WYKORZYSTANIE MODELU NLP DO ANALIZY EMOCJONALNYCH WZORCÓW W TREŚCIACH NA PLATFORMIE REDDIT

european student magazine
Numer 1 / 2025

Aleksander **Kurek**[1], Aleksandra **Sikora**[2*]

[1] Politechnika Świętokrzyska, Wydział Elektrotechniki, Automatyki i Informatyki, al. Tysiąclecia Państwa Polskiego 7, 25-314 Kielce, Polska, s092747@student.tu.kielce.pl
[2] Politechnika Świętokrzyska, Wydział Elektrotechniki, Automatyki i Informatyki, al. Tysiąclecia Państwa Polskiego 7, 25-314 Kielce, Polska, asikora@tu.kielce.pl

Streszczenie − Artykuł przedstawia oprogramowanie narzędzia umożliwiającego analizę nastrojów na podstawie metadanych dostępnych na platformie Reddit. Narzędzie pobiera przefiltrowane i posortowane dane o określonej frazie z serwisu, analizuje je pod kątem wiarygodność treści, a następnie wizualizuje wyniki w raporcie. Wyposażono je w zabezpieczenia chroniące przed nadmiernym obciążeniem serwera oraz wstrzykiwaniem złośliwego kodu. Wdrożono komunikaty walidacyjne dla danych wejściowych interfejsu użytkownika. Wykonane testy potwierdzają poprawność generowania raportu, skuteczność wdrożonych komunikatów oraz zgodność analizy nastrojów z podobnymi narzędziami.

Słowa kluczowe - Analiza sentymentu, Przetwarzanie języka naturalnego, Reddit, Web scraping

WSTĘP

Analiza sentymentu w mediach społecznościowych staje się coraz bardziej istotna w zrozumieniu nastrojów społecznych i trendów komunikacyjnych, co ma kluczowe znaczenie zarówno dla badań naukowych, jak i praktycznego wykorzystania w biznesie czy polityce. Rozwój przetwarzania języka naturalnego umożliwiający precyzyjną analizę sentymentu oraz rosnący wpływ platform społecznościowych na kształtowanie opinii publicznej zapoczątkowały chęć stworzenia narzędzia, które łączy nowoczesne technologie analityczne z intuicyjnym sposobem prezentacji wyników. [1]

Reddit to bogata w szeroki wachlarz opinii, zainteresowań i wiedzy platforma dyskusyjna tworzona przez użytkowników, którzy decydują i pilnują przestrzegania obowiązujących zasadach. Publikowane treści są stale monitorowane, co zwiększa wiarygodność zamieszczanych informacji oraz ogranicza występowanie fałszywych informacji, co sprawia, że Reddit, jako źródło danych oferuje rzetelne dane.

Analiza sentymentu znajduje zastosowanie w wielu dziedzinach, takich jak marketing, polityka, obsługa klienta oraz medycyna. Firmy korzystają z niej, aby śledzić opinie klientów i reagować na potrzeby rynku, natomiast w mediach społecznościowych wspomaga zarządzanie wizerunkiem marki w czasie rzeczywistym. Pomimo ciągłego postępu, analiza sentymentu wciąż zmaga się z wyzwaniami, takimi jak trudności w interpretacji kontekstu oraz identyfikacja sarkazmu i dwuznaczności. Jednak rozwój NLP i sztucznej inteligencji nieustannie zwiększa precyzję i efektywność tej technologii. [2]

Na stronach internetowych generowane są ogromne ilości zróżnicowanych danych, scrapowanie stron jest powszechnie uznawane za efektywną i potężną metodę zbierania dużych zbiorów informacji. Serwis pozwala na scrapowanie danych bez korzystania z oficjalnego API, poprzez protokół HTTP, pod warunkiem, że proces ten nie narusza zasad serwisu. Pobieranie nie obciąża serwerów Reddita, a także nie narusza praw prywatności użytkowników, w związku z tym, scrapowanie na serwisie jest legalne. W ramach projektu metadane pobierane z platformy Reddit obejmują informacje zawarte w kluczach JSON, które odnoszą się do treści i parametrów postów oraz ich komentarzy. Dodatkowo uwzględniają szczegóły dotyczące subreddita oraz kont autorów postów. [3]

Stworzone narzędzie nie korzysta z płatnych rozwiązań i technologii oraz nie wymaga zakładania konta. Koncentruje się wyłącznie na analizie sentymentu w postach i komentarzach na platformie Reddit. Rozwiązanie wprowadza ocenę zaufania źródeł, dedykowane indywidualnym użytkownikom, oferujące podstawową analizę sentymentu wraz z oceną zaufania źródeł.

I. PORÓWNANIE ORAZ WYBÓR MODELI ANALIZY SENTYMENTU

Tworzenie własnego modelu od podstaw wymaga dużych zasobów, takich jak czasu, danych i zaawansowanej wiedzy w zakresie przetwarzania języka naturalnego. Należy zebrać odpowiednio duży zbiór danych, który obejmie różne odmiany języka, sformułowania i wyrażenia emocji, a następnie dobrze go oczyścić i przygotować do analizy. Nawet po

odpowiednim przygotowaniu danych proces trenowania modelu jest czasochłonny i wymaga odpowiedniej infrastruktury obliczeniowej, szczególnie gdy dąży się do uzyskania wysokiej dokładności, co często wymaga wielokrotnego testowania i dostosowywania modelu. Natomiast gotowe, nauczone modele analizy sentymentu znacząco upraszczają proces wdrażania skutecznej analizy emocji i opinii z tekstu. Korzystając z gotowych modeli, można znacznie szybciej osiągnąć dokładne i skalowalne wyniki, minimalizując ryzyko błędów wynikających z niedoświadczonego trenowania. [4-6]

Ważnym aspektem analizy sentymentu jest odpowiedni dobór modelu, który nie tylko zapewnia wysoką precyzję klasyfikacji, ale również spełnia oczekiwania dotyczące efektywności czasowej. W przeprowadzonej analizie porównano skuteczność oraz czas działania kilku popularnych modeli analizy sentymentu na zbiorze danych ocenionym ręcznie, co pozwoliło na kompleksową ocenę ich przydatności. Precyzja w kontekście poprawności sentymentu analizowanych treści oznacza procent wszystkich poprawnie zaklasyfikowanych przykładów przez model. Wskazuje jaki odsetek danych został prawidłowo określony jako pozytywny, neutralny, lub negatywny zgodnie z rzeczywistą klasyfikacją.

Tabela 1. Porównanie precyzji modeli analizy sentymentu

Nazwa modelu	NLTK	VADER	Text Blob	BERT	RoBERTa	Distil BERT
Precyzja [%]	69	69	64	68	32	40

Tabela 2. Porównanie czasu wykonywania modeli analizy sentymentu

Nazwa modelu	NLTK	VADER	Text Blob	BERT	RoBERTa	Distil BERT
Precyzja [s]	0.018	0.014	0.022	4.352	4.333	2.172

Testy precyzji (Tabela 1) oraz pomiaru czasu (Tabela 2) wykonywania przeprowadzone na tym samym zbiorze wykazały zrównoważoną wydajność i szybkości działania modeli NLTK, Vader i TextBlob. Model BERT charakteryzujący się dobrą precyzją, nie spełnia wymagań związanych z wydajnością czasową, ogranicza to jego użyteczność w analizie dużych zbiorów danych. Modele RoBERTa i DistilBERT okazały się nieoptymalne w obu analizowanych aspektach, nie spełniają standardów koniecznych do uzyskania wiarygodnych i efektywnych wyników, co eliminuje je z dalszych rozważań.

Istotnym kryterium przy wyborze modelu jest jego zdolność do dalszego rozwoju i dostrajania. Modele NLTK, TextBlob, BERT i VADER oferują różne podejścia w tym zakresie. Proces dostrajania modeli NLTK oraz BERT jest czasochłonny i wymaga utworzenia własnych klasyfikatorów, co wiąże się z pozyskaniem dużego zbioru oznakowanych danych. Metody słownikowe takie jak TextBlob oraz VADER umożliwiają poprawę niskim nakładem pracy poprzez edytowanie słownika sentymentu, albo modyfikację przypisywania wag zwracanym wartościom składowym.

Innym podejściem jest zastosowanie kilku modeli jednocześnie, aby zrekompensować ich indywidualne ograniczenia. Przeprowadzono testy z wykorzystaniem macierzy pomyłek na zbiorze ponad dwustu samodzielnie ocenionych wpisów w celu określenia, czy różne modele uzupełniają swoje braki poprzez łączenie ich mocnych stron. Macierz pomyłek to narzędzie służące do oceny wydajności modelu klasyfikacyjnego, które przedstawia liczbę przypadków prawidłowo i błędnie zaklasyfikowanych jako pozytywne lub negatywne. Dzięki macierzy pomyłek możliwe jest ocenienie, jak model rozróżnia różne klasy decyzyjne. Testy nie ujęte w projekcie wykazały, że modele takie jak NLTK i TextBlob charakteryzują się podobnymi błędami klasyfikacyjnymi co VADER. W przeciwieństwie do nich model BERT wykazuje odmienny profil błędów. [7]

Model BERT wykazywał trudności w klasyfikacji danych negatywnych, co skutkowało większą liczbą błędnych decyzji w tej kategorii. Z kolei model VADER okazał się bardziej efektywny w rozpoznawaniu emocji pozytywnych i neutralnych, co przełożyło się na jego ogólną przewagę. Mimo zaawansowanej architektury, wyniki uzyskane przez BERT nie przewyższyły rezultatów VADER, co dowodzi, że większa złożoność modelu nie zawsze gwarantuje lepszą skuteczność w specyficznych zadaniach. Analiza pokazała, że modele nie uzupełniają swoich niedoskonałości, przez co ich połączenie nie przynosi dodatkowych korzyści. Brak istotnej różnicy w skuteczności BERT sugeruje, że dodatkowa złożoność wynikająca z integracji modeli nie jest opłacalna.

Ostatecznie spośród analizowanych modeli w testowanych warunkach na podstawie testów wybrano model VADER ze względu na szybkość dokonywanej analizy sentymentu, precyzję oraz zastosowanie modelu. Model ten wyróżnia się wysoką szybkością działania, co sprawia, że idealnie nadaje się do analizy dużych zbiorów krótkich tekstów, takich jak komentarze czy wpisy w mediach społecznościowych. VADER jest zoptymalizowany pod kątem języka codziennego, co umożliwia mu skuteczne rozpoznawanie emocji nawet w skrótowych lub kolokwialnych wypowiedziach. Dzięki swojej metodzie słownikowej, działa efektywnie i nie wymaga dużych zasobów obliczeniowych, co jest kluczowe w projektach z ograniczeniami czasowymi. Model ten w porównaniu z bardziej złożonymi modelami oferuje kompromis pomiędzy dokładnością a wydajnością.

VADER (Valence Aware Dictionary and sEntiment Reasoner) to słownikowy model analizy sentymentu, zaprojektowany specjalnie do analizy krótkich tekstów. Dzięki dostosowaniu do języka codziennego, emoji czy skrótów językowych, nadaje się do analizy mediów społecznościowych. Ocenia biegunowość tekstu na podstawie wcześniej zdefiniowanych słów i ich wartości sentymentu. Wartości wynikowe są podawane jako ułamkowe wartości, co pozwala na bardziej precyzyjną ocenę niż tylko etykiety. W celu dostrojenia modelu istotne jest zdefiniowanie i przypisanie etykiet określających relacje pomiędzy różnymi kategoriami

sentymentu. Dzięki analizie rozkładu sentymentów możliwe jest określenie wartości granicznych dla poszczególnych klas biegunowości sentymentu. [7]

Podczas przyporządkowania sentymentu postu jego opinia mogła być niepewna, przez co kategorie sentymentu zostały rozszerzone. Rozróżnienie to opiera się na fakcie, że jeden wpis może budzić mieszane emocje, co może prowadzić do różnorodnych interpretacji sentymentu.

Przyjęty podział klas sentymentu przedstawia się następująco:

- Positive dla zakresu <1.0; 0.47>,
- Neutral Positive dla zakresu (0.47; 0.18),
- Neutral dla zakresu <0.18; -0.26),
- Neutral Negative dla zakresu <-0.26; -0.43),
- Negative dla zakresu <-0.43; -1.0>.

Podczas oceny precyzji modelu analizy sentymentu VADER uwagę zwrócono na sytuacje, w których wyniki analizy plasują się w przedziałach o rozszerzonej biegunowości sentymentu. W takich przypadkach post przypisywany jest połowicznie do każdej z zawierających kategorii sentymentu, co pozwala na szersze ujęcie złożoności emocjonalnej danego wpisu. Dostrojenie modelu poprawia błędną klasyfikację postów negatywnych oraz precyzję modelu. Rozszerzenie biegunowości przy połowicznym podziale rozszerzonego sentymentu nie odzwierciedla w pełni liczby poprawnie sklasyfikowanych decyzji, a raczej ogólną dokładność odwzorowania emocji w tekście.

Dla dalszej walidacji, model VADER został przetestowany (Rys. 1) na ogólnodostępnym zbiorze danych Sentiment140. Zbiór zawiera krótkie wpisy w języku angielskim z przypisanymi już etykietami sentymentu. Do analizy wybrano ponad 20 tysięcy losowych próbek i uzyskano precyzję analizy sentymentu w wysokości 69%. [8]

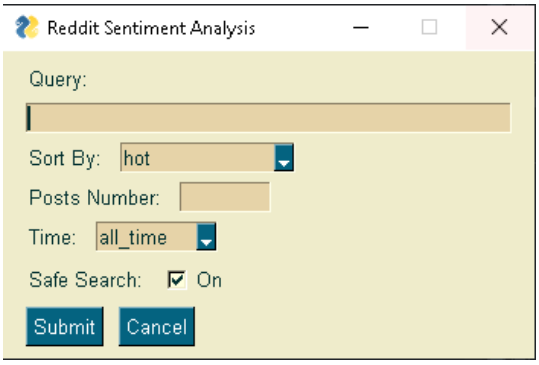
VADER					
Etykieta klasy:	Liczba próbek przyporządkowanych do klasy:			Liczba poprawnych decyzji:	Liczba błędnych decyzji:
	Pozytywne	Neutralne	Negatywne		
Pozytywne	5941	2041	642	5941	2683
Neutralne	961	6022	2476	6022	3437
Negatywne	116	542	2781	2781	658
Suma = 21522	7018	8605	5899	14744 (69%)	6778 (31%)

Rys. 1. Macierz pomyłek dla dostrojonego modelu VADER dla zewnętrznego zbioru danych

II. IMPLEMENTACJA NARZĘDZIA

Projekt wykorzystuje biblioteki języka programowania Python w celu manipulacji danymi, analizy tekstu, tworzenia wizualizacji oraz interakcji z użytkownikiem. Technologie HTML oraz CSS zostały użyte podczas generowania końcowego raportu. Odpowiadają za strukturę, organizację treści na stronie oraz stylizacji elementów. PyScript to narzędzie umożliwiające uruchamianie kodu Pythona bezpośrednio w przeglądarce internetowej, w ramach projektu został wykorzystywany w raporcie podczas dynamicznego obliczania liczby dni, które upłynęły od chwili wygenerowania raportu.

Interfejs użytkownika (Rys. 2) aplikacji stworzony z wykorzystaniem biblioteki PySimpleGUI przedstawia interaktywne komponenty, takie jak pola tekstowe, rozwijane listy i przyciski, które umożliwiają użytkownikowi dostosowanie parametrów wyszukiwania postów na platformie Reddit. Umożliwia sprawną komunikację między użytkownikiem, narzędziem, a serwerem.



Rys. 2. Okno dialogowe interfejsu użytkownika

Pobieranie danych z Reddita może odbywać się bez użycia API, które wymaga posiadania konta oraz kluczy dostępu. Dostęp do postów Reddita uzyskamy wykorzystując protokół HTTP z biblioteki Pythona httpx i zapytania w plikach JSON, co pozwala uprościć komunikację i zmniejszyć liczbę żądań do serwera. Dzięki temu unikamy pełnego scrapowania stron, uzyskując tylko potrzebne dane i optymalizując obciążenie serwera.

Pierwszym krokiem jest pobranie listy postów dla danego słowa kluczowego lub frazy. Dzięki niej uzyskujemy podstawowe informacje o każdym poście, w tym jego tytuł i główne parametry, co stanowi punkt wyjścia do dalszej analizy. Kolejnym etapem jest pobranie pełnych szczegółów postu oraz listy komentarzy. Dane te pozwalają dokładniej zbadać reakcje użytkowników na dany temat, a także analizować treść postu i komentarzy pod kątem sentymentu. Uzyskane informacje o subredditcie umożliwiają zrozumienie kontekstu społeczności, która angażuje się w dany post, oraz jej aktywności, co jest ważne dla oceny wiarygodności. Ostatnim etapem jest pozyskanie danych o autorze postu. Analiza jego aktywności, historii i reputacji jest niezbędna do oszacowania jego wiarygodności i wpływu w serwisie. Tak rozbudowany zestaw danych pozwala na przeprowadzenie wszechstronnej analizy. Dzięki analizie sentymentu możemy określić emocjonalny wydźwięk postów i komentarzy, zaś współczynnik zaufania, wyliczany na podstawie aktywności autora i subreddita daje wgląd w wiarygodność danego postu oraz społeczności, która go wspiera. Wyrażony jest jako wartość liczbowa, przy czym wyższe wartości oznaczają większą wiarygodność. Proces jego wyznaczania opiera się na analizie różnych czynników wpływających na wiarygodność. [9]

Podczas pobierania danych z Reddita istnieje możliwość sortowania i filtrowania wyników

wyszukiwania. Filtry i sortowanie te następują po stronie serwera Reddita, więc nie obciążają naszego sprzętu i upraszczają komunikację zmniejszając liczbę żądań wysyłanych do serwera. Parametry te pozwalają na dokładniejsze dopasowanie wyników do naszych potrzeb, ograniczając liczbę pobieranych postów tylko do tych najbardziej istotnych.

Aby przyspieszyć proces analizy danych, narzędzie wykorzystuje wielowątkowość do równoczesnego pobierania treści postu, informacji o autorze oraz o subredditcie. Równoległe pobieranie danych skraca czas oczekiwania na wyniki i pozwala szybciej przetwarzać dane, co jest przydatne przy analizie dużej liczby postów i ich komentarzy. W celu zwiększenia wydajności analizowanych informacji i zmniejszenie zużycia pamięci, narzędzie wyodrębnia i pobiera tylko konkretne klucze. W ramach analizy przypięte komentarze są pomijane, zazwyczaj są to komentarze dodawane przez boty moderacyjne Reddita i zawierają informacje techniczne lub przypomnienia regulaminowe, które nie wnoszą wartościowych treści do analizy. Pominięcie tych komentarzy pozwala skupić się na opiniach użytkowników, które są kluczowe dla oceny sentymentu dyskusji. Reddit ogranicza liczbę zapytań, pozwala to na równomierne obciążenie serwera i zapobiega nadmiernemu zużyciu zasobów. Reddit udostępnia wyniki stronicowane, jedno zapytanie zwraca do 100 postów lub 100 komentarzy, umożliwia to pobranie większej ilości danych przy mniejszej liczbie zapytań. Dzięki temu możemy efektywnie przeglądać większe zbiory danych przy mniejszym obciążeniu serwera.

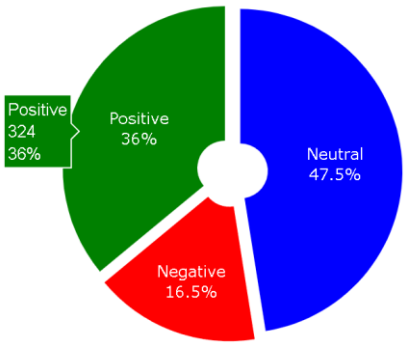
Moduł VADER wykorzystuje specjalnie skonstruowany słownik wyrazów o określonej polaryzacji, dzięki czemu potrafi wyznaczać poziom biegunowości tekstu. Każda analizowana fraza zwraca słownik z wynikami w czterech kategoriach: pozytywnej, neutralnej, negatywnej oraz związek średniej połączonych wskaźników polaryzacji. Związek ten to wartość zmiennoprzecinkowa z przedziału od -1 do +1, im wyższa wartość tym bardziej pozytywny tekst. [4]

Aby zwiększyć dokładność analizy sentymentu postu, narzędzie łączy tytuł i treść w jeden tekst, dzięki czemu analiza sentymentu uwzględnia kontekst zarówno z tytułu, jak i opisu, co pozwala na dokładniejsze odczytanie nastroju w pełnym kontekście postu. Program bada również sentyment poszczególnego komentarza w celu identyfikacji nastroju opinii w dyskusji. Analizowane są niepełne zbiory komentarzy składające się na jeden post, w celu zrozumienia czy redditorzy zgadzają się z przesłaniem głównego wpisu.

Narzędzie generuje interaktywne wykresy w celu zrozumienia przez użytkownika prezentowanych treści. Są one osadzane w formacie HTML, ułatwia to ich eksport. Wykresy kołowe i liniowe oferują podgląd trendów i rozkładu danych, a chmura słów pozwala na analizę popularnych tematów. Generowana tabela szczegółów postów z kolei dostarcza analizy jakościowej.

Program dynamicznie generuje raport za pomocą języka programowania Python w formacie HTML, który automatyzuje prezentację wyników analizy sentymentu. Proces implementacji obejmuje tworzenie tabeli z rozkładem sentymentów, która pokazuje pozytywne, neutralne i negatywne reakcje dla każdego postu, oraz uwzględnia szczegóły takie jak: tytuł postu, tekst, autor, opinia sentymentu i etykiety współczynnika zaufania. Raport wyświetla czas, jaki minął od jego wygenerowania i aktualizuje go przy każdym odświeżeniu strony. Zawiera także osadzone wykresy liniowe i kołowe rozkładu sentymentów, co umożliwia wizualne przedstawienie analizy oraz ułatwia zrozumienie wyników.

Raport końcowy wygenerowany jest w języku HTML przy użyciu stylów CSS oraz Python do automatyzacji tworzenia i formatowania jego zawartości. Wprowadzenie obejmuje dwie tabele zawierające informacje oraz filtry zastosowane podczas przeszukiwania i pobierania danych. Kolejna tabela przedstawia podział postów ze względu na nastrój. Raport zawiera dwa wykresy kołowe (Rys. 3), które przedstawiają rozkład sentymentu. Przedstawiają przykładową liczbę postów dla badanego zagadnienia, jeden z nich wprowadza do tego wagę opinii użytkowników.



Rys. 3. Wykres kołowy przedstawiający podział sentymentu dla przykładowego badanego zagadnienia

Raport podzielony jest na trzy kluczowe sekcje, które grupują wyniki według biegunowości sentymentu: Pozytywne, Neutralne i Negatywne. Każda z tych części zawiera zestawienie tabel i wykresów, podobne w strukturze, lecz oparte na danych specyficznych dla kategorii sentymentu. Jednym z elementów każdej sekcji jest tabela przedstawiająca podział komentarzy wraz z liczbą uzyskanych punktów upvotes przypisanych do trzech najważniejszych postów z danej kategorii sentymentu. Dostarcza informacji na temat emocjonalnego odbioru wiodących postów i popularności związanych z nimi komentarzy.

W kolejnym fragmencie raportu znajduje się interaktywny wykres liniowy, który ilustruje zależność liczby upvotes od czasu, rozłożoną w podziale na poszczególne kategorie sentymentu. Dzięki interaktywności wykresu, po najechaniu kursorem na dowolny punkt, użytkownik widzi szczegóły obejmujące dokładną datę oraz liczbę upvotes zdobytą przez posty w danym tygodniu. Taka wizualizacja umożliwia śledzenie trendów i zmian w reakcjach użytkowników w odniesieniu do postów o różnych sentymentach, pomaga zidentyfikować okresy wzmożonego zainteresowania lub

spadku popularności dla danego sentymentu. W kolejnej części raportu znajduje się tabela przedstawiająca metadane postu, pomaga w zrozumieniu kontekstu i wiarygodności analizowanych treści. Raport zawiera zestawienie trzech najlepiej ocenianych postów, ilustrujące ich parametry i komentarze, pozwala na analizę wypowiedzi pod kątem ich popularności oraz odbioru. Na końcu raportu znajduje się mapa myśli z najczęściej używanymi słowami, prezentowana jest w formie zbioru najpopularniejszych słów pojawiających się w analizowanych postach. Pozwala wyłapać powtarzające się motywy i słowa kluczowe odzwierciedlające główne zainteresowania społeczności w analizowanych wpisach.

Zabezpieczenia i komunikaty w programie są elementami zapewniającymi stabilność działania aplikacji oraz zapewnienie użytkownikowi informacji o stanie operacji. Projekt implementuje zarówno mechanizmy obsługi błędów, jak i kontroli liczby zapytań do serwera, co jest istotne w przypadku interakcji z zewnętrznymi usługami serwisu Reddit. Projekt obsługuje błędy w trakcie wykonywania zapytania HTTP i sprawdza status odpowiedzi serwera. Narzędzie wyposażone jest w walidację danych wejściowych. Przed wykonaniem głównego zapytania, program sprawdza, czy użytkownik podał frazę zapytania oraz czy liczba postów jest poprawna i nieprzekraczająca limitu. Jeżeli dane wejściowe są niepoprawne lub niekompletne, użytkownik zostaje poinformowany o błędzie za pomocą komunikatu pop-up. Ponadto, istnieje kontrola maksymalnej liczby zapytań, którą użytkownik może wysłać w danym czasie. Podczas generowania raportu końcowego narzędzie chroni przed atakami wstrzykiwania złośliwego kodu, takimi jak XSS (Cross-Site Scripting) oraz HTML Injection. Dane wejściowe wprowadzone przez użytkownika są zabezpieczane poprzez zamianę potencjalnie niebezpiecznych znaków specjalnych na ich bezpieczne odpowiedniki. Całość implementacji zapewnia stabilność i responsywność aplikacji, chroni użytkownika przed przekroczeniem limitów i wprowadza odpowiednie komunikaty, które pomagają w rozwiązywaniu ewentualnych problemów związanych z błędami w komunikacji z serwerem.

III. TESTY

Została przeprowadzona walidacja raportu HTML przy użyciu W3C Markup Validation Service, co zapewnia zgodność kodu ze standardami i prawidłowe wyświetlanie raportów w przeglądarkach.

Narzędzie przetestowano pod kątem wywołania komunikatów błędów. Przeprowadzono także analizę porównawczą działania z platformą Social Searcher, która mimo pewnych ograniczeń, takich jak brak dostępu do starszych postów czy ograniczone filtrowanie, umożliwia analizę sentymentu na podstawie różnych źródeł danych. Testy przeprowadzono na danych z platformy Reddit, stosując identyczne parametry dla obu narzędzi, co pozwoliło na rzetelne porównanie wyników. [10]

Rozwiązanie projektowe	Liczba wpisów			Liczba głosów użytkowników		
Sentyment	Pozytywne	Neutralne	Negatywne	Pozytywne	Neutralne	Negatywne
Średni udział nastrojów [%]	32,9	52,75	14,35	28,62	59,72	11,64
Social Searcher	Liczba wpisów			Wszystkie źródła		
Sentyment	Pozytywne	Neutralne	Negatywne	Pozytywne	Neutralne	Negatywne
Średni udział nastrojów [%]	25,8	62,4	11,8	16,4	73,9	9,7

Rys. 4. Porównanie narzędzia projektowego z platformą Social Searcher

Analiza wykazała różnicę w przypadku źródła Reddit w wysokości maksymalnie 14% pomiędzy obydwoma narzędziami, przy czym dominująca emocja była zawsze taka sama w obu przypadkach. Rozkład procentowy uwzględniający wagi użytkowników w niektórych przypadkach w narzędziu projektowym odpowiadał ogólnemu rozkładowi emocji prezentowanemu przez Social Searcher dla wszystkich analizowanych źródeł danych. Wskazuje to na potencjalną zgodność w identyfikacji kluczowych wzorców emocjonalnych, mimo odmiennych podejść i ograniczeń technicznych narzędzi.

IV. WNIOSKI

Projekt ma na celu stworzenie narzędzia umożliwiającego analizę sentymentu w postach i komentarzach na platformie Reddit. Aplikacja napisana w języku programowania Python pobiera dane z serwisu, dokonuje analizy sentymentu treści, wyznacza współczynnik zaufania dla postu, autora oraz subredditu, następnie wizualizuje wyniki w raporcie końcowym. Wyposażono ją w interfejs użytkownika w celu komunikacji z serwerem Reddit pozwalający na wpisanie wyszukiwanej frazy oraz dobór filtrów wyszukiwania. Interfejs użytkownika oraz raport końcowy zostały przygotowane w języku angielskim, co wynika z dominacji anglojęzycznej społeczności na platformie Reddit.

Głównym zadaniem programu jest gromadzenie postów zawierających określone frazy wraz z metadanymi dotyczącymi zarówno subredditu, na którym zostały opublikowane, jak i profilu autora. Dane te odbierane są ze strony internetowej poprzez scrapowanie jej, czyli automatyczne filtrowanie i pobieranie wybranych z niej treści. W celu zwiększenia wydajności ściągania rekordów zastosowano mechanizm równoczesnego pobierania informacji o poście, jego autorze oraz subredditcie do którego należy. Zebrane dane poddawane są analizie nastrojowej poprzez nauczony i dostrojony model VADER w celu określenia wzbudzanych emocji przez wpis. Biegunowość modelu uwzględnia nastroje: pozytywny, neutralny i negatywny. Zostały one rozszerzone o dodatkowe kategorie: pozytywno-neutralny i negatywno-neutralny. Program wyznacza współczynnik zaufania, aby określić wiarygodność analizowanych treści. Na tą wartość składają się takie elementy jak: aktywności autora i subreddita, zaangażowanie w tworzone treści oraz statusy. Wyniki analiz prezentowane są w formie

wykresów i tabel w interaktywnym raporcie końcowym generowanym w formacie HTML z arkuszem stylów CSS. Komunikacja z serwerem Reddita odbywa się za pomocą graficznego interfejsu, który pozwala na wpisanie wyszukiwanej frazy, dobór filtrów, sortowania i parametrów wpisów do przechwycenia. Aplikacja została wyposażona w szereg zabezpieczeń i komunikatów informujących użytkownika o błędach oraz poprawnym działaniu. Raport została wyposażony w zabezpieczenia chroniące przed atakami wstrzykiwania złośliwego kodu.

W ramach projektu opisano proces porównania dostępnych modeli analizy sentymentu, co umożliwiło wybór odpowiedniego modelu oraz jego dostrojenie, aby osiągnąć jak najlepsze wyniki. Opracowano optymalizację procesu pobierania danych dzięki zastosowaniu wielowątkowości w celu zwiększenia efektywności aplikacji. Zaprojektowano strukturę danych JSON z uwzględnieniem szybkości działania i łatwości modyfikacji, aby sprawnie filtrować oraz analizować duże zbiory danych. Zwrócono uwagę na funkcje odpowiedzialne za analizę sentymentu, które dostosowano za pomocą testów. Rozwiązano problemy związane z bezpieczeństwem aplikacji poprzez zaimplementowanie walidacji danych wejściowych oraz mechanizmów ochrony przed błędami.

Wybór technologii wymagał przeprowadzenia porównania dostępnych modeli analizy nastroju, zwrócono uwagę na precyzję, szybkość wykonywania jak i rozwój badanych modeli. Przeważający model VADER poddano procesowi dostrojenia, dla dalszej walidacji został przetestowany na zbiorze danych z ponad dwudziestoma tysiącami próbek, gdzie uzyskał bezbłędność 69%.

UTILIZING AN NLP MODEL FOR ANALYZING EMOTIONAL PATTERNS IN CONTENT ON THE REDDIT PLATFORM

The article presents a software tool that enables sentiment analysis based on metadata available on the Reddit platform. The tool retrieves filtered and sorted data related to a specific phrase from the platform, analyzes it for content reliability, and visualizes the results in a report. It is equipped with safeguards to prevent server overload and malicious code injection. Validation messages for user interface input data have been implemented. The tests conducted confirm the accuracy of report generation, the effectiveness of the implemented messages, and the consistency of sentiment analysis results with similar tools.

Key words: Natural Language Processing, Reddit, Sentiment analysis, Web scraping

BIBLIOGRAFIA

[1] Margarita R.,, Antonio C., Félix C., Pedro-Manuel C. (2023) “A review on sentiment analysis from social media platforms”. Expert Systems with Applications. Vol 223, doi: 10.1016/j.eswa.2023.119862

[2] Snehal S., Akshata B., Sumitra S., Arya C., Aryan S. (2024) “SENTIMENT ANALYSIS”. INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS). Vol. 4, Issue 4. pp 1542-1547, doi: 10.58257/IJPREMS33384

[3] Moaiad K. (2021) “Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application”. Int. J. Advance Soft Compu. Appl. Vol 13, Issue 3, doi: 10.15849/IJASCA.211128.11

[4] László N., Attila K. (2021) “Information Extraction and Named Entity Recognition Supported Social Media Sentiment Analysis during the COVID-19 Pandemic”. Applied Sciences. Vol 11, Issue 22, doi: 10.3390/app112211017

[5] Felix H., Karsten D. (2021) “NewsMTSC: A Dataset for (Multi-)Target-dependent Sentiment Classification in Political News Articles”. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. pp 1663–1675

[6] Lewis T., Leandro von W., Thomas W. (2024) “Przetwarzanie języka naturalnego z wykorzystaniem transformerów. Budowanie aplikacji językowych za pomocą bibliotek Hugging Face”, Helion S.A.

[7] C.J. H., Eric G. (2015) “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text”. Proceedings of the 8th International Conference on Weblogs and Social Media

[8] Go A., Bhayani R., Huang L. (2009) “Twitter Sentiment Classification using Distant Supervision”. CS224N Project Report, Stanford University. pp 1-12

[9] Xusen C., Shixuan F., Gert-Jan de V. (2017) “Understanding trust influencing factors in social media communication: A qualitative study”. International Journal of Information Management. Vol 37, Issue 2. pp 25-35, , doi: 10.1016/j.ijinfomgt.2016.11.009

[10] Social Searcher https://www.social-searcher.com [Dostęp 12.02.2025]