен резултат от проекта е че на база на анализа на този казус е създаден систематичен подход за количествено изследване на големи текстови данни. Подходът би могъл да се приложи по отношение на други обекти на изследване, включително и непряко свързани с академичната сфера.

Друг важен резултат е също и фактът, че реализацията на този контент-анализ е базирана само и единствено на софтуер за анализ на големи данни, който е със свободен достъп (Knime), което увеличава възможностите да бъде прилаган по-широко и от други изследователски екипи.

С помощта на софтуер със свободен достъп е реализиран и другият важен резултат от проекта: самото извличане на големи текстови данни. Наличието на тази база от данни би могло да се ползва в работата по бъдещи публикации и научни разработки, което е една от предстоящите цели на настоящия научен колектив.

Друг важен резултат на проекта е разработването на алгоритъм (поредица от стъпки) за извличане на големи тестови данни (big data) и респективно апробирането на този алгоритъм в процеса на работа по текстовия масив, използван в изследването на проекта. Във връзка с това извличане бяха разработени и основни насоки (инструкция) за работа по извличане на такива данни, които биха могли да бъдат използвани от други изследователи за техните проекти. Такива насоки липсват в справочници и учебници за контент-анализ, освен това, на практика липсва и структурирана документация на български език. Тези насоки биха могли да се използват не само в научно- изследователска работа, но и в процеса на обучение на множество курсове в Софийския университет, включително и специализираният курс по контент-анализ в специалност „Социология". Ето защо, изработването на тези насоки се оценява като изключително важно достижение на научния колектив. С цел максимална популяризация на насоките, те са качени със свободен достъп в Researchgate.

Разпространение на резултатите:

Приета заявка за участие в конференция на Международната социологическа асоциация, IV ISA Forum of Sociology (February 23-27), host ISA (конференцията ще бъде проведена онлайн заради пандемията от коронавирус). Докладът е на тема: On Media Representation of Sociology: A Case Study of the Bulgarian Digital Media и ще представи резултатите от проведеното по време на проекта основно изследване. Конференцията беше отложена във времето отново заради пандемията, първоначалните дати бяха определени за 2020 г.

Насоките (инструкцията) за извличане на големи текстови данни от интернет, която описва алгоритъма на такова извличане, използван и в работата по изследването в настоящия проект, са качени в научната мрежа Researchgate с цел максимална достъпност. Така и други автори ще могат да се възползват свободно и безплатно от постигнатите резултати от проекта.