**Uniwersytet Rzeszowski**

**Wydział Nauk Ścisłych i Technicznych**

**Instytut Informatyki**

**Krzysztof Motas**

***Analiza sentymentu postów na platformie X dotyczących transferów Realu Madryt***

**Projekt na ocenę 5 z przedmiotu Eksploracja danych internetowych**

Prowadzący: dr inż. Piotr Lasek

Rzeszów 2025

Spis treści

[1. Opis ontologii 3](#_Toc200150175)

[2. Proces zbierania danych 3](#_Toc200150176)

[3. Czyszczenie i przygotowanie danych 3](#_Toc200150177)

[4. Analiza sentymentu 4](#_Toc200150178)

[5. Wizualizacja wyników 6](#_Toc200150179)

[6. Wnioski 9](#_Toc200150180)

1. Opis ontologii

Celem projektu było przeprowadzenie analizy sentymentu postów użytkowników platformy X (dawniej Twitter), które odnoszą się do tematu transferów piłkarskich związanych z klubem Real Madryt. Głównym założeniem było stworzenie kompletnego procesu analitycznego: począwszy od zbierania danych, poprzez ich czyszczenie i analizę sentymentu, aż po wizualizację wyników oraz wyciągnięcie wniosków.

1. Proces zbierania danych

Dane zostały pozyskane z oficjalnego API platformy X, korzystając z wersji v2 i endpointu search/recent. W tym celu założono aplikację deweloperską na stronie developer.x.com oraz wygenerowano token autoryzacyjny, który umożliwia wykonywanie zapytań z poziomu aplikacji.

Aby pobrać posty dotyczące Realu Madryt i transferów, przygotowano specjalnie sformułowane zapytanie. Uwzględnia ono zarówno oficjalną nazwę klubu oraz jego hashtag, jak również popularne słowa i nazwiska związane z tematyką transferową. Zapytanie ograniczono do języka angielskiego, by ułatwić dalszą analizę sentymentu, i wykluczono retweety, które mogłyby zafałszować wyniki.

Użyte zapytanie wyglądało następująco:

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 1. Zapytanie do API X

Parametry tweet.fields zapewniły, że dla każdego posta otrzymano informacje takie jak treść tweeta (text), czas jego opublikowania (created\_at), język (lang) oraz identyfikator użytkownika (author\_id).

Dane zostały następnie zapisane w pliku JSON i posłużyły jako plik wejściowy do dalszego przetwarzania. Ograniczenia darmowego poziomu API (100 postów miesięcznie, 10 zapytań co 15 minut) wymusiły przemyślane zapytanie i selekcję tematów.

Cały kod odpowiedzialny za komunikację z API znajduje się w pliku fetch\_tweets.py.

1. Czyszczenie i przygotowanie danych

Przed przystąpieniem do analizy sentymentu, zebrane tweety zostały oczyszczone i  przygotowane do przetwarzani. Głównym celem tego etapu było usunięcie zbędnych lub zakłócających elementów, takich jak odnośniki, znaki interpunkcyjne czy tzw. słowa puste (ang. *stop words*), które nie niosą istotnej wartości informacyjnej.

Operacje czyszczenia wykonano w pliku clean\_tweets.py przy użyciu biblioteki re (do wyrażeń regularnych), string, nltk (do tokenizacji i usuwania słów pustych) oraz pandas. Proces obejmował następujące kroki:

* konwersję tekstu do małych liter,
* usunięcie adresów URL,
* usunięcie znaków interpunkcyjnych,
* tokenizację tekstu (podział na słowa),
* odfiltrowanie słów pustych z języka angielskiego.

Wynik oczyszczonego tekstu został zapisany w nowej kolumnie cleaned\_text. Ostateczny wynik czyszczenia zapisano do pliku tweets\_cleaned.json, który został użyty w dalszych etapach analizy.

Poniżej przedstawiono opisany fragment kodu:

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 2. Fragment pliku clean\_tweets.py

1. Analiza sentymentu

Analizę nastroju użytkowników przeprowadzono przy użyciu modelu VADER, który jest częścią biblioteki nltk.sentiment. VADER jest specjalnie przystosowany do pracy z krótkimi tekstami i  językiem potocznym, dzięki czemu bardzo dobrze sprawdza się w analizie treści pochodzących z mediów społecznościowych.

Dla każdego tweeta obliczana była wartość compound, czyli wskaźnik nastroju mieszczący się w  zakresie od -1 (bardzo negatywny) do +1 (bardzo pozytywny). Na jego podstawie przypisywano klasę nastroju:

* wynik ≥ 0.05 oznacza sentyment pozytywny,
* wynik ≤ -0.05 oznacza sentyment negatywny,
* wartości pośrednie interpretowano jako neutralne.

Analiza została zrealizowana w pliku analyze\_sentiment.py, a wyniki zapisano w dwóch kolumnach: sentiment (wartość liczbowa) oraz sentiment\_label (etykieta: positive, neutral, negative).

Opisywany fragment kodu został umieszczony poniżej:

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 3. Fragment pliku analyze\_sentiment.py

1. Wizualizacja wyników

W celu interpretacji wyników analizy przygotowano cztery wykresy z użyciem bibliotek matplotlib, seaborn oraz wordcloud. Każdy z nich przedstawia inne aspekty nastrojów użytkowników wobec transferów Realu Madryt.

* **Wykres słupkowy – Rozkład klas sentymentu**

Ten wykres przedstawia liczbę tweetów w każdej klasie nastroju: pozytywne, neutralne, negatywne. Pozwala on szybko określić, który typ emocji przeważał wśród użytkowników. W  analizowanym zbiorze dominowały tweety pozytywne, co sugeruje optymistyczne nastawienie fanów klubu wobec możliwych transferów.

Obraz zawierający tekst, zrzut ekranu, diagram, Prostokąt

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 4. Wykres słupkowy przedstawiający rozkład klas sentymentu

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 5. Kod wykresu słupkowego

* **Wykres liniowy – Średni sentyment w czasie**

Wykres ten przedstawia, jak zmieniała się wartość średniego sentymentu w czasie. Dane zagregowano godzinowo. W przypadku badanej próbki, tweety pochodziły z jednego dnia, co ograniczyło zakres czasowy analizy, ale pozwoliło zaobserwować niewielkie zmiany emocji w  ciągu dnia.

Obraz zawierający tekst, Czcionka, linia, pismo odręczne

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 6. Wykres liniowy przedstawiający zmianę nastrojów w czasie

Obraz zawierający tekst, Czcionka, zrzut ekranu

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 7. Kod wykresu liniowego

* **Heatmapa – Sentyment względem długości tweeta**

Mapa cieplna pokazuje, jak kształtował się średni sentyment w zależności od liczby słów w  tweetach. Dane pogrupowano na cztery przedziały długości (np. 0–5, 6–10 słów itd.). Wykres ten pozwala dostrzec, czy bardziej rozbudowane wypowiedzi miały inny charakter emocjonalny niż krótkie. Można zauważyć, że dłuższe tweety często miały bardziej pozytywny wydźwięk.

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, numer

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 8. Heatmapa przedstawiająca średni sentyment względem długości tweeta

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 9. Kod heatmapy

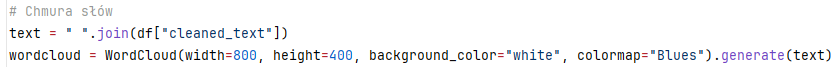
* **Chmura słów – Najczęściej używane słowa**

Chmura słów pokazuje, które słowa pojawiały się najczęściej w tweetach po czyszczeniu. Jest to przydatne narzędzie eksploracyjne, pozwalające zidentyfikować główne nazwiska, hasła i  tematy poruszane przez użytkowników. W chmurze dominowały wyrazy takie jak "real", "madrid", "window", a także nazwiska konkretnych piłkarzy.

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 10. Chmura słów



Rysunek 11. Kod chmury słów

1. Wnioski

Zrealizowany projekt pokazał, że analiza nastrojów na podstawie danych z platformy X może być przydatnym narzędziem do oceny ogólnego tonu wypowiedzi użytkowników. Pomimo ograniczeń w liczbie dostępnych danych, możliwe było zaobserwowanie podstawowych tendencji emocjonalnych. Projekt potwierdził także skuteczność prostych metod przetwarzania języka naturalnego i wizualizacji w pracy z treściami z mediów społecznościowych.