ВИКОРИСТАННЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ В ТВІТТЕРІ ДЛЯ ВИЗНАЧЕННЯ СУСПІЛЬНОЇ ДУМКИ ПІД ЧАС МАСОВОГО ЗАХВОРЮВАННЯ

ИСПОЛЬЗОВАНИЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В ТВИТТЕРЕ ДЛЯ ОПРЕДЕЛЕНИЯ ОБЩЕСТВЕННОГО МНЕНИЯ ВО ВРЕМЯ МАССОВОГО ЗАБОЛЕВАНИЯ

USE OF INTELLECTUAL ANALYSIS OF DATA ON TWITTER TO DETERMINE PUBLIC OPINION DURING A MASS DISEASE.

Науковій керівник - доц. каф. Інформаційних систем, канд. техн. наук
Шпинковський О.А., Шпинковский А.А., Shpinkovski O.A.
Студентка групи АІ-182 - Толмаченко Я.В., Толмаченко Я.В., Tolmachenko Y.V.

Анотація: Пропонується інформаційна системи аналізу настроїв у соціальних мережах, зокрема у твіттері, що актуально під час масових захворювань

Ключові слова: інтелектуальний аналіз даних, аналіз настроїв, потокова передача, візуалізація

Аннотация: Предлагается информационная система анализа настроений в социальных сетях, в частности в твиттере, что очень актуально во время массовых заболеваний.

Ключевые слова: : интеллектуальный анализ данных, анализ настроений, потоковая передача, визуализация

Abstract: An information system for analyzing moods in social networks, in particular on Twitter, is proposed. Now during mass diseases this is relevant.

Keywords: data mining, sentiment analysis, streaming, visualization

Сьогодні методи інтелектуального аналізу даних використовують у багатьох сферах життя. Динамічно покращуються алгоритми машинного навчання, що допомагають у вирішенні складних і часто не розв'язуваних раніше проблем [1,2]. Спортивна і медична галузі не стали винятками. Інтелектуальні технології використовуються у високоточному визначенні діагнозу хвороби, забезпеченні діяльності лікарів, донорів крові, спортсменів тощо [3-5].

З поширенням Covid-19 майже в усіх країнах світу діють карантинні обмеження. У суспільства виникають питання, на які самостійно не завжди можна знайти відповіді. Тому люди шукають альтернативні джерела інформації. Найчастіше таким джерелом стає Інтернет, а точніше соціальні мережі. Багато спірних питань обговорюються в соціальній мережі Twitter, як, наприклад, питання з необхідності використання медичної маски під час пандемії.

Використовуються методи інтелектуального аналізу даних та аналізу тональності тексту, щоб зібрати всі твіти на потрібну тему та проаналізувати їх смисловий зміст та настрій. Щоб завантажити дані твітів, знадобиться пакет NLTK в Python. Для використання власного набору даних, якщо треба збирати твіти з певного періоду часу, користувача або хештега, використовують Twitter API .

Твіттер надає API для взаємодії з їх сервісом. Для цього існує безліч бібліотек на основі Python, які можна використовувати. Зокрема, бібліотека Tweepy є найцікавішою та простою у використанні. Для пошуку твітів використовується розширення StreamListener (). Відбір твітів проходить за хештегом «#mask».

Мова в її первісній формі не може бути точно оброблена машиною, тому потрібно обробити контент додатково, для перетворення у формат зрозумілий комп'ютером.

Перша частина осмислення даних – процес токенізації, або розбиття рядків на більш дрібні частини. Токен - це послідовність символів в тексті, яка служить єдиним

цілим. Залежно від того, як створюються токени, вони можуть складатися з слів, смайликів, хештегів, посилань або навіть окремих символів. Основний спосіб розбити мову на токени - розділити текст на основі прогалин і знаків пунктуації.

Дані соціальних медіа неструктуровані, тому серед усіх твітів підійдуть лише деякі. Твіти слід далі фільтрувати. Це важливий крок, оскільки якість даних призведе до більш надійних результатів. Створюється словник, в який вводяться такі слова-шум як, наприклад, «facemask» і «masks», смайлики, формат зображень, гіперпосилання або твіти, коротші трьох символів. Також видаляються так звані стоп-слова, які не несуть в собі смислове навантаження, як, наприклад, артиклі “the”, “a”. Всі твіти, що містять ці обмеження, будуть видалятись із загального списку (рис.1)[6,7].

```
[[679]]
[1] "CGuy295: #AskingForAFriend, does frequent #mask wearing #flatten the human #n
ose ? 59 minutes and 21 seconds. #plasticsurgeryproblem #notwinethough"

[[680]]
[1] "acelinnice: @drsanjaygupta @CNN But other than the fact there is a shortage (p
oor planning), is there any harm in wearing a… https://t.co/AHdBxpBcJ8"

[[681]]
[1] "Nellari13545895: RT @BOsForWow: Here is the #Thunderhawk from the #Kenner #Cl
SK toyline. This is one of the iconic vehicles from the line and included loade…"

[[682]]
[1] "Droit_IA: Probably the best mask for #Coronavirus \n\n#COVID2019 #COVID19fran
ce #COVID #Covid_19 #COVID19\n#mask #healthcare. https://t.co/BiPWxlgOsV"
```

Рис. 1 Частина відібраних із загального пошуку твітів

Візуалізуються дані за допомогою платформи MonkeyLearn - це є проста у використанні платформа машинного навчання для аналізу тексту. MonkeyLearn надає різні статистичні дані для вимірювання настроїв зібраних твітів [7]. Тож користуючись цією платформою для візуалізації даних, наприклад за допомогою хмари слів, можна отримати показник у якому контексті використовується хештег «#mask» (рис 2) [8].

Рис. 2 Хмара слів з термінами

Не дивно, що твіти людей з хештегом «#mask» були присвячені пандемії Covid-19. Такі слова, як «потрібно», «важливо», «планування», дають нам уявлення про їх відношення. Вони також говорять про незбалансованість попиту та пропозиції щодо поточного ринку.

Діаграма нижче показує результати аналізу настроїв Slack оглядів (рис.3). Кожне слово має за собою емоцію людей, що фіксується за текстом повідомлень. Графік впорядковує результати аналізу тональності тексту і робить його простим для розуміння.

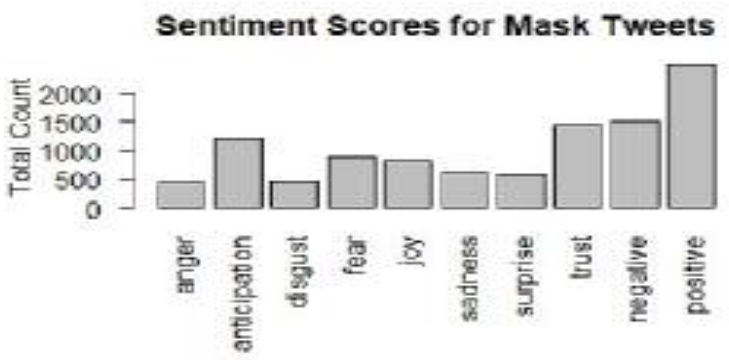

Рис.3. Аналіз тональності тексту для твітів з хештегом #mask

За проведеним аналізом можна свідчити, що твіти більше показують довіру суспільства до прийнятих карантинних мір та щодо використання медичних масок. Однак є і негативні тренди: близько 30% людей висловлюють свої побоювання, близько 50% людей демонструють негативні емоції. Але в цілому люди в Твіттері з оптимізмом сприймають використання масок [9].

ВИСНОВКИ

Спроектовано і опрацьовуються наступні основні модулі системи:

1. Збирання даних
2. Підготовки даних
3. Створення моделі аналізу настроїв
4. Візуалізація та збереження результатів

Отже, в умовах карантину, коли суспільство використовує інтернет для висловлення власної думки, аналіз тексту може використовуватися для визначення проблем, з якими стикаються люди на карантині. У даній роботі був використаний Python, а саме його бібліотека Tweepy для збору даних і аналіз тональності тексту за допомогою платформи MonkeyLearn для визначення громадської думки про використання медичних масок. Ці дослідження дозволять ефективніше аналізувати настрій користувачів у соціальних мережах.

Список використаних джерел

1. Прокопович, І. В. Використання інтелектуальних технологій у визначенні діагнозу хвороби / І. В. Прокопович, О. А. Шпинковський // Інформ. системи та технології в медицині (ISM-2018): І Міжнар. наук.-практ. конф., 28- 30 листоп. 2018 р. : зб. наук. пр. / ХНУРЕ. – Харків, 2018. – С. 127–129.

2. Шпинковська М.І. Застосування засобів машинного навчання у лікарській справі / Коваль Л. І. // І Міжн. наук.-прак. конф. «Інформаційні системи та технології в медицині» Зб. наук. праць. ХНУРЕ – Харків: «Друкарня Мадрид», 28-30 листопада 2018. – С.131–132.

3. О.А. Шпинковський, А.В. Цибулько. Інформаційна система ресурсного забезпечення діяльності донорів крові / Всеукраїнська науково-практична конференція молодих учених, спеціалістів, аспірантів «Проблеми енергоресурсозбереження в промисловому регіоні. Наука і практика»: Зб. тез доповідей. Маріуполь: ДВНЗ «ПДТУ», 2017. –159 с.

4. Шпинковська М.І. Засоби рекомендованого пошуку груп користувачів у соціальних мережах / Шпинковський О.А., Ус В.М. // Перспективні напрямки наукових

досліджень, XIV Міжнародна науково-практична інтернетконференція. – Вінниця, 24 листопада 2017 року. – ч.2, – С. 63-65

5. Shpinkovski A.A., Shpinkovska M.I., Korobova D.I. The automation system for accounting sporting activities. “Automation of technological and business-processes”, vol. 8, no. 4, pp. 49-54, december 2016.

6. CodeFlow [Електронний ресурс] Режим доступу до ресурсу: https://www.codeflow.site/ru/article/how-to-perform-sentiment-analysis-in-python-3-using-the-natural-language-toolkit-nltk

7. TowardsDataScience [Електронний ресурс] Режим доступу до ресурсу: https://towardsdatascience.com/creating-the-twitter-sentiment-analysis-program-in-python-with-naive-bayes-classification-672e5589a7ed

8. MonkeyLearn [Електронний ресурс] Режим доступу до ресурсу: https://monkeylearn.com/blog/sentiment-analysis-of-twitter/

9. TowardsDataScience [Електронний ресурс] Режим доступу до ресурсу: https://towardsdatascience.com/covid-19-outbreak-tweet-analysis