Anotacja korpusów oraz osadzenia słów i tekstów
Część I: Procedura anotacji

Autorzy: Oliwer Krupa, Adam Bednarski, Jan Masłowski, Łukasz Lenkiewicz

October 23, 2024

Rozdziały

1 Wybór Korpusu Tekstów

W celu przeprowadzenia zadania anotacji tekstów wybraliśmy korpus `poleval2019_cyberbullying` z HuggingFace Datasets. Korpus ten został opracowany w ramach konkursu PolEval 2019 i zawiera teksty w języku polskim dotyczące problematyki mowy nienawiści i cyberprzemocy. Zbiór został stworzony w celu oceny systemów do detekcji treści o charakterze nienawistnym i przemocowym w internecie. Składa się on z anonimowych postów i komentarzy z polskich mediów społecznościowych, które zostały ręcznie oznaczone pod kątem cyberprzemocy.

Korpus składa się z następujących elementów:

- **Posty i komentarze** - anonimowe wpisy pobrane z różnych platform internetowych.
- **Anotacje** - każdy post został ręcznie zaklasyfikowany jako zawierający lub niezawierający treści związane z mową nienawiści.

1.1 Specyfikacja Zbioru

Zbiór danych zawiera następujące cechy:

- Liczba przykładów: 10,000 wpisów i komentarzy.
- Struktura danych: Każdy wpis zawiera pole `text`, które reprezentuje zawartość tekstową, oraz pole `label`, które klasyfikuje wpis jako cyberprzemoc (`1`) lub brak cyberprzemocy (`0`).
- Język: Polski.

Więcej informacji na temat zbioru danych znajduje się na stronie projektu: `https://huggingface.co/datasets/poleval/poleval2019_cyberbullying`.

2 Wytyczne i Przeprowadzenie Anotacji Tekstów

2.1 Wprowadzenie

W celu oznaczenia i analizy danych dotyczących mowy nienawiści, zdecydowaliśmy się na wykorzystanie narzędzia `Docanno`. Docanno umożliwia intuicyjne i efektywne oznaczanie tekstu na różnych poziomach, co pozwala na realizację zadania zarówno na poziomie całych dokumentów, jak i ich fragmentów.

2.2 Notatka dla Anotatorów

Aby zapewnić spójność i jednolite podejście podczas procesu anotacji, przygotowaliśmy krótką notatkę zawierającą wytyczne dla anotatorów. Poniżej przedstawiamy instrukcje, które były stosowane przez wszystkie osoby zaangażowane w proces anotacji:

1. Oceniamy tweeta w taki sposób, że:
    - `0` - neutralny tweet,

- `1` - mowa nienawiści.

2. Oceniamy frazy tweeta, przypisując im odpowiednie etykiety słowne w zależności od ich wpływu na wydźwięk całego tweeta:
   - `4` - wzmacnianie,
   - `5` - odwracanie,
   - `6` - osłabianie.

Te wytyczne zapewniają, że anotatorzy zwracają uwagę zarówno na ogólny charakter wpisu, jak i na poszczególne frazy, które mogą mieć wpływ na ton całego tekstu.

2.3 Proces Anotacji

W ramach zadania anotacji, każdy anotator został poproszony o oznaczenie 100 wybranych postów, które zostały losowo wybrane z pełnego korpusu danych. Anotacja została przeprowadzona na dwóch poziomach:

- **Anotacja na poziomie całego tekstu** - ocena ogólnego wydźwięku tweeta jako neutralnego lub zawierającego mowę nienawiści.
- **Anotacja na poziomie poszczególnych fragmentów tekstu** - przypisanie odpowiednich etykiet frazom mającym wpływ na wydźwięk tweeta.

2.4 Dobre Praktyki Anotacji

Podczas procesu anotacji zastosowaliśmy następujące dobre praktyki:

- **Niezależność anotacji** - każdy anotator pracował niezależnie, co zapewnia brak wpływu innych osób na ocenę wpisów.
- **Losowe próbkowanie** - próbka 100 tweetów została losowo wybrana z pełnego zbioru danych, co zwiększa obiektywizm oceny.
- **Klarowne wytyczne** - dzięki jednoznacznym zasadom anotacji, anotatorzy mieli jasność co do sposobu oceny tweetów i ich fragmentów.

2.5 Podsumowanie

W wyniku procesu anotacji, każdy wpis w próbce został oznaczony zarówno na poziomie całego tekstu, jak i poszczególnych fraz. Wszystkie wyniki anotacji zostały zapisane w plikach JSONL, które zostaną wykorzystane do dalszej analizy. Załączone pliki obejmują oznaczone dane dla każdego anotatora.

3 Analiza Zgodności Anotatorów

3.1 Kappa Cohena

W celu oceny zgodności pomiędzy annotatorami w zadaniu klasyfikacji postów obliczono wartość Kappy Cohena. Kappa Cohena to statystyczna miara zgodności, która uwzględnia nie tylko zgodność rzeczywistą między annotatorami, ale także zgodność przypadkową. Jest to bardziej zaawansowana miara w porównaniu z procentową zgodnością, ponieważ eliminuje wpływ losowości w przypisywaniu kategorii, co czyni ją bardziej miarodajną w analizie wyników.

Macierz konfuzji: Na podstawie wyników oznaczania stworzono następującą macierz konfuzji, która pokazuje liczbę przypadków, w których annotatorzy przypisali takie same lub różne etykiety:

	Janek: Neutral	**Janek: Hate**
Adam: Neutral	89	4
Adam: Hate	1	6

Table 1: Macierz konfuzji dla oznaczeń Adama i Janka

Z tej macierzy wynika, że annotatorzy w 89 przypadkach przypisali kategorię "Neutral", natomiast w 6 przypadkach zgodnie przypisali kategorię "Hate". W 4 przypadkach annotator Adam przypisał "Hate", podczas gdy Janek przypisał "Neutral", natomiast w 1 przypadku było odwrotnie.

Obliczenie Kappy Cohena: Funkcja użyta do obliczenia Kappy Cohena bazuje na macierzy konfuzji, która pokazuje zgodności i niezgodności pomiędzy annotatorami. Na podstawie tej macierzy funkcja oblicza, jaka część zgodności wynika z rzeczywistych decyzji annotatorów, a jaka mogła być przypadkowa. Wynik, zwany Kappą Cohena, przedstawia stopień zgodności po uwzględnieniu przypadkowej zgodności.

Kappa Cohena w tym przypadku wynosi $\kappa = 0.679$, co wskazuje na umiarkowaną zgodność pomiędzy annotatorami. Część zgodności można przypisać przypadkowi, ale annotatorzy są w znacznym stopniu zgodni, co czyni wyniki wiarygodnymi, choć nie doskonałymi.

Wyniki: Wartości macierzy konfuzji oraz obliczona Kappa Cohena zostały podsumowane poniżej:

Wartość	**Opis**
89	Obaj annotatorzy przypisali "Neutral"
6	Obaj annotatorzy przypisali "Hate"
4	Adam przypisał "Hate", Janek "Neutral"
1	Adam przypisał "Neutral", Janek "Hate"
Kappa Cohena	0.679

Table 2: Wyniki obliczenia Kappy Cohena

Dyskusja wyników: Wartość Kappy Cohena wskazuje na umiarkowaną zgodność annotatorów. Obserwowana zgodność wynosi 95%, co sugeruje wysoką zgodność pomiędzy annotatorami. Jednak przewidywana zgodność losowa była na poziomie 87%, co oznacza, że część zgodności mogła wynikać z przypadku. Dlatego wynik $\kappa = 0.679$ wskazuje, że annotatorzy byli bardziej zgodni, niż wynikałoby to z przypadku, ale ich zgodność nie jest pełna. W związku z tym, aby zwiększyć spójność wyników, wskazane jest przeprowadzenie dalszej analizy oraz ewentualnie kolejnej iteracji anotacji, szczególnie w przypadkach, gdzie niezgodność była wyraźna.

3.2 Procent Częściowej Zgodności (PPA) dla pierwszej anotacji

Aby dokładniej ocenić zgodność między anotatorami, obliczono Procent Częściowej Zgodności (PPA). PPA to miara, która uwzględnia stopień częściowej zgodności między dwoma anotatorami podczas etykietowania tekstów lub ich fragmentów. W odróżnieniu od Kappy Cohena, która ocenia ogólną zgodność, PPA skupia się na częściowym pokrywaniu się anotacji, co jest szczególnie użyteczne, gdy mamy do czynienia z wieloma etykietami lub nakładającymi się zakresami.

3.2.1 PPA dla Całych Zdań

Na podstawie macierzy konfuzji możemy obliczyć PPA dla anotacji na poziomie zdań:

- Adam i Janek zgodzili się w 89 przypadkach co do etykiety „Neutral".
- Zgodzili się w 6 przypadkach co do etykiety „Hate" (mowa nienawiści).
- W 4 przypadkach Adam przypisał etykietę „Hate", podczas gdy Janek ocenił te same zdania jako „Neutral".
- W 1 przypadku Adam przypisał etykietę „Neutral", a Janek „Hate".

Wzór na obliczenie PPA dla zdań to:

$$PPA_{\text{zdania}} = \frac{2 \times \text{Zgoda na Neutral i Hate}}{2 \times \text{Zgoda na Neutral i Hate} + \text{Adam: Hate, Janek: Neutral} + \text{Adam: Neutral, Janek: Hate}} \times 100 \tag{1}$$

Podstawiając wartości:

$$\begin{aligned} PPA_{\text{zdania}} &= \frac{2 \times 89}{2 \times 89 + 4 + 1} \times 100 \\ &= 97.27\% \end{aligned} \tag{2}$$

Wskazuje to na wysoki poziom zgodności między Adamem a Jankiem na poziomie całych zdań.

3.2.2 PPA dla Frazy

Następnie przeanalizowano zgodność na poziomie fraz. Po odfiltrowaniu zdań oznaczonych jako „Hate" lub „Neutral", skupiono się na frazach, takich jak „Wzmacnianie" czy „Osłabianie", i oceniono, jak często Adam i Janek zgodzili się co do ich etykietowania:

- Obaj anotatorzy oznaczyli tę samą frazę: 6 przypadków.
- Janek oznaczył frazę, Adam tego nie zrobił: 4 przypadki.
- Adam oznaczył frazę, Janek tego nie zrobił: 7 przypadków.

Wzór na obliczenie PPA dla fraz to:

$$PPA_{\text{frazy}} = \frac{2 \times \text{Obaj zgodni}}{2 \times \text{Obaj zgodni} + \text{Janek oznaczył, Adam nie} + \text{Adam oznaczył, Janek nie}} \times 100 \tag{3}$$

Podstawiając wartości:

$$\begin{aligned} PPA_{\text{frazy}} &= \frac{2 \times 6}{2 \times 6 + 4 + 7} \times 100 \\ &= 52.17\% \end{aligned} \tag{4}$$

Ta niższa wartość odzwierciedla trudności w osiągnięciu pełnej zgodności na bardziej szczegółowym poziomie fraz.

3.2.3 Średni PPA

Aby podsumować zgodność na poziomie zarówno zdań, jak i fraz, obliczono średnią z obu wartości PPA:

$$\begin{aligned} PPA_{\text{średnie}} &= \frac{PPA_{\text{zdania}} + PPA_{\text{frazy}}}{2} \\ &= \frac{97.27 + 52.17}{2} \\ &= 74.72\% \end{aligned} \tag{5}$$

Ogólny PPA sugeruje stosunkowo wysoki poziom zgodności między anotatorami, z wyraźnie większą spójnością na poziomie całych zdań w porównaniu do bardziej szczegółowych anotacji fraz.

4 Powtórna Anotacja na Nowej Próbce Danych

W procesie powtórnej anotacji na nowej próbce danych wprowadzono zaktualizowane wytyczne, mające na celu precyzyjniejszą ocenę treści tweetów oraz ich fraz. Nowa wersja notatki dla anotatorów, oznaczona jako **Notatka dla Anotatorów 2.0**, zawiera następujące zasady:

1. Ocena ogólna tweetów:

- Tweet jest oceniany na dwóch poziomach:
  - **0** - tweet neutralny, nie zawierający treści związanych z mową nienawiści.
  - **1** - tweet zawierający mowę nienawiści.

2. Ocena fraz wewnątrz tweetów: Po zaklasyfikowaniu tweeta jako zawierającego mowę nienawiści (**1**), anotator ocenia poszczególne frazy tweeta, biorąc pod uwagę ich wpływ na wydźwięk tweeta:

- **Wzmacnianie (4)** - frazy, które wzmacniają negatywny ton tweeta.
- **Odwracanie (5)** - frazy, które zmieniają kierunek emocjonalny tweeta, łagodząc negatywny ton.
- **Osłabianie (6)** - frazy, które osłabiają negatywny ton tweeta.

3. Ograniczenia:

- Wzmacnianie, osłabianie i odwracanie dotyczy tylko tweetów, które zostały zaklasyfikowane jako zawierające mowę nienawiści. W przypadku tweetów neutralnych (0), nie oceniamy wpływu fraz na wydźwięk.

Zaktualizowane wytyczne w wersji 2.0 mają na celu bardziej precyzyjną ocenę wpływu poszczególnych fraz w tweetach zawierających mowę nienawiści, co pozwala na głębszą analizę treści i tonu wpisów.

5 Analiza Zgodności Drugiej Anotacji

5.1 Kappa Cohena

W celu oceny zgodności pomiędzy annotatorami w drugiej rundzie anotacji obliczono wartość Kappy Cohena. Anotacje zostały przeprowadzone na nowej próbce danych, a wyniki miały na celu sprawdzenie, czy zmiany w wytycznych dla anotatorów wpłynęły na zgodność ich ocen.

Macierz konfuzji: Na podstawie wyników anotacji stworzono macierz konfuzji, która przedstawia liczbę przypadków, w których annotatorzy przypisali zgodne lub różne etykiety:

	Janek: Neutral	**Janek: Hate**
Adam: Neutral	86	1
Adam: Hate	3	10

Table 3: Macierz konfuzji dla drugiej anotacji

Obliczenie Kappy Cohena: Kappa Cohena obliczona na podstawie powyższej macierzy wyniosła $\kappa = 0.811$. Wynik ten wskazuje na wysoką zgodność między annotatorami, co sugeruje, że zaktualizowane wytyczne były skuteczne w ujednoliceniu ocen.

Wyniki: Wynik Kappy Cohena na poziomie 0.811 oznacza znaczną poprawę w porównaniu do pierwszej anotacji. Obserwuje się wysoką zgodność, co może być efektem lepszego zrozumienia i stosowania się do nowych wytycznych przez annotatorów. .

5.2 Procent Częściowej Zgodności (PPA) dla Drugiej Anotacji

Dla drugiej rundy anotacji, PPA zostało ponownie obliczone na podstawie danych z plików `labeled_sample_adam_2.jsonl` oraz `labeled_sample_jan_2.jsonl`. Podobnie jak w przypadku pierwszej anotacji, obliczono PPA zarówno na poziomie zdań, jak i fraz.

5.2.1 PPA dla Całych Zdań

Na podstawie macierzy konfuzji poniżej możemy obliczyć PPA dla zdań:

$$\text{Macierz konfuzji} = \begin{bmatrix} 10 & 1 \\ 3 & 85 \end{bmatrix}$$

Macierz ta przedstawia następujące liczby przypisanych etykiet:

- 10 przypadków, w których zarówno Adam, jak i Janek przypisali etykietę "Hate".
- 85 przypadków, w których obaj anotatorzy przypisali etykietę "Neutral".
- 1 przypadek, gdzie Adam przypisał etykietę "Hate", a Janek "Neutral".
- 3 przypadki, gdzie Adam przypisał etykietę "Neutral", a Janek "Hate".

Wzór na obliczenie PPA dla zdań pozostaje taki sam:

$$\begin{aligned} PPA_{\text{zdania}} &= \frac{2 \times (10 + 85)}{2 \times (10 + 85) + 1 + 3} \times 100 \\ &= 97.27\% \end{aligned} \tag{6}$$

Wskazuje to na bardzo wysoki poziom zgodności między anotatorami na poziomie całych zdań, podobnie jak w pierwszej rundzie anotacji.

5.2.2 PPA dla Frazy

Po analizie fraz, PPA zostało obliczone na następującej podstawie:

- W 8 przypadkach obaj anotatorzy przypisali tę samą etykietę dla fraz.
- W 8 przypadkach Janek przypisał etykietę dla frazy, a Adam tego nie zrobił.
- W 6 przypadkach Adam przypisał etykietę dla frazy, a Janek tego nie zrobił.

Wzór na obliczenie PPA dla fraz to:

$$\begin{aligned} PPA_{\text{frazy}} &= \frac{2 \times 8}{2 \times 8 + 8 + 6} \times 100 \\ &= 53.33\% \end{aligned} \tag{7}$$

Wynik PPA dla fraz jest niższy niż w przypadku zdań, co sugeruje większą trudność w osiągnięciu zgodności na bardziej szczegółowym poziomie fraz.

5.2.3 Średni PPA

Średnią PPA dla zdań i fraz obliczono w następujący sposób:

$$\begin{aligned} PPA_{\text{średnie}} &= \frac{PPA_{\text{zdania}} + PPA_{\text{frazy}}}{2} \\ &= \frac{97.27 + 53.33}{2} \\ &= 75.3\% \end{aligned} \tag{8}$$

Średnia zgodność między anotatorami wynosi 75.3%, co jest wynikiem zbliżonym do pierwszej rundy anotacji, z wyższą zgodnością na poziomie zdań niż fraz.

6 Wnioski wynikające z przeprowadzonych analiz

Jednym z głównych problemów, który pojawił się podczas procesu anotacji, była różnica w liczbie zaznaczonych słów przez różnych anotatorów, mimo że ogólny sens anotacji był taki sam. Różnice te wynikały głównie z niejednolitego oznaczania zakresów słów, co utrudniało pełną zgodność.

Analiza wyników wskaźników PPA oraz Kappy Cohena pozwala na wyciągnięcie kilku kluczowych wniosków:

Przeprowadzenie drugiej iteracji anotacji przyniosło wyraźną poprawę zgodności między anotatorami, szczególnie w kontekście oznaczania fraz. Średnia wartość PPA wzrosła z 74,72% w pierwszej iteracji do 75,3% w drugiej, co wskazuje na większą spójność w etykietowaniu zarówno całych zdań, jak i ich fragmentów. Zmiany te sugerują, że uaktualnione wytyczne dla anotatorów pomogły lepiej zrozumieć kryteria anotacji, co wpłynęło na poprawę jakości ocen.

Warto podkreślić, że wskaźnik PPA dla anotacji na poziomie zdań pozostał na wysokim poziomie, wynosząc 97,27% zarówno w pierwszej, jak i drugiej iteracji. Świadczy to o dużej zgodności anotatorów w ocenie całościowego wydźwięku wpisów. Z kolei poprawa PPA dla fraz (z 52,17% do 53,33%) wskazuje na pewien postęp w bardziej szczegółowej analizie treści, choć pełna zgodność na poziomie fraz wciąż stanowi wyzwanie.

Podsumowując, wyniki sugerują, że wprowadzenie nowej iteracji anotacji oraz bardziej precyzyjnych instrukcji było krokiem we właściwym kierunku, poprawiającym spójność ocen między anotatorami, szczególnie w analizie bardziej złożonych fragmentów tekstu. Mimo to, dalsze udoskonalenie procesu anotacji może być konieczne, aby zminimalizować rozbieżności w ocenie poszczególnych fragmentów tekstu i osiągnąć jeszcze wyższą zgodność.

7 Podsumowanie Zbioru Danych za Pomocą Statystyk Opisowych

W celu lepszego zrozumienia struktury anotacyjnej i charakterystyki tweetów w zbiorze danych, obliczono szereg statystyk opisowych dla każdej grupy anotacji. Statystyki te obejmują liczbę tweetów, średnią i medianę długości wpisów, odchylenie standardowe, a także rozkład wpisów zaklasyfikowanych jako neutralne lub zawierające mowę nienawiści. Analiza tych danych umożliwia ocenę spójności, różnorodności i charakteru próbek danych przypisanych poszczególnym anotatorom.

7.1 Statystyki dla Anotacji Adama i Jana

Każdy z anotatorów pracował na niezależnej próbce danych, co oznacza, że ich zestawy tweetów różniły się zarówno pod względem treści, jak i charakterystyki klasyfikacji. W tabelach 4 i 5 przedstawiono szczegółowe statystyki dotyczące tweetów oznaczonych przez Adama i Jana, umożliwiające porównanie długości wpisów oraz przydzielonych etykiet.

Table 4: Statystyki dla Anotacji Adama

Statystyka	Wartość
Liczba tweetów	50
Średnia długość (słowa)	12.39
Mediana długości (słowa)	12.00
Odchylenie standardowe	4.44
Najkrótszy wpis (słowa)	6
Najdłuższy wpis (słowa)	23
Liczba tweetów neutralnych	0
Liczba tweetów z mową nienawiści	0
Wpisy z wieloma etykietami	0

Table 5: Statystyki dla Anotacji Jana

Statystyka	Wartość
Liczba tweetów	50
Średnia długość (słowa)	12.44
Mediana długości (słowa)	11.00
Odchylenie standardowe	5.58
Najkrótszy tweet (słowa)	6
Najdłuższy tweet (słowa)	25
Liczba tweetów neutralnych	45
Liczba tweetów z mową nienawiści	5
Wpisy z wieloma etykietami	3

Zarówno Adam, jak i Jan pracowali na różnych próbkach danych, co skutkuje różnicami w wynikach klasyfikacji. Zbiór anotacji Adama charakteryzuje się brakiem wpisów zaklasyfikowanych jako neutralne lub zawierające mowę nienawiści, co sugeruje, że jego próbka danych mogła składać się z bardziej jednolitych treści. Z kolei Jan, pracując na innej próbie, sklasyfikował 45 tweetów jako neutralne i 5 jako zawierające mowę nienawiści, co wskazuje na większą różnorodność treści w jego zbiorze.

Pomimo różnic w klasyfikacji, długość tweetów w obu próbkach była dość zbliżona, o czym świadczą podobne wartości średniej długości (12.39 słów u Adama i 12.44 słów u Jana). Odchylenie standardowe jest nieco wyższe w przypadku anotacji Jana, co sugeruje większą rozpiętość długości wpisów w jego próbce.

7.2 Podsumowanie

Mimo że Adam i Jan pracowali na różnych próbkach danych, można zauważyć kilka wspólnych elementów, takich jak podobna długość tweetów. Główna różnica dotyczy rozkładu

etykiet – Adam nie zaklasyfikował żadnych tweetów jako neutralnych lub zawierających mowę nienawiści, natomiast Jan zaklasyfikował znaczną liczbę tweetów jako neutralne i kilka jako zawierające mowę nienawiści.

Wyniki te sugerują, że charakterystyka próbek danych mogła mieć wpływ na decyzje anotatorów. Wprowadzenie lepszych wytycznych w kolejnych iteracjach anotacji mogło pomóc w bardziej spójnej klasyfikacji między anotatorami, jednakże różnice w próbkach danych nadal mogą wpływać na interpretację wyników. W przyszłych analizach warto dążyć do dalszej standaryzacji próbek oraz optymalizacji wytycznych, aby zminimalizować różnice w klasyfikacji wynikające z samego charakteru danych.