

EXACTDATA

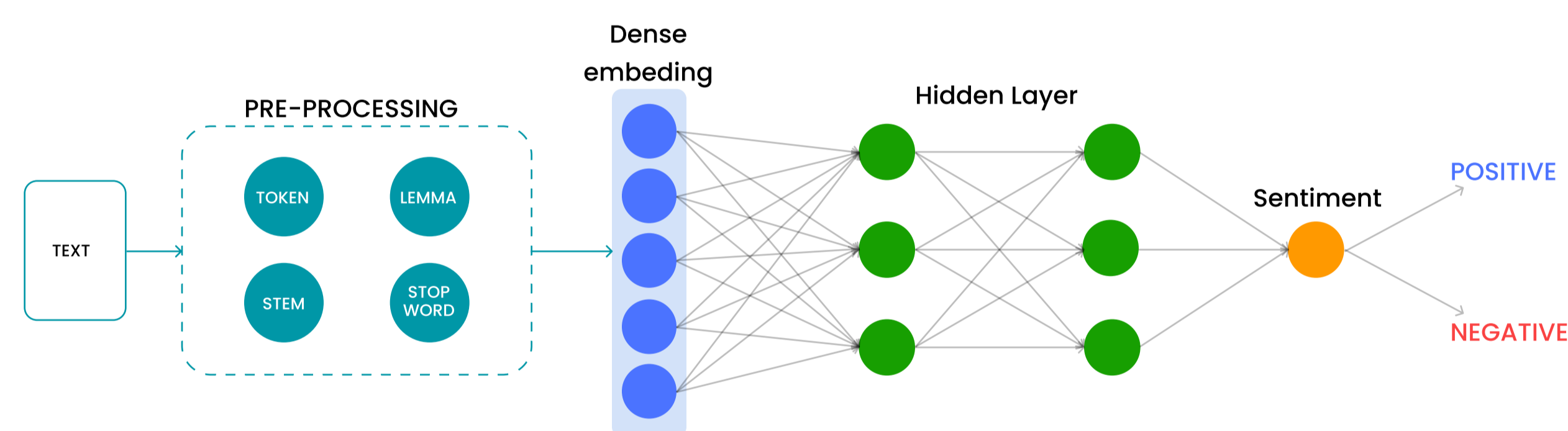
Aleksandra Gronowska, Aleksandra Jonas, Maciej Ścigacz, Szymon Jadczak
Uniwersytet im. Adama Mickiewicza w Poznaniu

O projekcie

- Projekt ExactData został stworzony z myślą o firmach, których celem jest poprawienie relacji z klientem poprzez dopasowanie do niego treści.
- System analizuje tworzone oraz już opublikowane posty w języku polskim dotyczące sprzedawanych produktów na platformach społecznościowych takich jak Facebook czy Instagram.
- Program po przeanalizowaniu tekstu zwraca statystyki wraz z wykresami dotyczącymi różnych aspektów posta - odbiór posta (w tym sprzedawanego produktu) przez użytkowników, sprawdza poprawność gramatyczną opublikowanej treści, analizuje komentarze pod kątem występowania ironii oraz interpretuje tekst reklamy pod kątem grup docelowych.

Analiza sentymentu w tekście

- Analiza sentymentu opiera się na problemie klasyfikacji tekstu. Jej zadaniem jest zaklasyfikowanie tekstu do odpowiedniej etykiety wydźwięku np. pozytywny, neutralny, negatywny.
- Model klasyfikujący tekst został oparty o uczenie głębokie, a dokładniej o MLP (wielowarstwowe sieci neuronowe).
- Sieć neuronowa uczy się z dostarczonych przykładów, dla których sentyment został już określony.



- Dane dostarczone do sieci należało odpowiednio przygotować. Zdania zostały oczyszczone z niepotrzebnych znaków. Usunięte zostały najczęściej występujące w języku słowa (nie wpływające na wydźwięk) oraz dokonano tokenizacji, lematyzacji i stemmingu. Następnie tokeny zostały poddane wektoryzacji.
- Wyuczony w ten sposób model sieci umożliwił klasyfikowanie tekstu, ze względu na jego wydźwięk emocjonalny.

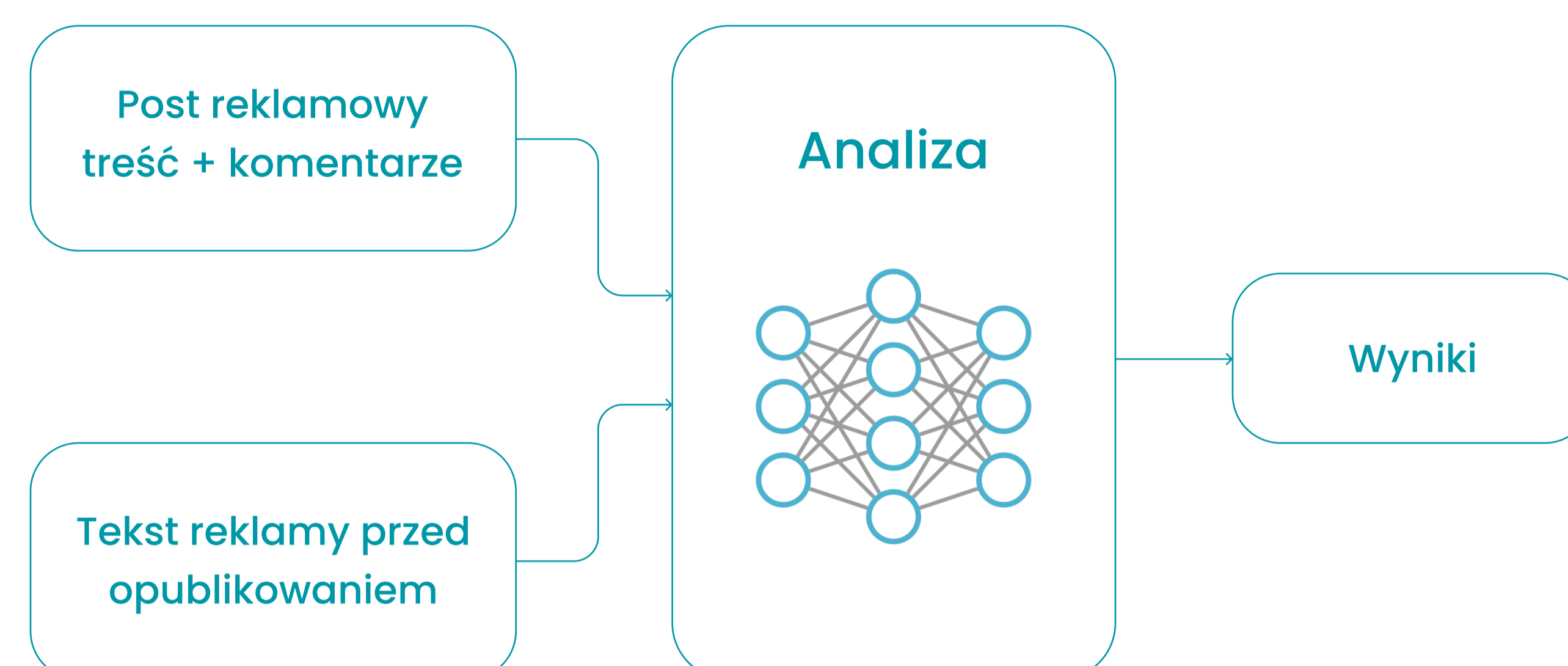
Wykrywanie ironii

- Wykrywanie ironii polega na klasyfikowaniu tekstów na ironiczne oraz nieironiczne.
- Model wykrywania ironii został wytrenowany w oparciu o zbiór treningowy zawierający ironiczne oraz nieironiczne komentarze wraz z oznaczeniem wydźwięku wypowiedzi.
- Podczas tworzenia modelu służącego do wykrywania ironii rozważono kilka metod klasyfikacji - między innymi: Naive Bayes, Regresja Logistyczna czy Drzewo Decyzyjne.



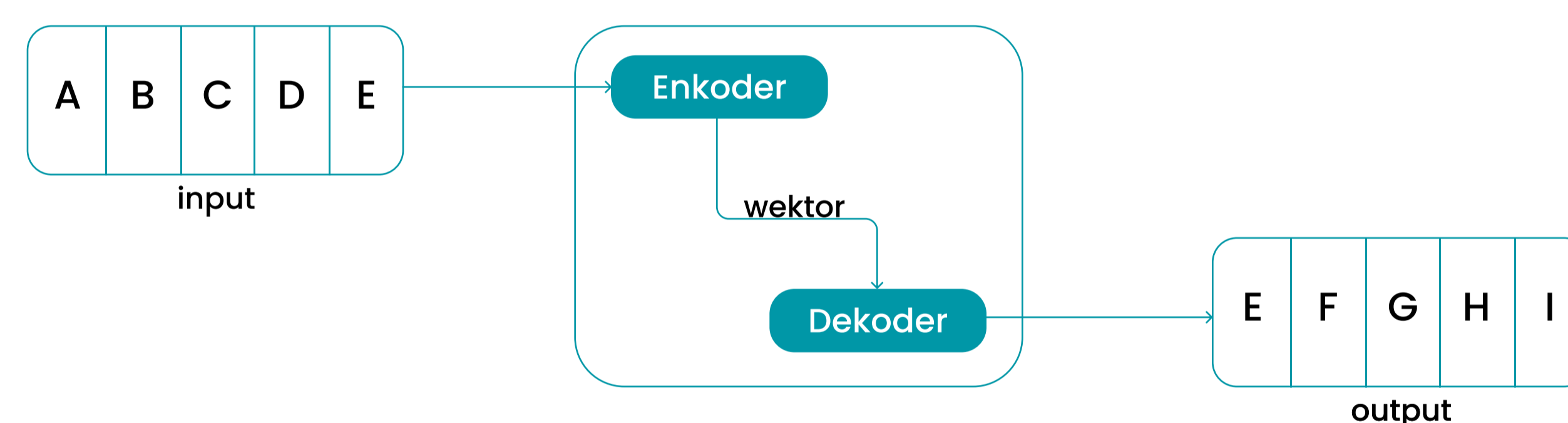
- Klasyfikacja ironii w badaniu reklamy jest potrzebna, aby zbadać nastawienie konsumenta do produktu i/lub firmy (producenta).

Schemat przepływu



Detekcja oraz poprawianie błędów w tekście

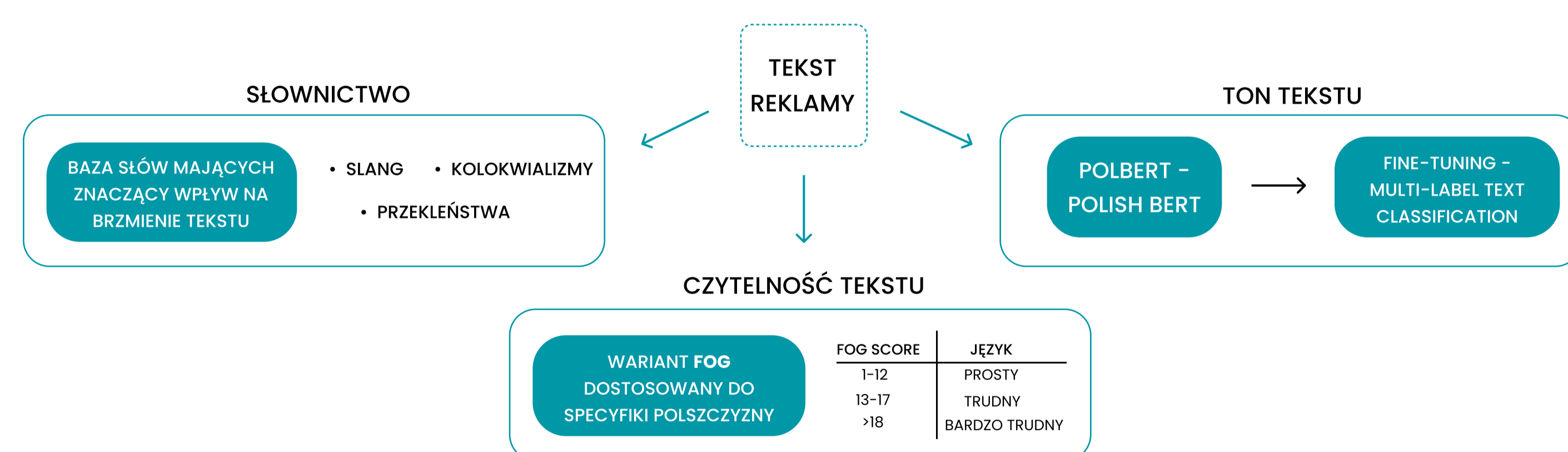
- Poprawianie błędów w naszym projekcie bazuje na podejściu seq2seq (sequence to sequence)
- Modele seq2seq to modele głębokiego uczenia osiągające wysokie wyniki w tłumaczeniu maszynowym, w streszczaniu tekstów czy klasyfikacji obrazów. Model seq2seq na wejściu przyjmuje sekwencję elementów (słów, liter, serii czasowych itp.) i wyprowadza inną sekwencję elementów.



- Enkoder przetwarza sekwencje wejściową, starając się "wylapać" wszystkie informacje o wejściu i tworzy z nich wektor kontekstu, następnie otrzymany wektor przekazywany jest do dekodera który tworzy sekwencje wyjściową.
- Model seq2seq w poprawianiu błędów w tekście wykorzystywany jest jako tłumaczenie z danego języka na ten sam język jednocześnie poprawiając błędy w danym tekście.

Analiza brzmienia tekstu

- Analiza brzmienia przeprowadzana jest pod kątem trzech kryteriów:



- Powyższe rozróżnienie składowych wpływających na odbiór tekstu pozwala nam na dokładniejszą charakterystykę reklamy.

JAKA JEST TWOJA REKLAMA?

- NIEFORMALNA
- INTRYGUJĄCA
- MŁODZIEŻOWA
- PRZYSTĘPNA