

Облікова картка ДіР



I. Загальні відомості

Державний обліковий номер: 0223U005421

Державний реєстраційний номер: 0123U104030

Особливі позначки: відкрита

Дата реєстрації: 22-12-2023

II. Етап виконання ДіР

Номер етапу: 1

Назва етапу: Створення анотованого багатомовного текстового корпусу достовірною урядовою інформацією (ТІС) та невизначеною інформацією Інтернет контенту з визначеної ПО.

Початок етапу: 08.2023

Закінчення етапу: 12.2023

Вид звітного документа: Остаточний звіт

III. Відомості про виконавця ДіР

Повне найменування юридичної особи (або ПІБ фізичної особи): Національний технічний університет "Харківський політехнічний інститут"

Код за ЄДРПОУ: 02071180

Місцезнаходження: вул. Кирпичова, буд. 2, м. Харків, Харківський р-н., Харківська обл., 61002, Україна

Форма власності:

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR: Не застосовується

Розмір організації:

Телефон: 380577076634

IV. Відомості про співвиконавців ДіР

V. Відомості про замовника ДіР

Повне найменування юридичної особи (або ПІБ фізичної особи): Національний фонд досліджень України

Код за ЄДРПОУ: 42734019

Місцезнаходження: вул. Бориса Грінченка, 1, м. Київ, 01001, Україна

Форма власності:

Сфера управління: Кабінет Міністрів України

Ідентифікатор ROR: Не застосовується

Розмір організації:

Телефон: 380442981622

VI. Джерела, напрями та обсяги фінансування ДіР

Підстава для проведення ДіР: 34 - договір (замовлення) з центральним органом виконавчої влади, академією наук (головними розпорядниками бюджетних коштів на проведення НДДКР)

Напрямок фінансування: 2.5 - програми і проекти у сфері міжнародного науково-технічного співробітництва

Джерела фінансування

7713 - кошти держбюджету

Код програмної класифікації видатків і кредитування (КПКВК): 2201380

Фактичний обсяг фінансування (тис. грн.): 199.000
--

VII. Відомості про ДіР

Назва роботи українською:

Визначення достовірності публічної інформації у соціальних медіа із застосуванням підходу текстової аналітики

Назва роботи англійською:

Determining the trustworthiness of public information on social media using linguistic-based text analytics approach

Реферат українською:

Об'єкт дослідження: методи та моделі контрольованого та неконтрольованого машинного навчання, що використовуються для класифікації достовірної та недостовірної текстової інформації. Метою проекту є розробка практичного підходу до визначення правдивості публічної інформації в соціальних мережах, який використовує методи обчислювальної та лінгвістичної аналітики, для (1) розширення ступеня розуміння і характеру достовірності публічної інформації соціальних медіа-платформ, та (2) визначення основних тем інформаційних потреб громадян у спілкуванні з урядом на державних цифрових платформах та у соціальних мережах. Отримані результати: (i) створено та анотовано багатомовний корпус достовірної інформації та інформації, надійність якої не визначена; корпус залучає тексти, щодо обраної тематики, української, польської та англійської мовами, які вилучено з соціальних мереж, новинних вебсайтів та офіційних урядових Інтернет-джерел; (ii) на базі корпусу розроблено анотований набір даних; (iii) проведено лінгвістичний препроцесорний аналіз текстів, що залучає крім традиційних етапів pipeline обробки природної мови, а саме, сегментацію речень, токенізацію та лематизацію слів, POS-тегування, та визначення синтаксичних залежностей, ще додаткову семантичну обробку корпусу: Latent Dirichlet Allocation, Topic Modeling та Sentiment Analysis. Основні конструктивні та техніко-економічні показники, ефективність: підвищення якості семантичного аналізу що до достовірної та недостовірної текстової інформації української, польської та англійськими мовами, створення нового анотованого набору даних, що може використовуватися для оцінки моделей класифікації правдивої та неправдивої інформації Галузі застосування: безпека та кібербезпека, комунікація держави з суспільством через соціальні мережі, створення додатків цифрового уряду, розвиток програм та проєктів з медіаграмотності, освіта, медіа та журналістика.

Реферат англійською:

The object of research: methods and models of supervised and unsupervised machine learning used for classifying reliable and unreliable textual information. The project aims to develop a practical approach to determine the truthfulness of public information in social networks using computational and linguistic analytics methods, for (1) expanding the understanding and nature of the credibility of public information on social media platforms, and (2) identifying the main topics of citizens' information needs in communicating with the government on state digital platforms and social media. Obtained results: (i) the multilingual corpus of reliable and undetermined reliability information has been created and annotated; the corpus includes texts in Ukrainian, Polish, and English languages extracted from social media, news websites, and official government Internet sources on selected topics; (ii) an annotated dataset has been developed based on the corpus; (iii) linguistic pre-processing analysis of texts, including traditional stages of natural language processing pipeline such as sentence segmentation, tokenization, word lemmatization, POS tagging, and syntactic dependency parsing, as well as additional semantic processing of the corpus: Latent Dirichlet Allocation, Topic Modeling, and Sentiment Analysis. Key constructive and techno-economic indicators, effectiveness: improvement in the quality of semantic analysis regarding reliable and unreliable textual information in Ukrainian, Polish, and English languages, creation of a new annotated dataset for evaluating models classifying true and false information. Areas of application: security and cybersecurity,

government communication with society through social networks, development of digital government applications, creation of programs and projects for media literacy, education, media, and journalism.

Індекс УДК: 002-052, 004.89:004.93, 316:002

Коди тематичних рубрик: 20.51.15, 28.23.24, 04.01.29

Керівники роботи

Власне Прізвище Ім'я По-батькові: Хайрова Ніна Феліксівна

Науковий ступінь: д. т. н.

Вчене звання: професор

Ідентифікатор ORCID ID:

Додаткова інформація:

VIII. Наукова (науково-технічна) продукція (НТП)

Назва НТП українською: Науково-технічний набір даних SOMTUME – довіра до соціальних медіа в контексті української міграції до Європи

Назва НТП англійською: Scientific and technical dataset SOMTUME – SOcial Media Trust: Ukrainian Migration into Europe

НТП, яку передбачалося створити:

Причини, через які НТП не було створено:

Отримані результати: Програмні продукти

Галузь застосування: Безпека та кібербезпека, комунікація держави з суспільством через соціальні мережі, розвиток програм та проєктів з медіа-грамотності, факт-чекінг та боротьба з дезінформацією

Реєстраційний номер картки технології:

Опис НТП: Набір даних SOMTUME (somtume.csv файл) залучає достовірну (TIC) та невизначену (UIC) текстову інформацію, що зібрана з соціальних мереж та новинних сайтів. Тематика текстів відноситься до міграції українців до Європейського Союзу за період з 22 лютого 2022 року до теперішнього часу. До TIC (Truth Information Content) достовірної інформації відносяться тексти та розпізнані аудіо файли з десяти урядових цифрових платформ України, Польщі, Канади та США, загалом 510 публікацій загальним розміром 630,188 токенів. До UIC відносяться тексти отримані з соціальної мережі Telegram та новинних веб-сайтів. Ця частина залучає 1,578,020 коротких текстових повідомлень, що є вилучені з 103 чатів 40 країн. З новинних сайтів до SOMTUME залучено 46,417 UIC текстової статті, загальним розміром 38729232 токенів. Набір даних є анотованим, тексти українською та англійською мовами. Набір даних має вільний доступ для наукових досліджень з ліцензією Massachusetts Institute of Technology (MIT)

Соціально-економічна спрямованість НТП: Забезпечення промисловості чи населення новим видом інформаційно-комунікаційних послуг

Вплив НТП на довкілля:

Впровадження НТП: Не впроваджено

Практична реалізація НТП

Початок етапу: 08.2023

Закінчення етапу: 12.2023

Споживачі продукції: Державні урядові організації, освітні організації, організації соціального та суспільного захисту

Перспективні ринки: Ринки України та міжнародні

Характер співробітництва з інвестором

Потрібний обсяг інвестицій, тис. грн.:

Права, що надаються інвестору після завершення роботи:

Наявність бізнес-плану:

Техніко-економічне обґрунтування:

Потенціальний обсяг продажу, тис. грн.:

Очікуваний термін окупності (років):

Додаткова інформація:

IX. Бібліографічний опис

1. N.Khairova, A. Shapovalova, O. Mamyrbayev, N. Sharonova, K. Mukhsina. Using BERT model to Identify Sentences Paraphrase in the News Corpus. CEUR Workshop Proceedings, 2022. – v.3171, p. 38-48
2. G.S Ybytayeva, N. F Khairova, O.Zh Mamyrbayev, K. Zh Mukhsina, B. Zh Zhumazhanov. Experimental Verification of Collocation Detection Methods. CS&SE@ SW 2022, 13p
3. AS Kolesnyk, NF Khairova. Justification for the Use of Cohen's Kappa Statistic in Experimental Studies of NLP and Text Mining. Cybernetics and Systems Analysis, 2022, v.58, iss.2, p.280-288.
4. G. Ybytayeva, O. Mamyrbayev, N. Khairova, N. Rizun, S. Berdali, K. Mukhsina. Creating a Thesaurus" Crime-Related Web Content" Based on a Multilingual Corpus. CEUR Workshop Proceedings, 2023, p. 77-87.
5. N. Sharonova, I. Kyrychenko, I. Gruzdo, G. Tereshchenko. Generalized Semantic Analysis Algorithm of Natural Language Texts for Various Functional Style Types. CEUR Workshop Proceedings, 2022, v. 3171, p. 16-26.
6. Oleksandr Karataiev, Dmytro Sitnikov, Nataliia Sharonova. A Method for Investigating Links between Discrete Data Features in Knowledge Bases in the Form of Predicate Equations. CEUR Workshop Proceedings, 2023.

X. Заклучні відомості

Керівник юридичної особи

Марченко Андрій Петрович

д. т. н., 05.05.03

Перелік осіб-виконавців

Бабкова Надія Вікторівна

(к.т.н.)

Лукашевський Арсеній Максимович

Редозуб Іван Андрійович

Шаронова Наталія Валеріївна

(д.т.н., 122)

Відповідальний за підготовку

Гуренко Ю.І.

облікових документів

Телефон

+38 (050) 609-83-92

Реєстратор

**Керівник відділу УкрІНТЕІ, що є
відповідальним за реєстрацію наукової
діяльності**

