PROJEKT
Analiza danych tekstowych

Dominika Schabowska
Mateusz Pompa

28.06.2023

Zbiór danych

W celu zdobycia danych i poddania ich analizie, skorzystaliśmy z platformy Kaggle, link do danych podajemy poniżej:

$$https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch$$

Zbiór danych jest zbiorem postów, kolekcjonowanych na platformie Reddit pod tematami "SuicideWatch" i "depression". Posty są zbierane za pomocą Pushshift API.

Pushshift to platforma do zbierania, analizy i archiwizacji danych w mediach społecznościowych, która od 2015 roku gromadzi dane Reddita i udostępnia je badaczom. Zestaw danych Reddit Pushshift jest aktualizowany w czasie rzeczywistym i zawiera dane historyczne z początków Reddita.

Wszystkie posty, które zostały wprowadzone do "SuicideWatch" od 16 grudnia 2008 (utworzenie) do 2 stycznia 2021, zostały zebrane, podczas gdy posty "depression" zostały zebrane od 1 stycznia 2009, do 2 stycznia 2021.

Nasz oryginalny zbiór danych, na którym przeprowadzamy analizę, składa się z 111342 obserwacji. Zestawione są w postaci tabeli o trzech kolumnach. Pierwsza kolumna zawiera indeksy danych, które usuwamy, ponieważ w naszym modelu będą nieprzydatne. Druga gromadzi wiadomości tekstowe, do których, w trzeciej kolumnie, przypisane są etykiety 'suicide', 'non-suicide'.

Suicide	Non-Suicide
I don't want anyone to find my bodyI just can't do it, because someone has to see it to confirm the death. :(	How do I change my reddit username? Is so cringe ._.
Feeling alone with my suicidal feelingsMy suicidal thoughts are slowly eating away at me. They never stop. I feel trapped. I feel scared and I feel alone	I believe in rebirth... Yes I do because various people die at young age so what about their deeds where will it get completed?.
Final postIf anyone wants to talk or something before it's too late	Does anyone else feel empty and sad even after they had a great day? bruh what's wrong with me
I just want to be happyI am too tired to try to kill myself but at the same time just want the pain to end.	Hey, im gonna sleep with socks Whatcha gonna do? Put them off?! Good luck ima gonna sleep with warm feet

Rysunek 1: Przykładowe dane tekstowe

Cel analizy

Warto zauważyć, że praca ze zbiorami danych związanych ze zdrowiem psychicznym może być delikatna i trudna, ponieważ wiąże się z przetwarzaniem informacji osobistych i potencjalnie wrażliwych. Ważne jest, aby podchodzić do tych danych z troską i szacunkiem oraz aby wszelkie badania i analizy były prowadzone w sposób etyczny i odpowiedzialny.

Potencjalny samobójca często świadomie, bądź nie, daje znaki swojemu otoczeniu o złym samopoczuciu i bezsilności. Jednym ze sposobów wyrażania emocji są wiadomości wysyłane do bliskich lub posty udostępniane na portalach społecznościowych. Często tego typu wiadomości są lekceważone, podczas gdy osoba, która je pisze potrzebuje prawdziwej pomocy. Oczywiście analiza tego typu tekstów nie jest łatwa, a odróżnienie tekstu samobójcy od tekstu osoby, która ma gorsze samopoczucie danego dnia, czy niepoprawnie żartuje, okaże się trudna techniką. Jednak w czasach, gdy technologia i machine learning się płynnie rozwija, oczekujemy, aby również pomagała w relacjach międzyludzkich.

Wiele organizacji oraz specjalistów szuka sposobu na różnorodną pomoc osobom z depresją, czy innymi chorobami lub zaburzeniami psychicznymi. Jedną z propozycji będzie aplikacja, czy ustawienia na forach społecznościowych, które wychwycą martwiące wiadomości tekstowe. Pojawienie się nowych technologii, które z jednej strony mogą pomóc rozpoznać zagrożonego samobójstwem Użytkownika, wymagającego natychmiastowej pomocy, a z drugiej mogą być nośnikami treści terapeutycznych i równocześnie narzędziem do monitorowania jego aktualnego stanu. Nasz model ma na celu wychwycenie takich tekstów, które świadczą o zagrożeniu samobójstwem.

Częste sygnały dotyczące myśli samobójczych, jak rozpoznać w wiadomościach:

- rozmowy kierowane są na tematykę śmierci i poczuciu braku sensu
- uczucie przepracowania, przepalenia i ciągły stres
- rezygnowanie z aktywności, które do tej pory sprawiały przyjemność
- strach przed przyszłością lub brak perspektywy przyszłości
- odrzucenie bliskich i izolowanie się
- działania pochopne i niebezpieczne, nadużywanie używek
- niska samoocena, brak motywacji
- żegnanie się w wiadomościach

Naiwna klasyfikacja Bayesowska

Do naszego modelu analizowania danych tekstowych, wykorzystaliśmy klasyfikator bayesowski. Metoda, która wykorzystuje twierdzenie Bayesa o prawdopodobieństwie warunkowym, wylicza prawdopodobieństwo wystąpienia pewnej cechy w danej klasie dla nowej obserwacji. Jej idea jest prosta i często stosowana do klasyfikacji tekstu.

Nasz model uczy się na części danych z naszej tabeli, tzw zbiór uczący. Wynajduje słowa najczęściej występujące, jako cecha i klasyfikuje je do wiadomości 'suicide' i 'non-suicide', czyli wylicza prawdopodobieństwo $P(cecha|klasa)$. Następnie dla nowych obserwacji wylicza prawdopodobieństwo przydzielenia danej klasy pod warunkiem cechy. Na podstawie słów i ich klasyfikacji, które poznał przy zbiorze uczącym, przydziela nowym wiadomością daną klasyfikacje, jeśli odpowiednie słowa w niej występują.

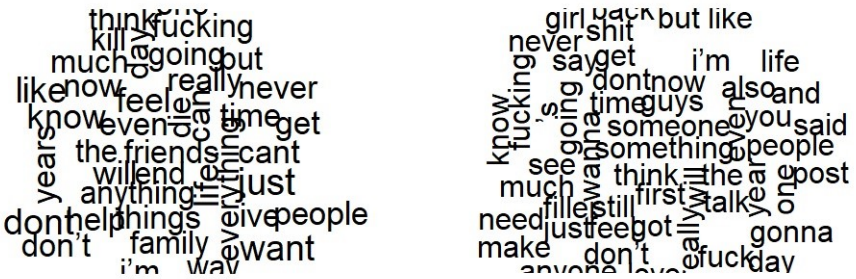
Przygotowanie danych oraz budowa modelu

Na początek, w celu zwiększenia sprawności kodu, będziemy rozważać pierwsze 15000 komórek danych. Analizę bazy danych rozpoczęliśmy od jej oczyszczenia. Pierwszym etapem obróbki będzie usunięcie pierwszej kolumny z indeksami, ponieważ będziemy chcieli analizować tylko wiadomości tekstowe.

	X	text	class
1	2	Ex Wife Threatening SuicideRecently I left my wife f...	suicide
2	3	Am I weird I don't get affected by compliments if it's...	non-suicide
3	4	Finally 2020 is almost over... So I can never hear "2...	non-suicide
4	8	i need helpjust help me im crying so hard	suicide
5	9	I'm so lostHello, my name is Adam (16) and I've bee...	suicide
6	11	Honetly idkI dont know what im even doing here. I j...	suicide
7	12	[Trigger warning] Excuse for self inflicted burns*I do...	suicide
8	13	It ends tonight.I can't do it anymore. I quit.	suicide
9	16	Everyone wants to be "edgy" and it's making me sel...	non-suicide
10	18	My life is over at 20 years oldHello all. I am a 20 ye...	suicide

Rysunek 2: Zbiór danych tekstowych

Skorzystamy z reprezentacji worka słów (ang. bag-of-words), czyli modelu w którym tekst jest reprezentowany jako worek słów, z pominięciem gramatyki czy kolejności słów. Model worka słów jest powszechnie stosowany w metodach klasyfikacji dokumentów, w których częstotliwość występowania każdego słowa jest wykorzystywana jako cecha do szkolenia klasyfikatora. Następnie konwertujemy zmienną jakościową *class*, która przyjmuje wartości *non-suicide* oraz *suicide* na czynnik. Nasze wiadomości to tekst złożony z słów, liczb, emotek, spacji i znaków interpunkcyjnych. Zanim zajmiemy się właściwym czyszczeniem wiadomości musimy stworzyć tzw. korpus danych, który jest zbiorem tekstów o odpowiedniej strukturze gotowej do dalszej analizy. Następnie oczyszczamy go ze zbędnych słów oraz znaków. Na tym etapie stworzyliśmy macierz dla naszego korpusu która przechowuje ilość wystąpień danego słowa w danym dokumencie tekstowym. Podzieliliśmy dane na 70% zbiór uczący i 30% na zbiór testowy. Stworzyliśmy zbiór uczący i testowy oraz etykiety dla tych zbiorów. Przeanalizowaliśmy chmury wyrazów, które prezentujemy poniżej:



Rysunek 3: Chmurka wyrazów dla klasyfikacji 'suicide' (po lewej), 'non-suicide' (po prawej)

A następnie zbudowaliśmy model oparty na metodzie naiwnej klasyfikacji bayesowskiej, który przewiduje dla nowej obserwacji do jakiej klasyfikacji należy; 'suicide' albo 'non-suicide'.

Ocena modelu

Otrzymaliśmy:
-macierz z testowymi wiadomościami (do sklasyfikowania suicide/non-suicide),
-wektor z etykietami do klasyfikacji wiadomości,
-zbudowany model do klasyfikacji

Po wykonaniu predykcji dla wiadomości otrzymaliśmy poniższe wyniki w postaci tabeli krzyżowej:

	actual		
predicted	non-suicide	suicide	Row Total
non-suicide	2031	802	2833
	0.451	0.178	
suicide	212	1455	1667
	0.047	0.323	
Column Total	2243	2257	4500

Rysunek 4: Tabela krzyżowa

Jak możemy zauważyć model znacznie lepiej klasyfikuje wyniki 'non-suicide', w przybliżeniu 90% skuteczności, niż wyniki 'suicide', których prawdopodobieństwo wynosi niespełna 65%. W najlepszym przypadku chcemy, aby model skupiał się na klasyfikowaniu wiadomości 'suicide', ponieważ to reakcja na nie jest ważna.

Nasz model uczy się przydzielać słowa do kategorii 'suicide' albo 'non-suicide'. Problem może pojawić się, gdy słowa z kategorii 'suicide' nie pojawią się w nowej obserwacji analizowanej przez nasz model. Rozwiązać tę wadę powinno wygładzenia Laplace'a. Zbudowaliśmy więc kolejny model z parametrem Laplace'a = 1 i otrzymaliśmy poniższe wyniki:

	actual		
predicted	non-suicide	suicide	Row Total
non-suicide	2036	811	2847
	0.452	0.180	
suicide	207	1446	1653
	0.046	0.321	
Column Total	2243	2257	4500

Rysunek 5: Tabela krzyżowa po wygładzeniu Laplace'a

Niestety wygładzenie Laplace'a nie polepszyło wyniku naszego modelu, ani zmiana parametrów, również zmniejszenie czy zwiększenie analizowanych danych tekstowych. W dalszej części raportu omówimy charakterystykę modelu i zbioru danych.

Podsumowanie

Możemy zauważyć, że dużą specyfikacją w naszych danych jest część wspólna zbioru słów, która należy do klasyfikacji 'suicide' oraz 'non-suicide'. W naszym modelu skupiliśmy się na oczyszczeniu danych z tych słów, tj. "don't", "got", "the" etc, ponieważ nie wnoszą żadnej wartości. Wiadomości są prywatne i nieoficjalne, więc autor tekstu nie zwracał uwagi na poprawność ortograficzną czy gramatyczną, pojawiały się emotikony, co spowodowało kłopot z dokładnym oczyszczeniem danych.

Skuteczność naszych wyników może zakłócać założenie, że cechy są niezależne i mają taki sam wpływ na nasze dane. Jest to założenie naiwnej klasyfikacji Bayesowskiej, która w tym przypadku może okazać się jej słabą stroną. Intuicyjnie analizując wiadomość skupimy się na kontekście zdania, a słowom dopasowujemy pewną wagę. W naszych danych pojawiają się wiadomości z cytatami piosenek lub losowo wypisane słowa niemające żadnego logicznego kontekstu. Wiadomości które nie są klasyfikowane 'suicide', ale pisane w żartach czy emocjach, przykładowo "im gonna die alone". Takie przypadki mogą zakłócać analizę naszego modelu.