Dokumentacja
mISIe
HackING Challenge

Wojciech Kosiuk
Politechnika Warszawska

Adam Majczyk
Politechnika Warszawska

Szymon Matuszewski
Politechnika Warszawska

Michał Mazuryk
Politechnika Warszawska

Damian Skowroński
Politechnika Warszawska

19 maja 2023

Spis treści

1 Wstęp

Poniższa dokumentacja stanowi opis efektów pracy zespołu **mISIe** podczas **HackING Challenge**. Zadanie polega na sklasyfikowaniu dokumentów na podstawie ich skanów.

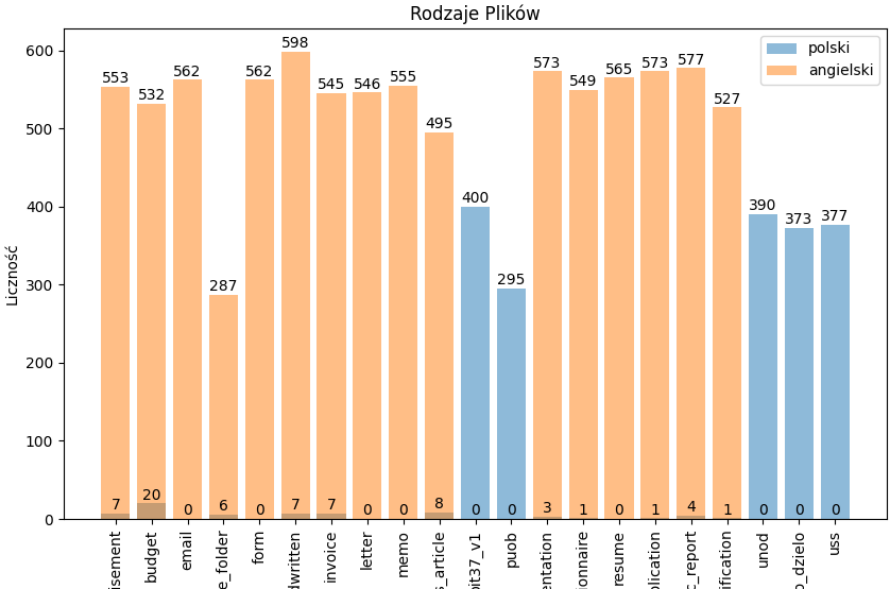
2 Rozwiązanie

2.1 Eksploracja Danych

Naszym pierwszym etapem rozwiązywania problemu była Eksploracja Danych. W tym celu przeanalizowaliśmy na początku **dane wyekstrachowane przez ING**. Oto najważniejsze spostrzeżenia:

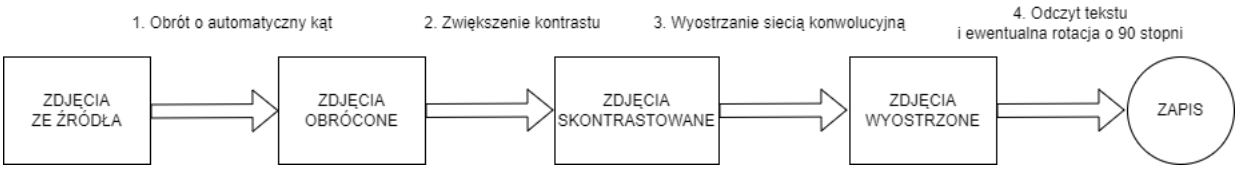
- **21** - liczba różnych kategorii,
- **10884** - liczba zdjęć, z których pobrano tekst,
- **10849** - liczba zdjęć posiadających etykiety,

- Są zdjęcia, z których pobrano tekst, jednak nie posiadają przypisanej etykiety.
- Wszystkie zdjęcia posiadające etykiety mają pobrany tekst.
- Część zdjęć to dokumenty polskie, część to dokumenty angielskie. Są również takie pliki, które nie możemy sklasyfikować jako polskie lub angielskie.



Rysunek 1: Wykres przedstawiający klasyfikacje plików pod względem języka. Jak można zauważyć dla danej etykiety język jest cechą. Są pewne wyjątki, w których klasyfikujemy plik jako polski w kategorii angielskiej.

2.2 OCR

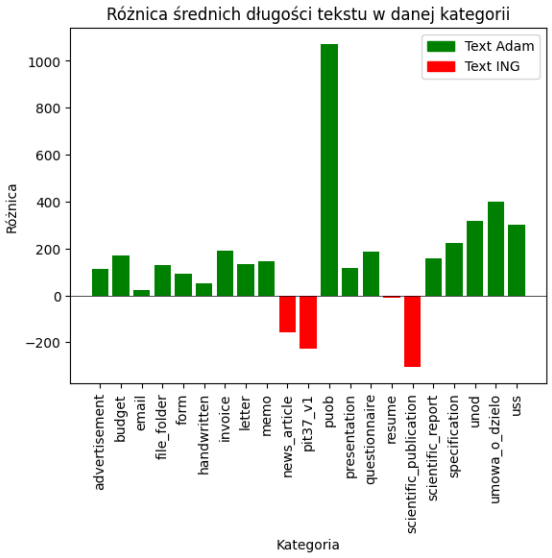


Rysunek 2: Diagram przedstawiający proces OCR. Jeżeli wykryto, że kąt obrotu jest $>= 30°$ to zdjęcie nie zostało obracane.

Zdecydowaliśmy się dodatkowo na przeprowadzenie Optical Character Recognition na dostarczonych w zadaniu plikach. Pozwoliło nam to skontrolować słuszność dostarczonych przez organizatorów tekstów oraz zliczać przy zapisie **liczbę nowych linii** w pliku. Mając większą kontrolę nad pobieraniem tekstów byliśmy w stanie zwiększych dokładność odczytu. Proces OCR przedstawiliśmy na diagramie 2 .

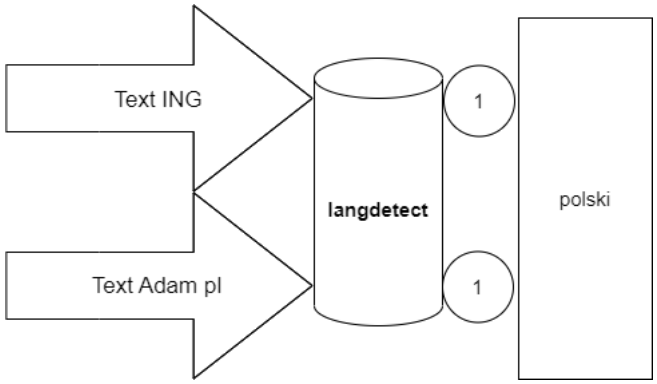
2.3 Kategoryzacja Tekstów

Po własnym pobraniu tekstów z plików zdecydowaliśmy się na wykorzystanie klasyfikacji dokumentów na angielskie i polskie. Nasze hipotezy o słuszności planu działania potwierdziły wyniki przedstawione w tabeli 1 przedstawiające prawdopodobieństwa, że sklasyfikowany dokument do poszczególnych



Rysunek 3: Wkres przedstawiający różnicę średnich długości odczytów dla poszczególnych kategorii pomiędzy naszym odczytem a odczytem organizatorów. Zauważmy, że nasz odczyt przeważnie jest dłuższy, co może (ale nie musi) sugerować, że skorzystanie z naszego odczytu jest słuszne.

kategorii jest w danym języku. Zauważmy, że jeśli dany dokument określimy jako polski musi należeć on do jednej z pięciu kategorii: *pit37_v1, pozwolenie_uzytkowanie_obiektu_budowlanego, umowa_na_odleglosc_odstapienie, umowa_o_dzialo* lub *umowa_sprzedazy_samochodu.*



Rysunek 4: Diagram przedstawiający proces decydowania, czy dany tekst jest klasyfikowany jako polski. Żeby sklasyfikować tekst jako polski porównujemy wyniki otrzymane z pakietu langdetect na tekstach pochodzących od organizatorów i zebranych przez nas, jeśli oba są sklasyfikowane jako polskie (1) to uznajemy tekst za polski.

Biorąc pod uwagę wcześniejsze spostrzeżenia zdecydowaliśmy się na utworzenie **2 modeli: jeden do dokumentów polskich, jeden do dokumentów angielskich** poprzedzone klasyfikacją dokumentu ze względu na język przy pomocy biblioteki **langdetect**. Proces wyboru języka prezentujemy na diagramie 4.

2.4 Ekstrakcja Cech

Na tym etapie musieliśmy rozdzielić nasze działania na dwa nurty: dokumenty polskie oraz dokumenty angielskie. Podział zbiorów na treningowy i testowy polega na tym, że w zbiorze testowym znajduje się 10% zdjęć z każdej kategorii.

Ekstrakcję cech zaczęliśmy od skomponowania danych do tabeli zawierającej:

KATEGORIA	JĘZYK	PSTWO ING	PSTWO MISIE
advertisement	eng	0.81	0.55
budget	eng	0.79	0.56
email	eng	0.97	0.93
file_folder	eng	0.54	0.1
form	eng	0.88	0.79
handwritten	eng	0.76	0.34
invoice	eng	0.82	0.6
letter	eng	0.95	0.92
memo	eng	0.98	0.94
news_article	eng	0.87	0.73
pit37_v1	pl	1.0	1.0
puob	pl	1.0	1.0
presentation	eng	0.78	0.64
questionnaire	eng	0.93	0.81
resume	eng	0.99	0.98
scientific_publication	eng	0.95	0.89
scientific_report	eng	0.85	0.77
specification	eng	0.96	0.83
unod	pl	1.0	1.0
umowa_o_dzialo	pl	1.0	1.0
uss	pl	1.0	1.0

Tabela 1: Tabela przedstawia prawdopodobieństwo, że dokument w danej kategorii jest w podanym języku. Ważnym odnotowania jest, że w przybliżeniu pojedyncze kategorie możemy traktować jako spójne językowo, tzn. że dokument należący do danej kategorii jest prawie na pewno w takim języku jak reszta dokumentów w tej kategorii. PSTWO ING - prawdopodobieństwo obliczone na podstawie tekstów dostarczonych przez organizatorów, PSTWO MISIE - prawdopodobieństwo obliczone na podstawie tekstów wyekstrachowany przez nasz zespół.

- **File** - nazwa pliku (bez ścieżki oraz bez rozszerzenia),
- **Text ING** - odczytany tekst z plików dostarczony przez organizatorów,
- **Text Adam** - odczytany tekst z plików utworzony przez nasz zaspół,
- **Nrow** - liczba nowych linii odczytanych w pliku,
- **Text Adam pl** - odczytany tekst z plików bez polskich znaków utworzony przez nasz zespół,
- **IsPolish** - flaga, czy dany tekst jest zaklysyfikowany jako polski,
- **IsEnglish** - flaga, czy dany tekst jest zaklasyfikowany jako angielski,
- **IsOther** - flaga, czy dany tekst jest zaklasyfikowany jako inny niż polski lub angielski.

2.4.1 Dokumenty po polsku

Na podstawie zgromadzonych danych stworzyliśmy tabelę, którą wykorzystujemy do utworznia modelu na podstawie dokumentów polskich. Składają się na nią kolumny:

- **Nrow** - liczba nowych linii odczytanych w pliku,
- **string_length** - długość wybranego przez nas odczytu z pliku (opcje: Text ING, Text Adam, Text Adam pl),
- **capitalletters_ratio** - stosunek wielkich liter do wszystkich znaków,
- **numbers_count** - liczba cyfr w ciągu znaków,
- **question_marks_count** - liczba znaków zapytania w ciągu znaków,

- **currency_signs_count** - liczba znaków pieniężnych w ciągu znaków,
- **flag_x** - flaga, czy występuje konkretny wybrany po przeszukaniu zdjęć z danej kategorii jeden z siedmiu patternów (x to cyfra od 1 do 7),
- **flag_fuzzx** - flaga, czy występuje konkretny wybrany po przeszukaniu zdjęć z danej kategorii jeden z siedmiu patternów z dodaniem fuzz.partial_ratio na poziomie 70% (x to cyfra od 1 do 7).

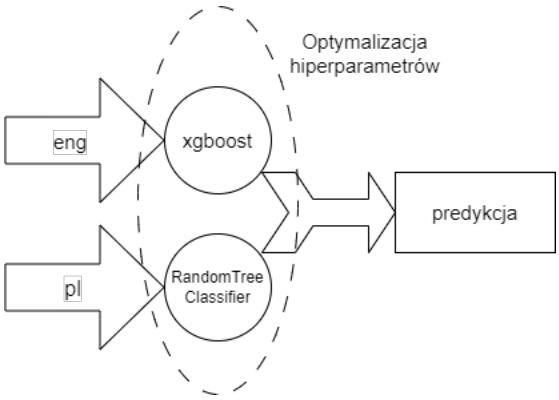
2.4.2 Dokumenty po angielsku

Do skonstruowania tabeli wykorzystywanej w modelowaniu zawierającej 55 kolumn posłużyliśmy się następującymi statystykami:

- Statystyki dotyczące tekstów takie jak w przypadków dokumentów angielskich (bez flag patternów i ich odpowiedników fuzz) - 7 kolumn,
- Flagi wyszukanych słów kluczowych - 43 kolumny,
- **numberEmpty** - procentowa ilość pustego tła na skanie,
- **nonWhiteFraction** - ilość niebiałych pikseli podzielona na ilość wszystkich pikseli w skanie,
- **possibleShapes** - liczba możliwych kształtów na skanie,
- **possibleImages** - liczba możliwych obrazów na skanie,
- **nonEmptySections** - liczba niepustych sekcji na obrazie po maskowaniu.

Pięć pogrubionych kolumn zostało przy pomocy naszej funkcji *generate_metrics*.

2.5 Model



Rysunek 5: Diagram przedstawiający proces modelowania przy podziale dokumentów na te w języku polskim i te w języku angielskim.

Modele jakie użyliśmy w naszym rozwiązaniu to:

- Dokumenty po polsku - **DecisionTreeClassifier**,
- Dokumenty po angielsku - **xgboost**.

Proces modelowania przedstawiony jest na diagramie 5.

3 Wyniki

3.1 Dokumenty po polsku

Accuracy:

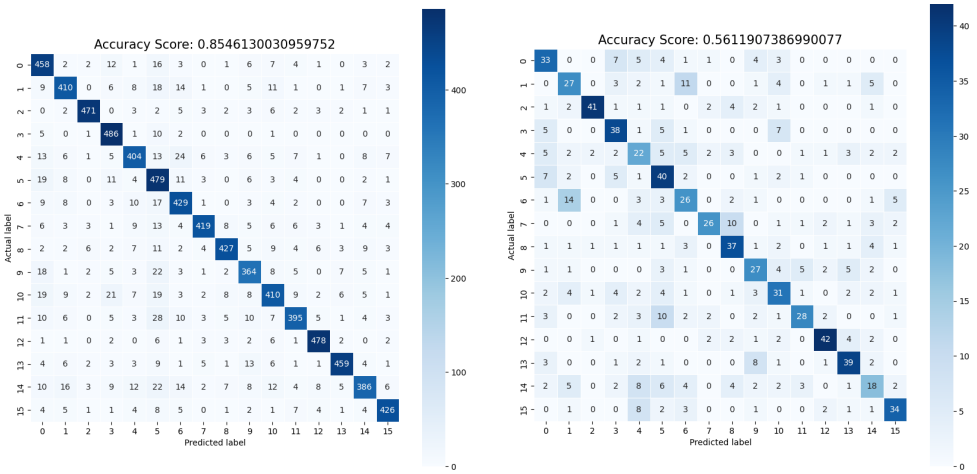
- Zbiór treningowy: 1.0
- Zbiór testowy: 0.99

Te wyniki potwierdziły słuszność podziału dokumentów według języków.

3.2 Dokumenty po angielsku

	Train	Test
Accuracy	0.85	0.56
F1 Score	0.86	0.56

Tabela 2: Tabela z wynikami zbioru treningowego i testowego.



Rysunek 6: Macierz błędów dla zbioru treningowego (na lewo) i zbioru testowego (na prawo).

4 Przyszłościowe Modyfikacje

Możliwości doskonalenia naszego rozwiązania w przyszłości:

- odczytywanie białego tekstu z czarnego tła,
- rozpoznanie większej ilości języków, a co za tym idzie, stworzenie większej ilości modeli.