УДК 614.8.013 DOI: 10.30929/1995-0519.2020.1.93-98

ЗАСТОСУВАННЯ ІНФОРМАЦІЇ З НОВИН ЗМІ ПРО НЕБЕЗПЕКИ ДЛЯ ПРОГНОЗУВАННЯ НАДЗВИЧАЙНИХ СИТУАЦІЙ В УКРАЇНІ

О. В. Долженкова, В. О. Тищенко
Дніпровський національний університет імені Олеся Гончара
просп. Гагаріна 72, м. Дніпро, 49000, Україна. E-mail: tischenko.vlada@gmail.com

Стрімке зростання кількості, точності та різноманіття алгоритмів машинного навчання спонукає дослідників знаходити застосування цим алгоритмам в задачах, які безпосередньо впливають на життя людей. Однією з таких задач є прогнозування надзвичайних ситуацій. Методи, що використовуються ДСНС, здебільшого ґрунтуються на статистичних оцінках та використовують обмежену кількість ознак, на основі яких будується прогнозуюча модель. Більш того, покращення точності прогнозу завжди матиме високий пріоритет, оскільки точність прямопропорційно впливає на успіх запобігання надзвичайних ситуацій. Дана стаття пропонує подивитись на проблему прогнозування надзвичайних ситуацій з іншого боку – з боку розширення множини ознак, що можуть використовуватись алгоритмом. Так, було виявлено, що потенційним джерелом ознак може бути архів новин про надзвичайні ситуації в Україні: журналісти, працюючи на матеріалом новин, часто відвідують місця аварій, консультуються з технічними спеціалістами, метеорологами та експертами, що встановлюють причину та передумови НС. Уся ця інформація міститься в текстовому вигляді на інтернет-джерелах ЗМІ. Алгоритми машинного навчання, зокрема обробки природньої мови (Natural Language Processing або NLP) можуть вилучати дану інформації та в належному форматі відправляти моделі, яка будуватиме прогноз на основі цієї додаткової інформації. Тож метою дослідження була перевірка репрезентативності вибірки новин про небезпеки, на прикладі одного з українських інтернет-джерел новин, відносно вибірки надзвичайних ситуацій, порівняння їх статистичних характеристик, аналіз аномалій за їх наявності. Задля досягнення даної мети було розглянуто класифікацію новин з надзвичайних ситуацій, автоматизовано побудову структурованої бази даних новин за період 8 років з використанням алгоритмів обробки природньої мови, проаналізовано тенденції зміни кількості новин у часі за чинниками надзвичайних ситуацій. Порівняно характеристики створеної бази даних зі статистикою ДСНС через залежності частоти появи надзвичайних ситуацій за категоріями та сезонами року. Відповіді на поставлені питання мають важливу практичну значимість, оскільки можуть бути впроваджені в реальних системах прогнозування надзвичайних ситуацій та виявити нові закономірності в даних НС.

Ключові слова: новини, надзвичайні ситуації, прогнозування, небезпека, машинне навчання, алгоритм, обробка тексту

ИСПОЛЬЗОВАНИЕ ИНФОРМАЦИИ ИЗ НОВОСТЕЙ СМИ О КАТАСТРОФАХ ДЛЯ ПРОГНОЗИРОВАНИЯ ЧРЕЗВЫЧАЙНЫХ СИТУАЦИЙ В УКРАИНЕ

Е. В. Долженкова, В. А. Тищенко
Днепровский национальный университет имени Олеся Гончара
просп. Гагарина - 72, г. Днепр, 49000, Украина. E-mail: tischenko.vlada@gmail.com

Стремительный рост количества, точности и разнообразия алгоритмов машинного обучения побуждает исследователей находить применение этим алгоритмам в задачах, которые непосредственно влияют на жизни людей. Одной из таких задач является прогнозирование чрезвычайных ситуаций. Методы, которые используются ДСНС, в основном базируются на статистических оценках и используют ограниченное количество признаков, на основе которых строится прогнозирующая модель. Более того, улучшение точности прогноза всегда будет иметь высокий приоритет, поскольку точность непосредственно влияет на успех предотвращения чрезвычайных ситуаций. Данная статья предлагает посмотреть на проблему прогнозирования с другой стороны – со стороны расширения множества признаков, которые могут использоваться алгоритмом. Так, было определено, что потенциальным источником признаков может быть архив новостей про чрезвычайные ситуации в Украине: журналисты, работая над материалом новостей, часто консультируются с техническими специалистами, метеорологами, экспертами, которые устанавливают причину и предпосылки возникновения ЧС. Вся эта информация хранится в текстовом виде на интернет-порталах СМИ. Алгоритмы машинного обучения, в частности обработки естественного языка (Natural Language Processing или NLP) могут извлечь данную информацию и в надлежащем формате передавать модели, которая строит прогноз. Таким образом, целью данного исследования являлась проверка репрезентативности выборки новостей о ЧС, на примере одного из украинских интернет-порталов новостей, относительно выборки чрезвычайных ситуаций, сравнение их статистических характеристик, анализ аномалий при их наличии. Для достижения данной цели была построена структурированная база новостей за период 8-ми лет с использованием алгоритмов обработки естественной речи, проанализировано тенденции изменения количества новостей во времени по факторам ЧС. Проведено сравнение построенной базы данных со статистикой ДСНС через зависимость частоты появления ЧС по категориям и сезонам года. Ответы на поставленные вопросы имеют важное практическое значение поскольку могут быть интегрированы в реальные системы прогнозирования чрезвычайных ситуаций и выявить новые закономерности в данных ЧС.

Ключевые слова: новости, чрезвычайные ситуации, прогнозирование, катастрофа, машинное обучение, алгоритм, обработка текста

АКТУАЛЬНІСТЬ РОБОТИ. Відповідно до Кодексу цивільного захисту України, одним з основних завдань державної системи цивільного захисту є забезпечення реалізації заходів щодо запобігання виникненню надзвичайних ситуацій (далі - НС) [1]. Першочерговим етапом запобігання виникненню НС є пошук закономірностей у їх появі, тривалості та силі розповсюдження та подальше прогнозування надзвичайних ситуацій. З метою виконання даної задачі ДСНС України починаючи з 2001 року щорічно надає Національну доповідь про стан техногенної та природної безпеки в Україні [2]. Одночасно з існуванням потреби у прогнозуванні надзвичайних ситуацій, з'являється все більше сучасних методів та алгоритмів, у тому числі машинного навчання, які здатні обробляти не лише кількісні дані, а й зображення та природню мову у текстовому та звуковому форматі і будувати прогноз на їх основі [3]. Постає питання чи можна застосувати сучасні алгоритми обробки інформації для покращення прогнозу надзвичайних ситуацій та пошуку додаткової інформації з альтернативних джерел саме для України.

Аналіз проблеми та існуючих методів. Дослідники вважають перспективним використання машинного навчання для запобігання, прогнозу та покращення менеджменту під час появи надзвичайних ситуацій. Так, були побудовані моделі ідентифікації повідомлень щодо НС у соціальних мережах, зокрема Twitter, для отримання додаткової інформації про знаходження і стан постраждалих та майна [4, 5, 6].

Тож з одного боку, існують дослідження, що доводять доцільність створення моделей для пошуку додаткової інформації про наслідки НС. З іншого боку, на відміну від західних країн, в Україні не розповсюджено поширення інформації про НС у соціальних мережах. Але існують альтернативні джерела інформації, які включають в себе дані про НС. Одними з таких джерел є інтернет-ресурси новин. Більш того, дані ресурси слугують об'єднаним джерелом предметів діяльності декількох служб України: ДСНС, Національної поліції, ДАІ та інших, робота яких дуже часто пов'язана. До аналогічного висновку дійшли у дослідженні [7], в якому була побудована система, що знаходить новини іспанською мовою про надзвичайні ситуації природного чиннику та виділяє корисну інформацію з них, а саме місце появи надзвичайної ситуації, дату, кількість постраждалих. Аналогічні дослідження для українських новин не було знайдено, тож дана тема є одночасно перспективною, оскільки досліджується іноземними вченими та в той же час і відносно новою для українських даних.

Метою даної статті є створення структурованої бази новин ЗМІ з надзвичайних ситуацій та перевірка її повноти і доцільності шляхом порівняння з трендом та сезонною складовою статистичних даних, наданих ДСНС. Для досягнення поставленої мети необхідно вирішити наступні задачі: проаналізувати потенційні джерела новин, зберегти новини шляхом веб-обробки джерел, структурувати набір новин, знайти і вилучити аномальні значення за наявності, проаналізувати тренд та сезонну складову появи НС у новинах.

МАТЕРІАЛ І РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ**.** Після проведення порівняльного аналізу джерел новин було обрано інтернет-ресурс «24tv.ua», оскільки структура сторінок є простою і водночас зручною для автоматичної обробки [8]. Також дані є частково структурованими, оскільки присутня система тегів або фільтрів. Так, було обрано фільтр «надзвичайні новини», що показує новини про НС у всьому світі. Очевидно, що даний набір новин хоч і містить міжнародні новини, все ж таки орієнтується на надзвичайні новини України і не лише національного, а й регіонального масштабу. Це припущення було доведено експериментальним шляхом під час перевірки набору новин.

Таким чином, збір новин був проведений у 2 етапи: отримання посилань на новини, оскільки кожна веб-сторінка містить 36 новин, та безпосереднє отримання новин за найденими посиланнями. До веб-ресурсу було виконано 126 запитів на етапі 1 та 4626 запитів на етапі 2. Отримана база даних містила 4626 новин починаючи з 2011 року до 2018-го року, включаючи в себе світові новини з надзвичайних ситуацій. У табл. 1 наведена структура завантажених даних та типовий приклад новини.

Таблиця 1 – Структура бази даних новин та типовий приклад

Унікальний номер новини	4149
Перелік тегів	Новини України; Надзвичайні ситуації; Пожежа; ДСНС
Текст статті	Полум'я охопило близько 200 квадратних метрів будівлі. При евакуації постраждали від чадного газу один із співробітників міліції та …
Анотація	На Одещині внаслідок пожежі в Овідіопольському райвідділі внутрішніх справ постраждали міліціонер і шестеро затриманих
Дата і час появи новини	2011-04-26 20:25

Наведена вище структура дозволила вилучити саме надзвичайні новини України, яких виявилось 1561. Також для подальшого аналізу необхідно було класифікувати новини за чинником надзвичайної ситуації за допомогою тегів інтернет-джерела, побудована схема ієрархії тегів наведена на рис. 1.

Завдяки наявності дати і часу появи новин можна розглядати новини як часовий ряд і порівнювати кількість і частоту подій хронологічно впродовж 8 років. Тож, наведемо загальну кількість новин з надзвичайних ситуацій щорічно, додаючи розгалуження за чинником НС. Також, оскільки антропогенний чинник включає в себе багато підтипів, окремо розглянемо 3 найбільш впливові антропогенні чинники та їх зміну у часі.

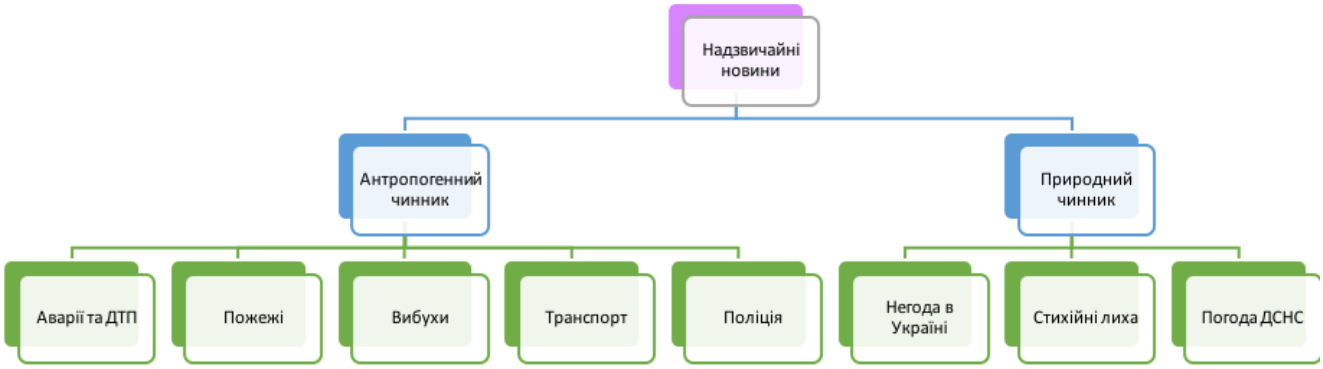

Рисунок 1 – Схема ієрархії тегів за чинником НС

На рис. 2 наведено історичну кількість новин за чинником НС та окремо розподіл кількості НС антропогенного характеру.

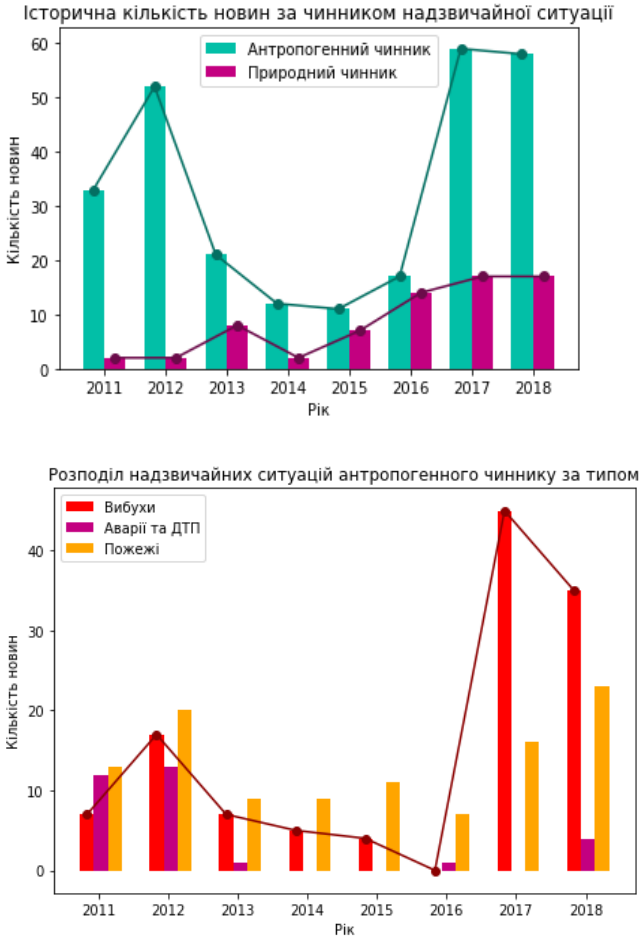

Рисунок 2 – Графіки історичної кількості новин за чинниками НС

Цікавим фактом, що виходить з графіків, є стрімке зростання кількості новин антропогенного характеру у 2017-2018 роках. Причина такого росту зображена на правій частині рисунку – стрімко зросла кількість вибухів. Для розуміння чи є така поведінка аномальним значенням або закономірністю, був розроблений наступний алгоритм:

1. Зроблено запит до створеної бази даних для отримання усіх новин з тегом «вибухи» за 2017 рік.
2. Шляхом автоматизованого виділення іменованих сутностей з тексту отримано «Місце» виникнення надзвичайної ситуації.
3. Порівняно кількість новин кроку 1 з кількістю унікальних значень кроку 2 та на основі їх різниці зроблено висновок про аномальність.
4. Виконано аналогічні дії для даних 2018-го року.

Пункт 2 вищезазначеного алгоритму виконувався за допомогою NER (Named Entity Recognition) моделі для української мови, що є розширенням бібліотеки Mitie [9] і є єдиною відкритою лінгвістичною моделлю для української мови. Дана модель на вхід приймає текст новини, а на виході видає, наприклад, наступний результат: {'Loc': 'Одесі', 'score':0.74}. Оскільки алгоритм видає назви населених пунктів у різних відмінках, їх було додатково пропущено через лематизатор, який представив їх у називному відмінку.

Після групування населених пунктів, виявилось що усі новини про вибухи 2017-го та 2018-го року – сталися у трьох населених пунктах: Калинівці (Вінницька обл.), Балаклії (Харківська обл.) та Ічні (Рівненська обл.). А кількість новин, що описують дане явище становить 120 статей, хоча надзвичайних ситуацій при цьому лише 3. Таким чином для правильності подальшого аналізу така кількість статей була вилучена і залишено по одній новині на кожну надзвичайну ситуацію з цієї групи.

Наступним етапом аналізу, було порівняння трендової складової динаміки появи новин про надзвичайні ситуації, зібраних у базі даних, та динаміки виникнення НС на території України за статистичними даними ДСНС у 2017-му році [10]. Було розглянуто статистичні дані з 2011 по 2017 рік включно. Для оцінки залежності двох часових рядів, використовували коефіцієнти кореляції Спірмана та Кендала, оскільки вони не накладають жодних обмежень на розподіл даних та є робастними відносно аномальних значень.

Таким чином, коефіцієнти Спірмана та Кендала становили 0,35 та 0,33 відповідно, що свідчить про існування прямої залежності між кількістю новин про НС та виникненням НС. Але ця залежність є досить слабкою, тобто можна зробити висновок, що

певні НС не висвітлюються в новинах і в той же час деякі масштабні, наприклад, у випадку вибухів на військових складах, містяться у більш ніж 30 новинах на одну НС. Також хочеться підкреслити різницю в розмахах значень між роками, оскільки рангові коефіцієнти кореляції не залежать від нього. Так, у випадку виникнення НС це середній розмах складає 10,2% відсотки, від середнього щорічного значення, а у випадку новин про НС – 54,2% відсотки.

Останнім етапом перевірки доцільності новин про НС як джерела додаткової інформації була перевірка наявності сезонної складової. Так, із звітів ДСНС відомо, що НС природного характеру, а саме гідрометеорологічні, відбуваються з певною сезонністю, залежно від місяця і пори року. Було перевірено наявність даної характеристики у новин про НС.

На рис. 3 зображено кількість таких новин кожного місяця впродовж 8 років. З графіку видно, що середнє значення за місяцями не є постійним, наприклад у весняні місяці (березень, квітень, травень) воно набагато нижче ніж в осінні (вересень, листопад). З іншого боку, очікувалось отримати певні аномальні значення взимку, але цей ефект не присутній на графіку.

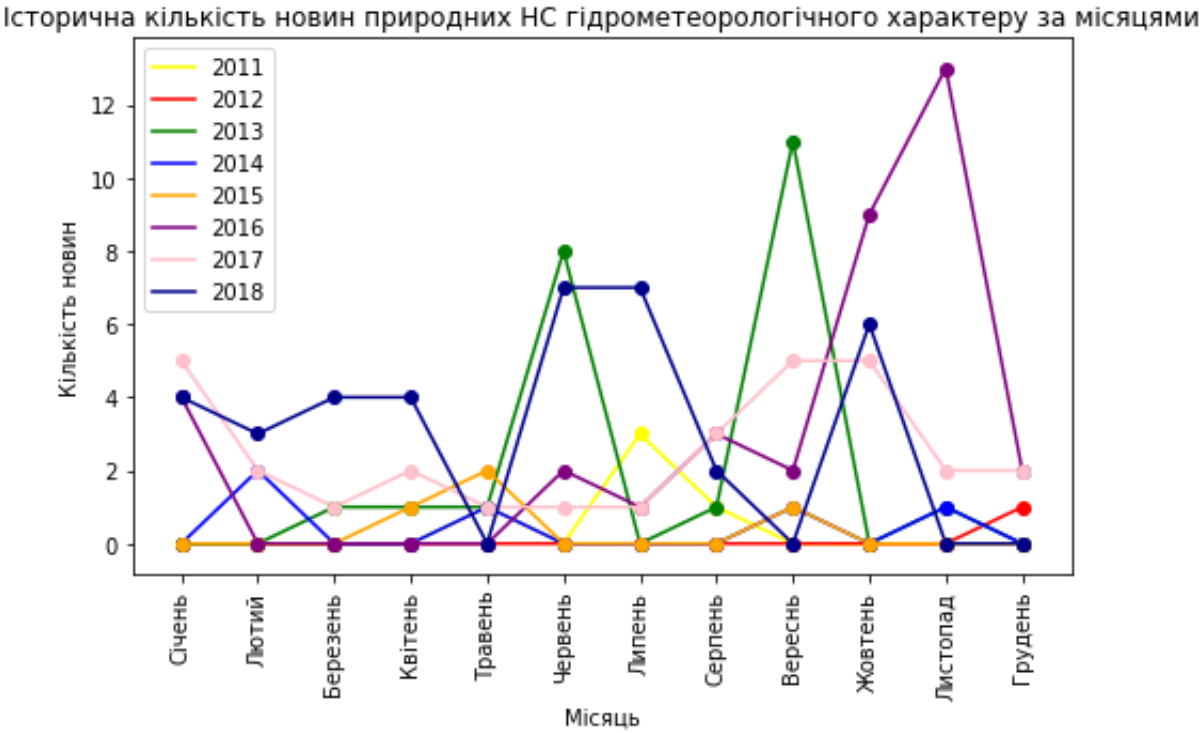

Рисунок 3 – Графік кількості новин про гідрометеорологічні НС за місяцями року

ВИСНОВКИ. У ході дослідження створено базу даних новин з НС, яка складає більше ніж 1,5 тисячі записів. Ці дані структуровані, з використанням розробленої класифікації за тегами. Проаналізовано кількість новин з НС впродовж 8-ми років, враховуючи чинник НС. Знайдені аномальні значення окремо проаналізовано за допомогою алгоритму виділення іменованих сутностей з тексту українською мовою та вилучено із статистичного аналізу.

Проведено порівняння тенденцій виникнення НС за даними ДСНС та появи новин з НС за допомогою двох коефіцієнтів кореляції. Базу даних новин перевірено на наявність сезонної складової. Додаток до дослідження у вигляді програмного коду та csv файлу з новинами знаходиться за посиланням [11].

Таким чином, новини ЗМІ про НС є джерелом інформації, з якого можна виділити такі характеристики як дата, місце, чинник та уточнений чинник НС. З іншого боку, новини про НС мають достатньо низький коефіцієнт збігу тенденцій і сезону, які присутні у часових рядах виникнення самих НС. Дане явище пояснюється тим, що деякі НС висвітлюються в новинах з великою кількістю уваги та присвячених статей, а інші можуть бути не загадані зовсім. Тож, новини про НС не можуть бути джерелом додаткової інформації для усіх випадків, але вони можуть містити цінну інформацію про певні масштабні НС, яким ЗМІ приділяють особливу увагу. Для виділення додаткової інформації з таких новин, можна використовувати існуючу лінгвістичну модель української мови [9] та додавати до неї більше видів іменованих сутностей шляхом тренування моделі на новинах з бази даних. Наприклад, з коментарів очевидців можна виділити стан погоди до НС у випадку дії природних чинників або коментарів експертів промисловості під час техногенних НС. Іменовані сутності, виділені з таких новин, можуть служити додатковими чинниками (ознаками) подій при побудові регресійних моделей прогнозування, точність яких пропорційно збільшується при збільшенні множини ознак об'єктів.

ЛІТЕРАТУРА

1. Кодекс цивільного захисту України від 2 жовтня 2012 року № 5403–VI станом на 0.4.11.2018 / Законодавство України: відповідає офіц. тексту.

2. Аналітичний огляд стану техногенної та природної безпеки в Україні за 2016 рік / Державна служба України з надзвичайних ситуацій. URL: http://www.dsns.gov.ua/ua/Analitichniy-oglyad-stanu-tehnogennoyi-ta-prirodnoyi-bezpeki-v--Ukrayini-za-2015-rik.html. (дата звернення: 13.04.2019)

3. Mitchell M. T. Machine Learning. Columbus, OH: McGraw-Hill Science/Engineering/Math, 1997. 392 p.

4. H. To, S. Agrawal, S. H. Kim and C. Shahabi, "On Identifying Disaster-Related Tweets: Matching-Based or Learning-Based?," 2017 IEEE Third International Conference on Multimedia Big Data (BigMM), Laguna Hills, CA, 2017, pp. 330-337.

5. AIDR: Artificial Intelligence for Disaster Response / Imran et al. Proceeding of the 23rd Conference on World Wide Web, Seoul, Korea, 2014, pp. 159-162

6. Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency / Verma et al. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 2011.

7. Alberto Tйllez-Valero, Manuel Montes-y-Gymez, Luis Villasecor Pineda Using Machine Learning for Extracting Information from Natural Disaster News Reports. Computaciyn y Sistemas, 2009.

8. Надзвичайні ситуації/ Òåëåêàíàë íîâèí «24» URL: https://24tv.ua/nadzvichayni_situatsiyi _tag350/

9. NER models for MITIE / Спільнота lang-uk. – Режим доступу: http://lang.org.ua/en/models/ (дата звернення: 06.04.2019)

10. Звіт про основні результати діяльності Державної служби України з надзвичайних ситуацій у 2017 році/ Державна служба України з надзвичайних ситуацій. – URL: http://www.dsns.gov.ua/files/2018/1/26/Zvit%202017(%D0%9A%D0%9C%D0%A3).pdf (дата звернення: 03.04.2019)

11. Analysing disaster news in Ukraine / Github – URL: https://github.com/lovelyscientist/ disaster_news _ua_ner (дата звернення: 25.04.2019)

USAGE OF DISASTER NEWS INFORMATION FOR PREDICTING THE EMERGENCY SITUATIONS IN UKRAINE

O. Dolzhenkova, V. Tyshchenko

Oles Honchar Dnipro National University

prosp. Haharyn, 72, Dnipro, 49000, Ukraine. E-mail: tischenko.vlada@gmail.com

Now we can observe rapid growth of accuracy, diversity and number of machine learning algorithms while many of those can be applied to the human-related problems. One of those problems is disaster prediction. Methods used by State Emergency Service of Ukraine are mostly based on statistical analysis and use limited number of features for building their prediction model. Moreover, improvement of the accuracy of disaster prediction will always have a high priority as such accuracy is in direct ratio with the success of disaster prevention operations. This article proposes to take another view on the problem of disaster prediction from the side of extension of the features set, that can be used by prediction models. Thus, it was outlined that potential source of those features is disaster news archive in Ukraine: journalists, while working on news material, often visit places where accidents happened, consult with technical specialists, meteorologists and experts, which point to the reason and prerequisites of disaster. All this information is stated in the text of disaster related news. Machine learning algorithms, in particular Natural Language Processing (NLP) algorithms are capable to extract such information and send it to the prediction model in a proper format. **Purpose** of this research was to test the hypothesis of disaster news being a representative dataset for building additional features for disaster prediction models, to compare the statistical characteristics of disaster news dataset and disasters dataset provided by State Emergency Service of Ukraine, to analyze outliers if such detected. **Methodology** of the research contains 2 main parts: news dataset gathering and comparison of built dataset to original disaster events dataset. While gathering the news dataset articles were classified by reason of disaster which allowed splitting dataset on functional categories. News dataset was gathered by automatic script, which scrawled one of the online news portals in Ukraine and filled dataset with news for last 8 years. NLP algorithms allowed extracting the place and date of the news. Having disaster news and disaster events yearly allowed to compare changes in their number by categories of disaster. Characteristics of gathered dataset were compared to original disaster events dataset by frequency of events by categories of disasters and seasonality. This research brings a **practical value** as discovered feature set from news can be added to real-world disaster prediction models, improve the accuracy of those and outline new patterns in analyzed data. This type of features was never before extracted from Ukrainian language texts and no ready and fitting Ukrainian news dataset was found which proves the **originality** of the idea and results of the research. The results **conclude** that although not all disasters are stated in the news and at the same time size and details of the event can also vary a lot, still, most of the disasters are covered by news and detailed description, prerequisites and facts, which can be extracted by NLP and used as potential features.

Key words: disaster, prediction, NLP, machine learning, news, extraction.

REFERENCES

1. The Code of Civil Protection of Ukraine (2012), No. 5403-VI.

2. Analytical review of the state of man-made and natural safety in Ukraine for 2016. *The State Emergency Service of Ukraine*, [Online], available at: http://www.dsns.gov.ua/ua/Analitichniy-oglyad-stanu-tehnogennoyi-ta-prirodnoyi-bezpeki-v--Ukrayini-za-2015-rik.html. (Accessed: 13.04.2019)

3. Mitchell, M. T. (1997), "Machine Learning", Columbus, OH: McGraw-Hill Science /Engineering /Math, 392 p.

4. To, H., Agrawal, S., Kim, S. H., Shahabi, C. (2017), "On Identifying Disaster-Related Tweets: Matching-Based or Learning-Based?", *IEEE Third International Conference on Multimedia Big Data* (BigMM), Laguna Hills, CA, 2017, pp. 330-337.

5. "AIDR: Artificial Intelligence for Disaster Response" (2014), *Proceeding of the 23rd Conference on World Wide Web*, Seoul, Korea, pp. 159-162.

6. "Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency", (2011), *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain .

7. Tyllez-Valero, A., Montes-y-Gymez, M., Pineda, L. V., (2009), Using Machine Learning for Extracting Information from Natural Disaster News Reports. Computaciyn y Sistemas.

8. Emergencies/ News Channel «24», [Online], available at: https://24tv.ua/nadzvichayni_situatsiyi_tag350/ (Accessed: 06.04.2019)

9. NER models for MITIE / lang-uk Group, [Online], available at: http://lang.org.ua/en/models/ (Accessed: 02.04.2019)

10. Report on the main results of the State Service of Ukraine for Emergencies in 2017/ The State Emergency Service of Ukraine, [Online], available at: http://www.dsns.gov.ua/files/2018/1/26/Zvit%202017(KMУ).pdf (Accessed: 03.04.2019)

11. Analysing disaster news in Ukraine / Github, [Online], available at: https://github.com/lovelyscientist/disaster_news_ua_ner (Accessed: 25.04.2019)

Стаття надійшла 09.01.2020.