

POLITECHNIKA KRAKOWSKA

Dashboard z funkcjami analizy sentymentu oraz generowania tweetów i komentarzy dla platform Twitter oraz Reddit

Projekt z przedmiotu Przetwarzanie Języka Naturalnego

MK, BK

Projekt składający się z dashboardu umożliwiającego analizę sentymentu dla tweetów i komentarzy oraz generowanie własnych tweetów oraz komentarzy

1. Wprowadzenie

W dobie mediów społecznościowych, analiza treści publikowanych na takich platformach jak Twitter czy Reddit stanowi cenne źródło informacji, zarówno dla badaczy, jak i dla firm. W niniejszym projekcie skupiliśmy się na stworzeniu dashboardu, który umożliwi zarówno analizę sentymentu na ww. platformach, jak i generowanie nowych treści w postaci tweetów i komentarzy. Dashboard będzie zawierał funkcjonalności analizy sentymentu oraz generowania treści, dostarczając narzędzie przydatne dla szerokiego zakresu użytkowników.

Dane

Dane wykorzystane do uczenia modeli pochodzą z platform Twitter oraz Reddit. Poniżej przedstawiono fragment przykładowego zestawu danych gdzie w kolumnie “category” możliwe są 3 wartości i są to:

- 1 - pozytywny sentyment,
- 0 - neutralny,
- -1 - negatywny

Twitter:

	A	B
1	clean_text	category
2	when modi promise	-1
3	talk all the nonsense	0
4	what did just say vot	1
5	asking his supporter	1
6	answer who among t	1
7	kiya tho refresh maa	0
8	surat women	0
9	this comes from cab	0
10	with upcoming electi	1
11	gandhi was gay does	1
12	things like demoneti	1
13	hope tuthukudi peop	1
14	calm waters wheres	1
15	one vote can make a	0
16	one vote can make a	0
17	vote such party and l	-1
18	vote modi who has n	0
19	through our vote ens	0
20	dont play with the	1
21	didnâ€™ write chowk	-1
22	was the one who reci	1
23	with firm belief the le	-1
24	crush jaws those wh	0
25	sultanpur uttar pradi	-1
26	thiugh nehru not aliv	-1

Reddit:

	A	B	C
1	clean_comment	category	
2	family mormon have	1	
3	buddhism has very n	1	
4	seriously don say thi	-1	
5	what you have learne	0	
6	for your own benefit	1	
7	you should all sit dow	-1	
8	was teens when disc	1	
9	jesus was zen meets	0	
10	there are two varietie	-1	
11	dont worry about tryi	1	
12	recently told family t	1	
13	unto others you wou	1	
14	first understand that	1	
15	recently heard simila	1	
16	different times differe	1	
17	does evil include the	-1	
18	our campaign has tw	1	
19	technically you coul	-1	
20	zarus	0	
21	blood and souls for l	0	
22	always liked vecna t	1	
23	though don have any	-1	
24	cliche but you can w	-1	
25	zon kuthon and his s	1	
26	homebrew setting th	-1	
27	his name shall lump	0	
28	for one campaign pro	-1	
29	demogorgon becaus	-1	
30	you want badasss ev	-1	
31	pelor the burning hat	-1	
32	fyi the traditional cht	1	
33	far evil goes big fan	-1	

Dane te zawierają teksty publikacji, które następnie zostały wykorzystane do trenowania modeli analizy sentymentu oraz generowania nowych treści w ramach niniejszej pracy.

2. Cel projektu

Celem projektu jest stworzenie interaktywnego dashboardu, który pozwoli na:

1. Analizę sentymentu tweetów i komentarzy z platform Twitter i Reddit.
2. Generowanie nowych tweetów i komentarzy w oparciu o wytrenowane modele.

3. Zakres

Zakres pracy obejmuje przygotowanie analizy danych wykorzystanych do wytrenowania modeli użytych w ramach niniejszej pracy, przygotowanie narzędzia do analizy sentymentu dla tweetów oraz komentarzy z serwisów Twitter oraz Reddit, a także przygotowanie narzędzia do generowania komentarzy oraz tweetów.

4. Technologie (Metodyka)

Do wykonania projektu zastosowano następujące technologie:

- **Python** – Główny język programowania
- **Jupyter Notebook** – Wykorzystany do kodu trenującego wykorzystywane modele w projekcie
- **Streamlit framework** – Wykorzystany do stworzenia interaktywnego dashboardu
- **Biblioteki** Pandas, Numpy, matplotlib, wordcloud, tensorflow, nltk, sklearn – główne biblioteki/narzędzia do realizacji analizy sentymentu oraz generatora komentarzy i tweetów

5. Komponenty projektu

System składa się z trzech głównych komponentów:

- Moduł nr. 1 – Analiza danych wykorzystanych do uczenia modeli
- Moduł nr. 2 – Narzędzie do analizy sentymentu
- Moduł nr. 3 – Generator tweetów oraz komentarzy

Komponenty te współpracują ze sobą w celu dostarczenia finalnego rozwiązania: dashboardu z funkcjami analizy sentymentu oraz generowania tweetów i komentarzy dla platform Twitter oraz Reddit.

5.1. Moduł nr. 1

Moduł ten obejmuje analizę wstępną danych, które posłużą do trenowania modeli. Analiza danych jest kluczowym etapem, który wpływa na jakość wyników uzyskiwanych przez modele. Proces ten obejmuje następujące kroki:

5.1.1. Oczyszczanie danych

Oczyszczanie danych to proces usunięcia lub poprawiania nieprawidłowych, uszkodzonych lub niekompletnych danych. W kontekście tego projektu, oczyszczanie danych obejmuje:

- Usuwanie duplikatów: Powielone wpisy są usuwane, aby uniknąć zniekształcenia wyników analizy.
- Usuwanie spamowych i niewłaściwych treści: Komentarze i tweety, które nie wnoszą wartości merytorycznej (np. reklamy, spam), są eliminowane.
- Korekta literówek i błędów gramatycznych: Poprawa błędów językowych, które mogą wpływać na skuteczność modeli językowych.

5.1.2. Eksploracja danych

Eksploracja danych pozwala na lepsze zrozumienie zbioru danych przed przystąpieniem do dalszej analizy. Kluczowe działania w ramach eksploracji danych obejmują:

- Analiza rozkładu danych: Sprawdzenie, jak dane są rozłożone w różnych kategoriach (np. pozytywne, negatywne, neutralne).
- Identyfikacja kluczowych cech: Wykrywanie najważniejszych atrybutów, które będą miały wpływ na trenowanie modeli.
- Wykrywanie anomalii: Znalezienie i oznaczenie danych, które są nietypowe lub odstające.

5.1.3. Wizualizacja danych

Wizualizacja danych pomaga w intuicyjnym zrozumieniu i prezentacji wyników eksploracji danych. W projekcie zastosowano następujące techniki wizualizacji:

- Histogramy i wykresy słupkowe: Do przedstawienia rozkładu różnych kategorii danych.
- Chmury słów: Do wizualizacji najczęściej występujących słów w danych.
- Wykresy korelacji: Do identyfikacji związków między różnymi cechami danych.

5.2. Moduł nr. 2

Moduł ten wykorzystuje wytrenowane modele do analizy sentymentu tweetów i komentarzy. Analiza sentymentu pozwala na określenie, czy dane wyrażają pozytywne, negatywne czy neutralne emocje. Moduł ten integruje różne klasyfikatory, które należą do różnych kategorii algorytmów uczenia maszynowego, zapewniając elastyczność i dokładność analizy. Proces ten obejmuje następujące etapy:

5.2.1. Przygotowanie modelu

Model analizy sentymentu jest trenowany na dużych zbiorach danych, aby nauczyć się rozpoznawać różne emocje i opinie w tekstach. Kluczowe działania obejmują:

- Wybór algorytmu: Decyzja o wyborze odpowiedniego algorytmu (np. regresja logistyczna, drzewa decyzyjne) w oparciu o charakterystykę danych.
- Trenowanie modelu: Proces nauki modelu na bazie zgromadzonych danych treningowych
- Walidacja modelu: Ocena skuteczności modelu na zestawach walidacyjnych i testowych, aby upewnić się, że model działa poprawnie.

5.2.2. Implementacja narzędzia

Narzędzie do analizy sentymentu jest implementowane jako część dashboardu. Główne elementy implementacji to:

- Integracja modelu: Włączenie wytrenowanego modelu do aplikacji, umożliwiające analizę nowych danych.
- Interfejs użytkownika: Proste UI, które umożliwia użytkownikom łatwe wprowadzenie danych do analizy.
- Prezentacja wyników: Wyświetlanie wyników poniżej w przystępny sposób.

5.2.3. Testowanie i optymalizacja

Testowanie narzędzia obejmuje sprawdzenie jego działania na rzeczywistych danych, w celu zapewnienia wysokiej dokładności i wydajności. Optymalizacja polega na:

- Udoskonalaniu modelu: Modyfikacja parametrów modelu w celu poprawy jego skuteczności.
- Optymalizacji kodu: Poprawa efektywności kodu, aby narzędzie działało szybciej i bardziej niezawodnie.

5.3. Moduł nr. 3

Moduł ten umożliwia generowanie nowych tweetów i komentarzy na podstawie wytrenowanych modeli językowych. Generowanie treści odbywa się z uwzględnieniem kontekstu i jest wspierane przez model AI, którego nazwa to LSTM. Proces ten obejmuje następujące etapy:

5.3.1. Przygotowanie modelu generującego

Model generujący treści jest trenowany wymienionych na początku pracy zbiorach danych tekstowych. Kluczowe działania obejmują:

- Trenowanie modelu: Proces uczenia modelu na zgromadzonych danych tekstowych

- Walidacja modelu: Testowanie modelu na nowych danych w celu oceny jego zdolności generowania sensownych i płynnych tekstów oraz utworzenie w wyniku testu macierzy błęd (ang. confusion matrix).

5.3.3. Testowanie i optymalizacja

Testowanie generatora obejmuje sprawdzenie jakości i sensowności wygenerowanych treści. Optymalizacja polega na:

- Poprawie modelu: Ulepszanie modelu na podstawie wyników testów, aby generowane treści były bardziej spójne i odpowiednie.
- Optymalizacji kodu: Poprawa efektywności kodu, aby proces generowania treści był szybszy i bardziej niezawodny.

6. Szczegóły implementacyjne oraz prezentacja działania

W oparciu o rozwiązania opisane we wcześniejszych sekcjach niniejszej pracy, zaimplementowano program. Poniżej zawarto szczegóły implementacyjne oraz prezentacje działania każdego modułu tworzonego w ramach niniejszej pracy.

6.1. Moduł nr. 1

Moduł ten, zaimplementowany w pliku Module1.py, wykorzystuje bibliotekę Streamlit do tworzenia interaktywnego dashboardu. Poniżej przedstawiono główne funkcje i ich role:

6.1.1. plot_total_comments_comparison()

Funkcja ta porównuje całkowitą liczbę komentarzy z różnych platform: Reddit, Twitter oraz ich połączenie. Tworzy wykres słupkowy, który wizualizuje liczby komentarzy, ułatwiając porównanie.

- Oblicza liczbę komentarzy dla każdej platformy.
- Rysuje wykres słupkowy z oznaczeniami liczbowymi na słupkach.

6.1.2. plot_sentiment_comparison()

Funkcja ta porównuje liczbę komentarzy w trzech kategoriach sentymentu: negatywne, neutralne i pozytywne dla platform Reddit, Twitter oraz ich połączenia. Używa wykresów słupkowych do wizualizacji danych.

- Oblicza liczbę komentarzy dla każdej kategorii sentymentu i platformy.
- Rysuje wykres słupkowy z oznaczeniami liczbowymi na słupkach.

6.1.3. `plot_comment_length_average()`

Funkcja ta pokazuje średnią długość komentarzy dla każdej kategorii sentymentu na różnych platformach. Pomaga to zrozumieć, jak długość komentarzy różni się w zależności od sentymentu i platformy.

- Oblicza średnią długość komentarzy dla każdej kategorii sentymentu i platformy.
- Rysuje wykres słupkowy z oznaczeniami liczbowymi na słupkach.

6.1.4. `plot_sentiment_pie_charts()`

Funkcja ta tworzy wykresy kołowe, które pokazują procentowy rozkład kategorii sentymentu dla platform Reddit, Twitter oraz ich połączenia. Wizualizacja ta ułatwia zrozumienie dominujących sentymentów na różnych platformach.

- Oblicza procentowy rozkład kategorii sentymentu dla każdej platformy.
- Rysuje wykresy kołowe z oznaczeniami procentowymi.

6.1.5. `plot_word_clouds()`

Funkcja ta generuje chmury słów dla każdej kategorii sentymentu, pokazując najczęściej występujące słowa w komentarzach i tweetach. Wizualizacja ta pozwala na szybkie zidentyfikowanie kluczowych tematów i słów w danych.

6.1.6. `run_module()`

Funkcja ta jest punktem wejścia modułu, wywołującym wszystkie powyższe funkcje w celu wyświetlenia wyników na dashboardzie.

Podsumowanie:

Powyższe funkcje składają się na moduł analizy danych, który zapewnia kompleksową eksplorację, analizę i wizualizację danych z platform Twitter i Reddit.

6.2. Moduł nr. 2

Moduł ten, zaimplementowany w pliku Module2.py, wykorzystuje bibliotekę Streamlit do tworzenia interaktywnego narzędzia do analizy sentymentu. Poniżej przedstawiono główne funkcje i ich role:

6.2.1. Klasyfikatory

Moduł integruje różne klasyfikatory, które należą do różnych kategorii algorytmów uczenia maszynowego:

- **Logistic Regression**: Model liniowy do klasyfikacji binarnej, który przewiduje prawdopodobieństwo wystąpienia jednej z dwóch klas.
- **LinearSVC**: Liniowy model klasyfikacji oparty na wektorach nośnych, skuteczny w zadaniach klasyfikacyjnych.
- **KNeighborsClassifier**: Algorytm k-Najbliższych Sąsiadów, który klasyfikuje nowe przypadki na podstawie podobieństwa do znanych przypadków.
- **GradientBoostingClassifier**: Technika zespołowa, która buduje modele sekwencyjnie, a każdy nowy model koryguje błędy poprzedniego.
- **AdaBoostClassifier**: Metoda zespołowa, która łączy wiele słabych klasyfikatorów w celu stworzenia silnego klasyfikatora.
- **LGBMClassifier**: Framework gradientowego wzmocnienia opartego na drzewach decyzyjnych.
- **RandomForestClassifier**: Wszechstronny algorytm uczenia maszynowego, który tworzy wiele drzew decyzyjnych i używa średniej do poprawy dokładności predykcji.
- **DecisionTreeClassifier**: Prosty, ale potężny algorytm, który tworzy strukturę drzewiastą do podejmowania decyzji na podstawie atrybutów.

6.2.2. Funkcje implementacyjne

1. `__init__()` - Inicjalizuje moduł, ładując wektor TF-IDF, który jest używany do przekształcania tekstów wejściowych do formatu odpowiedniego do analizy przez modele.
2. `run_module()` - Funkcja główna, która jest odpowiedzialna za interakcję z użytkownikiem. Wywołuje inne funkcje w celu analizy sentymentu wprowadzonego tekstu.
 - Wybór klasyfikatora: Umożliwia użytkownikowi wybór jednego z dostępnych klasyfikatorów.
 - Wprowadzenie tekstu: Umożliwia użytkownikowi wprowadzenie tekstu do analizy.
 - Predykcja: Po kliknięciu przycisku "Predict", wybrany klasyfikator jest ładowany i używany do analizy sentymentu wprowadzonego tekstu.
 - Prezentacja wyników: Wynik predykcji jest wyświetlany w formie kolorowego tekstu, wskazując sentyment (negatywny, neutralny, pozytywny).
 - Opis klasyfikatora: Wyświetla opis wybranego klasyfikatora, aby użytkownik mógł zrozumieć, jak działa model.
 - Macierz konfuzji: Jeśli jest dostępna, wyświetla macierz konfuzji dla wybranego modelu, pokazując jego wydajność na zbiorze testowym.

Podsumowanie:

Powyższe funkcje składają się na moduł analizy sentymentu, który zapewnia elastyczność, dokładność i intuicyjność analizy danych z platform Twitter i Reddit. Dzięki integracji różnych klasyfikatorów, użytkownik może wybrać jeden z dostępnych modeli.

6.3. Moduł nr. 3

Moduł ten, zaimplementowany w pliku `Module3.py`, wykorzystuje bibliotekę Streamlit do tworzenia interaktywnego narzędzia do generowania tekstu. Poniżej przedstawiono główne funkcje i ich role:

1. `__init__()` - Inicjalizuje moduł, ładując modele do generowania tekstu dla Reddita i Twittera oraz ich odpowiednie tokenizery.
 - Modele: Ładuje wytrenowane modele LSTM dla platform Reddit i Twitter z plików .h5.
 - Tokenizery: Ładuje tokenizery dla platform Reddit i Twitter z plików .pkl.
 - Długość sekwencji: Ustawia maksymalną długość sekwencji na podstawie kształtu wejścia modelu lub tokenizera
2. `generate_text()` - Funkcja generująca tekst na podstawie podanego przez użytkownika początkowego tekstu (seed text), liczby słów do wygenerowania oraz wybranego modelu i tokenizera.
 - Tokenizacja: Przekształca podany tekst w sekwencję tokenów.
 - Pad sequences: Dopasowuje sekwencję do maksymalnej długości sekwencji modelu.
 - Predykcja: Generuje następne słowo, wybierając je na podstawie rozkładu prawdopodobieństwa przewidywanego przez model.
 - Budowanie tekstu: Dodaje wygenerowane słowo do tekstu początkowego i powtarza proces, aż do wygenerowania określonej liczby słów.
3. `run_module()` - Funkcja główna, która jest odpowiedzialna za interakcję z użytkownikiem i wywoływanie funkcji generujących teksty.
 - UI dla Reddita:
 - Wprowadzenie tekstu: Pole tekstowe dla wprowadzenia początkowego tekstu.
 - Suwnica: Slider do wyboru liczby słów do wygenerowania.
 - Przycisk generowania: Po kliknięciu, wywołuje funkcję `generate_text` i wyświetla wygenerowany komentarz.
 - UI dla Twittera:
 - Wprowadzenie tekstu: Pole tekstowe dla wprowadzenia początkowego tekstu.
 - Pasek wyboru (slider): Slider do wyboru liczby słów do wygenerowania.

- Przycisk generowania: Po kliknięciu, wywołuje funkcję `generate_text` i wyświetla wygenerowany tweet.

Podsumowanie:

Powyższe funkcje składają się na moduł generowania tekstów, który umożliwia użytkownikom tworzenie spójnych i kontekstowo odpowiednich tweetów i komentarzy na podstawie wprowadzonego tekstu początkowego. Dzięki zastosowaniu algorytmu LSTM, moduł zapewnia generowanie treści, co zwiększa jego użyteczność w różnych zastosowaniach, np. tworzenie treści.

7. Podsumowanie

W ramach niniejszego projektu zaimplementowano moduł realizujący analizę danych wykorzystanych do uczenia modeli oraz utworzono dwa narzędzia realizujące analizę sentymentu oraz generowanie komentarzy i tweetów. Dzięki narzędziom utworzonym za pomocą niniejszej pracy możliwe jest lepsza analiza sentymentu poprzez takie media społecznościowe jak Twitter oraz Reddit i w efekcie podejmowanie np. lepszych decyzji finansowych w tradingu.

8. Perspektywy dalszego rozwoju

Przyszłe wersje projektu zakładają takie unowocześnienia jak:

- Automatyzacja pozyskiwania danych przez web scraper w celu zwiększenia danych do uczenia maszynowego
- Zbudowanie bota do handlu na rynkach finansowych, podejmującego decyzje w czasie rzeczywistym na podstawie sentymentu panującego na Twitterze oraz Reddicie
- Stworzenie modułu realizującego filtrowanie spamowych komentarzy oraz tweetów
- Stworzenie bota wykrywającego fałszywą aktywność na portalach Twitter oraz Reddit

BIBLIOGRAFIA

- [1] Widhi Winata Sakti, *Twitter and Reddit Sentiment Analysis*,
<https://kaggle.com/code/widhiwinata/twitter-and-reddit-sentiment-analysis/notebook> [dostęp 21.05.2024]
- [2] Chanchal Alam, *Twitter and Reddit Sentimental analysis Dataset*,
<https://www.kaggle.com/code/chanchal24/twitter-and-reddit-sentimental-analysis-dataset> [dostęp 21.05.2024]
- [3] Yusup Ibrahim Nursiddiq, *Reddit Text Generation*,
<https://www.kaggle.com/code/yusupibrahim/reddit-text-generation/notebook>
[dostęp 21.05.2024]