Projekt IO 2

Mateusz Kurowski

June 2024

Spis treści

1 Wstęp

Jako temat mojego projektu nr 2 wybrałem analizę tekstu z popularnego portalu społecznościowego x.com (dawniej twitter.com). Dokładniej opisując, zdecydowałem się na zebranie statystyk i analizę wpisów dotyczących wojny toczącej się obecnie (stan na 20.06.2024 r.) na terenie Bliskiego Wschodu, a dokładniej państw Izrael oraz Palestyna.

2 Dane wejściowe

Pierwszym napotkanym przeze mnie problemem był brak oficjalnego zbioru danych, z którego mógłbym skorzystać. Mojego zadania nie ułatwiło to, że obecny CEO firmy Twitter, Elon Musk, postanowił wprowadzić opłatę wynoszącą około 100 USD miesięcznie za możliwość korzystania z API serwisu.

Toteż zdecydowałem się na skorzystanie z narzędzia **twscrape**, które pozwala na pobieranie wpisów z owego portalu. W tym celu zdobyłem dodatkowe

2 numery telefonów i za ich pomocą założyłem 10 kont email oraz 10 kont na portalu **x.com**.

Poszukiwania tweetów rozpocząłem od konfiguracji programu, by szukał postów o następujących parametrach:

1. W języku angielskim
2. Zawierających tagi lub słowa ze zbioru: [israel, palestine, gaza, rafah, hamas, middleeast, netanyahu]

Następnie przez około 4 godziny pobierały się tweety ze względu na limity pobierania wynoszące bliżej nieokreśloną liczbę, która odnawia się co 15 minut.

W rezultacie otrzymałem następujący zbiór danych, który można logicznie podzielić na 2 części:

1. Dane sprzed wybuchu wojny - od 2020-05 do przedednia wybuchu konfliktu zbrojnego w tym:
   (a) 907 postów z tagiem #**israel**
   (b) 981 postów z tagiem #**palestine**
   (c) 429 postów z tagiem #**gaza**
   (d) 1108 postów z tagiem #**rafah**
   (e) 1115 postów z tagiem #**hamas**
   (f) 1046 postów z tagiem #**middleeast**
   (g) 1189 postów z tagiem #**netanyahu**
2. Dane po wybuchu wojny - po 2023-10 w tym:
   (a) 1221 postów z tagiem #**israel**
   (b) 1180 postów z tagiem #**palestine**
   (c) 1232 postów z tagiem #**gaza**
   (d) 275 postów z tagiem #**rafah**
   (e) 1195 postów z tagiem #**hamas**

Co w rezultacie przyniosło zbiór danych o długości 11878 rekordów oraz 9 kolumn.

Rzeczywiście wykorzystane przeze mnie kolumny ze zbioru danych to:

1. **raw_content** - alfanumeryczna zawartość wpisu
2. **hash_tags** - tagi, którymi został oznaczony wpis

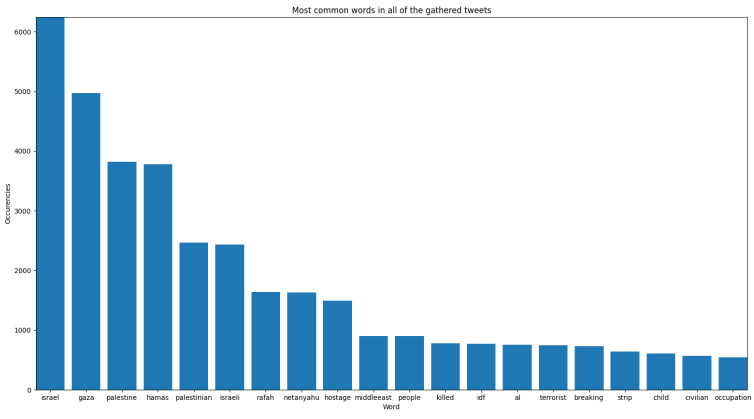
3 Preprocessing

By dokonać preprocessingu danych, pierw połączyłem dwa wyżej wymienione zbiory danych w jeden, a następnie przystąpiłem do:

1. usunięcia pustych wpisów
2. usunięcia linków (hiperłączy) ze wpisów
3. poprawienia literówek za pomocą paczki **Speller**[1]
4. podzielenia wpisów na tokeny
5. usunięcia stop-słów (stopwords) z pomocą różnych ich zbiorów[2, 3] oraz własnych spostrzeżeń
6. lematyzacji tokenów

4 Bag of words oraz Term Frequency

Kolejnym etapem zadania było przygotowanie Bag of words, którego rezultaty raczej nie powodują żadnych zaskoczeń. Najczęściej występującymi słowami były określenia państw, regionów, obywateli oraz ogólnych zagadnień związanych z operacjami wojennymi tych obydwu krajów ogarniętych wojną.

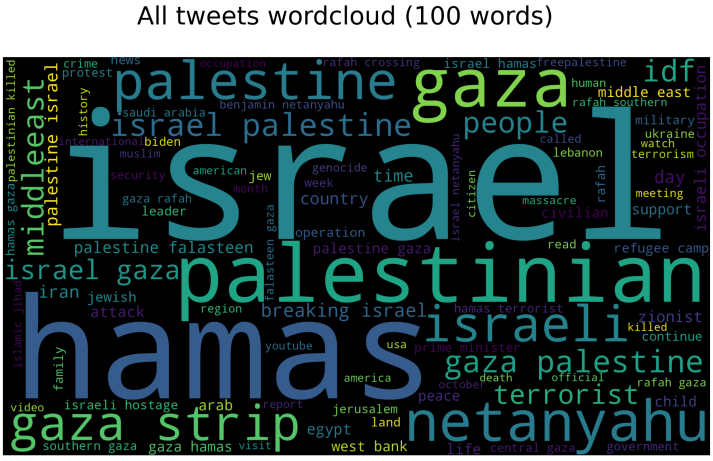


Rysunek 1: Wykres 20 najczęściej występujących słów w tweetach

Word	**Count**	**TF (%)**
israel	6245	3.20
gaza	4973	2.55
palestine	3823	1.96
hamas	3783	1.94
palestinian	2464	1.26
israeli	2430	1.24
rafah	1635	0.84
netanyahu	1627	0.83
hostage	1489	0.76
middleeast	901	0.46
people	897	0.46
killed	777	0.40
idf	767	0.39
al	758	0.39
terrorist	747	0.38
breaking	727	0.37
strip	641	0.33
child	610	0.31
civilian	571	0.29

Tabela 1: Słowa, liczba ich wystąpień i ich udział w całym bag of words

5 Word clouds



Rysunek 2: Chmura słów dla wszystkich tweetów

Wordcloud - tweets including #israel tag (300 words)

Rysunek 3: Chmura słów dla tweetów związanych z Izraelem

Rysunek 4: Chmura słów dla tweetów związanych z Palestyną

6 Analiza wydźwięku

Przeanalizujmy po kolei wszystkie wydarzenia przedstawione na wykresie:

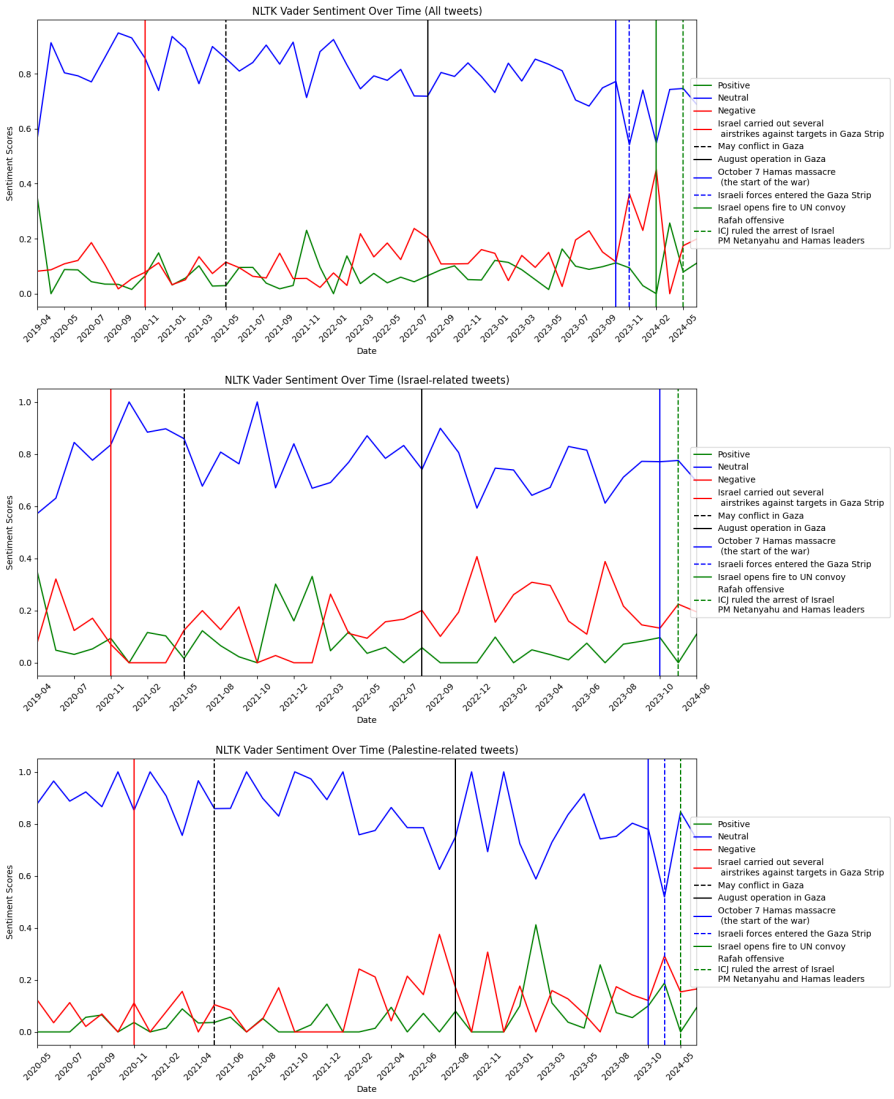
1. Ostrzał strefy Gazy przez Izrael w listopadzie 2020 r. - znaczny spadek w ilości racjonalnych oraz neutralnych wypowiedzi, a wzrost skrajnie negatywnych odczuć względem wpisów.

2. Wzrost popularności Hamasu oraz jego pierwsze ataki w 2021 r. - następuje odwrócenie trendów i skrajne emocje ustępują miejsca racjonalnym rozmowom na temat powstańczego charakteru tej organizacji.

3. Przejęcie władzy w strefie Gazy przez Hamas w marcu 2023 r. - powtarza się trend obserwowany w punkcie pierwszym.

4. Okres względnego spokoju trwający aż do przełomu września i października 2023 r.

5. Wystrzelenie przez Hamas 2000 rakiet w kierunku Izraela, 7 października 2023 r. oraz masowe ataki terrorystyczne (zabijanie dzieci oraz branie cywilów jako zakładników).



Rysunek 5: Analiza wydźwięku w zależności od czasu oraz powiązanych wydarzeń

7 Modelowanie tematów (Topic Modelling)

Topic modelling to metoda, która pozwala na automatyczne wykrywanie tematów w dużych zbiorach dokumentów. W naszym przypadku zastosowaliśmy model Latent Dirichlet Allocation (LDA) do analizy tweetów związanych z konfliktem na Bliskim Wschodzie.

Lp.	Tematy	Liczba dokumentów dot. tego tematu
1	gazaunderattack, warplane, southern, airstrikes, aircraft	202
2	crossing, egypt, border, egyptian, rafah	186
3	netanyahu, hitler, elonmusk, musk, elon	181
4	palestine, freepalestine, alla, palestinewillbefree, freedom	162
5	rubble, child, mother, gazaunderattack, house	145
6	child, childkillers, baby, genocidal, woman	135
7	normalization, saudi, arabia, saudiarabia, deal	96
240	bringthemhomenow, focus, lose, attacked, fighting	10

7.1 Wizualizacja tematów

Poniżej przedstawiamy wizualizację tematów uzyskanych za pomocą modelu LDA. Wizualizacja pozwala lepiej zrozumieć, jakie tematy dominują w analizowanych tweetach.

8 Wnioski

Z przeprowadzonej analizy wynikają następujące wnioski:

1. Większość wpisów jest związana z bieżącymi wydarzeniami i odzwierciedla ich dynamikę.
2. Słownictwo używane w tweetach zależy od strony konfliktu, której dotyczą.
3. Skrajne emocje i negatywne wydźwięki są wyraźnie związane z konkretnymi wydarzeniami wojennymi.
4. Modelowanie tematów pozwala na wyodrębnienie głównych tematów dyskusji, które dominują w analizowanych tweetach.

Literatura

[1] Speller: Python package for spelling correction, `https://pypi.org/project/speller/`

[2] Stopwords ISO: Collection of stopwords for multiple languages, `https://github.com/stopwords-iso/stopwords-iso`

[3] NLTK Stopwords: Stopwords collection provided by the Natural Language Toolkit, https://www.nltk.org/nltk_data/

[4] https://www.datacamp.com/tutorial/wordcloud-python

[5] https://spacy.io/

[6] https://matplotlib.org/stable/plot_types/index.html

[7] Wykłady Pana Grzegorza Madejskiego