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

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

October 15, 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.

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. .

6 Podsumowanie Zbioru Danych za Pomocą Statystyk Opisowych

W celu zwięzłego podsumowania pozyskanego zbioru danych, obliczono szereg statystyk opisowych, które pozwalają na lepsze zrozumienie struktury danych anotacyjnych oraz charakterystyki tweetów. Statystyki te obejmują liczbę wpisów, średnią i medianę długości wpisów, odchylenie standardowe, a także liczby tweetów zaklasyfikowanych jako neutralne oraz zawierające mowę nienawiści.

6.1 Statystyki dla Anotacji Łukasza

Dla anotacji Łukasza zebrano 100 tweetów. W tabeli 4 przedstawiono statystyki opisowe dotyczące długości wpisów oraz klasyfikacji.

6.2 Statystyki dla Anotacji Adama

Anotacje Adama obejmują 49 tweetów. W tabeli 5 przedstawiono statystyki opisowe dotyczące długości wpisów oraz klasyfikacji.

Table 4: Statystyki dla Anotacji Łukasza

Statystyka	Wartość
Liczba tweetów	100
Średnia długość (słowa)	12.66
Mediana długości (słowa)	12.50
Odchylenie standardowe	4.37
Najkrótszy wpis (słowa)	4
Najdłuższy wpis (słowa)	22
Liczba tweetów neutralnych	0
Liczba tweetów z mową nienawiści	0
Wpisy z wieloma etykietami	0

Table 5: Statystyki dla Anotacji Adama

Statystyka	Wartość
Liczba tweetów	49
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

6.3 Statystyki dla Anotacji Jana

W przypadku anotacji Jana zebrano 50 tweetów. W tabeli 6 przedstawiono statystyki opisowe dotyczące długości wpisów oraz klasyfikacji.

6.4 Podsumowanie

Z zebranych statystyk wynika, że dane anotacyjne różnią się pod względem liczby wpisów, długości tweetów oraz klasyfikacji na wpisy neutralne i zawierające mowę nienawiści. Największy zbiór anotacji pochodzi od Łukasza, jednak w jego danych oraz w danych Adama nie zaklasyfikowano żadnych tweetów jako neutralnych lub zawierających mowę nienawiści. Z kolei zbiór anotacji Jana wykazuje różnorodność pod względem klasyfikacji, z przewagą tweetów neutralnych.

Table 6: 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