Projekt eksploracji danych

Mateusz Bartos (122437)
Mikołaj Rozwadowski (127205)

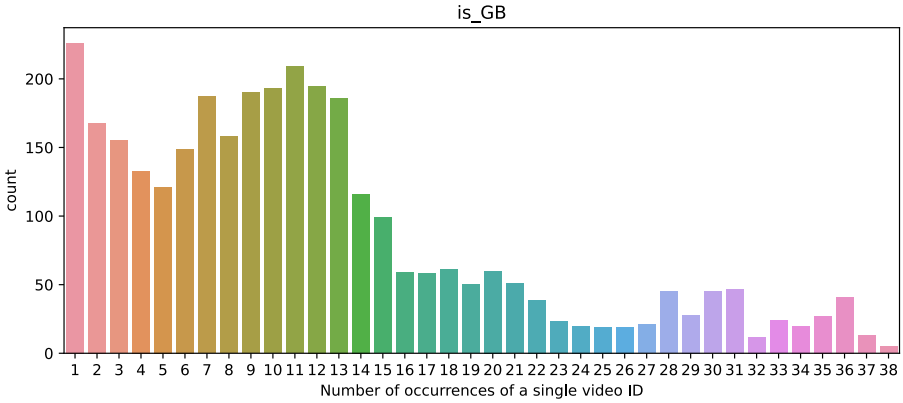
23 marca 2020

0 Analiza eksploracyjna

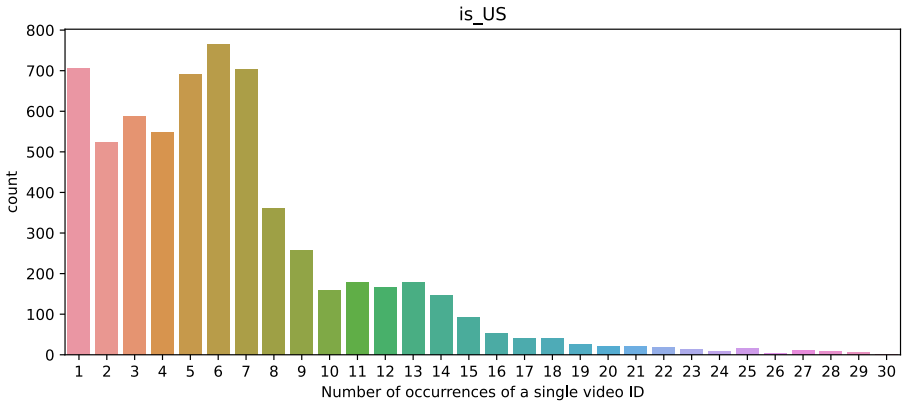
Zbiór danych zawiera 38916 rekordów o filmach w serwisie YouTube, które trafiły do kategorii „Na czasie” (ang. *trending*) w Wielkiej Brytanii i 40949 rekordów ze Stanów Zjednoczonych. Analizę wykonano po połączeniu ze sobą obu zbiorów, przy zachowaniu informacji, którego kraju dotyczy każdy rekord. Dane są opisane szesnastoma atrybutami, które zostaną omówione poniżej.

0.1 Identyfikator filmu (video_id)

W 727 wierszach był niedostępny i zastąpiony kodem `#NAZWA?`. Brakujące wartości można było jednak łatwo odtworzyć, korzystając z adresu URL miniatury. Unikalnych filmów pojawiło się 8607, a powtórzone wystąpienia interpretujemy jako utrzymanie się w sekcji Trending przez wiele dni.



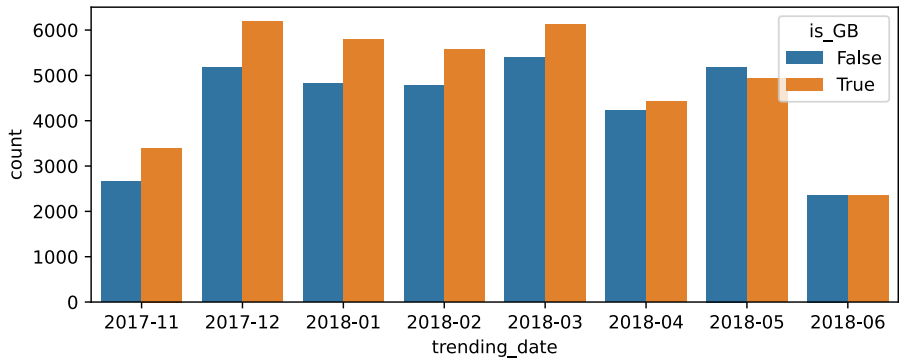
Rysunek 1: Rozkład liczby wystąpień pojedynczego filmu w Wielkiej Brytanii



Rysunek 2: Rozkład liczby wystąpień pojedynczego filmu w Stanach Zjednoczonych

0.2 Data obecności w sekcji Trending (trending_date)

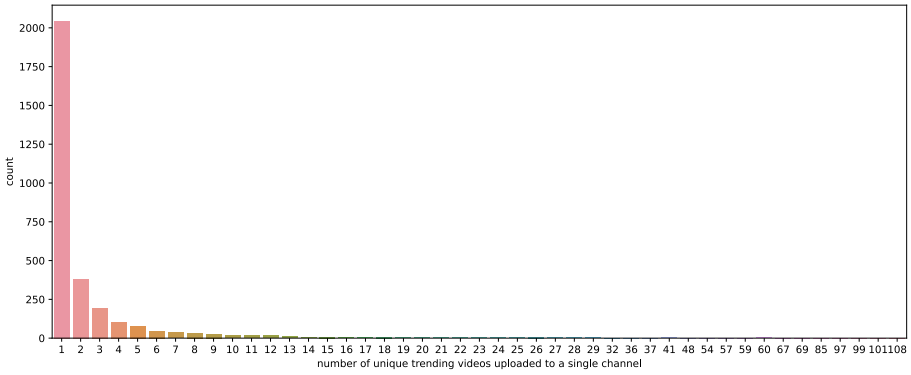
Wyrażona jest w amerykańskim formacie RR.DD.MM i obejmuje okres od połowy listopada 2017 do połowy czerwca 2018, czyli łącznie 7 miesięcy.



Rysunek 3: Rozkład liczby rekordów w poszczególnych miesiącach (niebieski – USA, pomarańczowy – Wielka Brytania)

0.3 Kanał (channel_title)

Dane obejmują 3099 różnych kanałów, z których ok. $\frac{2}{3}$ udało się trafić do Trending z tylko jednym filmem. Z drugiej strony najpopularniejsze kanały (TheEllenShow i The Tonight Show Starring Jimmy Fallon) opublikowały ponad 100 takich nagrań.



Rysunek 4: Rozkład liczby filmów na czasie opublikowanych przez jeden kanał

0.4 Kategorie (category_id)

Większość (87%) rekordów nie zawiera informacji o kategorii. Pozostałe (rys. 5) to głównie filmy rozrywkowe i muzyczne.

0.5 Czas publikacji (publish_time)

Po sprawdzeniu rozkładu czasu publikacji w ciągu doby, okazuje się, że większość filmów została opublikowana w godzinach popołudniowych (rys. 6).

0.6 Filmy błędne lub usunięte (video_error_or_removed)

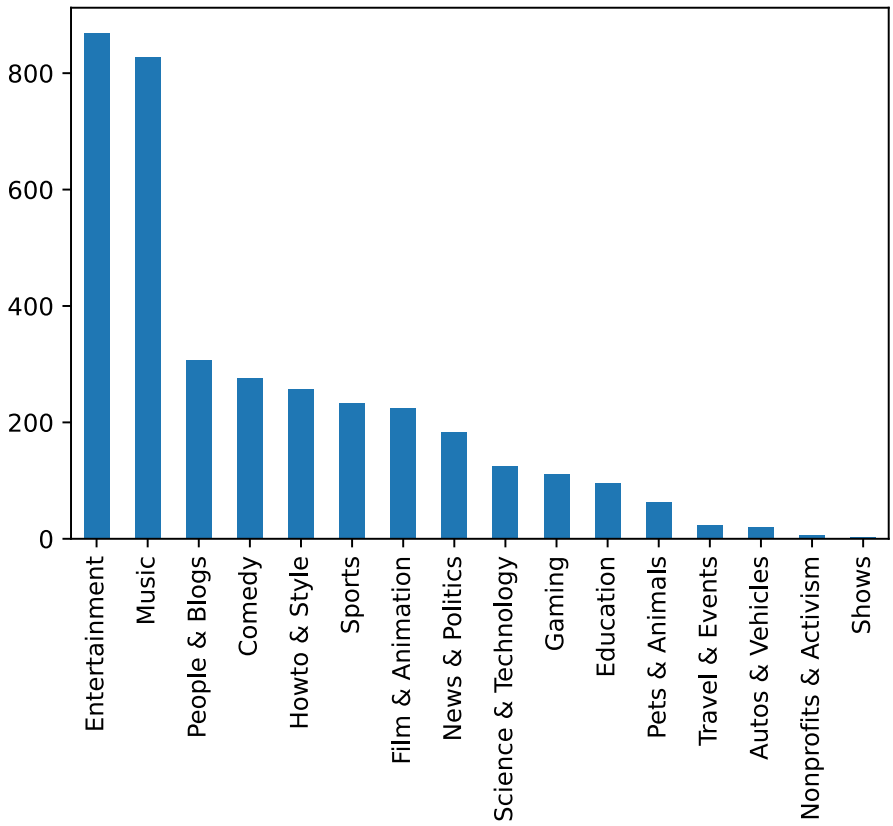
Najbardziej zagadkowy atrybut, który przyjmuje wartość `True` jedynie dla ułamka filmów (81 wystąpień, $< 1\%$). Nasza interpretacja zakłada, że atrybut ten oznacza nagranie, które po trafieniu do listy „Trending" zostało usunięte przez autora lub moderatora.

Jeśli konieczne jest stwierdzenie, że któryś z przedstawionych atrybutów nie był pomocny do analizy to powinien zostać wskazany atrybut **Filmy błędne lub usunięte (video_error_or_removed)**.

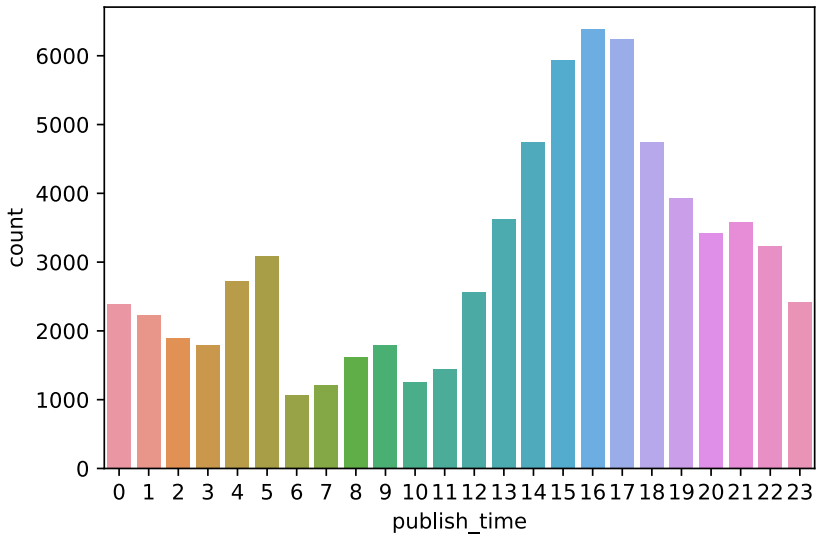
0.7 Wyłączenie komentarzy i/lub ocen

Niektóre filmy korzystają z funkcji wyłączenia możliwości wyłączenia komentarzy i ocen. Spośród `73369` nagrań:

- `1212 (1,6%)` ma wyłączoną możliwość komentowania
- `403 (0,5%)` ma wyłączoną możliwość oceny
- `297 (0,4%)` ma wyłączone obie funkcjonalności



Rysunek 5: Rozkład kategorii (z wyłączeniem brakujących wartości)

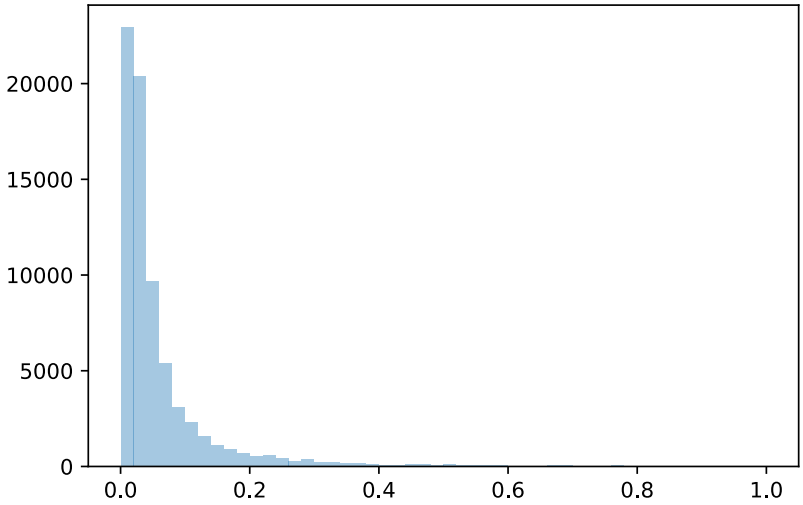


Rysunek 6: Rozkład godzin publikacji

0.8 Rozkład liczby ocen oraz komentarzy

Poniżej przedstawiono rozkłady ocen (`like` i `dislike`) oraz liczby komentarzy. Ciekawą obserwacją jest proporcja `dislike` do `like`:

- **1** nagranie miało **wyłącznie** negatywne oceny:
  - „Kelly Oubre Punches John Wall in the Lead during warriors wizards scuffle"
- Większość nagrań miała mniej niż **20%** `dislike` (rys. 7)



Rysunek 7: Rozkład proporcji ocen

1 Atrybuty tekstowe

1.1 Tytuł (title)

1.1.1 Najczęstsze słowa i bigramy (po stemmingu)

słowo	liczba wystąpień
offici	738
video	530
2018	522
trailer	443
new	268
2017	255
live	250
ft	239
vs	236
music	219

bigram	liczba wystąpień
offici video	219
offici trailer	189
star war	133
music video	133
offici music	108
trailer hd	105
last jedi	93
lyric video	75
offici audio	71
super bowl	56

Najczęstszym słowem pojawiającym się w tytułach filmów jest „official", zwykle w zbitkach „official video" i „official trailer". Materiały na czasie często podkreślają swoją aktualność przez podanie roku. Oprócz skupienia na muzyce, widoczne jest też nawiązywanie do trendów w popkulturze (film *Gwiezdne Wojny: Ostatni Jedi* miał premierę w grudniu 2017 r.)

1.2 Najczęstsze tagi

Filmy mają zazwyczaj kilkanaście a nawet kilkadziesiąt tagów. Oto najpopularniejsze z nich:

słowo	**liczba wystąpień**
video	22963
music	18302
new	14843
2018	11670
trailer	11581
funni	11564
show	11500
movi	11084
makeup	10446
news	9441

1.3 Opisy

Zauważono, że opisy bardzo często zawierają odnośniki do innych mediów społecznościowych a także wezwania do akcji (ang. call to action) - „follow", „subscribe", „like"

słowo	**liczba wystąpień**
video	6592
follow	4068
twitter	3985
subscrib	3726
show	3691
instagram	3578
music	3578
facebook	3570
watch	3216
get	3186

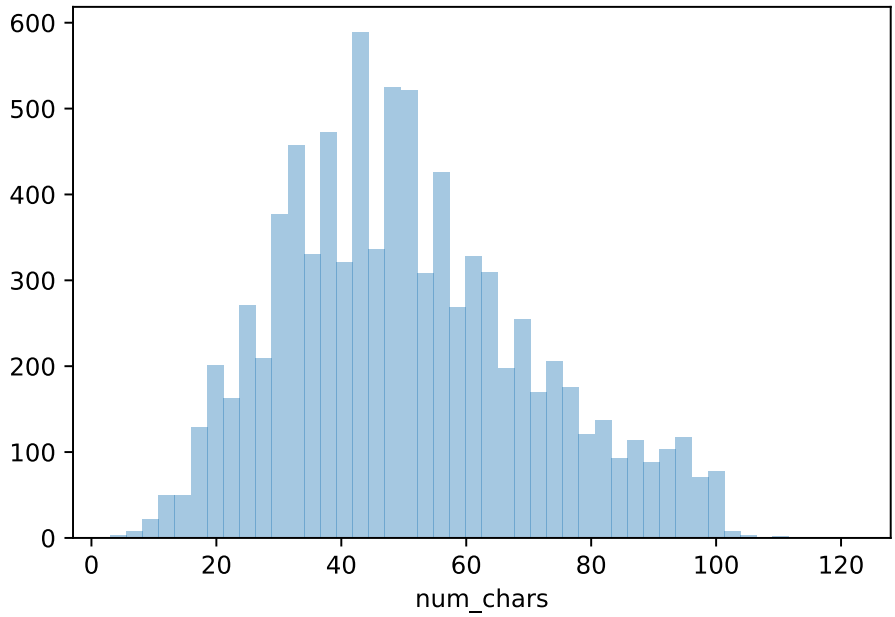
1.4 Cechy

Tytuły popularnych filmów są krótkie. Długość mieści się zwykle między 30 a 50 znakami (dla porównania, pierwsze zdanie tego akapitu ma 41 znaków), a najczęstsza liczba słów to 9.

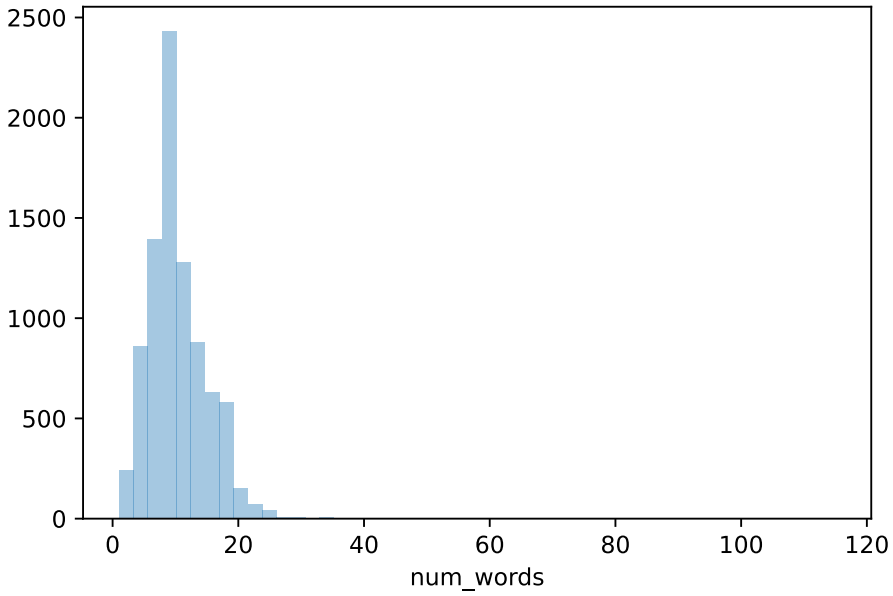
Bardzo częstym zabiegiem jest pisanie całych słów wielkimi literami. Ponad $frac$13 wszystkich tytułów zawiera przynajmniej jedno (trzyliterowe lub dłuższe) słowo zapisane w ten sposób. 5% tytułów było pisanych w całości wielkimi literami.

Najbardziej przyciągającymi uwagę znakami interpunkcyjnymi są wykrzykniki i pytajniki (co najmniej jeden pojawił się w odpowiednio 10% i 5% tytułów). Wykrzyknik przekazuje więcej emocji, ale trafnie postawione pytanie budzi ciekawość odbiorców i tak samo może zachęcać do obejrzenia materiału.

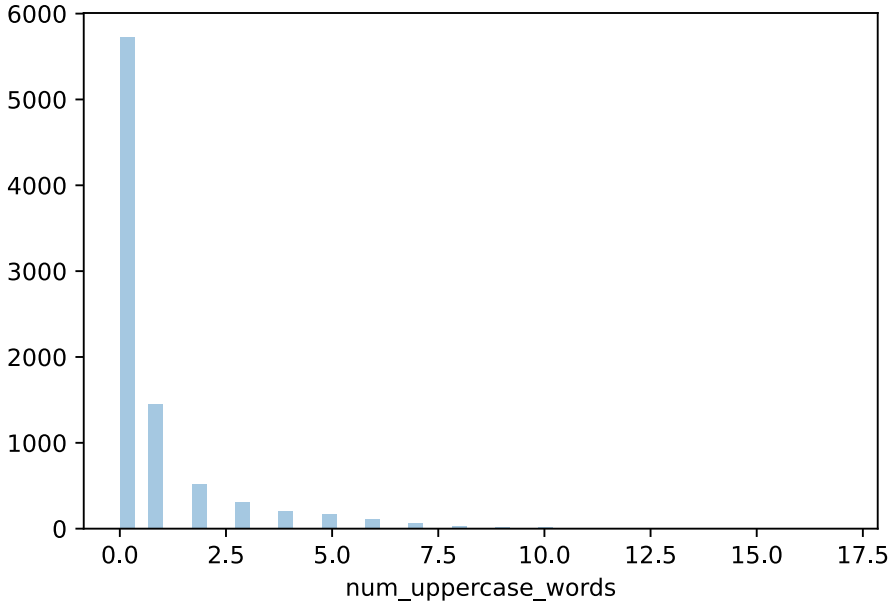
Rekordowy tytuł zawierał 6 wykrzykników (w dwóch grupach po 3), a jednym ciągiem pojawiło się ich maksymalnie 4. Znaków zapytania było najwięcej 3.



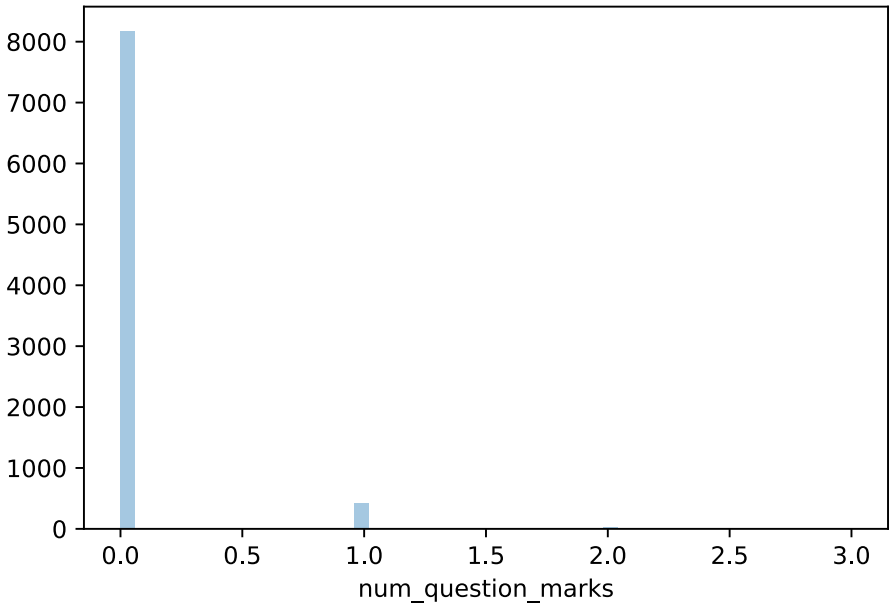
Rysunek 8: Liczba znaków w tytule



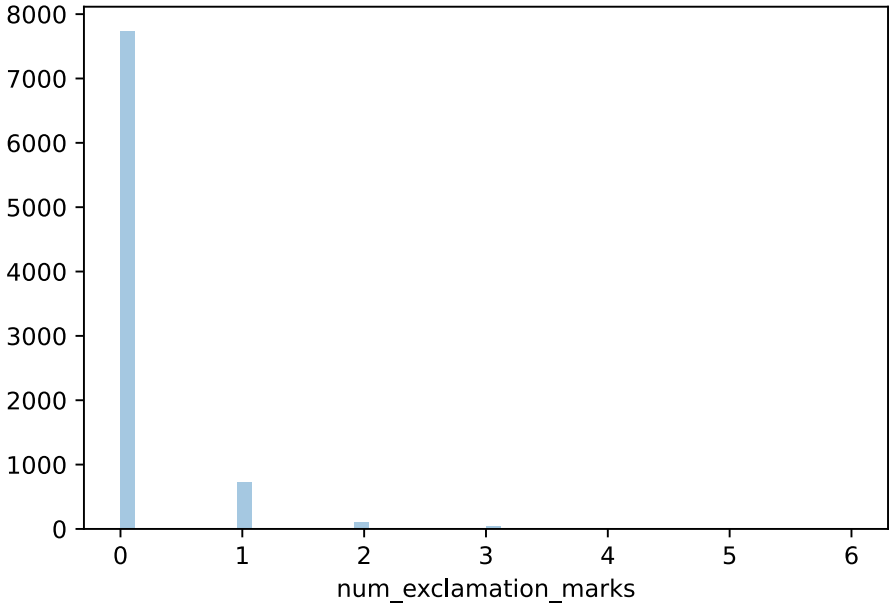
Rysunek 9: Liczba słów w tytule



Rysunek 10: Obecność słów pisanych WIELKIMI LITERAMI



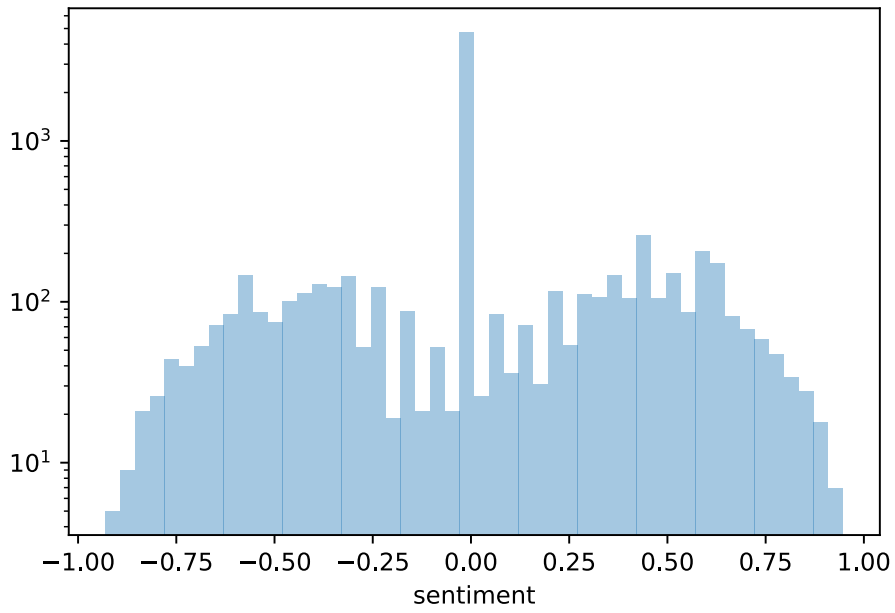
Rysunek 11: Obecność pytajników



Rysunek 12: Obecność wykrzykników

1.4.1 Analiza wydźwięku

Wynik tytułów systemem analizy wydźwięku Vader (dostępnym w bibliotece `nltk`) przedstawia wykres 13. Przeważająca większość tytułów została zaklasyfikowana jako neutralna. W pozostałych możemy zaobserwować, że najrzadsze są skrajne i bardzo delikatne emocje, a przeważają te umiarkowane. Wśród tytułów niosących mniejszy ładunek uczuciowy częściej mamy do czynienia z wydźwiękiem pozytywnym.



Rysunek 13: Średnia zawartość emocjonalna (*compound polarity score*) tytułów