Praca domowa nr 3

Analiza opinii z Twittera

Zadanie

Celem pracy domowej nr 3 jest przećwiczenie technik przetwarzania tekstu poznanych na zajęciach: modelu Bag of Words, analizy opinii oraz pobierania tweetów. Należy pobrać dużą bazę danych tweetów na wybrany temat. Następnie przeanalizować jaką opinię (nastawienie emocjonalne) mają użytkownicy twittera na ten temat.

Alternatywnie do Twittera, można pobrać posty/opinie/komentarze z innych mediów społecznościowych. Należy wówczas znaleźć odpowiednie narzędzie do ściągania danych, lub użyć tzw. Web Scraperów.

Jak w poprzedniej pracy domowej, całe rozwiązanie należy umieścić w notebooku jupyterowym, który w przejrzysty sposób będzie prezentował rozwiązanie. Notebook powinien zawierać wstawki kodu, ich outputy, Twoje komentarze do kodu i wyników. Notebook powinien być sensownie podzielony na rozdziały:

- Wstęp – jaki temat, wybrane hashtagi. Sposób pobierania danych.
- Preprocessing – czy tweety były w jakiś sposób przetworzone (bag of words?) lub przefiltrowane
- Analiza opinii – co ludzie sądzą na dany temat?
- Podsumowanie, interpretacja wyników, wnioski
- Bibliografia

Co należy zrobić?

1) **Wybieramy tematykę.** Należy wybrać temat tweetów i związany z nim #hashtag lub kilka #hashtagów. Zidentyfikowanie hashtagów pasujących do tematu nie zawsze jest takie łatwe. Temat powinien być w miarę popularny: baza danych musi mieć minimum 10 000 tweetów, a najlepiej parędziesiąt do nawet 100 tysięcy tweetów.
   Temat powinien budzić emocje (pozytywne vs negatywne, zwolennicy vs przeciwnicy).
   Kilka podpowiedzi, jakie tematy są warte rozważenia:

   - **Tematy polityczne**, bo polityka zawsze budzi emocje. 😉 Można wziąć wybraną partię polityczną lub polityka i przeanalizować, czy ma poparcie na Twitterze. Nawet ciekawsze może być porównanie dwóch rywalizujących partii lub dwóch polityków i porównanie opinii na ich temat. Kusząca jest analiza polskiej polityki (ze względu na nadchodzące wybory parlamentarne), ale proszę pamiętać, że polskie tweety są trochę trudniejsze do obróbki (patrz punkt 3) niż angielskie. Można szukać też tematów na arenie światowej, np. przeanalizować opinie ludzi na temat dwóch kandydatów na prezydenta w Turcji w kontekście wyborów 14 maja.

   - **Tematy społeczne** to wręcz kopalnia emocji. Nie tak dawno temu w Polsce była duża dyskusja na temat aborcji, odbywały się strajki kobiet (czarne piątki) z powodu zaostrzenia przepisów. Na świecie, w kontekście wielu przypadku nadużyć

seksualnych popularny stał się hashtag #metoo. W USA cały czas trwa walka z powodu nierówności na tle rasowym/etnicznym (np. #blacklivesmatter). Obecnie dość „gorące" tematy związane są z tożsamością płciową (czy dzieci powinny mieć możliwość tranzycji płciowej? Czy powinniśmy wprowadzić specjalne zaimki dla osób niebinarnych? Czy osoby transpłciowe/transseksualne powinny mieć dostęp do rywalizacji sportowej zgodnej z nową płcią?). Jeszcze inny temat to wprowadzanie feminatywów do języka polskiego (np. psycholożka, naukowczyni).

- **Wojny i konflikty**. Można sprawdzić jaką opinię miała Rosja i Ukraina przed rozpoczęciem wojny w lutym 2022, i po rozpoczęciu wojny. Albo przeanalizować tylko polityków (Putin vs Zelensky). Wojna na Ukrainie to dość szeroki temat. Można też przeanalizować różne jej aspekty: migrację Ukraińców do Polski i Europy, pomoc militarną, konkretne wydarzenia (bitwa o Mariupol, wybuch Nordstream, zniszczenie mostu na Krym, itp.). Nie zapominajmy też o innych konfliktach na świecie (np. Chiny vs Tajwan, Armenia vs Azerbejdżan).

- **Rozrywka**. Można wybrać budzące różne emocje seriale, filmy, gry komputerowe, książki, itp. Na przykład, jakie opinie miał serial „Gra o Tron" na przestrzeni ostatnich sezonów? Czy „Wiedźmin" na Netflix ma pozytywne opinie? A z „Pierścieniami Władzy" na Amazon? Czy gra CyberPunk straciła w oczach graczy po premierze? W ogóle, co już zahacza o tematy społeczne, sporo kontrowersji wzbudza wprowadzanie do seriali obsady aktorskiej o zróżnicowanym tle etnicznym. Najnowszym tego przykładem jest czarnoskóra Kleopatra na Netflix (premiera 10 maja). Można też zbadać opinie nt. osób/celebrytów, albo zbadać tematy sportowe (MŚ lub ME w piłce nożnej).

- **Finanse**. Tutaj można badać opinie użytkowników na temat akcji jakiejś firmy, albo kursu waluty czy kryptowaluty na przestrzeni czasu. A następnie sprawdzić jaki był faktyczny kurs tych akcji/walut. Czy występuje korelacja?

- **Inne tematy**. Jest mnóstwo innych tematów, które budzą emocje: szczepionki na COVID, zmiany klimatyczne, różne teorie spiskowe, inflacja, itp. Być może warto rozeznać się w Internecie, co jest aktualnie popularne.

2) **Potwierdzenie i rezerwacja wybranego problemu.** W obrębie grupy laboratoryjnej, każdy powinien mieć unikalny temat. Po wyborze tematu, zarezerwuj go publicznie, tak by inni wiedzieli, że temat jest zajęty. Robimy to, dodając komentarz w konwersacji na Teams, założonej przez prowadzącego zajęcia. Komentarz powinien zawierać nazwę tematu i ewentualnie przykładowe badane #hashtagi.
Obowiązuje zasada: kto pierwszy ten lepszy.

3) **Preprocessing bazy danych**. Tweety należy przygotować do analizy opinii. Być może potrzebne będzie przeprowadzenie kroków dla Bag of Words (tokenizacja, stopwords, lematyzacja). Jeśli zdecydujesz się na bazę danych polskich tweetów, to trzeba się zastanowić jak z nimi postąpić:

- Tłumaczę je na angielski, przeprowadzam preprocessing i analizę opinii. Wyniki (wizualizacje) robię albo w języku angielskim albo polskim (po ponownym tłumaczeniu na polski).
- Cały czas operuję na polskich tweetach, ale wówczas dobieram narzędzia, które pozwalają na preprocessing i analizę opinii dla języka polskiego. To podejście jest pewnie trudniejsze, ale wpłynie to pozytywnie na ocenę projektu.

4) **Analiza opinii.** Każdy tweet można przeanalizować pod kątem nastawienia emocjonalnego autora. Należy to zrobić przynajmniej dwoma narzędziami poznanymi na zajęciach:
   - NLTK Vader (pozytywne, neutralne, negatywne)
   - Text2Emotion (5 emocji)

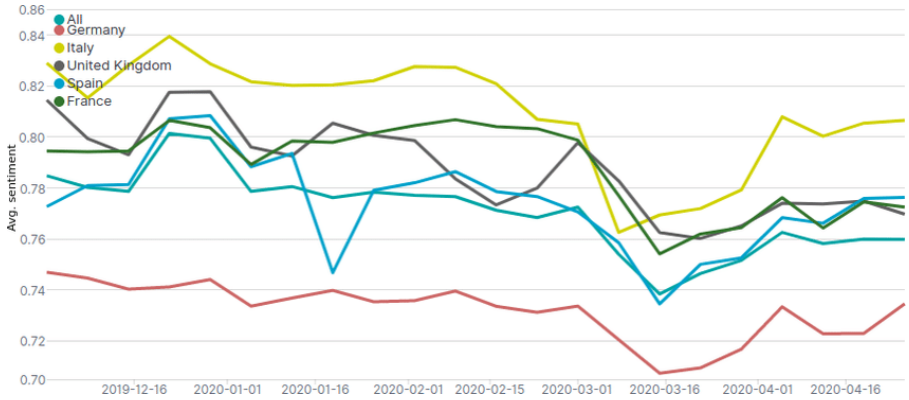
   Można dodać też jakieś inne narzędzie, ale nie jest to obowiązkowe. Należy się zastanowić jak interpretować wyniki. Np. czy ignorować „neutralność"? Od jakiego poziomu negatywności uznać tweeta za tekst negatwny, a od jakiego poziomu pozytywności za pozytywny? A może nie wprowadzać takiego rozróżnienia binarnego, tylko zachować oceny numeryczne, które zwraca nam narzędzie, i potem je uśrednić dla wszystkich tweetów?

   Jakie grupy tweetów badać pod kątem opinii?

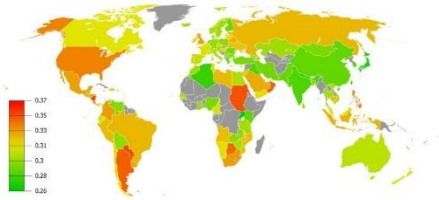
   - **Analiza całościowa**. Jeśli temat na dwa podtematy (np. Politycy: Kaczynski vs Tusk), to oczywiście trzeba dla każdego podtematu zrobić oddzielną analizę opinii (osobno Tusk, osobno Kaczyński). Dla każdego podtematu należy też zrobić chmurę tagów (np. chmura dla Kaczyńskiego i chmura dla Tuska). Dodatkowo, należy spróbować w sensowny sposób podzielić tweety na pozytywne i negatywne i zrobić dodatkowo chmury tagów dla tweetów pozytywnych i negatywnych (u nas byłyby to 4 chmury: pozytywne Tusk, negatywne Tusk, pozytywne Kaczyński, negatywne Kaczyński). W przypadku narzędzie Text2Emotion tych chmur będzie więcej. Je również trzeba dodać. Można je zestawić w taki sposób, żeby nie rozpychały prezentacji. Do analizy proszę dołączyć swoje komentarze i spostrzeżenia, zwłaszcza gdy w chmurkach znajdą się jakieś zaskakujące słowa.

   - **Analiza czasowa**. Oprócz całościowej analizy opinii należy zrobić też analizę czasową. Grupy tweetów należy rozbić na takie z różnych okresów. I zbadać jak zmienia się opinia w czasie. Dla niektórych tematów, dzieje się to naturalnie (np. dla filmu mogą to być 3 okresy: przed trailerem, po trailerze ale przed premierą, po premierze). Dla innych tematów (zwłaszcza finansowych) warto po prostu zbadać zmiany opinii na przestrzeni dni, tygodni lub miesięcy. Czyli ściągamy tweety równomiernie (np. po 300 dziennie, przez okres 4 miesięcy = około 4*30*300=36 000 tweetów). Następnie patrzymy jaka była średnia opinia w każdym ze 120 dni, lub w każdym z 17 tygodni. Takie dane warto nanieść na wykres liniowy (przykładowy rysunek poniżej, następna strona) i go przeanalizować. Czy zmiany w sentymencie są spowodowane jakimiś wydarzeniami? Czy potrafisz zidentyfikować te wydarzenia?
     Uwaga 1: Wykres dla Vadera dość łatwo zrobić. Dla Text2Emotion trzeba by zrobić kilka wykresów dla różnych emocji.

Uwaga2: Jeśli analiza czasowa ma mało okresów czasu (2 lub 3) to dla każdego można zrobić osobną chmurę tagów. W innym przypadku, chmury tagów można sobie darować.



- **Analiza przestrzenna (dla chętnych)**. Można spróbować pobrać tweety z różnych Państw lub województw i sprawdzić różnice w opiniach dla różnych miejsc. Różnice można zilustrować na mapce. To zadanie jest dla chętnych, można za nie dostać mały bonus punktowy.



Terminy i ocenianie

Czas na zrobienie zadania (termin oddania) zostanie napisany przez prowadzącego na Teams (tydzień lub dwa tygodnie).

Jeśli prowadzący stworzy zadanie na Teams, z możliwością załączenia plików, to należy załączyć notebooki Jupyter w podanym terminie. Jeśli to możliwe, to proszę załączyć też wersję HTML lub PDF notebooka.

Praca domowa oceniana jest na **maks 5 punktów**. Oceniane będą takie aspekty jak:

- Czy baza tweetów jest odpowiednio duża i czy temat jest sensowny?
- Czy sprawozdanie jest przejrzyste i kompletne?
- Jak przeprowadzono preprocessing bazy danych?
- Czy przeprowadzono wymagane eksperymenty tzn. analizę całościową, czasową, obie zilustrowane chmurami tagów lub wykresami liniowymi?
- Czy wyniki opatrzono komentarzami?

Praca domowa może być oceniona na trzy sposoby. Dla części osób PD może być sprawdzona jednym sposobem, dla części innym.

- Notebook będzie pobrany przez prowadzącego zajęcia i sprawdzony zdalnie/osobiście.

- Prowadzący poprosi o odpalenie notebooka na zajęciach i krótką prezentację pracy domowej.
- Prowadzący poprosi o prezentacje notebooka na projektorze i prezentację rozwiązania przed całą grupą (w przypadku wybranych osób, lub szczególnie ciekawych prac).