Wykrywanie mowy nienawiści (Cyberbullying) w polskich komentarzach z wykorzystaniem modelu HerBERT

Eryk Zerbin

Streszczenie

Celem projektu było stworzenie i wytrenowanie modelu uczenia maszynowego zdolnego do automatycznej detekcji hejtu (cyberbullyingu) w języku polskim. Wykorzystano architekturę Transformer (model HerBERT) oraz technikę Fine-tuningu. Ostateczny model osiągnął wynik F1-score na poziomie 0.93 na zbiorze testowym, co potwierdza jego wysoką skuteczność.

Spis treści

1 Opis Projektu

Wraz z rosnącą liczbą treści generowanych przez użytkowników w Internecie, problem mowy nienawiści (hate speech) oraz cyberprzemocy staje się coraz poważniejszy. Ręczna moderacja komentarzy jest czasochłonna i kosztowna, dlatego celem niniejszego projektu jest automatyzacja tego procesu przy użyciu nowoczesnych metod Przetwarzania Języka Naturalnego (NLP).

Projekt polega na wytrenowaniu modelu klasyfikacyjnego, który na podstawie treści komentarza określi, czy jest on neutralny, czy też nosi znamiona hejtu.

2 Definicja Problemu

Problem detekcji hejtu definiujemy jako zadanie **binarnej klasyfikacji tekstu** z nadzorem.

Niech X oznacza przestrzeń wszystkich możliwych ciągów znaków (komentarzy), a $Y = \{0, 1\}$ zbiór etykiet, gdzie:

- $y = 0$ oznacza komentarz neutralny,
- $y = 1$ oznacza hejt/cyberbullying.

Naszym celem jest znalezienie funkcji mapującej $f_\theta : X \to Y$, sparametryzowanej wektorem wag θ, która dla zadanego wejścia $x \in X$ przewiduje etykietę $\hat{y}$ tak, aby zminimalizować błąd predykcji względem etykiety rzeczywistej y.

Formalnie, model uczy się rozkładu prawdopodobieństwa warunkowego $P(Y|X)$. Decyzja klasyfikacyjna podejmowana jest na podstawie:

$$\hat{y} = \underset{c \in \{0,1\}}{\operatorname{argmax}} P(Y = c|x) \tag{1}$$

3 Dane Wejściowe i Wyjściowe

3.1 Źródło danych

Wykorzystano zbiór danych **BAN-PL**, zawierający autentyczne polskie komentarze z mediów społecznościowych. Zbiór ten jest reprezentatywny dla języka potocznego, zawierając błędy ortograficzne, slang oraz wulgaryzmy.

3.2 Przygotowanie danych (Preprocessing)

Dane zostały poddane wstępnemu przetwarzaniu:

1. Usunięcie zbędnych znaków białych (wielokrotne spacje, nowe linie).
2. Zachowanie znaczników specjalnych takich jak `{USERNAME}` czy `{URL}`. Jest to celowy zabieg, mający na celu zapobieganie przeuczeniu modelu na konkretne nazwy użytkowników (anonimizacja) i skupienie atencji modelu na kontekście wypowiedzi.

Dane podzielono na trzy rozłączne podzbiory (z zachowaniem stratyfikacji klas):

- **Zbiór treningowy (80%)**: używany do aktualizacji wag modelu.

- **Zbiór walidacyjny (10%)**: używany do monitorowania postępów podczas treningu.
- **Zbiór testowy (10%)**: używany wyłącznie do ostatecznej ewaluacji.

4 Opis Algorytmu i Metodyka

4.1 Architektura Modelu

W projekcie wykorzystano model **HerBERT** (`allegro/herbert-base-cased`), który jest polską adaptacją architektury BERT (Bidirectional Encoder Representations from Transformers).

Kluczowym elementem tej architektury jest mechanizm **Self-Attention** (atencja), który pozwala modelowi ważyć ważność poszczególnych słów w kontekście całego zdania równolegle, niezależnie od ich odległości w tekście. Mechanizm ten można zapisać wzorem:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Gdzie:

- Q (Query), K (Key), V (Value) to macierze powstałe z liniowej transformacji wejścia,
- d_k to wymiar wektora kluczy (czynnik skalujący).

4.2 Proces treningu (Fine-tuning)

Zastosowano technikę **Fine-tuningu** (dostrajania). Polega ona na wzięciu modelu wstępnie wytrenowanego na ogromnym korpusie języka polskiego i dotrenowaniu go na specyficznym zadaniu klasyfikacji hejtu. Do wyjścia modelu dodano warstwę liniową klasyfikującą, a całość trenowano przy użyciu funkcji kosztu **Cross-Entropy Loss**:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{3}$$

Gdzie $\hat{y}_i$ to prawdopodobieństwo klasy 1 zwrócone przez funkcję aktywacji Softmax.
Parametry treningu:

- Optymalizator: AdamW
- Learning rate: $2e-5$
- Batch size: 8
- Liczba epok: 3

5 Wyniki i Wnioski

5.1 Ewaluacja ilościowa

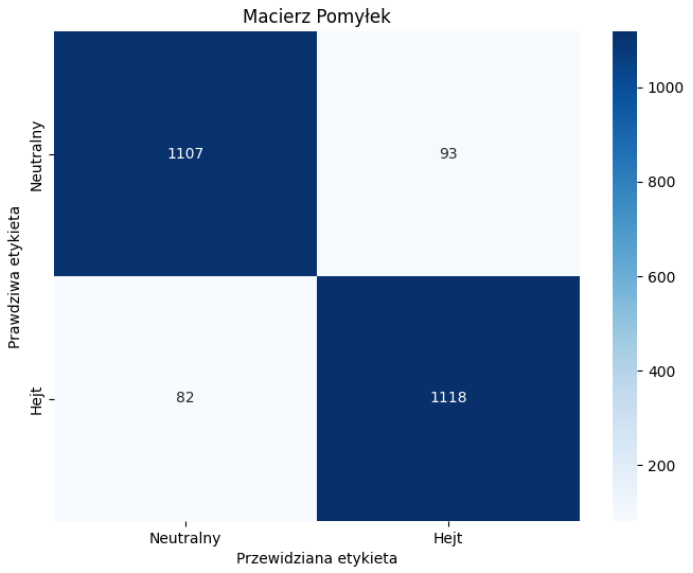
Model został przetestowany na niezależnym zbiorze testowym liczącym 2400 próbek. Uzyskano następujące wyniki:

Metryka	**Wartość**
Accuracy (Dokładność)	92.71%
Precision (Precyzja)	0.92
Recall (Czułość)	0.93
F1-Score	**0.93**

Tabela 1: Wyniki modelu na zbiorze testowym.

5.2 Macierz Pomyłek (Confusion Matrix)

Analiza macierzy pomyłek wskazuje na zrównoważone działanie klasyfikatora.



Rysunek 1: Macierz pomyłek wygenerowana na zbiorze testowym.

Szczegółowe dane z macierzy:

- **True Positives (Prawdziwy Hejt):** 1118
- **True Negatives (Prawdziwy Neutralny):** 1107
- **False Positives (Błędy fałszywego alarmu):** 93
- **False Negatives (Niewykryty hejt):** 82

5.3 Wnioski i analiza błędów

Osiągnięty wynik (F1 = 0.93) znacząco przewyższa założenia początkowe (0.5-0.7). Model wykazuje wysoką zdolność generalizacji.

Analiza błędnych predykcji wykazała, że model jest wrażliwy na specyficzne słowa kluczowe. Przykładowo, słowo "lala" (użyte w zdaniu "latam jak ta lala") zostało błędnie sklasyfikowane jako hejt. Wynika to z faktu, że w zbiorze treningowym słowo to występowało głównie w kontekście obraźliwym (np. "głupia lala"). Wskazuje to na lekkie przeuczenie modelu na poziomie tokenów (token overfitting), mimo poprawnego rozpoznawania kontekstu w większości przypadków (np. poprawne zignorowanie slangu "Xd" jako neutralnego).

6 Podsumowanie

Projekt zakończył się sukcesem. Zastosowanie modelu Transformer (HerBERT) pozwoliło na osiągnięcie wysokiej skuteczności w detekcji cyberbullyingu, spełniając z nawiązką wymagania postawione w projekcie.

Literatura

[1] Kołos, A., Okulska, I., Głąbińska, K., Karlińska, A., Wiśnios, E., Ellerik, P., Prałat, A. (2024). *BAN-PL: a Novel Polish Dataset of Banned Harmful and Offensive Content from Wykop.pl web service*. arXiv preprint arXiv:2308.1059. Dostępne pod adresem: `https://arxiv.org/abs/2308.1059`

[2] Mroczkowski, R., Rybak, P., Wróbel, M., Gawlik, I. (2021). *HerBERT: Efficiently Pretrained Transformer-based Language Model for Polish*. 8th Workshop on Balto-Slavic Natural Language Processing.

[3] Wolf, T., et al. (2020). *Transformers: State-of-the-Art Natural Language Processing*. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.