

Оцінка міграційного потоку на основі аналізу текстових повідомлень (Migration Flow Assessment based on Messenger Texts Analysis)

► виконав
студент КН-4
Кондратенко Дмитро
Науковий керівник: доцент Швай Н. О.

Мета дослідження

Провести огляд літератури та аналіз існуючих методів дослідження міграційних потоків.

Запропонувати реалізацію алгоритму збору і фільтрації даних текстових повідомлень

Провести емпіричне дослідження міграційних потоків з використанням розробленого алгоритму.

Оцінити ефективність запропонованої методології та порівняти результати з традиційними методами аналізу міграційних потоків.

План

Реалізація алгоритму збору і фільтрації даних з текстових повідомлень

Обґрунтування вибору відповідного методу машинного навчання

Експериментальне визначення оптимальних значень параметрів тренування моделі

Проведення оцінки міграційного потоку за допомогою розробленого алгоритма

Порівняння отриманих даних з уже інсуючими

Реалізація алгоритму

Алгоритм складається з 3 етапів:



1) Виконання завдань NLP з текстом повідомлень (вилучення стоп-слів, лематизація, розпізнавання мовних сутностей NER)



2) Виявлення місцезнаходження для кожної з розпізнаної сутності



3) Фільтрація відповідно до встановлених умов

- У результаті роботи алгоритму отримано датасет, в якому зібрано десятки тисяч повідомлень про перетин кордону з Україною
- Для формування тестового набору даних необхідно було самостіно анотувати частину з них таким чином, що :
 - 1 – повідомлення стосується в'їзду в Україну
 - 0 – повідомлення стосується виїзду з України
 - 2 – повідомлення не стосується міграції

text	label
hello ! to transfer parcel from Kharkov ? city in Germany travel from ?	2
hello . transfer from Leipzig to Kharkov ? ?	1
good evening , need to send parcel from Thuringia , Eisfeld to Ukraine , Kherson . possible ?	2
good morning . tell carrier from Flensburg / Hamburg to Kharkov ?	1
excuse , . - topic . maybe need . guy friend go tomorrow 21.04 from Lviv to Stuttgart car . need to .	0
hello , travel from Mannheim to Ukraine today ?	1
good evening . see message . carrier ? need to send small parcel to 3 kg from Odessa to Germany .	2
good afternoon ! need carrier from Germany to Kyiv .	1
driver from Munich to Lviv late tonight , people parcel for 1.5 v,- kg , in Ukraine tomorrow , send parcel mail in Ukraine !	1
June 27 , travel from Berlin to Chernivtsi route show . travel car . 2 passenger . write in personal	1
hello , advise , need to transfer parcel from Germany to Simferopol .	2
guy from Ukraine to Saarbrücken , need to pass parcel , thank	2
good evening , tell get from Odessa to Munich , option ?	0

Обґрунтування вибору відповідного методу машинного навчання

- ▶ Для тренування моделі було обрано метод машинного навчання few-shot learning
- ▶ Не потребує великої кількості анотованих даних для навчання
- ▶ Показує високий перформанс на невеликому обсязі натренованих даних

Експериментальне визначення оптимальних значень параметрів тренування моделі

All models

Model Name	Performance Sentence Embeddings (14 Datasets)	Performance Semantic Search (6 Datasets)	Avg. Performance	Speed	Model Size
all-mpnet-base-v2	69.57	57.02	63.30	2800	420 MB
multi-qa-mpnet-base-dot-v1	66.76	57.60	62.18	2800	420 MB
all-distilroberta-v1	68.73	50.94	59.84	4000	290 MB
all-MiniLM-L12-v2	68.70	50.82	59.76	7500	120 MB
multi-qa-distilbert-cos-v1	65.98	52.83	59.41	4000	250 MB
all-MiniLM-L6-v2	68.06	49.54	58.80	14200	80 MB
multi-qa-MiniLM-L6-cos-v1	64.33	51.83	58.08	14200	80 MB
paraphrase-multilingual-mpnet-base-v2	65.83	41.68	53.75	2500	970 MB
paraphrase-albert-small-v2	64.46	40.04	52.25	5000	43 MB
paraphrase-multilingual-MiniLM-L12-v2	64.25	39.19	51.72	7500	420 MB
paraphrase-MiniLM-L3-v2	62.29	39.19	50.74	19000	61 MB
distiluse-base-multilingual-cased-v1	61.30	29.87	45.59	4000	480 MB
distiluse-base-multilingual-cased-v2	60.18	27.35	43.77	4000	480 MB

All-MiniLM-L12-v2 була обрана базовою моделлю для навчання нашої

<u>batch_size</u>	Час, витрачений на тренування моделі, хв	Ефективність моделі (точність)
16	31	0.713534
32	52	0.735963
64	84	0.735963

Таблиця 2.4.1. Залежність ефективності моделі та часу її тренування від параметра *Batch_size*

Кількість ітерацій	Час, витрачений на тренування моделі, хв	Ефективність моделі (точність)
20	31	0.713534
40	42	0.8
60	51	0.8

Таблиця 2.4.2. Залежність ефективності моделі та часу її тренування від кількості ітерацій

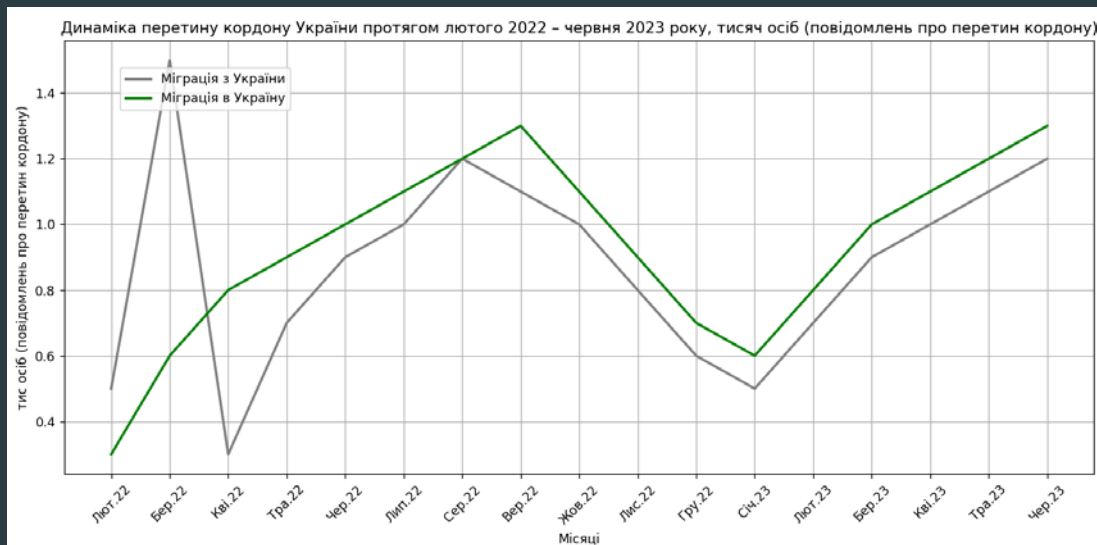
Кількість епох для навчання	Час, витрачений на тренування моделі, хв	Ефективність моделі (точність)
1	31	0.713534
2	62	0.764667
3	94	0.815574
4	105	0.857142
5	139	0.763553

Таблиця 2.4.3. Залежність ефективності моделі та часу її тренування від кількості епох для тренування

Модель, натренована на нелематизованих даних		Модель, натренована на лематизованих даних	
Час тренування, хв	Ефективність	Час тренування, хв	Ефективність
110	0.817293	105	0.857142

Таблиця 2.4.3. Порівняння двох підходів до тренування моделей

Проведення оцінки міграційного потоку за допомогою розробленого алгоритма

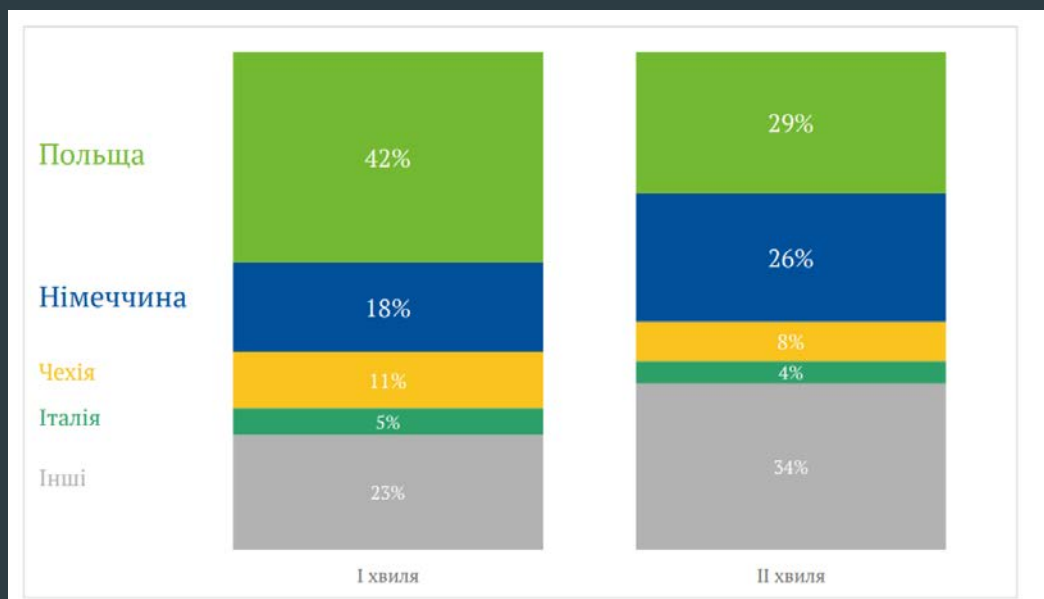
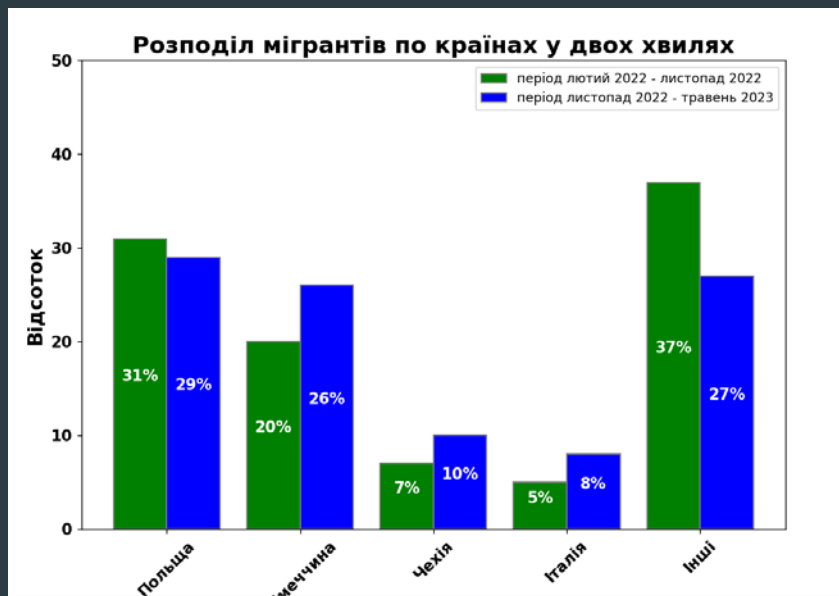


СКІЛЬКИ ЗАРАЗ УКРАЇНСЬКИХ БІЖЕНЦІВ?



- Станом на кінець червня 2023 за кордоном перебували **5,6 – 6,7 млн** українців (оцінки за червень, це на 300-500 тис. осіб більше ніж станом на кінець 2022 року)
- ~ **1,3 млн** із них в росії та Білорусі, 4,3 – 5,4 млн в інших країнах
- Німеччина** вийшла на перше місце по кількості українських біженців в ЄС (1,1 млн осіб), Польща – друга (976 тис.)





Висновки

- ▶ Розробка алгоритму збору і фільтрації даних
- ▶ Тренування мовної моделі за допомогою few-shot-learning.
- ▶ Класифікацію та аналіз міграційних потоків.
- ▶ Практичне значення та перспективи використання.

Дякую за увагу