

Raport badawczy - Wybory prezydenckie 2025

Analiza tekstu z mediów społecznościowych

Antoni Gawron

1. Wstęp	2
2.Pobieranie danych	3
2.1.Kluczowe momenty	3
2.2.Całkowita liczba tweetów	4
2.3.Rozkład pobranych tweetów w czasie	5
2.4.Pobieranie tweetów	6
2.4.1.Struktura zapytania	6
2.4.2.Symulacja zachowania użytkownika	7
2.4.3. Działanie algorytmu	8
3.Analiza treści	10
3.1.Preprocessing	10
3.2.Analiza treści kluczowych momentów - Rafał Trzaskowski	12
3.2.1.Debata TVP w Końskich	12
3.2.2.NASK	14
3.2.3.Obietnica	17
3.3.Analiza treści kluczowych momentów - Karol Nawrocki	19
3.3.1.Mieszkanie	19
3.3.2.Rozmowa ze Sławomirem Mentzenem	22
3.3.3.Snus na debacie	24
3.4.Wnioski	26
4.Analiza emocji	26
4.1.Preprocessing	26
4.2.Analiza modelem dkleczek/bert-base-polish-cased-v1	26
4.2.1.Pierwsza analiza modelem bert.	27
4.2.2.Druga analiza modelem bert	28
4.2.3.Trzecia analiza modelem bert	30
4.2.4.Wnioski	31
4.3.Analiza modelem eevvgg/PaReS-sentimenTw-political-PL	31
4.3.1.Pierwsza analiza	33
4.3.2.Druga analiza	34
4.3.3.Trzecia analiza	36
4.3.4.Wnioski	37

4.4.Analiza modelem twitter-xlm-roberta-base-sentiment-finetunned	38
4.5.Analiza za pomocą text2emotion	41
4.5.1.Preprocessing	41
4.5.2.Wyniki	41
5.Porównanie emocji z sondażami wyborczymi	43
6.Wnioski	45

1. Wstęp

Celem tego projektu była analiza treści publikowanych w mediach społecznościowych w kontekście wyborów prezydenckich w Polsce w 2025 roku. Wpisy były pobierane z platformy X (dawniej Twitter), będącej jednym z głównych kanałów komunikacji politycznej oraz źródłem opinii społecznych w czasie kampanii wyborczej.

Analizie poddano wpisy opublikowane w okresie: **01.03.2025 - 10.06.2025** (100 dni). Głównym przedmiotem badania były treści dotyczące dwóch kandydatów, którzy uzyskali największe poparcie w tegorocznych wyborach prezydenckich: Rafała Trzaskowskiego oraz Karola Nawrockiego.

W ramach analizy szczególną uwagę poświęcono kluczowym momentom kampanii wyborczej dla każdego z wymienionych kandydatów. Skupiono się na tym, jak wydarzenia te były komentowane przez użytkowników platformy X oraz jakie reakcje wywoływały w przestrzeni medialnej.

2.Pobieranie danych

2.1.Kluczowe momenty

Na potrzebę projektu zostało pobranych **14158** tweetów. W tym celu wykorzystana została paczka **twikit** (<https://twikit.readthedocs.io/en/latest/index.html>). Wpisy były pobierane z konkretnych momentów kampanii wyborczej. Były to momenty, które

były najbardziej "medialne" i które mogły mieć największy wpływ na wynik wyborów. Dla obu kandydatów zostały wybrane 3 kluczowe momenty.

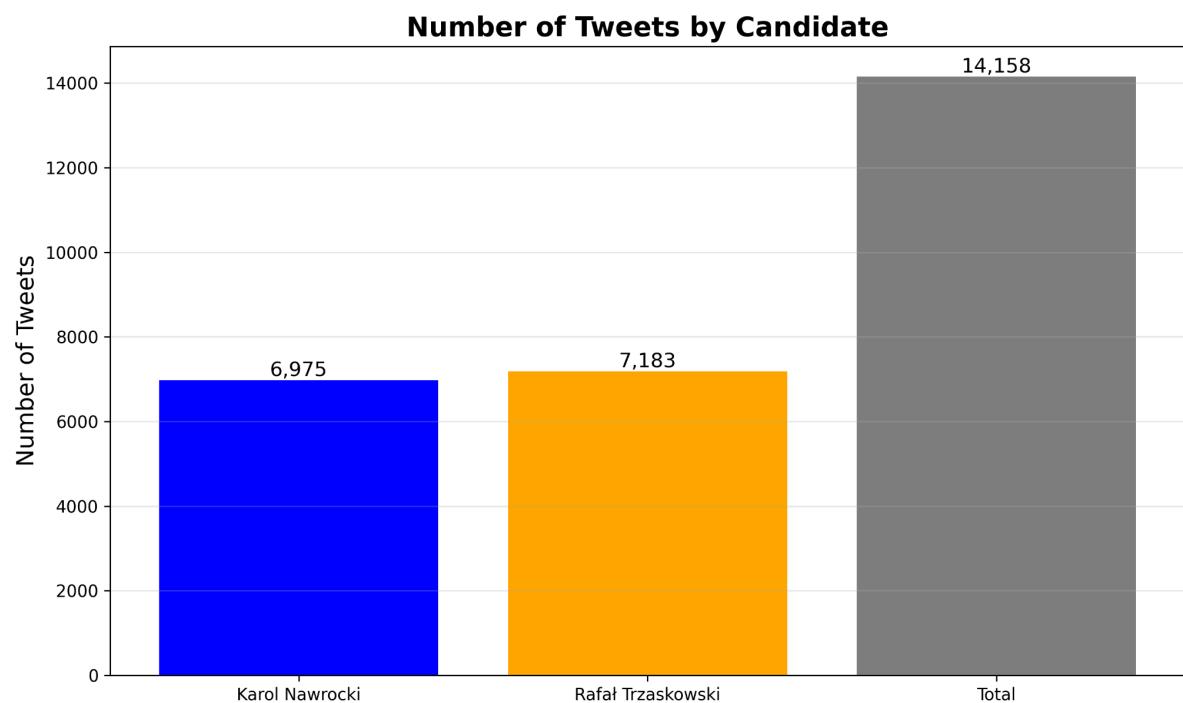
Rafał Trzaskowski:

1. **Oświadczenie NASK** - 14.05.2025 NASK ogłasza możliwą ingerencję poprzez zagraniczne reklamy na Facebooku. Dzień później Wirtualna Polska ujawniła, że za kontami stoją osoby powiązane z fundacją Akcja Demokracja, a także podano podejrzenie, że środki pochodzą spoza kraju.
2. **Cóż szkodzi obiecać** - 20.05.2025 w programie "Debata Gozdyry" Przemysław Witek zapytany o to czy nie wprowadzą żadnych nowych podatków odpowiedział: "Cóż szkodzi obiecać". Wypowiedź ta była wywołała skrajne emocje w mediach społecznościowych i do końca kampanii wyborczej była cytowana.
3. **Debata TVP w Końskich** - 11.04.2025 odbyła się debata w Końskich. Na tej debacie zabrakło niektórych kandydatów (między innymi Sławomira Mentzena czy Adriana Zandberga) z powodu zbyt późnego zaproszenia na tę debatę (2 godziny przed). Wydarzenie to wywołało wiele kontrowersji dotyczących organizacji a także finansowania, co stało się przedmiotem badań PKW.

Karol Nawrocki:

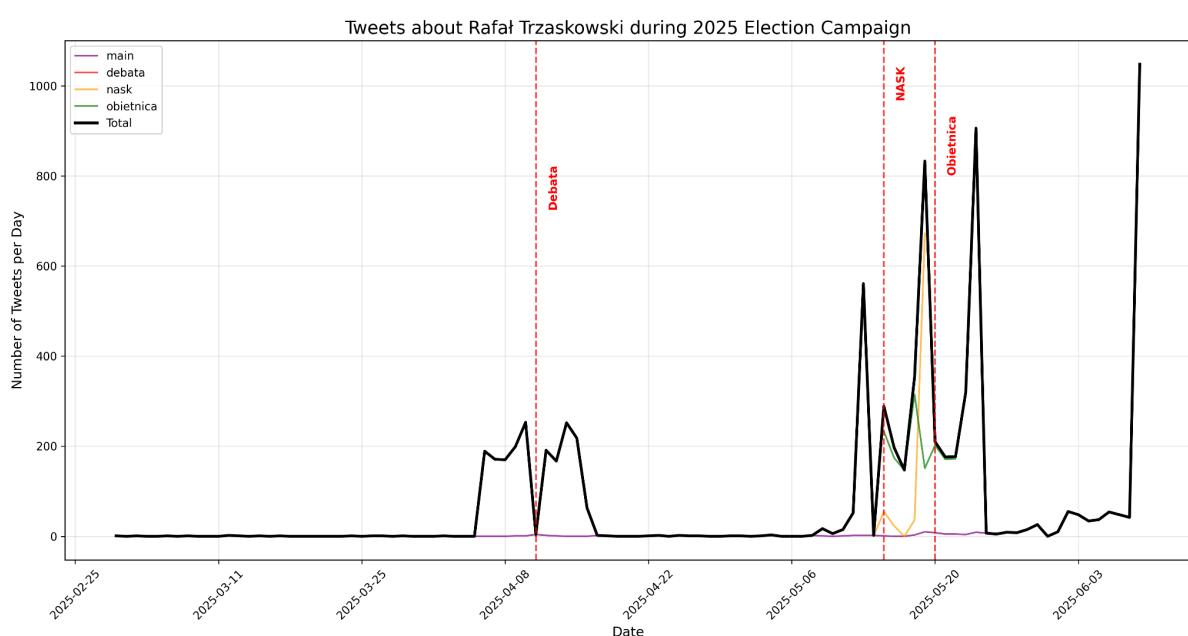
1. **Sprawa z mieszkaniem** - 05.05.2025 po debacie w Super Express Onet ujawnił, że Karol Nawrocki ma drugie mieszkanie, które przejął od Jerzego Ż. w zamian za opiekę. Ujawniono także, że pan Jerzy przebywa w DPS. Sprawa wywołała ogromne emocje w opinii publicznej i była tematem debaty aż do końca wyborów.
2. **Snus na debacie** - 23.05.2025 na debacie w TVP Karol Nawrocki zasłonił usta a drugą ręką umieścił coś pod wargą. Początkowo sztab Nawrockiego mówił, że to guma natomiast przyznali później, że był to snus. Stało się to jednym z najbardziej viralowych momentów tej kampanii wyborczej.
3. **Rozmowa z Mentzenem** - 22.05.2025 odbyła się rozmowa między Karolem Nawrockim a Sławomirem Mentzenem. W rozmowie tej Nawrocki skrytykował wiele rzeczy, które robił PiS (partia która popierała Nawrockiego), zgodził się z punktami Sławomira Mentzena a także przyznał się do udziału w ustawkach kibicowskich. Rozmowa ta mogła mieć potencjalnie pozytywny wpływ na jego wynik wyborczy.

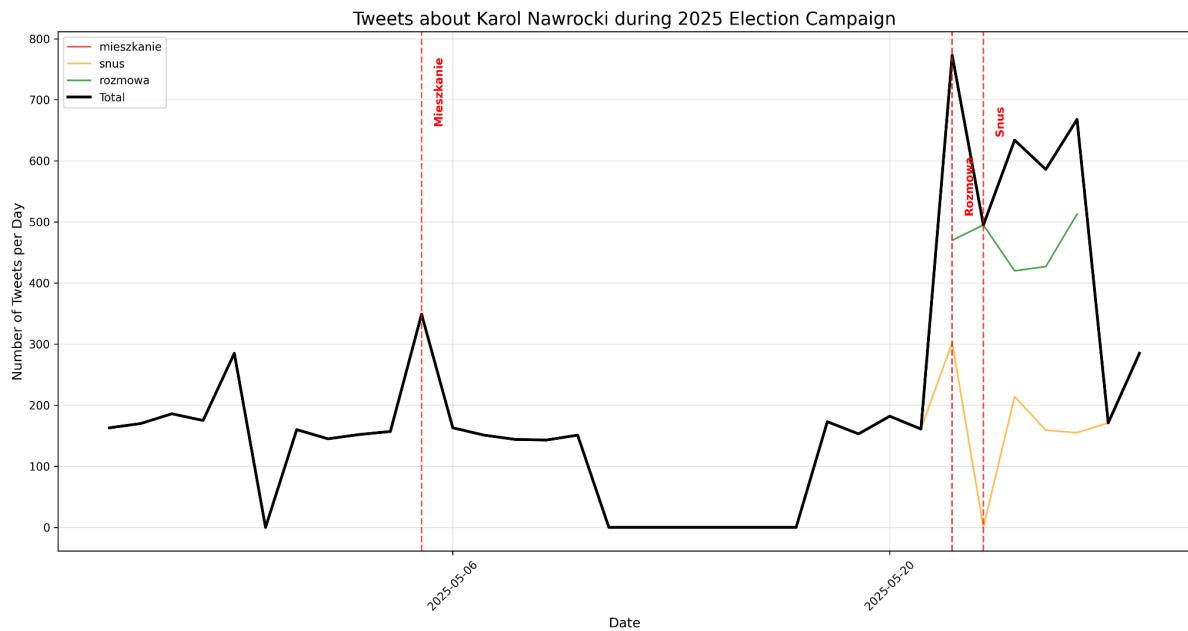
2.2.Całkowita liczba tweetów



Wykres przedstawia liczbę tweetów pobranych dla każdego z kandydatów. Pobrano 6975 tweetów o Karolu Nawrockim i 7183 tweety na temat Rafała Trzaskowskiego, co daje łącznie 14158 tweetów.

2.3.Rozkład pobranych tweetów w czasie





Wykresy przedstawiają rozkład tweetów na temat Rafała Trzaskowskiego i Karola Nawrockiego w czasie. Zaznaczono na nich czerwonymi przerywanymi liniami wyżej wymienione kluczowe momenty. Dla Karola Nawrockiego wykres przedstawia tweety zebrane od dnia 25.04.2025. Wcześniej nie pobierano żadnych wpisów na jego temat. Dla obu kandydatów wpisy były pobierane z 2 “kategorií” Top - tweety z największymi liczbami polubień, komentarzy itd. oraz Latest - tweety ostatnio opublikowane tweety.

2.4.Pobieranie tweetów

Aby pobrać dane wcześniej należy pobrać token autoryzacyjny logując się do X używając danych prawdziwego konta.

```
async def login():
    client = Client(language='pl')
    await client.login(auth_info_1=username, auth_info_2=email, password=password)
    client.save_cookies('cookies.json')
```

Wszystkie potrzebne do autoryzacji dane zapisywane są w pliku cookies.json, z którego później korzysta scraper:

```
client = Client(language='en-US')

client.load_cookies('cookies.json')
```

Biblioteka twikit pobiera tweety batchami - domyślnie pobiera ich 20 w każdym batchu, lecz często zdarza się, że zostanie pobrane więcej lub mniej tweetów. Po każdym batchu tweety zapisywane są do pliku csv. Dla każdego tweeta zapisywany jest autor, treść tweeta, data utworzenia, liczba odpowiedzi, liczba polubień i liczba wyświetleń. Po każdym batchu należy odczekać przed wykonaniem kolejnego zapytania.

2.4.1. Struktura zapytania

X umożliwia wyszukiwanie tweetów na podstawie specjalnie skonstruowanego zapytania. Pozwala to na filtrowanie wyników zanim pobierze się tweety. Struktura takiego zapytania wygląda następująco:

all of these "exact phrase" (any OR of OR these) -none -of -these (#hash1 OR #hash2) lang:pl until:2025-06-20 since:2025-01-01

Elementy zapytania:

- all of these - wszystkie z tych słów muszą być w treści tweeta
- “exact phrase” - szukana fraza musi wystąpić w treści tweeta
- (any OR of OR these) - jedno z tych słów musi wystąpić w treści tweeta
- -none -of -these - żadne z tych słów nie może wystąpić
- (#hash1 OR #hash2) - hasztagi które mają wystąpić w tweecie
- lang:pl - język tweeta
- until, since - ustalenie przedziału czasowego z którego mają pochodzić tweety

Ponadto można filtrować, z jakich kont, do kogo lub o kim mają być dane tweety

- (from:account) - tweety użytkownika account
- (to:account) - tweety do użytkownika account
- (@account) - tweety które wspominają account

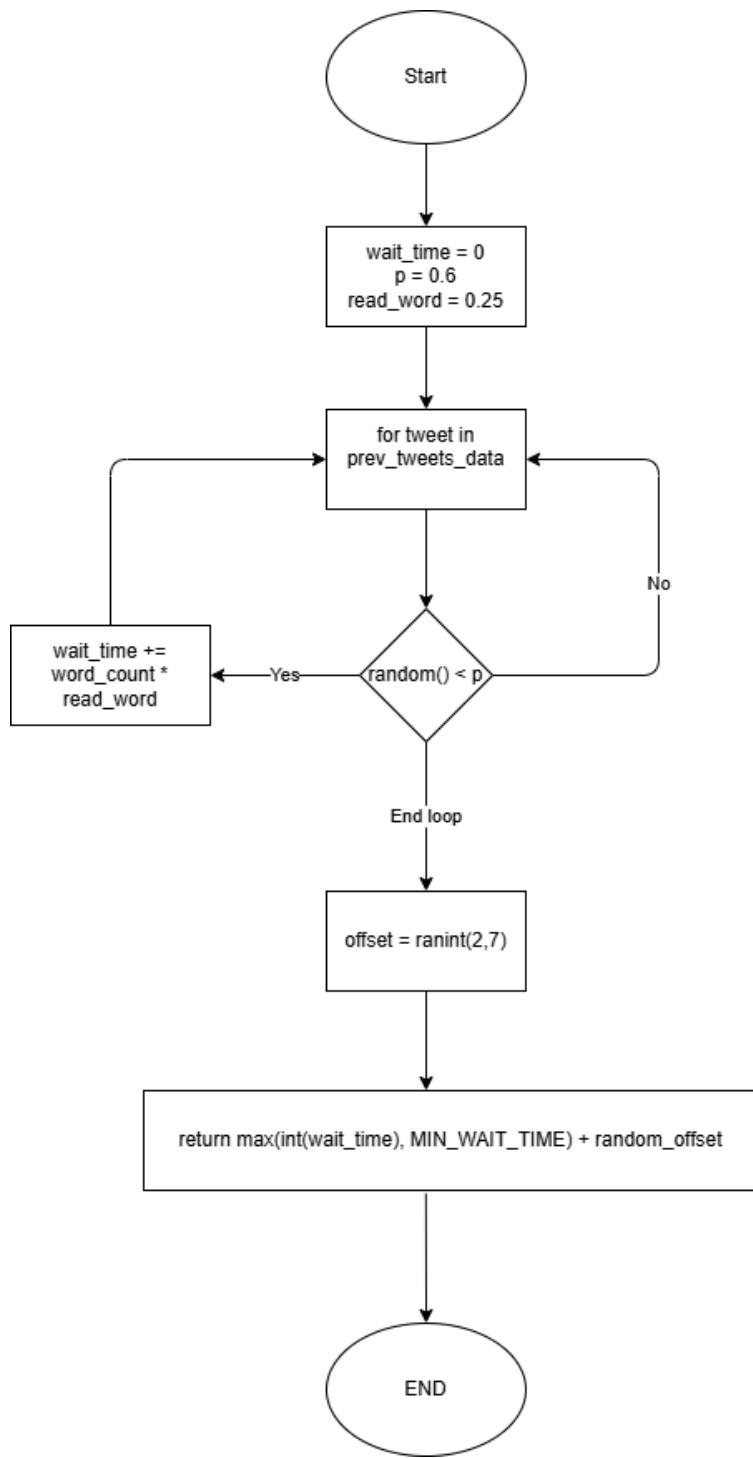
W celu szukania tweetów najpopularniejszych, najczęściej lubianych/repostowanych można używać dodatkowo takich parametrów:

- min_replies:100 - minimalna liczba odpowiedzi

- min_faves:101 - minimalna liczba polubień
- min_retweets:102 - minimalna liczba retweetów

2.4.2. Symulacja zachowania użytkownika

Aby zasymulować ile sekund odczekać przed wysłaniem kolejnego zapytania stworzony został następujący algorytm do symulacji ludzkiego zachowania. Dla każdego pobranego tweeta sprawdzane jest czy użytkownik ma przeczytać tweeta (nie każdy tweet jest czytany). Jeśli czytamy tweeta to liczona jest liczba słów w danym tweecie i mnożona jest ona razy czas potrzebny na przeczytanie słowa i dodajemy ten iloczyn do wyniku. Na koniec wyliczane jest losowe przesunięcie między 2 a 7 sekund. Zwracana jest całkowita liczba sekund potrzebna do oczekania - nie mniejsza niż ustalony próg. Poniżej znajduje się schemat działania tego algorytmu.

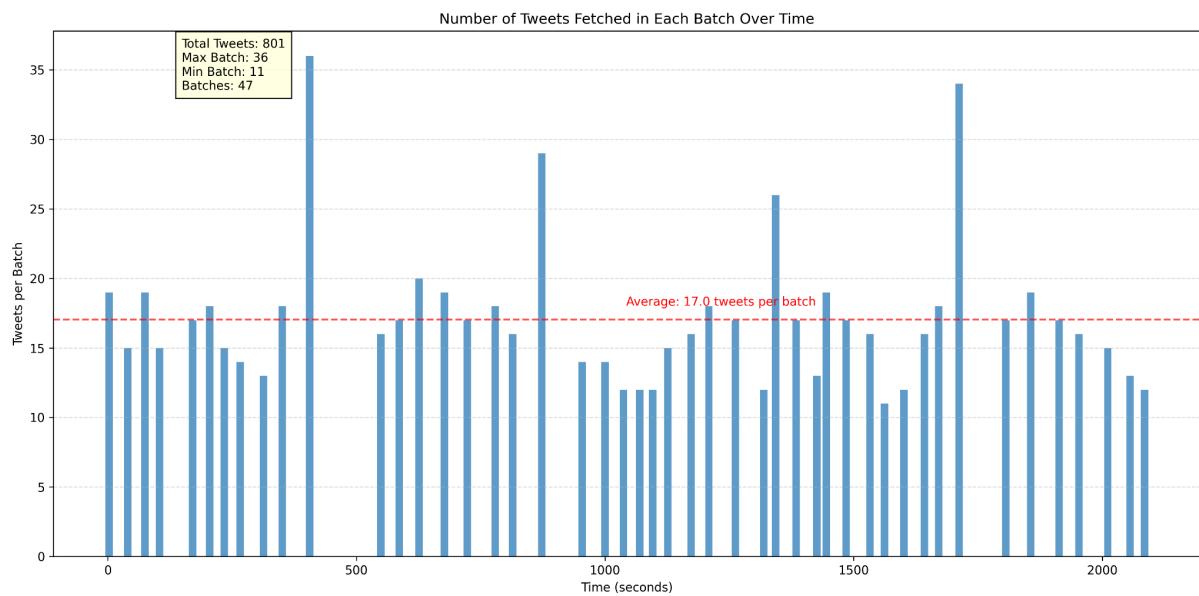


2.4.3. Działanie algorytmu

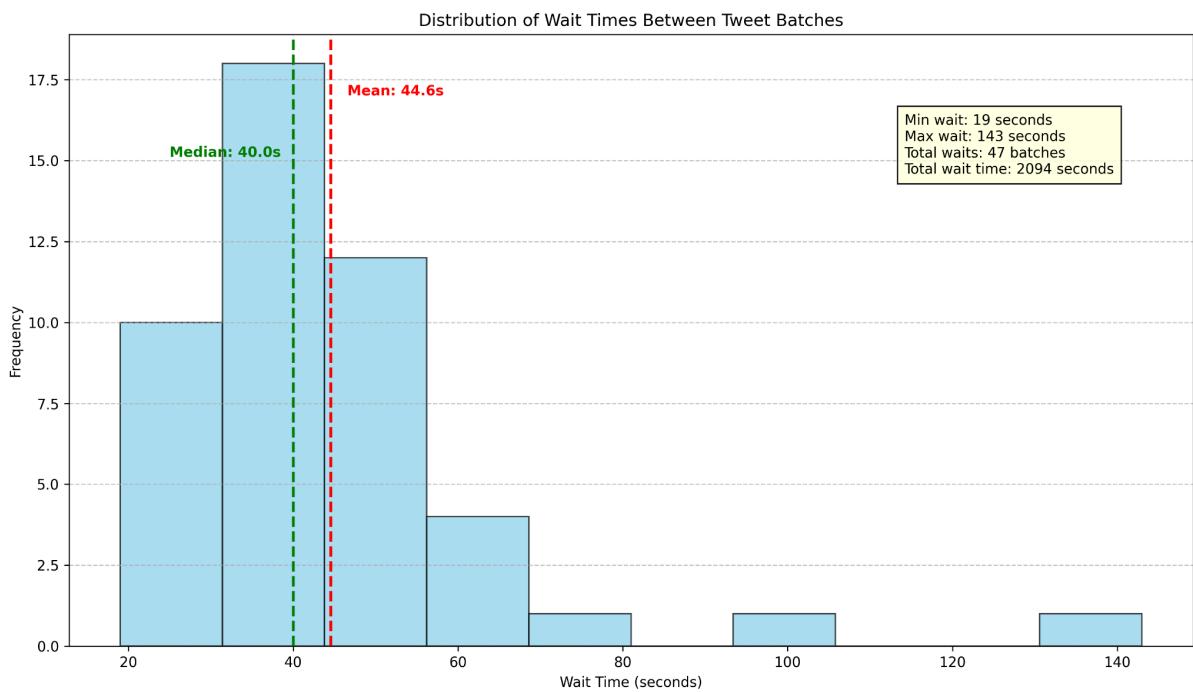
Dla “sesji” pobierania tweetów zbadane zostało działanie algorytmu. Dla każdego batcha zapisano liczbę pobranych tweetów oraz czas jaki należało odczekać przed

pobraniem kolejnych tweetów. Na poniższym wykresie widać zależność pomiędzy liczbą pobranych tweetów a czasem oczekiwania. Im więcej tweetów zostało pobranych w danym batchu tym dłużej trzeba było czekać aby pobrać następne. Dodatkowo na histogramie widać jak rozkładają się czasy oczekiwania.

Liczba tweetów w każdym batchu w czasie.



Histogram czasów oczekiwania



3.Analiza treści

3.1.Preprocessing

Przed analizą treści tweety zostały oczyszczone z linków, symboli wspomnień (@) oraz symboli hasztagów (#). Rzeczy te nie mają znaczenia w kontekście analizowania treści tweetów.

Popularna paczka **nltk** nie ma wsparcia dla języka polskiego, w związku z tym nie można skorzystać z popularnych rozwiązań takich jak lematyzer WordNetLemmatizer czy też stemmer SnowballStemmer. W tym celu wykorzystana została paczka **Morfeusz** (<http://morfeusz.sjip.pl/>). Narzędzie to dokonuje analizy morfologicznej dla języka polskiego. Określa ona dla danego słowa wszystkie formy wszystkich leksemów (abstrakcyjna jednostka leksykalna, wyraz słownikowy).

Lista stop words dla języka polskiego została pobrana ze strony (<https://countwordsfree.com/stopwords/polish>). Znajdują się tak 274 słowa, które powszechnie wykorzystywane są w języku polskim i zostały usunięte przed dalszą analizą. Jako narzędzie do tokenizacji wykorzystano paczkę **nltk.tokenize**.

Poniżej przedstawiono kod który posłużył do preprocessingu.

```
def preprocess_tweets(tweets):
    cleaned_tweets = []

    stop_words = load_stop_words()
    morf = morfeusz2.Morfeusz()

    for tweet in tweets:
        tweet = clean_tweet(tweet)
        tokens = word_tokenize(tweet.lower())

        lemmatized_tokens=[]
        for word in tokens:
            if word.isalpha():
```

```
lemma = morf.analyse(word)[0][2][1]

if ":" in lemma:
    lemma = lemma.split(":")[0]
if lemma not in stop_words:
    lemmatized_tokens.append(lemma)

cleaned_tweets.extend(lemmatized_tokens)

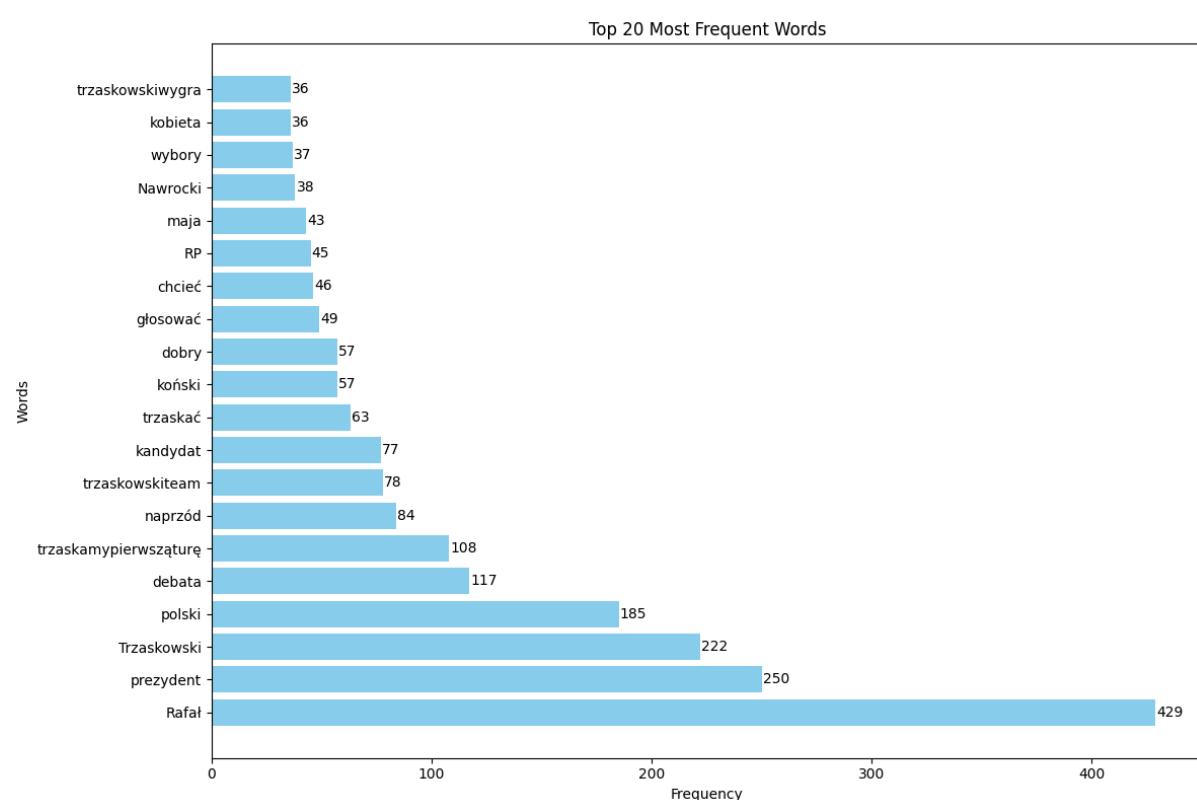
# removed_tags_tokens = remove_morf_tags(cleaned_tweets)
return cleaned_tweets
```

Na początek ładowane są stop words z pliku w formacie **.json**. Następnie tworzony jest obiekt za pomocą konstruktora Morfeusz() z paczki morfeusz2. Dla każdego tweeta usuwane są linki, symbole używając funkcji clean_tweets. Następnie wykorzystana jest funkcja word_tokenize aby przerobić tweeta na tokeny. Dla każdego tokenu sprawdzane jest, czy jest to słowo (czy składa się z liter alfabetu). Jeżeli tak to wykonywana jest analiza morfologiczna używając funkcji analyse. Domyślnie funkcja ta zwraca wiele informacji: formę tekstową, lemat, znacznik morfoskładniowy, listę informacji o „pospolitości” rzeczownika (np. nazwa pospolita, marka, nazwisko), listę kwalifikatorów stylistycznych (np. daw., pot., środ., wulg.) i dziedzinowych (np. bot., zool.). Na potrzebę preprocessingu wybierany jest lemat (forma bazowa/hasłowa). Funkcja analyse w niektórych przypadkach dodaje do lematu tagi (po znaku :). Tagi te są usuwane. Następnie sprawdzane jest, czy wyciągnięty lemat nie należy do stop words - jeśli tak to jest on pomijany.

3.2.Analiza treści kluczowych momentów - Rafał Trzaskowski

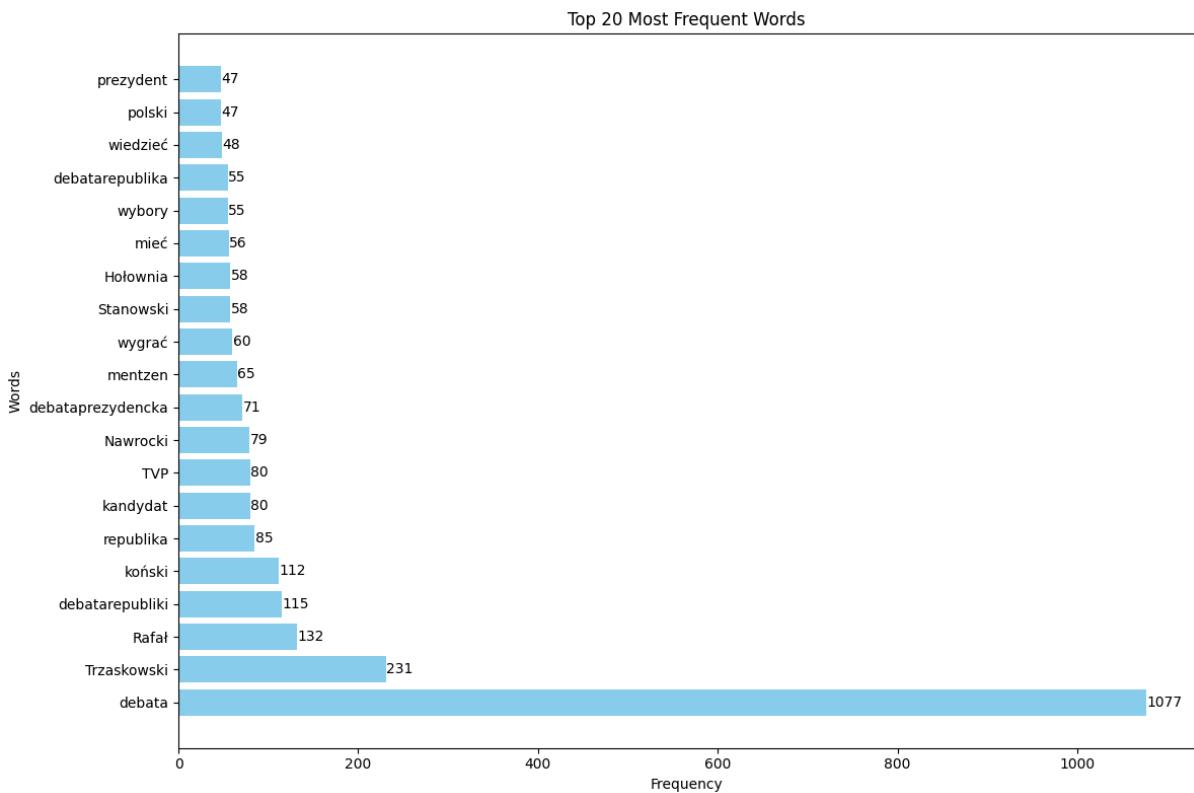
3.2.1.Debata TVP w Końskich

Poniżej przedstawiono wykresy częstotliwości występowania słów oraz chmurę słów dla próbki 980 tweetów z okresu 06.04.2025 - 10.04.2025 (5 dni przed debatą)





Poniżej przedstawiono wykresy częstotliwości występowania słów oraz chmurę słów dla próbki 878 tweetów z okresu 12.04.2025 - 16.04.2025 (5 dni po debacie)



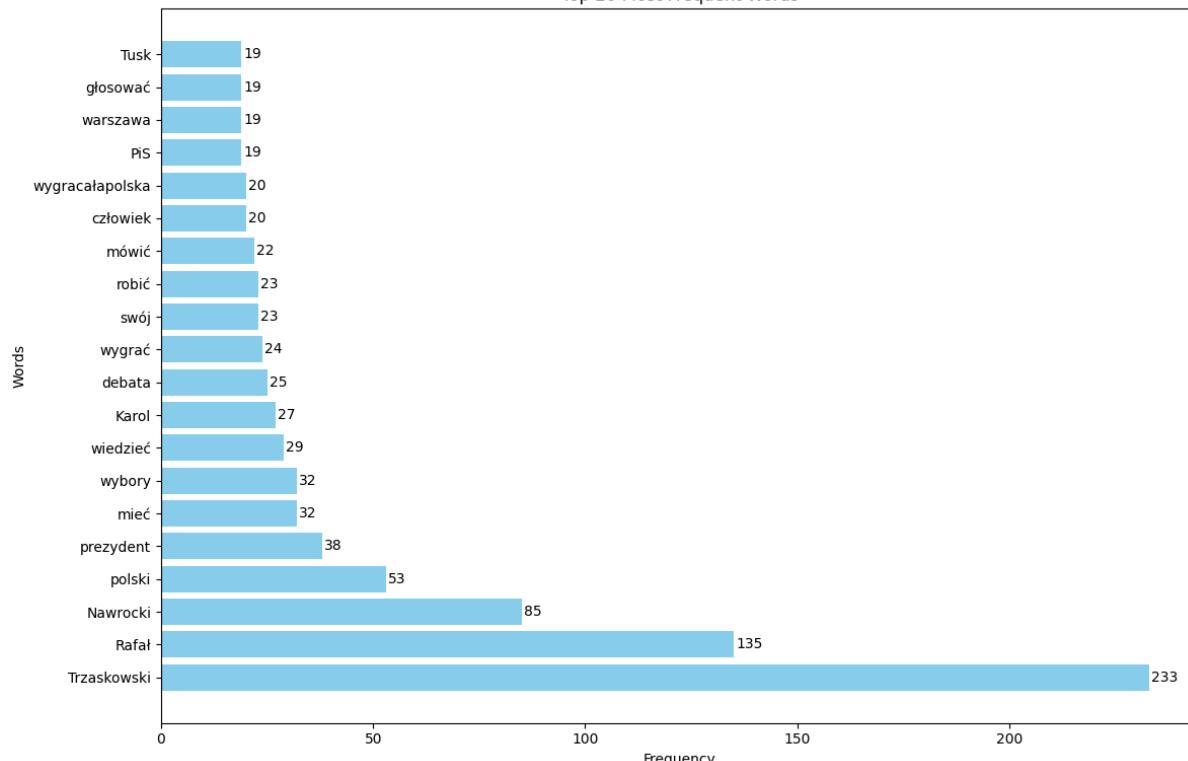


Na wykresach częstotliwości przed debatą widać, że wśród tweetów o Rafałowi Trzaskowskim dominowały słowa takie jak: Rafał, prezydent, Trzaskowski. Popularne były także hasztagi takie jak: #trzaskamypierwszature, #trzaskowskiteam. Na wykresie słów po debacie widać drastyczny wzrost częstotliwości słowa “debata” co pokazuje, że debata w Końskich była wydarzeniem, które było głośne medialnie. Widać również wzrost innych słów/haseł/hasztagów takich jak: debatarepubliki, koński czy też debataprezydencka. Ponadto widać, że w tweetach zaczęły pojawiać się nazwiska innych kandydatów: Nawrocki, Mentzen, Hołownia, Stanowski. Pokazuje to, że tweetach o Rafałowi Trzaskowskim wspominani byli także inni kandydaci a sama debata w Końskich sprawiła, że mniej popularni/rozpoznawalni kandydaci tacy jak Krzysztof Stanowski zyskali rozgłos medialny.

3.2.2.NASK

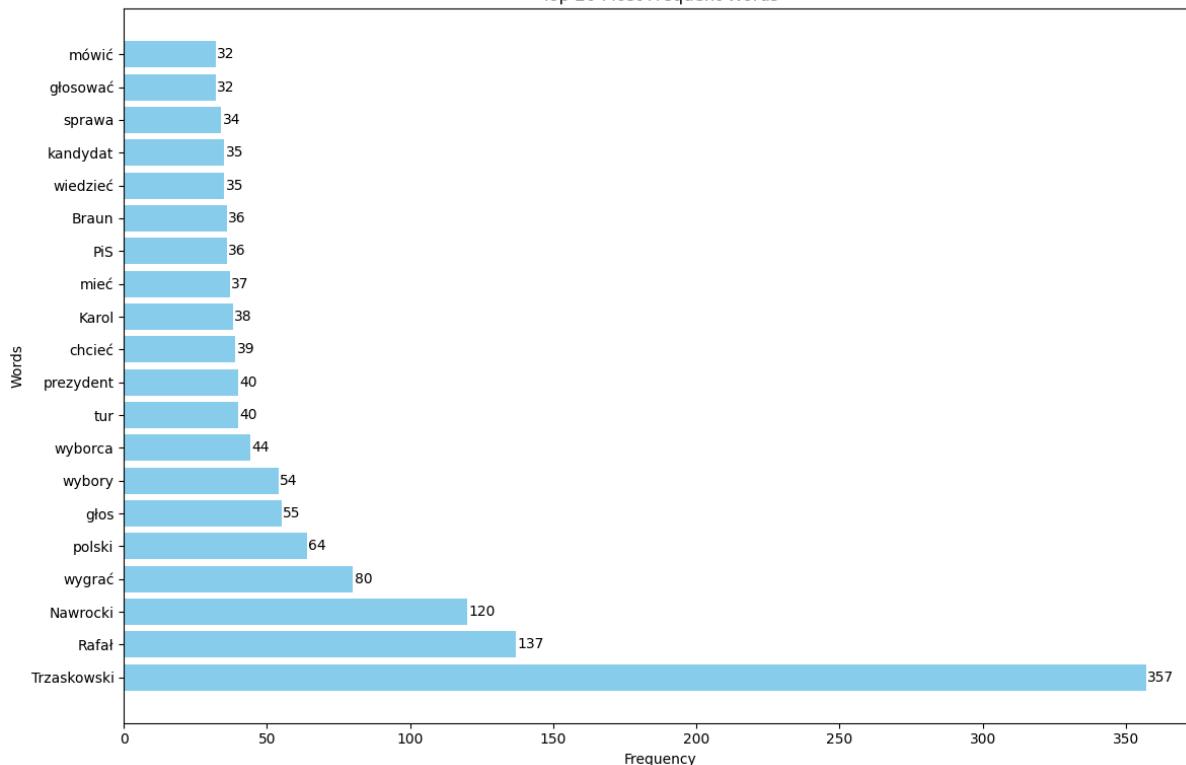
Poniżej przedstawiono wykresy częstotliwości występowania słów oraz chmurę słów dla próbki 645 tweetów z okresu 11.05.2025 - 14.05.2025 (4 dni przed oświadczeniem NASK)

Top 20 Most Frequent Words



Poniżej przedstawiono wykresy częstotliwości występowania słów oraz chmurę słów dla próbki 786 tweetów z okresu 15.05.2025 - 19.05.2025 (5 dni po oświadczenie NASK)

Top 20 Most Frequent Words

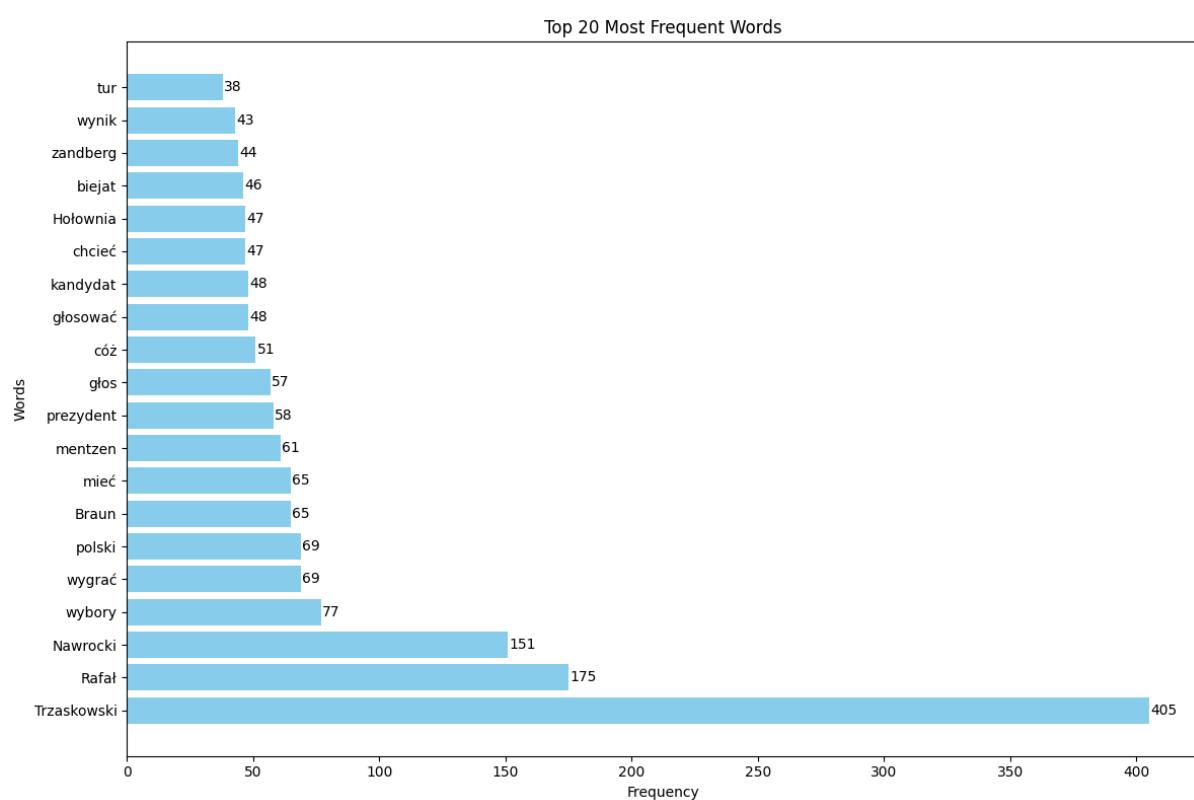


Na powyższych wykresach częstotliwości słów nie widać aby słowa/hasła związane z oświadczeniem NASK w sprawie spotów wyborczych na Facebooku. Wykresy te obejmują jedynie 20 najczęściej występujących słów. Natomiast na chmurach słów, które obejmują znacznie więcej słów widać, że pojawiały się hasła związane z tą sprawą: NASK, finansować, demokracja. Ponadto, tweety zbierane po oświadczeniu NASK tj. 15.05.2025 - 19.05.2025 zawierają w sobie wpisy po pierwszej turze wyborów (18.05.2025). Zarówno na wykresie częstotliwości słów jak i chmurze słów widać, że oprócz Rafała Trzaskowskiego i Karola Nawrockiego pojawiało się

nazwisko Grzegorza Brauna. Jest to spowodowane wynikiem wyborczym tego kandydata - 6,34% - co dla większości społeczeństwa było ogromnym zaskoczeniem.

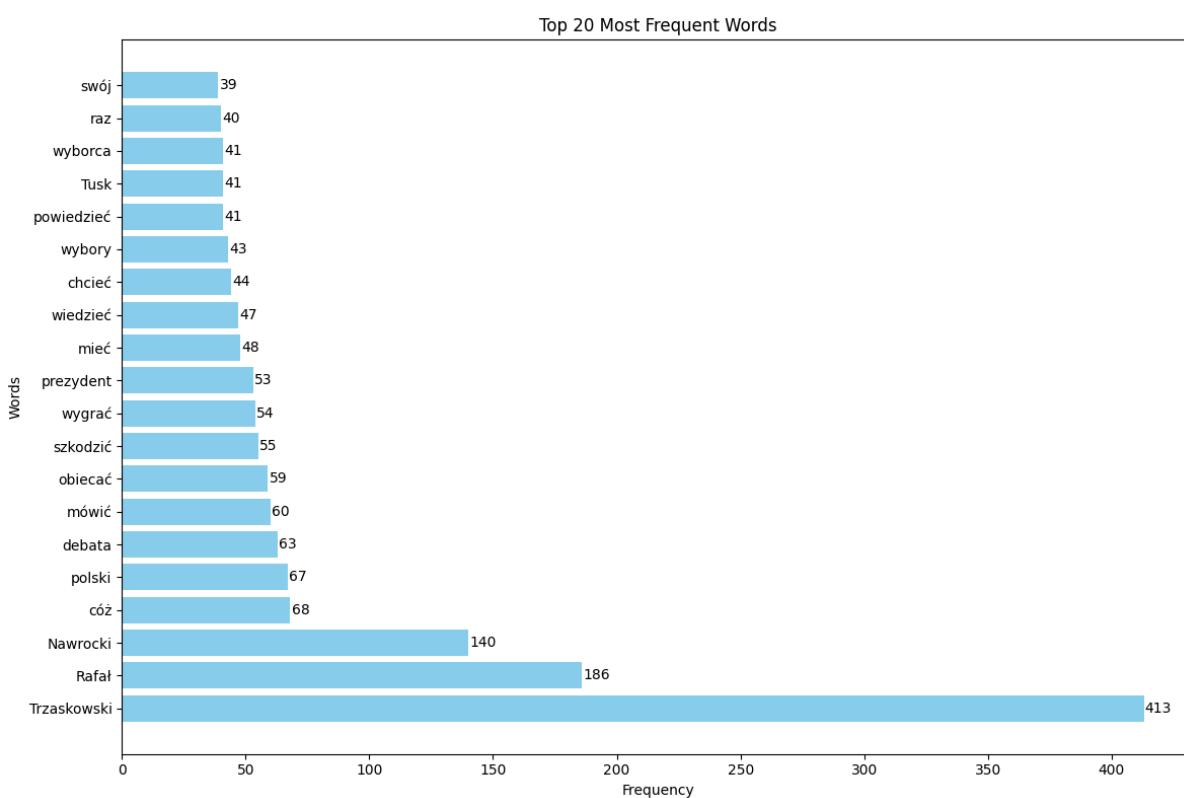
3.2.3. Obietnica

Poniżej przedstawiono wykresy częstotliwości występowania słów oraz chmurę słów dla próbki 1020 tweetów z okresu 15.05.2025 - 19.05.2025 (5 dni przed słowami Przemysława Witka)





Poniżej przedstawiono wykresy częstotliwości występowania słów oraz chmurę słów dla próbki 1003 tweetów z okresu 20.05.2025 - 24.05.2025 (5 dni po słowach Przemysława Witka)





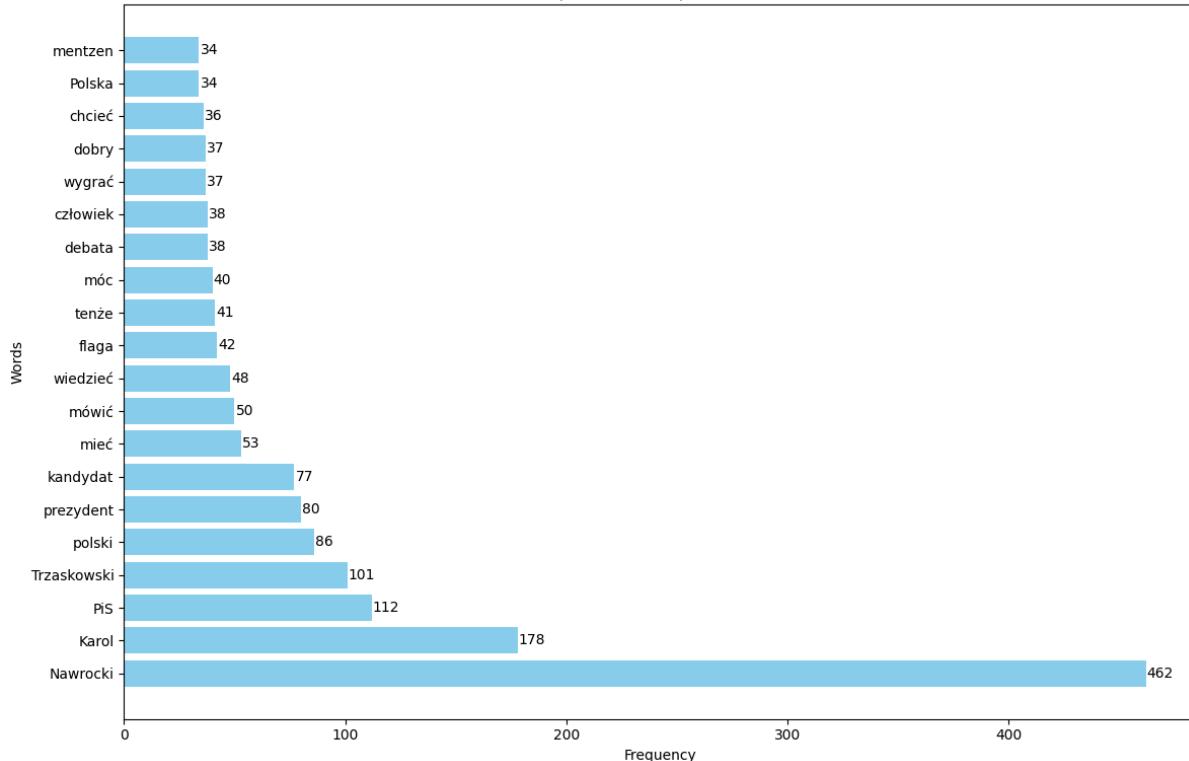
Na wykresie częstotliwości słów przed słowami Przemysława Witka widać, że pojawiały się nazwiska kandydatów a także, inne słowa związane z wyborami. Jest to okres kilku dni przed pierwszą turą a także 1 dzień po niej. Widać, że 3 najczęściej pojawiającym się nazwiskiem jest Braun, co jest spowodowane jego zaskakującym wysokim wynikiem. Natomiast na wykresie po słowach Przemysława Witka można zauważyć, że nazwiska innych kandydatów występowały rzadziej. Wszystkie słowa z frazy “cóż szkodzi obiecać” występowały wśród 8 najbardziej popularnych słów. Na wykresie tym widać także, że częściej pojawiało się nazwisko Donalda Tuska. Słowa posła KO miały duży wydźwięk medialny i pojawiały się w wielu wpisach. Ponadto mogły mieć one duży wpływ na ostateczny wynik wyborów.

3.3. Analiza treści kluczowych momentów - Karol Nawrocki

3.3.1. Mieszkanie

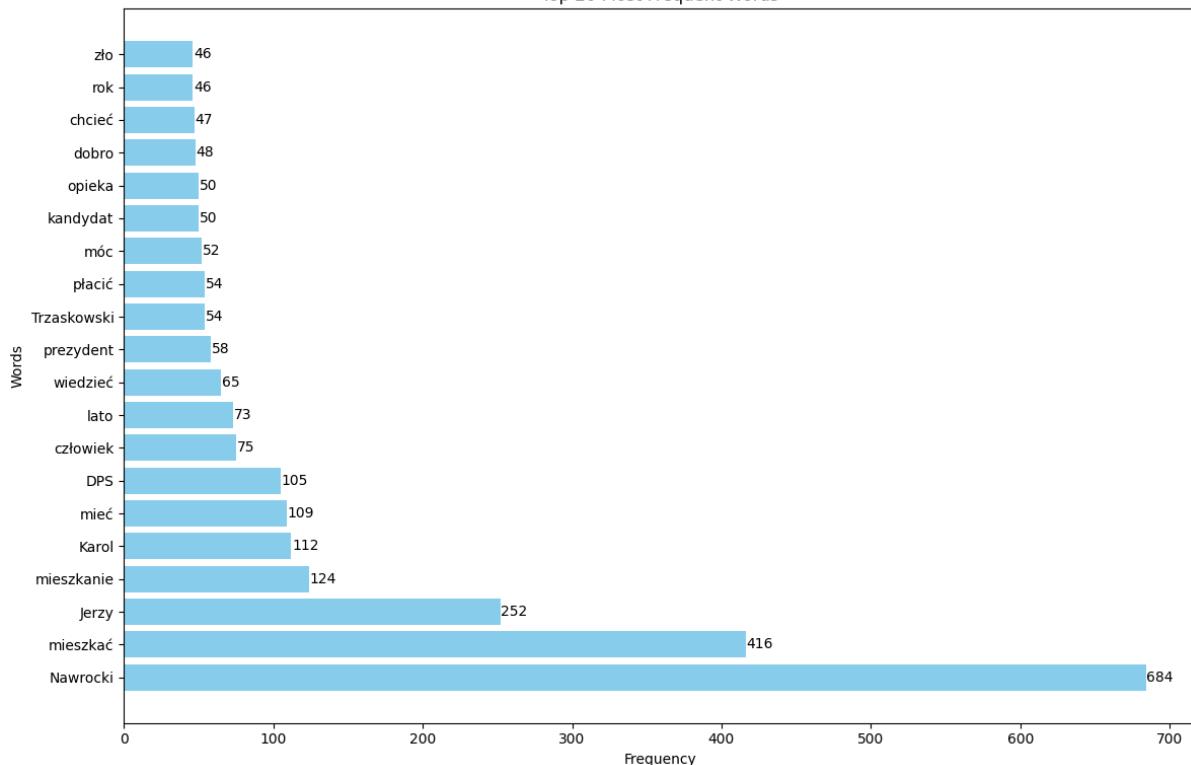
Poniżej przedstawiono wykresy częstotliwości słów oraz chmurę słów dla próbki 979 tweetów z okresu 25.04.2025 - 29.04.2025 (okres 5 dni przed opublikowaniem przez Onet artykułu)

Top 20 Most Frequent Words



Poniżej przedstawiono wykresy częstotliwości słów oraz chmurę słów dla próbki 954 tweetów z okresu 05.05.2025 - 10.05.2025 (okres 6 dni po opublikowaniu przez Onet artykułu)

Top 20 Most Frequent Words

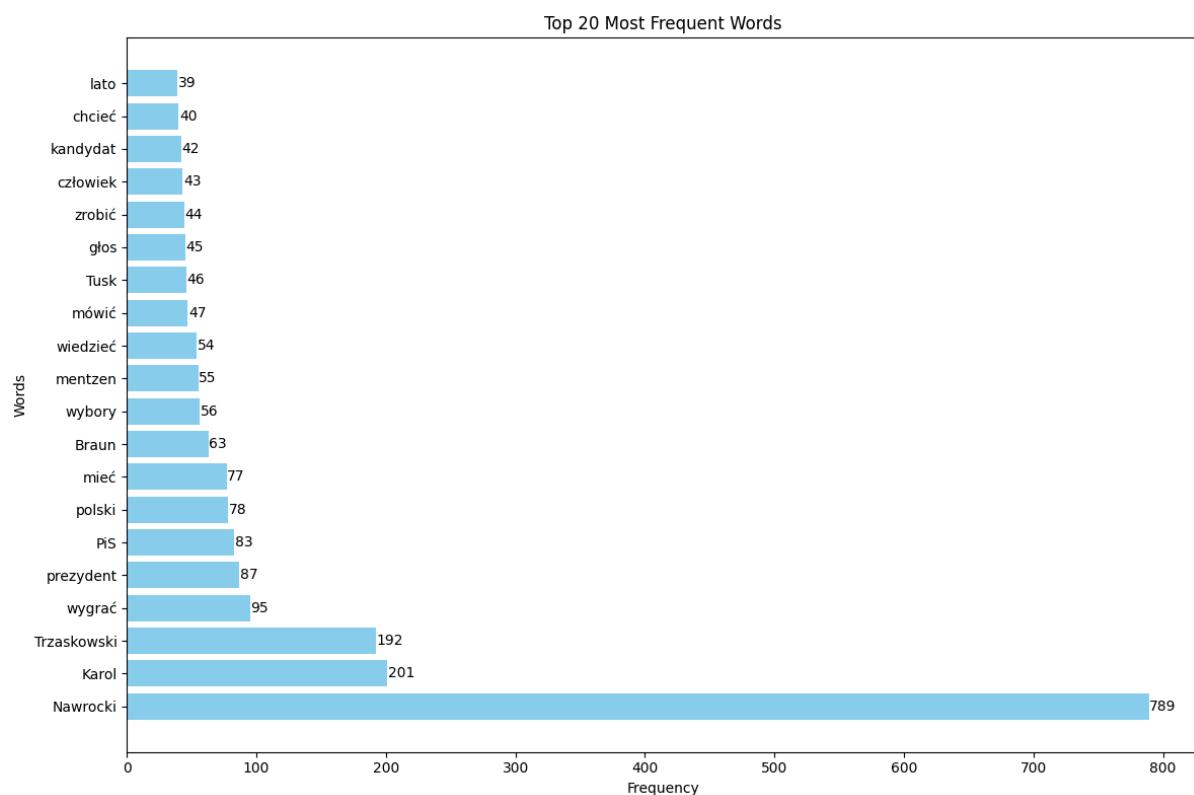


Na wykresie częstotliwości słów przed opublikowaniem artykułu przez Onet oprócz nazwiska Nawrockiego 3 najczęściej padającym hasłem jest PiS. Jest to spowodowane tym, że 27.04.2025 odbyła się konwencja Karola Nawrockiego, na której byli obecni członkowie PiS między innymi prezes PiS Jarosław Kaczyński. Dla wielu ludzi był to moment w którym Karol Nawrocki przestał być odbierany jako kandydat obywatelski, a zaczął być traktowany jak kandydat partyjny. Często pojawiającym się słowem była "flaga" co może odnosić się do sytuacji z debaty TVP w Końskich podczas której Magdalena Biejat odebrała Rafałowi Trzaskowskiemu

flagę LGBT. Na wykresie częstotliwości słów po artykule Onetu widać, że sprawa drugiego mieszkania Karola Nawrockiego całkowicie zdominowała debatę publiczną. Wśród wpisów na X dominowały słowa takie jak: mieszkanie, mieszkać, Jerzy, DPS, opieka itd. Sprawa drugiego mieszkania Karola Nawrockiego była jedną z najgłośniejszych w kampanii wyborczej 2025 i mogła mieć duży wpływ na poparcie tego kandydata.

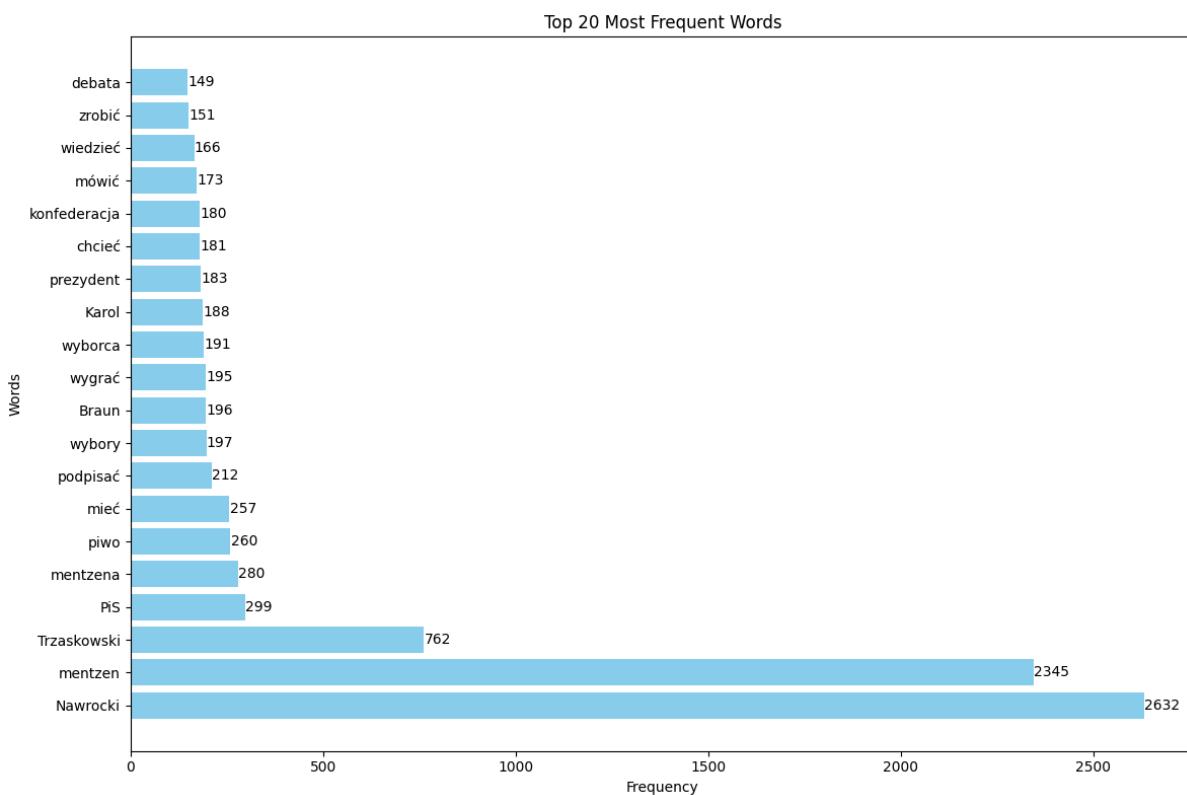
3.3.2. Rozmowa ze Sławomirem Mentzenem

Poniżej przedstawiono wykres częstotliwości słów oraz chmurę słów dla próbki 972 wpisów z okresu 18.05.2025 - 22.05.2025. Jest to ta sama próbka, która została pobrana w celu analizy treści następnego kluczowego momentu - snusa na debacie. Wykorzystano te same tweet, ponieważ pochodziły one z tego samego okresu i wyszukiwane były takimi samymi zapytaniami.





Poniżej przedstawiono wykresy częstotliwości słów dla próbki 2325 wpisów z okresu 23.05.2025 - 26.05.2025. Jest to okres po rozmowie Karola Nawrockiego ze Sławomirem Mentzenem i po debacie w TVP w drugiej turze.



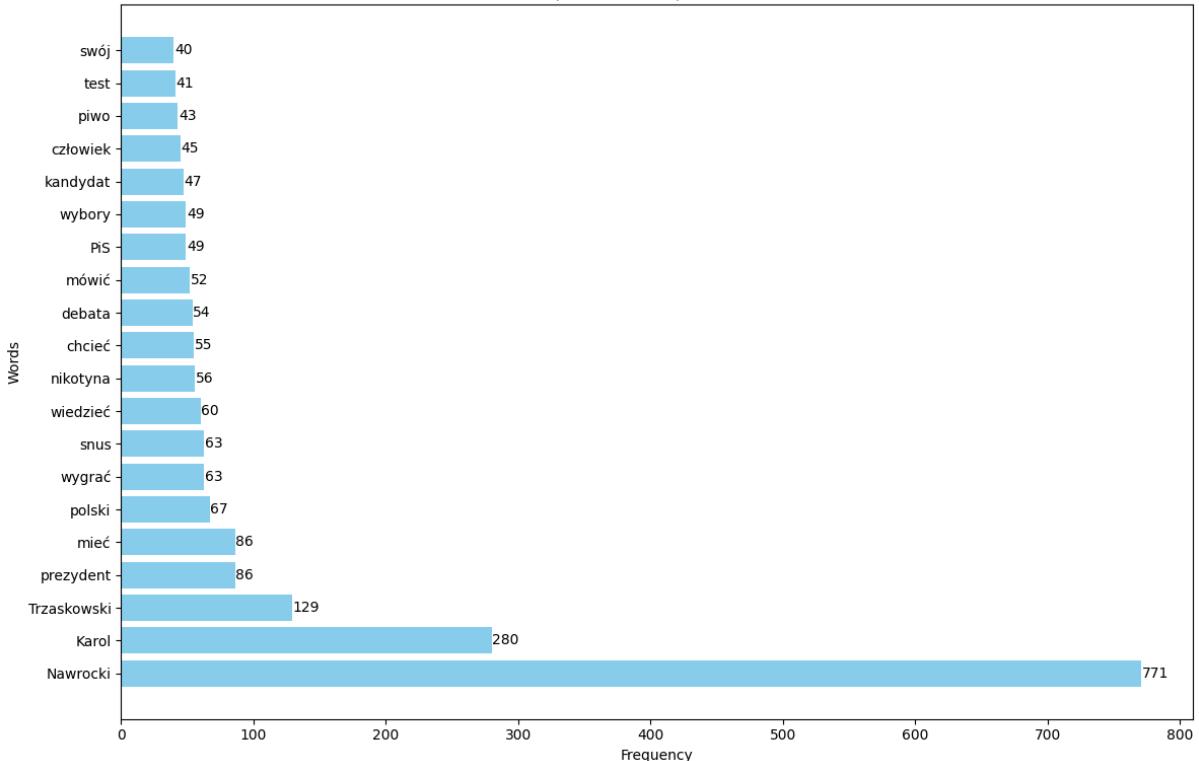


Na wykresie częstotliwości słów przed rozmową ze Sławomirem Mentzenem widać, że głównym tematem były wyniki wyborów 1 tury. Pojawiają się tam nazwiska kandydatów a także słowa związane z wyborami takie jak: głosować, wygrać, prezydent. Natomiast na wykresie słów po rozmowie, najczęściej padającymi nazwiskami były Nawrocki oraz Mentzen. Można też zauważać słowa/hasła które nawiązują do tej rozmowy między innymi: podpisać (Karol Nawrocki podpisał postulaty Mentzena) a także PiS (Karol Nawrocki skrytykował wiele rzeczy, które zrobił PiS). Pojawia się też słowo piwo, które nawiązuje do spotkania Mentzena, Trzaskowskiego i Sikorskiego w pubie Mentzen 25.05.2025.

3.3.3.Snus na debacie

Wykres częstotliwości słów występujących przed debatą został przedstawiony w punkcie 3.3.2. Poniżej przedstawiono wykres częstotliwości słów oraz chmurę słów dla próbki 984 tweetów z okresu 24.05.2025 - 28.05.2025.

Top 20 Most Frequent Words



Na wykresie częstotliwości słów widać, że po debacie w drugiej turze (23.05.2025) we wpisach występuły takie słowa jak: nikotyna, snus. Na chmurze słów widać, że pojawiały się też inne słowa, które odnosiły się do tej sytuacji: ćpun, woreczek, dziąsło, zażyć, narkotyk itd. Sprawa ta była komentowana na X używając wielu różnych słów, co sprawia, że nie ma jednego hasła/słowa które znacząco wyróżniało się wśród innych. W tym samym okresie tj. 25.05.2025. miało miejsce spotkanie Rafała Trzaskowskiego, Radosława Sikorskiego i Sławomira Mentzena w pubie, podczas którego spożywali piwo. Sytuacja ta również była głośna w mediach społecznościowych i mogła przyczynić się do przyciszenia sprawy woreczka

nikotynowego.

3.4.Wnioski

Analiza treści przeprowadzona za pomocą wykresów częstotliwości słów oraz chmur słów okazała się skutecznym narzędziem do identyfikacji najważniejszych tematów pojawiających się w debacie publicznej podczas kampanii wyborczej. Dzięki tym metodom możliwe było uchwycenie zarówno ogólnych trendów, jak i reakcji użytkowników platformy X na konkretne wydarzenia polityczne.

Chmury słów pozwalały dostrzec szersze spektrum haseł, które pojawiały się rzadziej, ale były istotne z perspektywy kontekstu politycznego. Z kolei wykresy częstotliwości słów umożliwiały identyfikację najczęściej powtarzających się terminów i ich zmian w czasie. Użycie obu metod w zestawieniu pozwoliło uchwycić jakie tematy dominowały w debacie publicznej oraz jakie wątki zyskiwały lub traciły na znaczeniu.

Na podstawie analizy można stwierdzić, że takie narzędzia jak chmura słów i wykres częstotliwości są przydatne do badania tematyki kampanii wyborczej. Umożliwiają szybkie zorientowanie się w dominujących wątkach, pozwalając wykrywać zmiany w nastrojach społecznych oraz identyfikować wpływ poszczególnych wydarzeń na sposób, w jaki kandydat jest postrzegany w mediach społecznościowych.

4.Analiza emocji

4.1.Preprocessing

W celu analizy emocji we wpisach zastosowano część preprocessingu z części 3. Z tweetów usuwane zostały linki, symbole wspomnień (@) oraz symbole hashtagów (#). W odróżnieniu od preprocessingu dla analizy treści wypowiedzi nie były usuwane stop words.

4.2.Analiza modelem dkleczek/bert-base-polish-cased-v1

Jest to polska wersja modelu BERT - modelu opracowanego przez Google, który ma na celu lepsze zrozumienie kontekstu w zapytaniach i tekście. Model dostępny jest poprzez HuggingFace (<https://huggingface.co/dkleczek/bert-base-polish-cased-v1>).

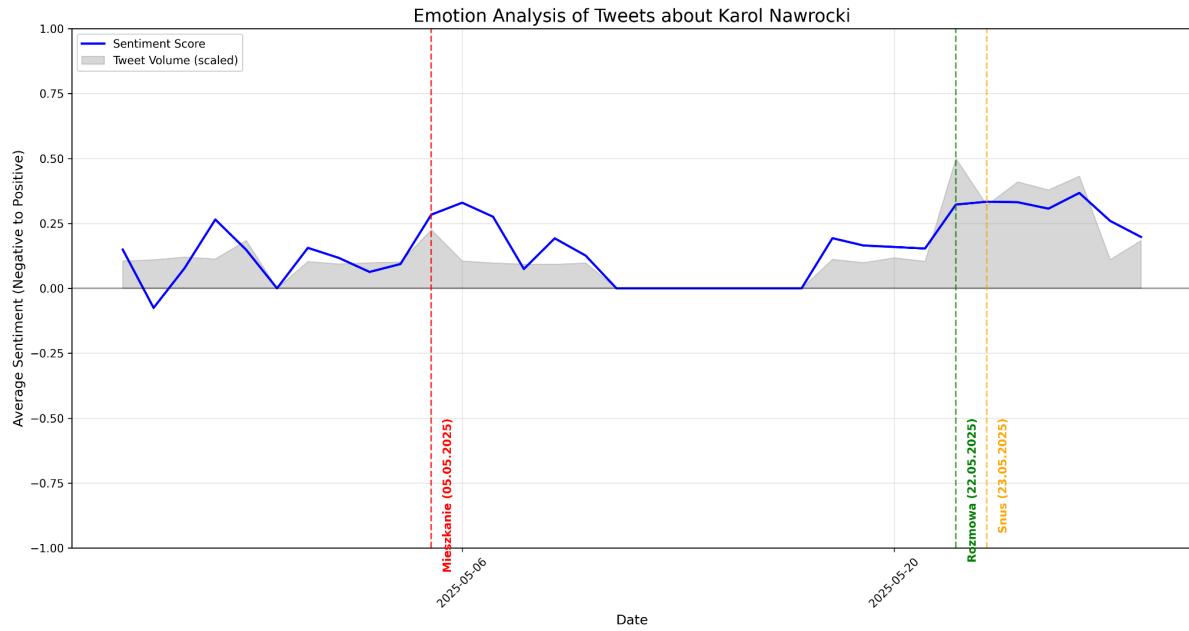
Wykorzystana została rekomendowana przez autora wersja cased modelu, który poprawnie tokenizuje polskie znaki. Poniżej przedstawiono konfigurację modelu. Wczytywany jest tokenizer oraz model a na samym końcu tworzony jest pipeline do analizy sentymentu.

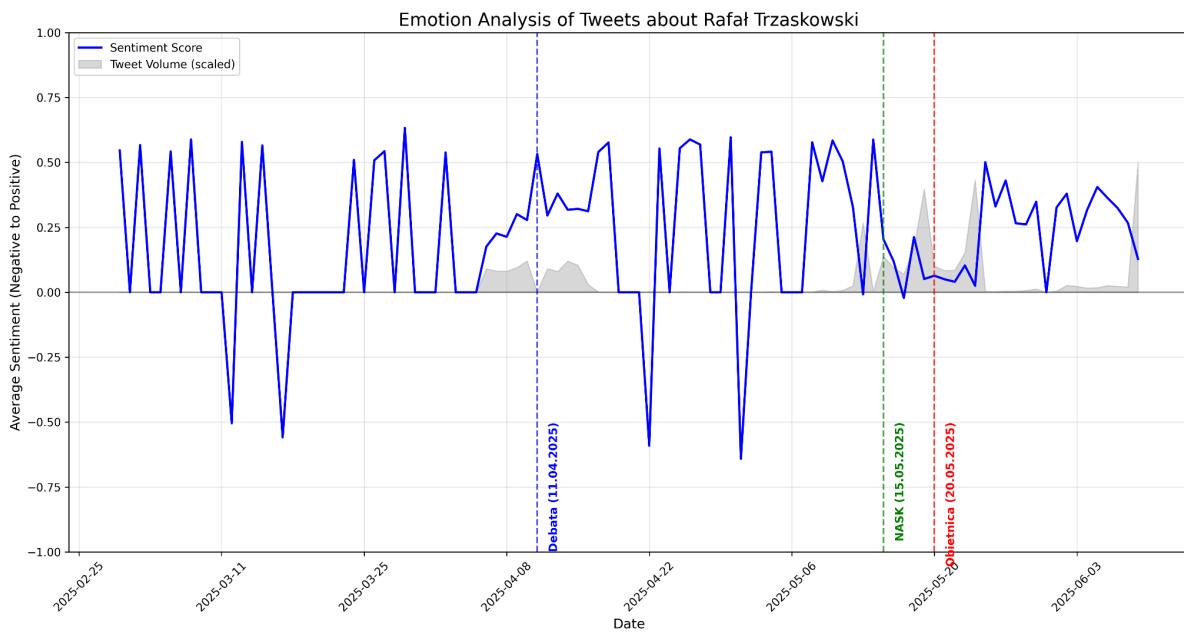
```
model_name = "dkleczek/bert-base-polish-cased-v1"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model =
AutoModelForSequenceClassification.from_pretrained(model_name)
sentiment_analyzer = pipeline("sentiment-analysis", model=model,
tokenizer=tokenizer)
```

Model ten został wykorzystany 3 razy w celu analizy emocji we wpisach na temat obu kandydatów.

4.2.1. Pierwsza analiza modelem bert.

Pierwsza analiza została przeprowadzona używając domyślnej konfiguracji. Brakujące wartości emocji w tweetach wypełniane były 0 (emocja neutralna). W tej analizie nie ustawiano seed'u w celu uzyskania powtarzalności analizy.

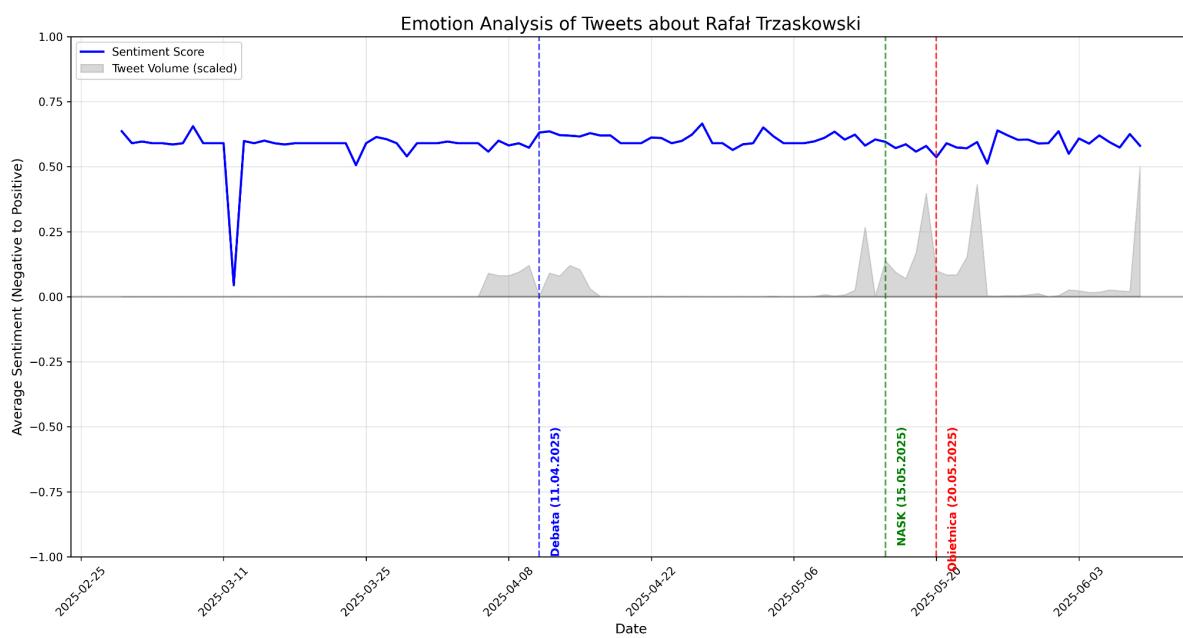
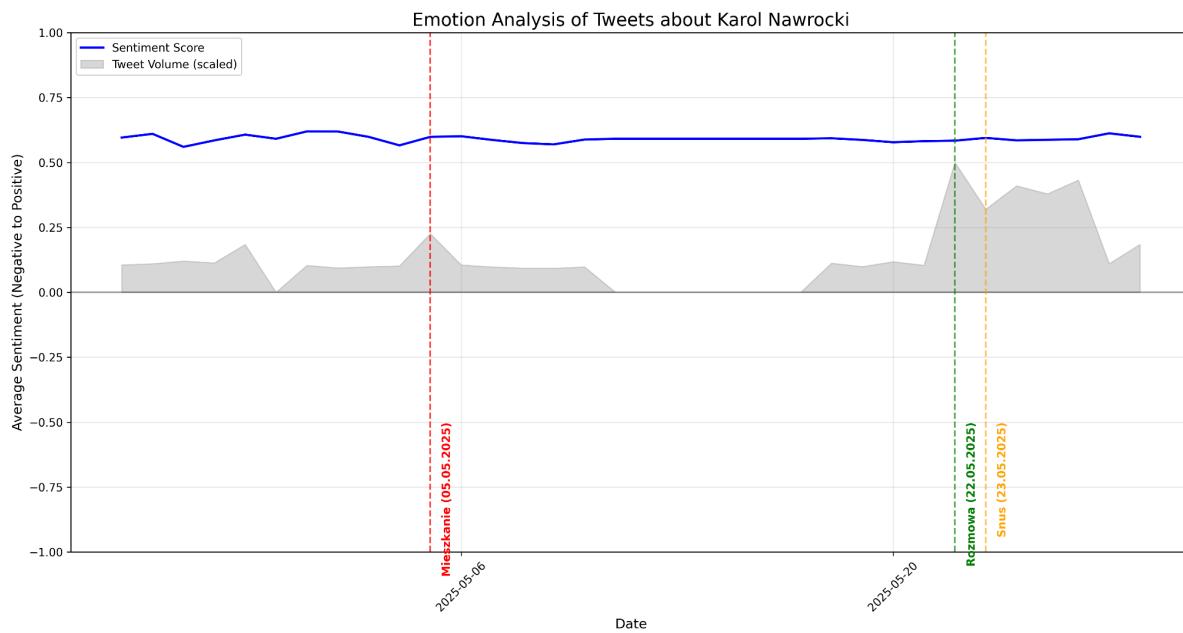




Powyższe wykresy przedstawiają zmianę emocji (sentymentu) w czasie. Na wykresach widać, że kluczowe momenty kampanii nie miały wpływu na wynik emocji we wpisach. W przypadku Karola Nawrockiego linia sentymenu pokrywa się z linią zebranych tweetów (przeskalowanych do wartości 0-1). Świadczy to o tym, że model oceniał wpisy o Karolu Nawrockim jako umiarkowanie pozytywne - sentymenit nie przekraczający 0.4. W przypadku Rafała Trzaskowskiego wykres jest chaotyczny. Skoki sentymenu z poziomu 0.5 do 0 wynikają z tego, że w tych dniach nie zebrano żadnych wpisów na temat Rafała Trzaskowskiego i domyślnie stosowano 0 jako neutralną wartość sentymenu. Wpisy na temat Rafała Trzaskowskiego pochodzą z dłuższego okresu - co sprawia, że jest więcej dni gdzie nie zebrano żadnych wpisów. W odróżnieniu od Karola Nawrockiego, wpisy na temat Trzaskowskiego częściej były oceniane jako bardziej pozytywne (sentymenit w okolicach 0.5). W przypadku Trzaskowskiego były też tweety, które zostały ocenione jako negatywne (sentymenit w okolicach -0.5)

4.2.2. Druga analiza modelem bert

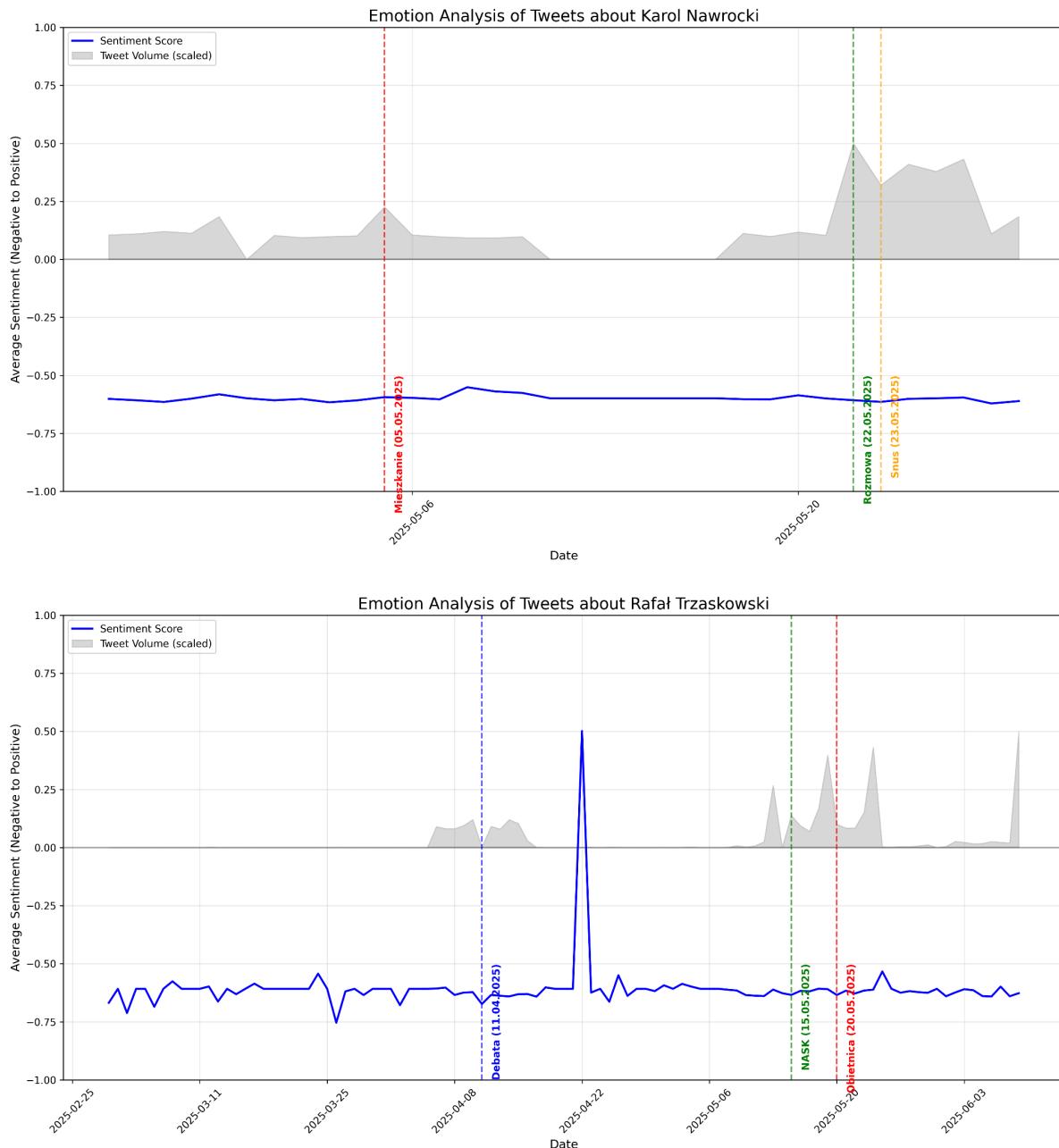
W drugiej analizie modelem bert brakujące wartości był wypełnianie 0 w celu zmniejszenia "skoków" linii sentymenu na wykresie.



Na powyższych wykresach widać, że linia sentymentu dla obu kandydatów utrzymuje się na stałym poziomie. Dla Karola Nawrockiego odchylenia są minimalne i nie ma ich wiele. W przypadku Rafała Trzaskowskiego odchylenia są większe i występują częściej z jednym szczególnym przypadkiem, gdzie wartość sentymentu spadła prawie do 0. Ponownie nie widać na wykresach wpływu kluczowych momentów kampanii wyborczej na emocje we wpisach. Pomimo tego, że jedynie zmieniła się wartość, którą zostały wypełnione puste dni, to wpisy o Karolu Nawrockim był oceniane bardziej pozytywnie (wartość sentymentu ponad 0.5). Może to świadczyć o tym, że model nie jest dobry do analizowania emocji/sentymentu w tekście.

4.2.3. Trzecia analiza modelem bert

W tej analizie dodano seed do wartości losowych w celu uzyskania odtwarzalności analiz.



Na powyższych wykresach widać, że wpisy były oceniane negatywnie dla obu kandydatów. Podobnie jak w analizie 2, wpisy na temat Trzaskowskiego mają więcej zmian w wartościach sentymenu. W przypadku Nawrockiego wartości sentymenu są jednolite. Podobnie jak w poprzednich analizach, nie widać wpływu kluczowych momentów kampanii na emocje we wpisach.

4.2.4. Wnioski

Na podstawie przeprowadzonych analiz, widać że kluczowe momenty kampanii wyborczej nie miały wpływu na wartości sentymentu we wpisach dotyczących obu kandydatów. Między poszczególnymi analizami wprowadzono zmiany, które nie powinny mieć żadnego wpływu na wartości sentymentu. Do zmian tych należało: ustawienie średniej sentymentu jako wartości domyślnej, w przypadku braku wpisów w danym dniu czy też ustawienie seed'u. Pomimo tych zmian, wartości sentymentu znacząco różniły się od siebie. Świadczy to o tym, że model [dkleczek/bert-base-polish-cased-v1](#) nie jest dobry do analizowania sentymentu w tekście.

4.3. Analiza modelem eevvgg/PaReS-sentimenTw-political-PL

Wpisy poddano analizie modelem eevvgg/PaReS-sentimenTw-political-PL, który również można znaleźć na HuggingFace (<https://huggingface.co/eevvgg/bert-polish-sentiment-politics>). Jest to model, który został rozbudowany bazując na modelu [dkleczek/bert-base-polish-cased-v1](#) (model z poprzedniej analizy). Jak twierdzą autorki, jest to model, który został wytrenowany na danych pochodzących z Twittera i jest dostosowany do wpisów w języku polskim na temat polityki.

Przed użyciem modelu do analizy tweetów, został on przetestowany na 10 przykładowych wypowiedziach dotyczących polityki. Dla każdej wypowiedzi podano etykietę (label) którą przewidział model a także wynik (score).

Text	Label	Score
W końcu ktoś zabrał się za realną reformę sądownictwa. Brawo dla rządu za odwagę!	POSITIVE	0.999419093132019
Dzięki programom socjalnym wielu rodzinom żyje się po prostu lepiej. To jest realna pomoc, a nie tylko obietnice.	POSITIVE	0.9992684721946716
Po raz pierwszy od lat czuję, że Polska ma konkretną strategię energetyczną. Inwestycje w atom i OZE idą w dobrym kierunku.	POSITIVE	0.9990071654319763

Duży plus za walkę z wykluczeniem komunikacyjnym. Pociągi wracają do małych miejscowości. Tak trzymać!	POSITIVE	0.9993748068809509
Fajnie widzieć, że Polska potrafi prowadzić niezależną politykę zagraniczną i stawiać własne interesy na pierwszym miejscu.	POSITIVE	0.9941157102584839
Obiecali transparentność, a mamy jeszcze większy chaos i układy niż wcześniej. Zero zaufania.	NEGATIVE	-0.9998077750205994
Kolejna afera i żadnych konsekwencji. Czy ktokolwiek jeszcze wierzy w uczciwość tej władzy?	NEGATIVE	-0.9998970031738281
Młodzi wyjeżdżają, bo nie widzą tu przyszłości. Gdzie są reformy, które miały zatrzymać emigrację?	NEGATIVE	-0.9998941421508789
Rolnicy protestują, a rząd udaje, że wszystko jest OK. Ignoracja wobec wsi to katastrofa.	NEGATIVE	-0.9998974800109863
Politycy przejmują media publiczne jak swoją własność. To już nie informacja, tylko propaganda.	NEGATIVE	-0.999903678894043

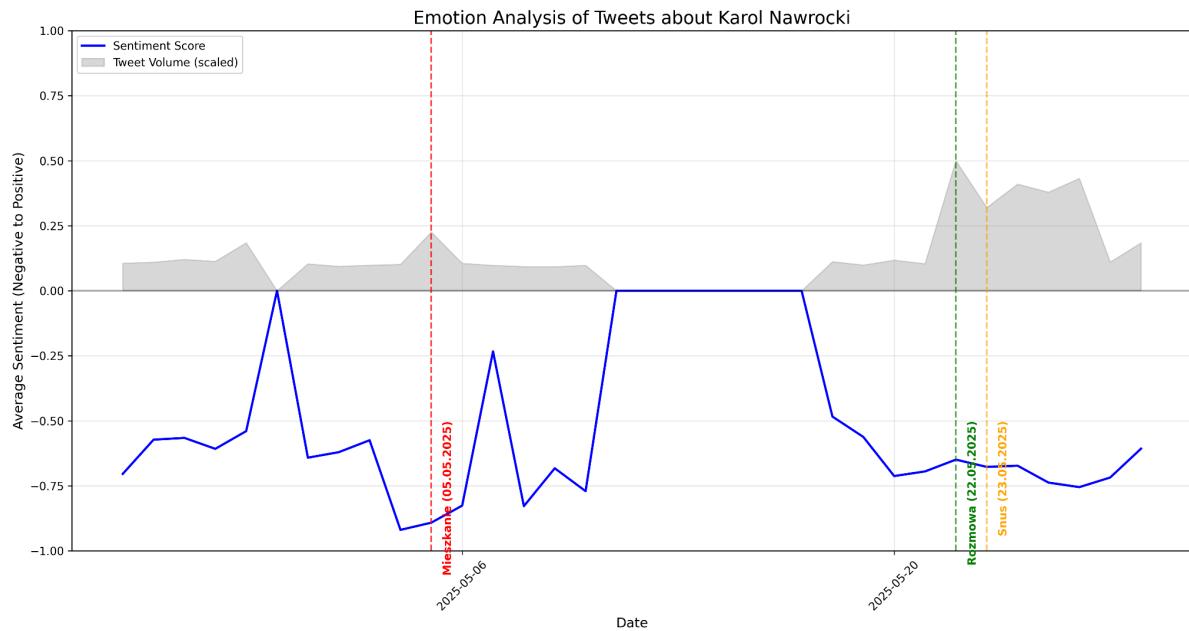
Model poprawnie przewidział wszystkie 10 wypowiedzi. W odróżnieniu od poprzedniego modelu, wartości score są skrajne. Dla każdej pozytywnej próbki score jest większy niż 0.99. Z kolei dla każdej próbki negatywnej wartości score są mniejsze niż -0.99. Oznacza to, że nie można bezpośrednio używać wartości score do określenia emocji we wpisach. W analizach z użyciem tego modelu, wykorzystano różne sposoby do określenia emocji/sentymentu w tweetach.

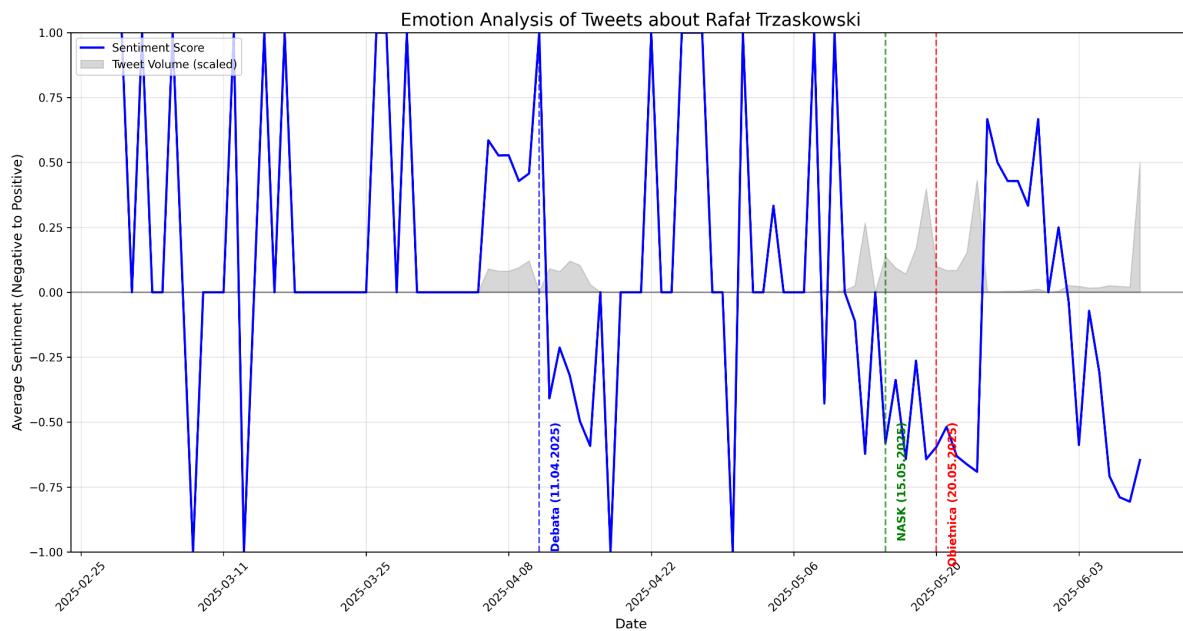
4.3.1.Pierwsza analiza

W tej analizie w celu wyznaczenia wartości sentymenu dla każdego dnia wykorzystany został następujący wzór:

$$\frac{\sum \text{numeric-score}}{\text{sentimen-count}}$$

gdzie *numeric-score* to wartości: -1 dla opinii negatywnych, 1 dla opinii pozytywnych *sentiment-count* to liczba wpisów które są pozytywne lub negatywne - brak wpisów neutralnych.



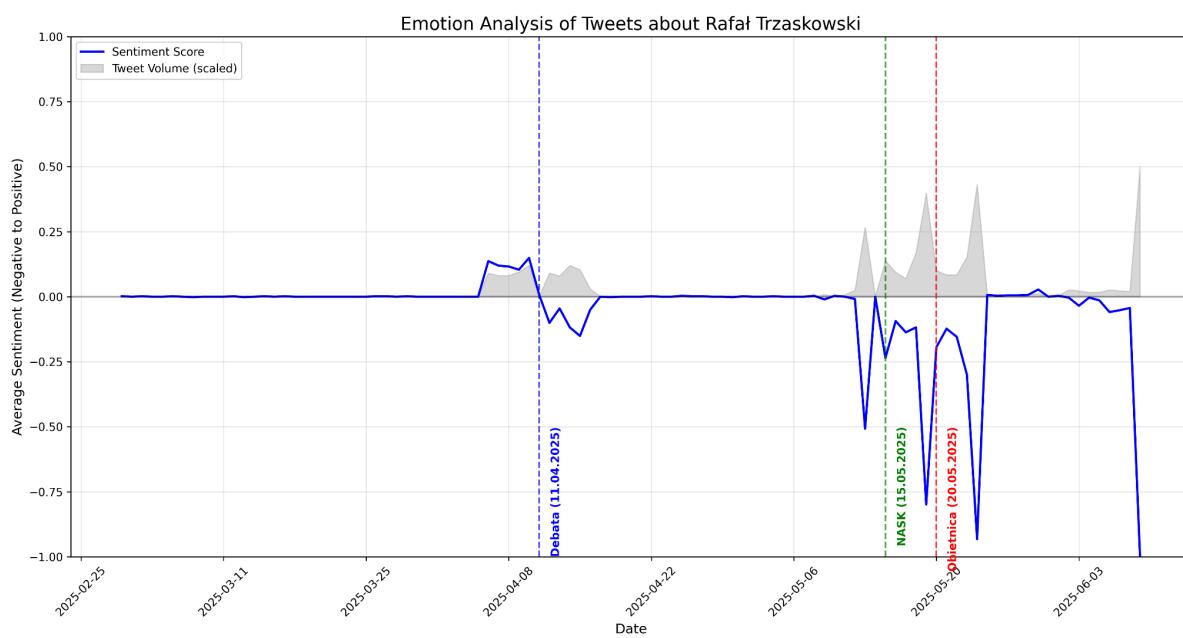
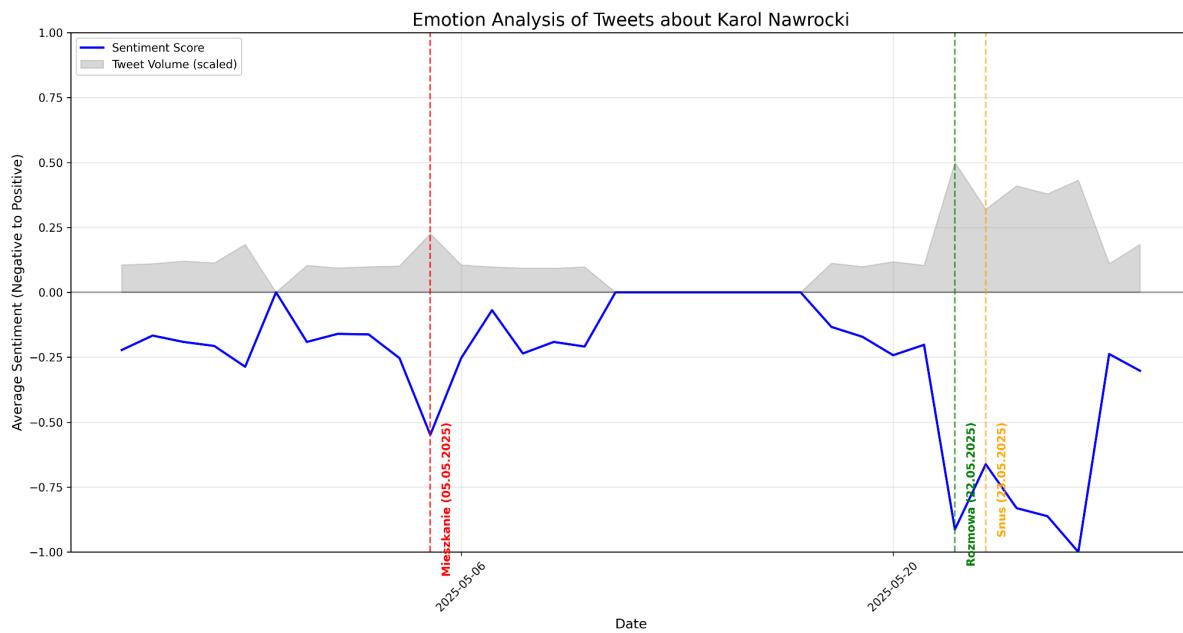


Na poniższych wykresach widać, że w przypadku Karola Nawrockiego, wpisy były oceniane jako negatywne. Nie widać większego wpływu kluczowych momentów na linię sentymentu. W przypadku Rafała Trzaskowskiego po debacie w TVP widać, że nastąpił duży spadek sentymentu i utrzymywał się on przez kilka dni. W przypadku wydarzeń z maja (NASK oraz obietnica) również można zauważać mały spadek. W przypadku Rafała Trzaskowskiego było wiele dni, w których zebrano pojedyncze tweety co skutkuje tym, że jest wiele dni ocenianych w sposób skrajnie negatywny (-1) i pozytywny (1).

4.3.2. Druga analiza

W tej analizie do wyznaczenia linii sentymentu, wykorzystana została suma *numeric-score* (znaczenie to samo co we wcześniejszej analizie). Dodatkowo wykonana została normalizacja (Max-abs normalization)

$$x' = \frac{x}{\max(\text{abs}(x))}$$

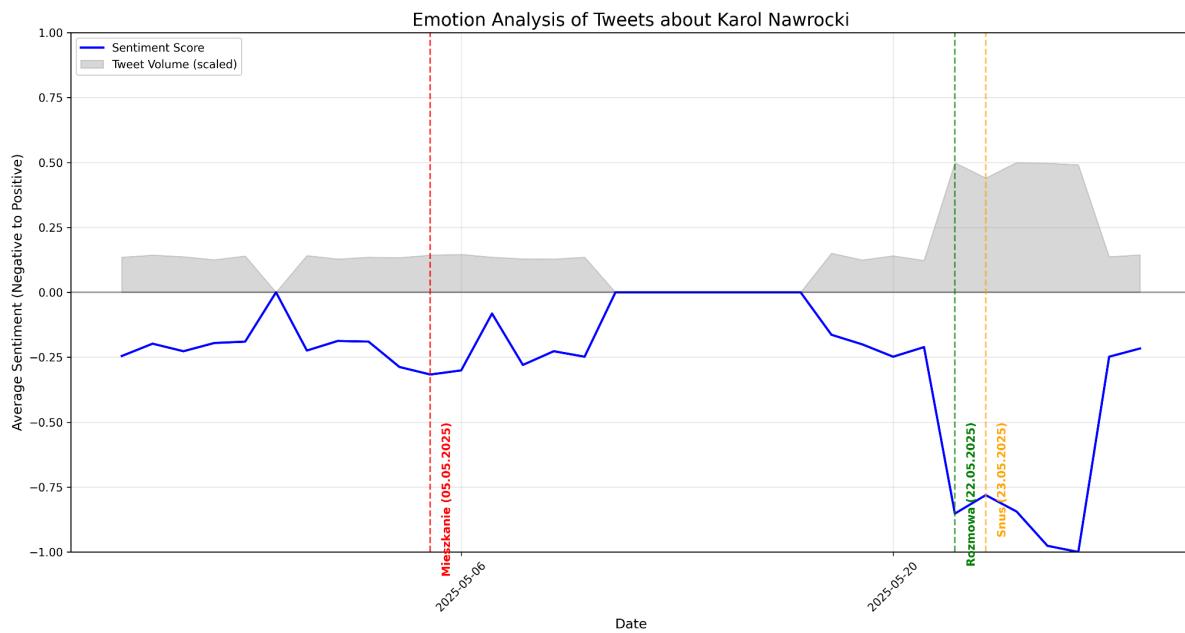


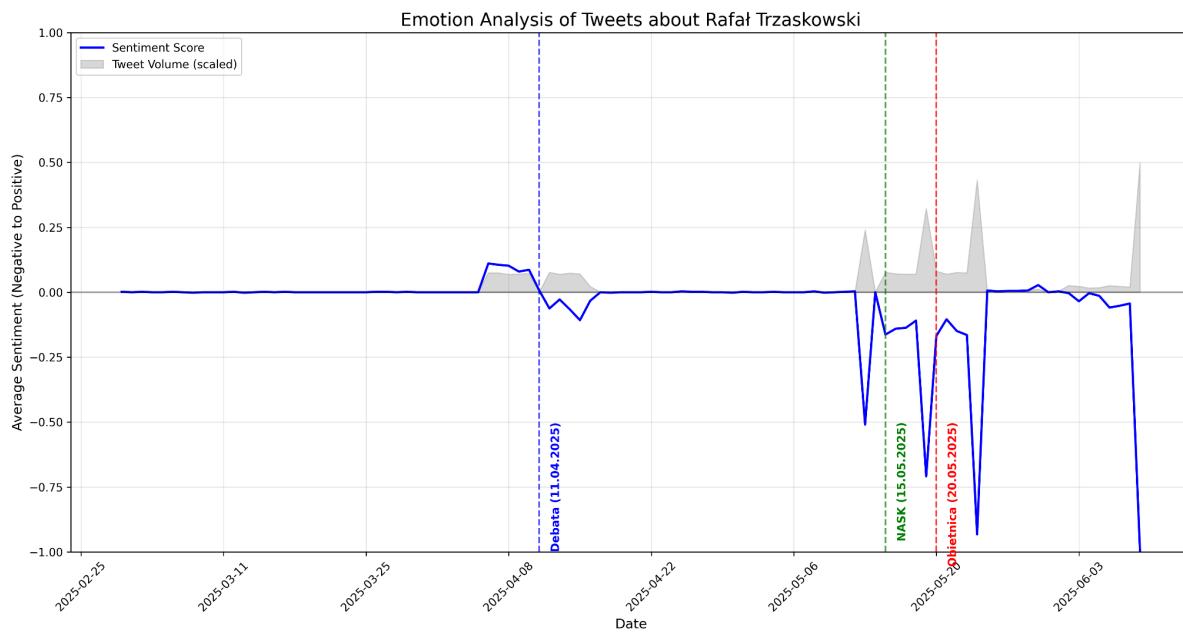
W przypadku Karola Nawrockiego i momentu opublikowania artykułu na temat jego drugiego mieszkania przez Onet emocje we wpisach miały trend wzrostowy - co jest sprzeczne z oczekiwaniami. Emocje we wpisach drastycznie spadły 21.05.2025 - przed rozmową ze Sławomirem Mentzenem. Może być to spowodowane tym, że 20.05.2025 Karol Nawrocki zgodził się podpisać postulaty Mentzena - co przez wielu internautów zostało odebrane negatywnie. Widać także, że po tym jak Nawrocki zażył snusa na debacie, sentyment zaczął spadać aż do skrajnie negatywnej wartości (-1).

W przypadku Rafała Trzaskowskiego i debaty w TVP sentyment “odwrócił się”. Z wartości pozytywnych oscylujących w granicach 0.15 spadł do wartości około -0.15. W przypadku kolejnych kluczowych momentów dla Rafała Trzaskowskiego emocje gwałtownie spadały na dzień a następnie rosły do momentu sprzed spadku. Model wykrył też spadek emocji do skrajnie negatywnych po 2 turze wyborów. Może być to spowodowane tym, że niektóre osoby związane ze środowiskiem Trzaskowskiego zaczęły popularyzować teorię o tym, że wybory były sfałszowane. Przez większość opinii publicznej ta teoria jest uznawana za nieprawdziwą co spowodowało duży wzrost negatywnych emocji.

4.3.3. Trzecia analiza

W tej analizie wykluczono tweety z kategorii Top. Są to głównie wpisy popularnych osób tj. polityków, dziennikarzy itd. które niekoniecznie dobrze odwzorowują emocje zwykłych ludzi. Wykluczenie tych wpisów sprawiło też, że zmalała liczba dni, w których zebrano mało wpisów - na wykresie powinno to zmniejszyć liczbę skoków emocji w poszczególnych dniach. Podobnie jak we wcześniejszej analizie zastosowano normalizację Max-abs.





W przypadku Karola Nawrockiego widać jedną zmianę - tweety z okresu opublikowania sprawy mieszkania były oceniane mniej negatywnie. Nie ma tutaj spadku emocji do poziomu -0.5 jak w przypadku analizy drugiej. W przypadku Rafała Trzaskowskiego nie ma większych zmian w linii sentymantu.

4.3.4. Wnioski

W przypadku modelu eevvgg/PaReS-sentimenTw-political-PL oceny sentymantu we wpisach były dokładniejsze i odwzorowywały spadek w kluczowych momentach dla obu kandydatów. Szczególnie dobrze widać to w przypadku debaty TVP 11.04.2025. Ponadto, po zastosowaniu normalizacji oraz usunięciu tweetów z kategorii Top udało się usunąć poszczególne skoki sentymantu w dniach, w których zebrano niewielką ilość wpisów. Wyniki w analizie drugiej oraz trzeciej były zbliżone do siebie, co pokazuje, że predykcje modelu nie są chaotyczne.

4.4.Analiza modelem twitter-xlm-roberta-base-sentiment-finetunned

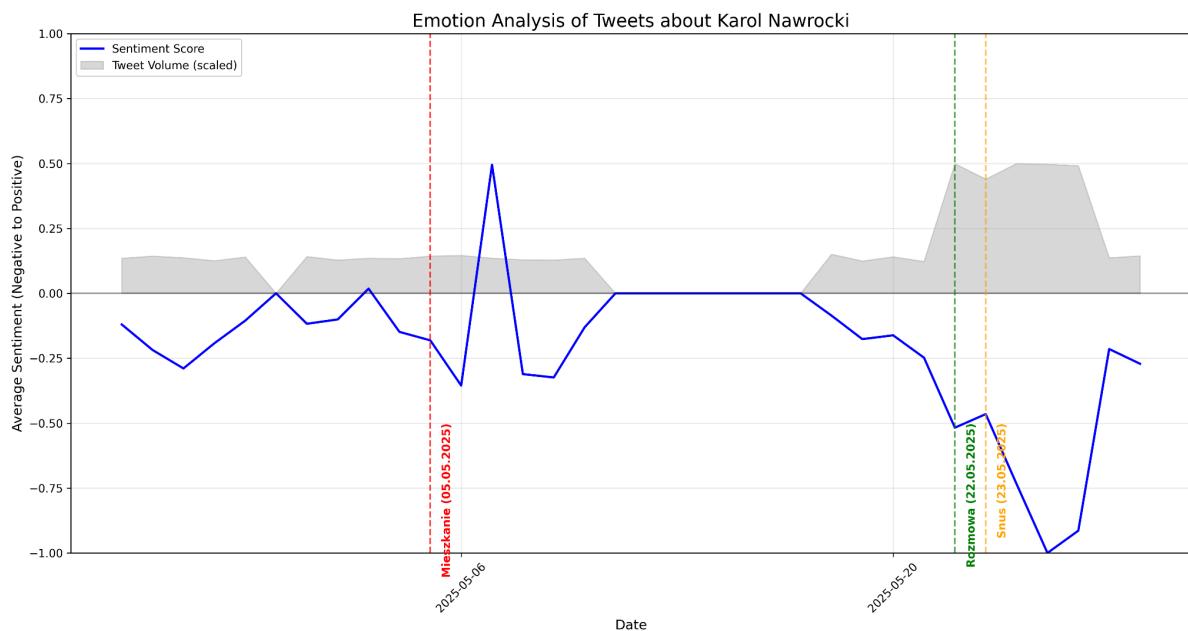
Jest to model wielojęzykowy, oparty na modelu XLM-Roberta. Posiada on wsparcie dla języka polskiego. Dostępny jest on poprzez HuggingFace (<https://huggingface.co/citizenlab/twitter-xlm-roberta-base-sentiment-finetunned>).

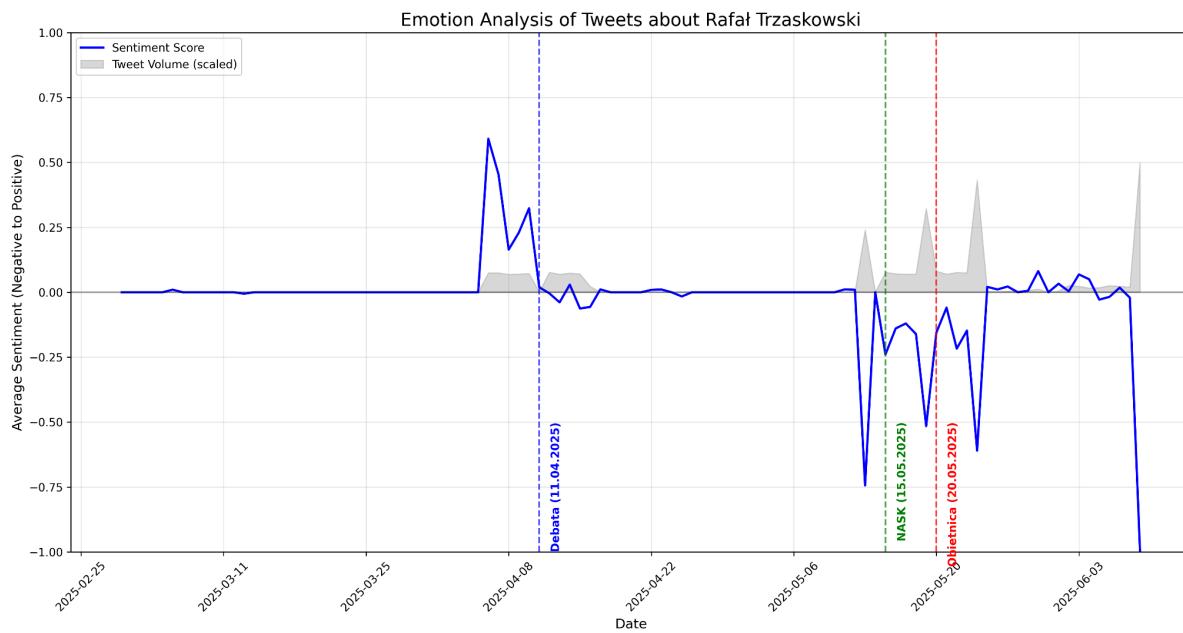
Model ten posiada wsparcie dla analizy sentymentu w tekście. Przed analizą wpisów, model przetestowany został na tych samych wpisach co model z analizy 4.3

Text	Label	Score
W końcu ktoś zabrał się za realną reformę sądownictwa. Brawo dla rządu za odwagę!	POSITIVE	0.9796779155731201
Dzięki programom socjalnym wielu rodzinom żyje się po prostu lepiej. To jest realna pomoc, a nie tylko obietnice.	NEUTRAL	0.9756248593330383
Po raz pierwszy od lat czuję, że Polska ma konkretną strategię energetyczną. Inwestycje w atom i OZE idą w dobrym kierunku.	POSITIVE	0.8240476250648499
Duży plus za walkę z wykluczeniem komunikacyjnym. Pociągi wracają do małych miejscowości. Tak trzymać!	POSITIVE	0.9256331920623779
Fajnie widzieć, że Polska potrafi prowadzić niezależną politykę zagraniczną i stawiać własne interesy na pierwszym miejscu.	POSITIVE	0.9554429054260254
Obiecali transparentność, a mamy jeszcze większy chaos i układy niż wcześniej. Zero zaufania.	NEGATIVE	0.9251587986946106
Kolejna afera i żadnych konsekwencji. Czy ktokolwiek jeszcze wierzy w uczciwość tej	NEGATIVE	0.9095222353935242

władzy?		
Młodzi wyjeżdżają, bo nie widzą tu przyszłości. Gdzie są reformy, które miały zatrzymać emigrację?	NEUTRAL	0.6846190690994263
Rolnicy protestują, a rząd udaje, że wszystko jest OK. Ignorancja wobec wsi to katastrofa.	NEGATIVE	0.9485732316970825
Politycy przejmują media publiczne jak swoją własność. To już nie informacja, tylko propaganda.	NEUTRAL	0.6700894832611084

Model ten, w odróżnieniu od poprzedniego, częściej zwraca etykiety neutralne. Wyniki score są bardziej rozproszone. Wyniki score są zawsze dodatnie, niezależnie od etykiety. Aby wyliczyć wartości sentymentu sumowane są wartości score etykiet pozytywnych oraz etykiet negatywnych ze zmienionym znakiem (pomnożone przez -1). Całość jest normalizowana z użyciem max-abs.





W przypadku Karola Nawrockiego po sprawie z mieszkaniem sentyment spadł nieznacznie, po czym zaliczył chwilowy wzrost do poziomu 0.5, a następnie powrócił do poziomu ok. -0.25. Po rozmowie z Mentzenem i debacie w drugiej turze sentyment spadł do wartości skrajnie negatywnych (-1).

W przypadku Rafała Trzaskowskiego sentyment po debacie w Końskich spadł z do wartości neutralnych (ok. 0). Po raporcie NASK oraz obietnicy wartości sentymentu zaliczyły spadek do wartości (-0.5) lecz chwile potem powróciły do wartości neutralnych. Dopiero po wynikach w 2 turze, sentyment we wpisach spadł do wartości skrajnie negatywnych.

Model twitter-xlm-roberta-base-sentiment-finetunned umożliwił wykrycie spadków sentymentu we wpisach na temat danego kandydata. W porównaniu do poprzedniego modelu, stosunek wpisów pozytywnych do negatywnych jest większy. Można to zauważyć na wykresach, gdzie wartości sentymentu w niektórych momentach sięgają 0.5, co nie miało miejsca podczas analiz poprzednim modelem.

4.5.Analiza za pomocą text2emotion

Dla każdego z kandydatów wybrany został jeden kluczowy moment, w którym odnotowano spadek sentymetu we wpisach na jego temat. Dla Karola Nawrockiego była to debata w drugiej turze (zażycie snusa) a dla Rafała Trzaskowskiego debata w TVP. Wpisy z tego okresu z kategorii Latest zostały przeanalizowane za pomocą text2emotion.

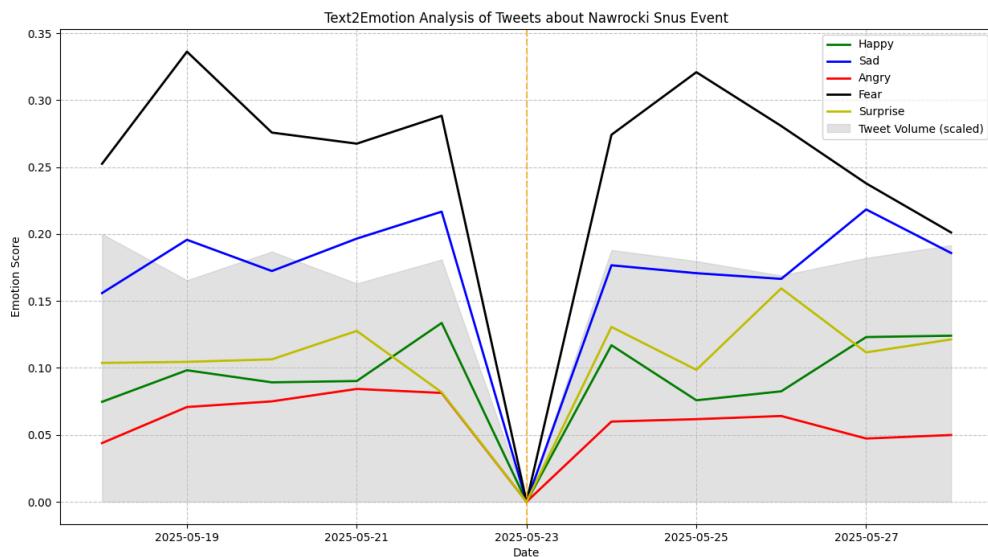
4.5.1.Preprocessing

Paczka ta wymaga tekstu w języku angielskim aby działała dobrze, więc wpisy zostały wcześniej przetłumaczone za pomocą modelu Helsinki-NLP/opus-mt-pl-en, który dostępny jest poprzez HuggingFace (<https://huggingface.co/Helsinki-NLP/opus-mt-pl-en>). Model ten jest przystosowany do tłumaczenia tekstu z języka polskiego na język angielski.

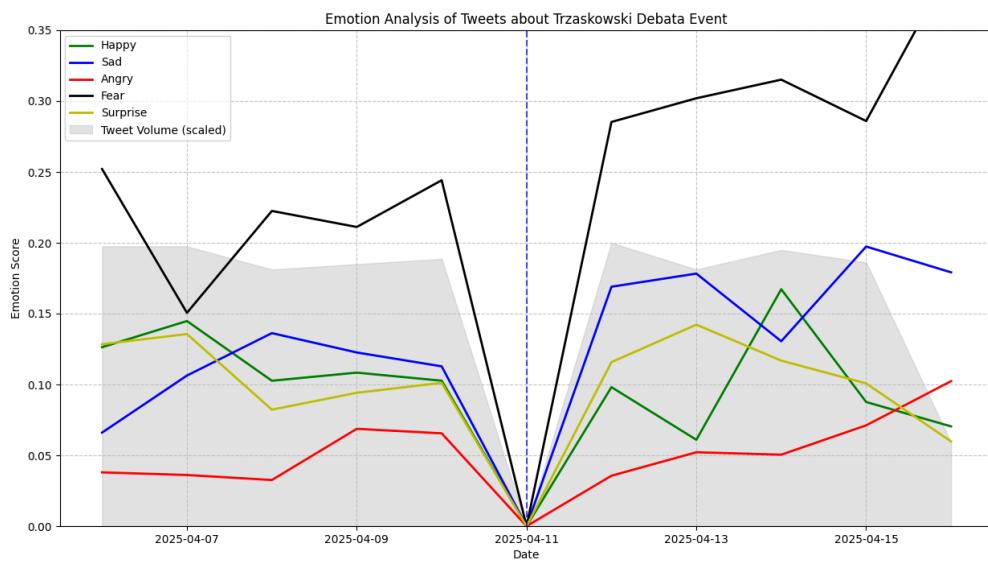
Tweety zostały pozbawione symboli oznaczeń (@), hashtagów (#) oraz linków. Są to rzeczy, które nie mają wpływu na treść wpisu a mogłyby jedynie pogorszyć proces tłumaczenia.

Paczka text2emotion dla każdego tekstu zwraca 5 emocji, są to: happy (szczęście), angry (złość), surprise (zaskoczenie), sad (smutek), fear (strach). Wyniki tych emocji przedstawiono na wykresie.

4.5.2.Wyniki



Na powyższym wykresie przedstawiono wyniki emocji dla Karola Nawrockiego. W dniu debaty (23.05.2025) nie zebrano wpisów więc wyniki wszystkich emocji są równe 0. Wszystkie emocje utrzymują się na podobnym poziomie, przed i po debacie. Nie ma znaczących skoków/spadków jakiekolwiek emocji.



Na powyższym wykresie przedstawiono wyniki emocji dla Rafała Trzaskowskiego. W dniu debaty w Końskich (11.04.2025) nie zebrano wpisów dlatego wartość wszystkich emocji jest równa 0. Większość emocji utrzymuje się na podobnym

poziomie przed i po debacie. Można zauważyć, że jedynie wartość emocji fear i sad (negatywne emocje) wzrosły.

Na obu wykresach widać, że wartości emocji nie zmieniają się znacząco (poza pojedynczymi przypadkami). W przypadku modeli deep-learningowych zmiany sentymentu były zauważalne natomiast w przypadku text2emotion zmiany są nieznaczne (ok. 0.05 - 0.10). Świadczy to o tym, że narzędzie to nie jest skuteczne w wykrywaniu emocji we wpisach na temat polityki. Dodatkowo wpisy musiały zostać przetłumaczone, co znacząco obniżyło skuteczność oceny emocji.

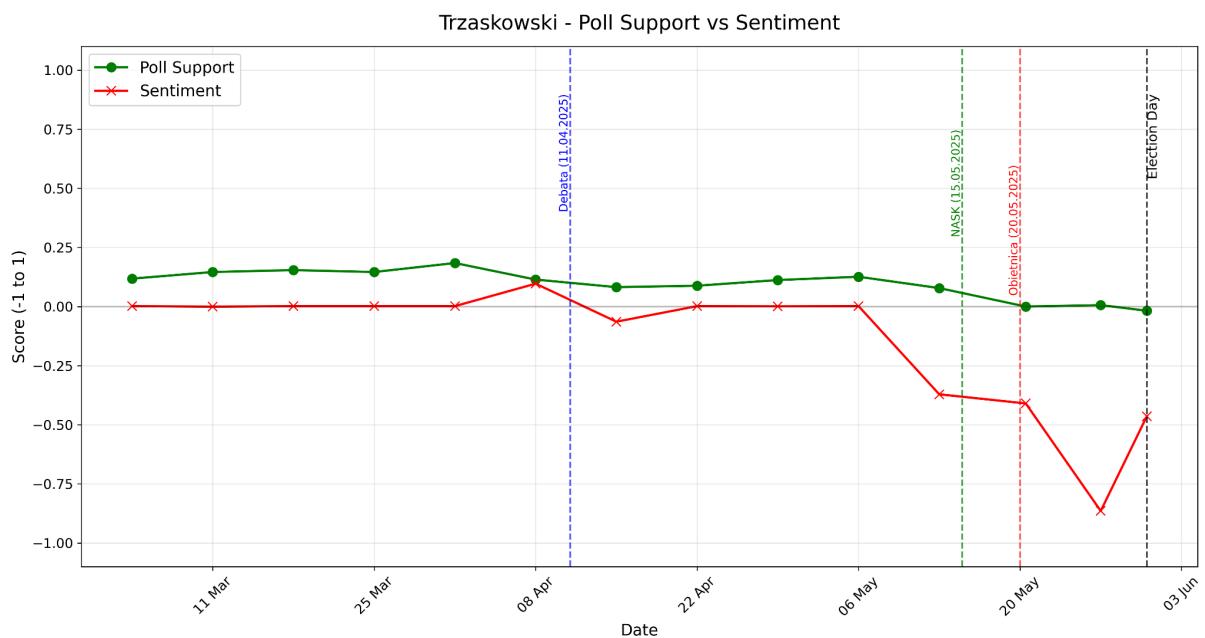
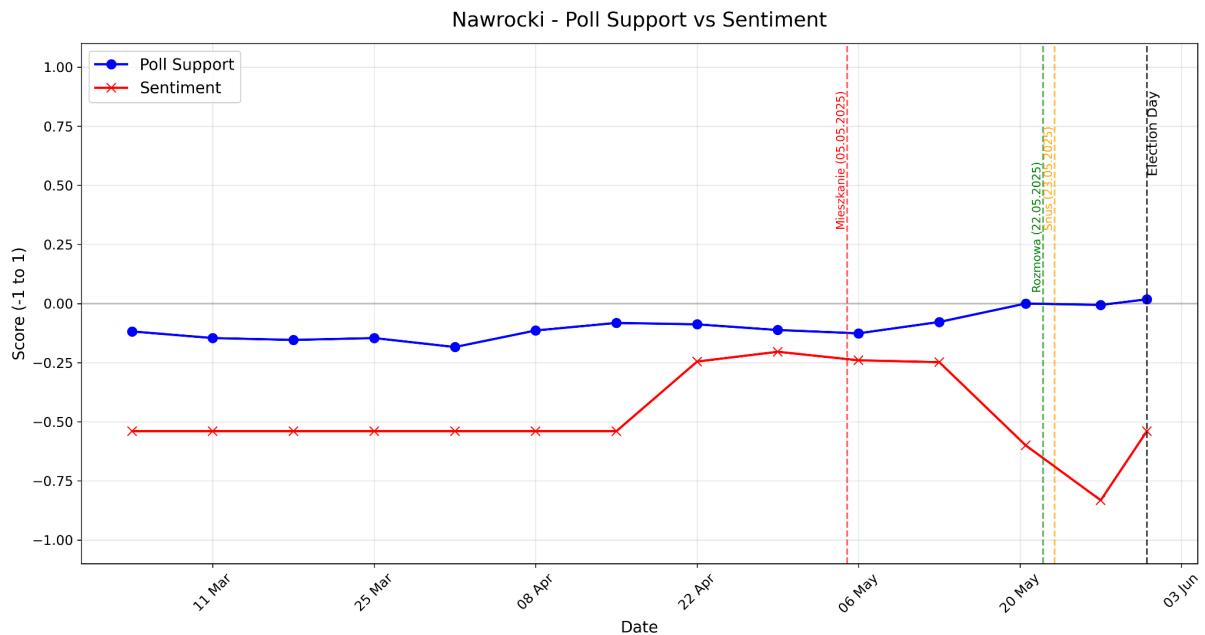
5. Porównanie emocji z sondażami wyborczymi

W tej części wykorzystano dane z sondaży wyborczych (https://en.wikipedia.org/wiki/Opinion_polling_for_the_2025_Polish_presidential_election) przygotowanych przez [ewybory.eu](#). Pobrano dane dotyczące poparcia w drugiej turze dla obu kandydatów z okresu 01.03.2025 - 01.06.2025. Aby poparcie danego kandydata porównać z sentymentem we wpisach na jego temat poparcie zostało przeskalowane do wartości [-1, 1] używając następującego wzoru:

$$p' = \frac{p}{50} - 1$$

gdzie p' to nowa wartość poparcia, p to poparcie.

Wykorzystano dane z analizy modelem eevvgg/PaReS-sentimenTw-political-PL (4.3). W odróżnieniu od tej analizy, wartości sentymentu są tutaj średnią ze wszystkich dni. Nie poddawano ich procesowi normalizacji max-abs. Dodatkowo na wykresach zaznaczono kluczowe momenty kampanii dla każdego kandydata. Dla dni w których nie udało się pobrać żadnych wpisów i ustalić sentymentu zastosowano średnią wartość sentymentu.



W przypadku Nawrockiego w momencie kiedy wzrastał sentyment we wpisach, jego poparcie zanotowało nieznaczny spadek. Z kolei kiedy wartości sentymentu spadły do wartości skrajnie negatywnych, jego poparcie wzrosło. W przypadku Trzaskowskiego oba większe spadki poparcia następowały w tym samym czasie, w którym spadała wartość sentymentu we wpisach na jego temat.

6. Wnioski

1. Tematyka kampanii wyborczej
 - a. Kluczowe wydarzenia w kampani wyborczej miały wpływ na treści pojawiające się w mediach społecznościowych
 - b. Hasła/słowa dotyczące tych wydarzeń bardzo często pojawiały się w analizowanych wpisach i utrzymywały się tam przez kilka dni. Przy takich sprawach jak [m.in.](#) afera z drugim mieszkaniem Karola Nawrockiego, słowa związane z tą sprawą często znajdowały się wśród najczęściej używanych.
 - c. Zdarzało się, że analizując dane wydarzenie, pojawiały się słowa nie związane z danym momentem np. często pojawiało się nazwisko Braun. Świadczy to o szerokiej tematyce tej kampanii wyborczej.
2. Wybór modelu NLP istotnie wpływa na wyniki
 - a. dkleczek/bert-base-polish-cased-v1 okazał się zbyt „chaotyczny” – wrażliwy na dobór seeda i wypełnianie braków, przez co trudno było powiązać sentyment z wydarzeniami.
 - b. PaReS-sentimenTw-political-PL okazał się najlepszym modelem do wychwytywania zmian emocji. Jest to spowodowane tym, że model ten wytrenowany został na wpisach dotyczących polityki. Użycie go umożliwiło wykrycie spadku emocji w większości przypadków
 - c. Wielojęzykowy twitter-xlm-roberta-base-sentiment zapewnił czytelne skoki dodatnie/ujemne, lecz wymagał transformacji wyników. W porównaniu do poprzedniego modelu częściej oceniał on wpisy jako pozytywne.
 - d. Narzędzie text2emotion okazało się nieskuteczne w ocenie emocji - nie wykrywało znaczących zmian emocji we wpisach. Ponadto proces tłumaczenia wpisów mógł przyczynić się do pogorszenia oceny emocji.
3. Emocje a sondaże wyborcze
 - a. Nie wykazano bezpośredniej korelacji między sentymentem we wpisach na temat danego kandydata z jego poparciem
 - b. W przypadku Karola Nawrockiego, jego poparcie zachowywało się w odwrotny sposób niż emocje. Negatywny szum medialny wokół drugiego mieszkania czy też woreczka nikotynowego nie przełożył się na trwał spadek poparcia. Może to świadczyć, że elektorat PiS jest mniej podatny na krytykę w mediach społecznościowych, a na jego ostateczny wynik duży wpływ miało zjednoczenie elektoratu po 1 turze.
 - c. W przypadku Rafała Trzaskowskiego, spadek sentymentu odbywał się w tych samych okresach co spadek poparcia. Może to świadczyć o tym, że elektorat Rafała Trzaskowskiego jest bardziej wrażliwy na krytykę, a sprawy takie jak słowa Przemysława Witka mogły zniechęcić ludzi do głosowania na tego kandydata.