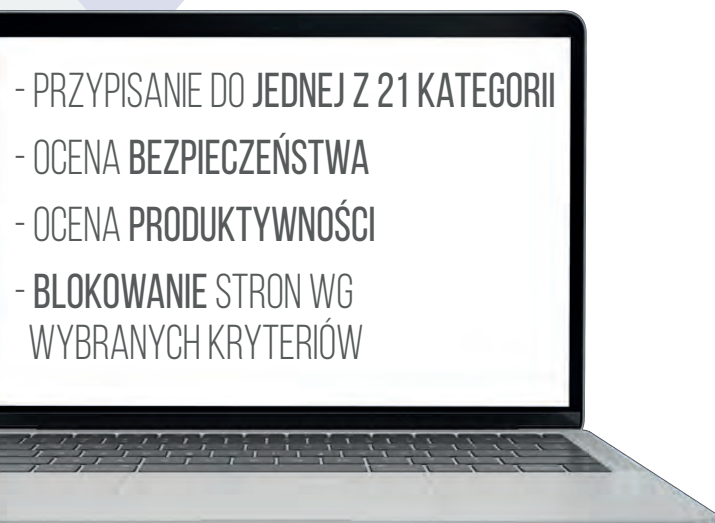




21 KATEGORII

Strony klasyfikowane są w oparciu o rzeczywistą ich treść w sposób powtarzalny oraz przypisywane są do jednej z 21 kategorii.

SKUTECZNOŚĆ > 95%

Zastosowanie algorytmów sztucznej inteligencji oraz zbudowanej bazy wzorców słów kluczowych odpowiada za wysoką skuteczność.

API

Dla podmiotów zewnętrznych dostępne jest API.

24H / DOBĘ, 365 DNI W ROKU

Klasyfikator działa cały czas w czasie rzeczywistym.

MACHINE LEARNING W KLASYFIKACJI STRON INTERNETOWYCH

Klasyfikacja stron internetowych może być przydatna w każdym podmiocie, gdzie nadzór i kontrola aktywności użytkowników może mieć **realny wpływ na bezpieczeństwo**. Zaimplementowanie algorytmu uczenia maszynowego pozwala na **sprawne i szybkie klasyfikowanie każdej strony internetowej** pod kątem jej zawartości i przypisania (sklasyfikowania) do odpowiedniej kategorii. Moduł klasyfikacji stron www w systemie eAuditor V7 WEB przygotowany jest na występowanie różnych zdarzeń losowych w taki sposób, aby mimo błędu po stronie serwera lub wygaśnięcia strony internetowej nie przerywał działania i poprawnie wykonywał swoje zadanie, **przypisując strony www do odpowiednich kategorii**.

KLASYFIKATOR STRON INTERNETOWYCH

Klasyfikator bayesowski, który bazuje na twierdzeniu Bayesa, nadaje się w szczególności do rozwiązywania problemów o bardzo wielu wymiarach na wejściu. Mimo prostoty metody, często działa ona lepiej od innych, bardzo skomplikowanych metod klasyfikujących. Wspomniany klasyfikator można uczyć w trybie uczenia z nadzorem. Oznacza to, że do poprawnego i jeszcze lepszego działania algorytmu konieczny jest nadzór człowieka, który na bieżąco będzie analizował i poprawiał ewentualne błędy algorytmu. **Klasyfikacja jest tak długo poprawna, jak długo poprawna kategoria jest bardziej prawdopodobna od innych. Wsparcie dla działania algorytmu zapewnia BTC.**

W praktyce zdarza się, że algorytm wskaże inną kategorię, niż się tego spodziewamy. Dzieje się to zwłaszcza na stronach informacyjnych, które składają się z wielu artykułów o wielu tematykach i branżach. Wtedy algorytm może wskazać chybioną kategorię.

CZAS KLASYFIKACJI STRONY INTERNETOWEJ

Klasyfikacja pojedynczego URL trwa **od 1 do 2 sekund**. W praktyce osiąga się wyższą wydajność z uwagi na wielowątkową obsługę procesów klasyfikacji (jednoczesne klasyfikowanie kilkudziesięciu lub kilkuset stron).

POPRAWNOŚĆ KLASYFIKACJI STRON INTERNETOWYCH

W ramach testu machine learningu w eAuditor V7 WEB **skategoryzowano 1000 losowych** i mało popularnych stron internetowych. **Poprawność** przypisania kategorii dla tych stron **wynosi około 90%**. Problem z osiągnięciem lepszych rezultatów nie stoi po stronie algorytmu, gdyż ten stwierdza największe prawdopodobieństwo wystąpienia danej kategorii. Problematyczny jest fakt, że jedna strona internetowa może zawierać się w kilku kategoriach na raz i każda z kategorii może być poprawna.

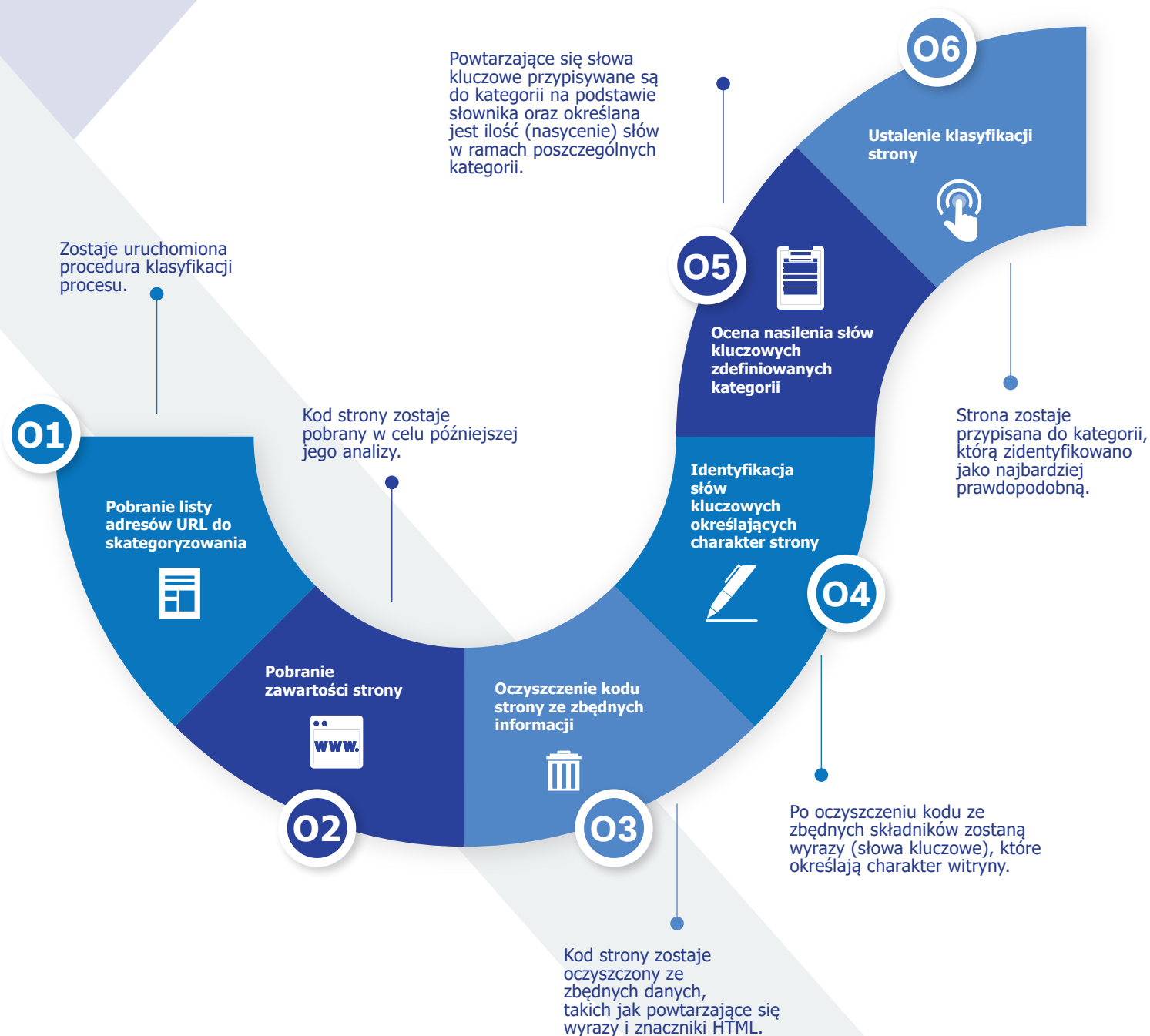
Przykładowo www.onet.pl może być skategoryzowany zarówno jako wiadomości i media, jak również jako rozrywka czy prawo i polityka.

WYDAJNOŚĆ

Silnik klasyfikatora jest umiejscowiony w datacenter i obsługiwany jest przez skalowalny zestaw serwerów, co zapewnia **nieograniczoną wydajność**.

BTC MACHINE LEARNING WEB CLASSIFICATION

JAK DZIAŁA MACHINE LEARNING W KLASYFIKACJI STRON INTERNETOWYCH?



DLACZEGO WPROWADZILIŚMY MACHINE LEARNING DO SYSTEMU EAUDITOR V7 WEB?

Oto kilka powodów, dla których zastosowaliśmy Machine Learning w systemie eAuditor V7 WEB zamiast bazy danych klasyfikacji stron WWW:

- ✓ baza danych stron WWW z przypisanymi kategoriami zajmuje dużo miejsca (pow. 1 TB),
- ✓ ilość stron WWW, to nie kilka tysięcy czy nawet milionów, obecnie jest to liczba trudna do oszacowania,
- ✓ strony WWW mogą zmieniać swoją kategorię szybciej niż gotowe bazy danych kategorii stron,
- ✓ bazy danych wymagają stałej aktualizacji, co jest kosztowne oraz pochłania mnóstwo czasu,
- ✓ machine learning kategoryzuje strony WWW indywidualnie, pod potrzeby każdego użytkownika,
- ✓ klasyfikator doskonale radzi sobie ze specyfiką języka polskiego.