ZPP Murmuras - HLD

Gustaw Blachowski Szymon Kozłowski
Natalia Junkiert Kamil Dybek

Wprowadzenie

Celem projektu jest stworzenie uniwersalnego rozwiązania do procesowania danych uzyskanych z ekranu smartfonu (wspierany system: Android 9 wzwyż), takich jak lokalizacja elementu na ekranie, jego typ (tekst, zdjęcie itp), jego zawartość. Dane te reprezentują zawartość widzianą przez użytkownika - posty na mediach społecznościowych, reklamy na witrynach, itp. System ten ma umożliwić analizę informacji dotyczących treści obserwowanych przez użytkowników, wspierając tym samym badania o charakterze komercyjnym i społecznym.

Przykładowymi zastosowaniami danych przetowrzonych przez system są (1) analiza danych o oglądanych przez użytkownika reklamach oraz (2) badanie poglądów politycznych i społecznych.

(1) System będzie umożliwiał wygodne badanie zasięgów kuponów promocyjnych. Dzięki temu będziemy mogli sprawdzić, jak dany kupon radzi sobie na tle innych, oraz jak radzi sobie konkurencja. Może to być pomocne w planowaniu przyszłych akcji marketingowych.

(2) System będzie umożliwiał lokalną analizę danych prywatnych takich jak konwersacje użytkowników. Dostarczy to wiarygodnych danych trudno dostępnych innymi metodami, na przykład poglądów politycznych i społecznych.

Istniejące rozwiązania

DOZRO

Rozwiązanie Murmurasa

Istniejące prototypowe rozwiązanie korzysta z danych dostarczonych w postaci tekstu wykrytego na zrzutach ekranu w połączeniu z metadanymi, takimi jak lokalizacja pola tekstowego na ekranie. Następnie dane te są poddawane bardzo podstawowej obróbce (usuwanie pustych kolumn itp), a następnie są przetwarzanie przez ChatGPT4o-mini (za pomocą prompt-engineeringu). Rozwiązanie to ma dwa zasadnicze problemy: nie działa ono lokalnie na urządzeniu mobilnym (model jest za duży), a dane są często opisywane niepoprawnie (np. objętość jest traktowana jak cena produktu).

Specyfikacja skończonego projektu

Jako podstawową część projektu planujemy zaimplementowanie wspomnianego rozwiązania jedynie do usecase z reklamami. Problem z wykrywaniem poglądów pozostawiamy jako opcjonalny kierunek rozwojowy. 1. Wymagane jest narzędzie, które przeprocesuje dane wyekstraktowane z urządzenia w postać nadającą się do użycia przez model.
2. Wymagane jest użycie narzędzia z dziedziny Machine Learningu do ekstrakcji interesujących z naszej perspektywy danych.
3. Opcjonalne jest narzędzie do postprocessingu danych wyjściowych narzędzia z punktu 2 do wspólnego formatu.
4. Wymagany jest deployment powyższych trzech narzędzi na urządzenie mobilne.

Wyzwania

Zapewnianie prywatności użytkowników

DOZRO (do doprezyzowania z Murmurasem;)

Wymagania sprzętowe

Moc obliczeniowa i pamięć operacyjna

Zarówno współczesne LLMy jak i algorytmy preprocessingu danych wymagają często dużej ilości zasobów; jednoczesnie chcemy aby wszystko działało lokalnie na urzadzeniu mobilnym. Wyzwaniem więc będzie dobór narzędzi które nie będą zbyt zasobożerne.

pamięć dyskowa

Funkcjonowanie aplikacji będzie wymagać użycia dużej przestrzeni dyskowej. Zakładamy że nie będzie to problemem dla uzytkownika ze względu na model biznesowy firmy (uzytkownicy są wynagradzani za zainstalowanie rozwiązania na telefonie)

Benchmarkowanie

W celu oceny jakości naszego rozwiązania obecnie planujemy posługiwać się benchmarkiem zapewnionym nam przez Murmuras. Bazuje on na obliczaniu funkcji podobieństwa między wynikiem użytego modelu a wynikiem modelu wzorcowego (obecnie jest to GPT4o-mini). Benchmark ten może okazać się niewystarczająco dokładny i miarodajny także może pojawić się konieczność zaproponowania alternatywy, przykładowo testowania systemu na sztucznie wygenerowanych i poetykietowanych danych.

Propozycja rozwiązania

W implementacji naszego rozwiązania wyróżniamy następujące 4 główne moduły:

Preprocessing danych

Na ten moment rozpatrujemy trzy możliwe rozwiązania: użycie rozwiązań z zakresu uczenia maszynowego, w szczególności klastryzacji, algorytmów niezwiązanych z MLem lub pominięcie jakiegokolwiek preprocessingu.

Processing danych

W tym celu wykorzystamy prawdopodobnie wybrany LLM, ale technicznie rzecz biorąc nie jesteśmy w tym temacie ograniczeni. Na tem moment prawdopodobnie będziemy korzostać z narzędzi HuggingFace; zapewniają one proste i wygodne użycie wielu ogólnodostępnych modeli.

Wybór modelu Po wstępnym researchu postanowiliśmy skupić się na modelach typu transformer o liczbie parametrów z zakresu 10 do około 100 milionów. Większośc wybranych przez nas opcji to pochodne modelu BERT[**?**].
3 główne podtypy to:
1. Bert 100mln parametrów
2. DistilBert 65mln parametrów
3. AlBert 11 mln parametrów

Postprocessing

DOZRO

Deployment na urządzeniu mobilnym

Stworzymy aplikację mobilną bądź dodamy funkcjonalność do istniejącej w ramach której zaimplementujemy powyższe punkty. Utworzymy service działający w tle i przetwarzający nadpływające dane w czasie rzeczywistym. Jeśli okaże się że przetwarzanie w czasie rzeczywistym jest zbyt kosztowne zaimplementujemy przechowywanie danych (DOZRO: ile danych per day) z ekranu i ich analizę w nocy, gdy użytkownik nie korzysta z urządzenia. Planujemy wykorzystać framework TensorFlow Lite (TensorFlow dla urządzeń mobilnych), ewentualnie PyTorch bądź ONNX ze względu na łatwą integrację z aplikacjami rozwijanymi w Android Studio. Chcemy by aplikacja była kompatybilna z Androidem 9+. Nie jest wymagane by aplikacja działała na wszystkich urządzeniach.
DOZRO - gdzie wysyłamy zebrane dane

Kamienie Milowe

Research

Planowane ukończenie 30.11
Do końca listopada chcecmy mieć wybraną architekturę i konkretny model, chcemy mieć propozycje algorytmów odpowiedzialnych za preprocessing i postprocessing.

Proof of Concept

Planowane ukończenie 31.12
Chcemy stworzyć prototypową aplikację demonstrującą całościową funkcjonalność.

Sesja/zbieranie pomysłów na ulepszenia

Planowane ukończenie 31.01
Styczeń będzie miesiącem w trakcie którego nie planujemy bardzo intensywnej pracy nad projektem ze względu na sesję. Przeznaczymy ten czas na ewentualne dokończenie poprzednich kamieni milowych oraz na przemyślenie kierunku projektu.

Rozwój docelowego rozwiązania

Planowane ukończenie 30.04
Na tym etapie zajmiemy się ulepszeniem rozwiązania, usunięciem błędów i testowaniem.

Praca Licencjacka

Planowane ukończenie 30.06
Skupimy się na napisaniu i dopracowaniu pracy licencjackiej.