[NLP 25L] The Food Hazard Detection Challenge
Dokumentacja wstępna projektu

Michał Kaczmarczyk
Piotr Kitłowski
Hubert Podlewski

Kwiecień 2025

1 Wstęp

W ramach projektu podjęliśmy się wyzwania dziewiątego przedstawionego na tegorocznych (2025) warsztatach SemEval. Zadanie to znajduje się w kategorii „Fact Checking and Knowledge Verification". Zgodnie z opisem organizatorów, celem zadania jest utworzenie wyjaśnialnych systemów klasyfikacji tytułów raportów o wypadkach gastronomicznych. Wyjaśnialność systemu ma tu grać kluczową rolę, umożliwiając człowiekowi szybką weryfikację poprawności klasyfikacji.

Bezpieczeństwo żywności jest kluczowym zagadnieniem zdrowia publicznego – każdego roku na świecie odnotowuje się miliony przypadków zatruć i chorób wywołanych przez skażoną żywność. Tradycyjne metody monitorowania takich zagrożeń często opierają się na ręcznej analizie raportów przez ekspertów, co jest czasochłonne i podatne na błędy ludzkie. Z pomocą odpowiedniego modelu, który wykona żmudną część tego procesu, eksperci mogą skupić się na czysto merytorycznym procesie weryfikacji predykcji i wyjaśnień modelu.

Opracowanie zautomatyzowanych systemów wspomagających klasyfikację i weryfikację faktów może, między innymi, znacząco zminimalizować opóźnienie reakcji na kryzysy żywnościowe i ograniczyć ich skutki.

2 Zadania

W ramach projektu do wykonania są 2 podzadania:

- **ST1** - tekstowa klasyfikacja przewidująca kategorię zagrożenia i produktu, których jest kolejno 10 i 22.
- **ST2** - klasyfikacja przewidująca konkretne zagrożenie i konkretne produkty, których jest kolejno 128 i 1142.

3 Dane

3.1 Dostępne atrybuty

Atrybuty decyzyjne dostępne w dostarczonym przez organizatorów zbiorze danych to między innymi data i miejsce zgłoszenia, a także jego tytuł i treść. Domyślnie, w ramach wyzwania, uczestnicy wybierali, na podstawie których atrybutów dokonywać będą klasyfikacji i w zależności od tego trafiali do odpowiedniego rankingu. W naszym przypadku, wybór atrybutów będzie częścią eksperymentów, które będziemy prowadzić w ramach wykonywania projektu.

3.2 Etykiety

Punkty w zbiorze danym zawierają etykiety określające jakiego zagrożenia i jakiego produktu konkretny raport dotyczy. Te etykiety natomiast są dalej pogrupowane w szersze kategorie opisujące grupy zagrożeń lub produktów.

3.3 Ocena modelu

Do ogólnej oceny modelu wykorzystamy, tak jak w oryginalnym konkursie, makrouśrednioną metrykę F1, w skrócie - makro-F1. Taki wybór sposobu uśredniania najprawdopodobniej wynika z tego, że zbiór danych jest wysoce niezrównoważony. W takiej sytuacji mikrouśrednianie faworyzowałoby klasy częściej występujące w zbiorze danych.

Poniżej przykładowe użycie funkcji bibliotecznej w celu obliczenia tej metryki:

```
from sklearn.metrics import f1_score

def compute_score(haz_true, prod_true, haz_pred, prod_pred):
    f1_h = f1_score(haz_true, haz_pred, average='macro')
    f1_p = f1_score(prod_true[haz_pred==haz_true],
                    prod_pred[haz_pred==haz_true], average='macro')
    return (f1_h + f1_p) / 2
```

4 Metodyka

W ramach projektu zastosowaliśmy, do tej pory, trzy podejścia aby ocenić skuteczność coraz bardziej zaawansowanych metod klasyfikacji.

4.1 Majority Classifier

Najprostszą użytą metodą bazową był klasyfikator większościowy, który zawsze przewiduje najczęściej występujące klasy zagrożenia i produktu w danych treningowych. Choć nie jest to użyteczna strategia predykcyjna, stanowi istotny punkt odniesienia dla porównania z bardziej złożonymi modelami, szczególnie gdy klasy są niezbalansowane.

4.2 TF-IDF + Regresja Logistyczna

Drugie podejście opiera się na klasycznej metodzie reprezentacji tekstu jako n-gramów. Wykorzystaliśmy jednowyrazowe i dwuwyrazowe n-gramy z przetworzeniem TF-IDF (z parametrem `min_df=2`).

Następnie zbudowaliśmy dwa oddzielne modele regresji logistycznej: jeden dla klasyfikacji kategorii zagrożenia, drugi dla klasyfikacji kategorii produktu. Ocenę skuteczności przeprowadziliśmy za pomocą metryki makro-F1. Dla klasyfikacji produktu, metryka liczona była wyłącznie dla próbek, w których przewidywana kategoria zagrożenia była poprawna (`hazard_pred == hazard_true`).

4.3 BERT z dwiema głowami klasyfikującymi

Trzecim podejściem był model językowy `bert-base-uncased`, z dodatkowymi warstwami klasyfikacyjnymi dla przewidywania kategorii zagrożenia i produktu. Wspólna baza BERT była trenowana równolegle z dwiema głowami klasyfikacyjnymi. Model trenowany był przez 6 epok, z użyciem optymalizatora AdamW, współczynnika uczenia (ang. learning rate) `2e-5` oraz rozmiaru partii (ang. batch size) `16`. Dane wejściowe stanowiły stokenizowane tytuły raportów (z ograniczeniem długości `max_length=128`).

Po każdej epoce model był ewaluowany przy użyciu wspomnianej wcześniej metryki makro-F1.

5 Uzyskane wyniki

Rysunek 1: Porównanie wyników metod klasyfikacji

Podejście	F1 (hazard)	F1 (product)	Średnia
Majority Classifier	0.060	0.016	0.038
TF-IDF + Logistic Regression	0.513	0.464	0.489
BERT dual-head	0.667	0.618	0.642

Majority Classifier: Bardzo niska skuteczność (średnia 0.038) potwierdza, że metody uczące są konieczne dla osiągnięcia akceptowalnych wyników.

TF-IDF + Logistic Regression: Pomimo prostoty, podejście to osiągnęło solidną skuteczność (średnia 0.489). Model był w stanie dobrze rozróżniać klasy zagrożeń na podstawie klasycznej reprezentacji bag-of-words.

BERT: Najlepszy rezultat (średnia 0.642) osiągnięto dzięki modelowi BERT z podwójną głowicą klasyfikacyjną. Zauważono jednak fluktuacje wyników w epoce 5 i 6, co może sugerować ryzyko przeuczenia — konieczne może być dalsze monitorowanie lub wcześniejsze zatrzymanie treningu.

6 Dalsze kroki / Plan działania

Później, chcemy skorzystać z klasyfikatora XGBoost lub LLM Llama 3.1 od firmy Meta (poprzednio Facebook). Obecnie, nasze ogólne plany na następne kroki to:

- Badania klasyfikatorów zespołowych, „model ensembling”
- Implementacja i analiza wyjaśnialności modeli (ang. „explainability”) , np. LIME oraz SHAP