Czy Instagram to dobre miejsce do pozyskiwania danych i wiedzy dla problemu klasyfikacji sentymentu

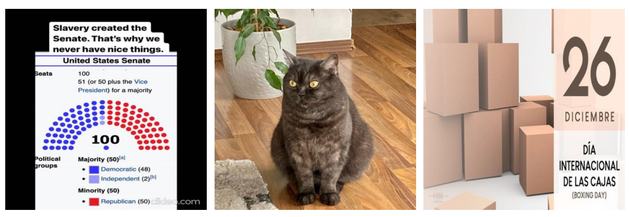
Filip Drewnowski, Vladimir Zaigrajew
PWr WIT, Sztuczna Inteligencja

Wstęp

- W branży danologa bardzo ważnym aspektem są dane, które posiadamy. Dzięki danym można tworzyć modele uczenia maszynowego bądź przeprowadzać analizy ukazujące informacje o świecie.
- Jednym z największych problemów danologów w przypadku widzenia komputerowego jest częsty brak adnotacji i potrzeba dużego zbioru danych, aby nauczyć AI do sprecyzowanych zadań. Mimo przeróżnych technik jak transfer learning liczba zdjęć i ich adnotacja przekracza rozmiary tysiąca przypadków dla większości zadań.
- Zagadnienie wykrywania sentymentu jest bardzo znane w dziedzinie nauki języka, jednak w przypadku widzenia komputerowego jest to nadal wymagający i nie rozwikłany problem.
- Problemy związane z wykrywaniem sentymentu na obrazkach można rozdzielić głównie na dwa pod problemy. Pierwszy z nich to problem małej dostępności potrzebnych zbiorów danych, natomiast drugi to trudność związana z adnotacją zdjęć.
- Adnotacja jest czasochłonnym i skomplikowanym zadaniem ze względu na stronniczość ludzką i ciężkość wyciągnięcia kontekstu z samego zdjęcia, ponieważ to samo zdjęcia może wpływać na różnych ludzi inaczej.

Dane i Adnotacja

Rys. 1:(a) Pozytywny sentyment (b) Neutralny sentyment (c) Negatywny sentyment



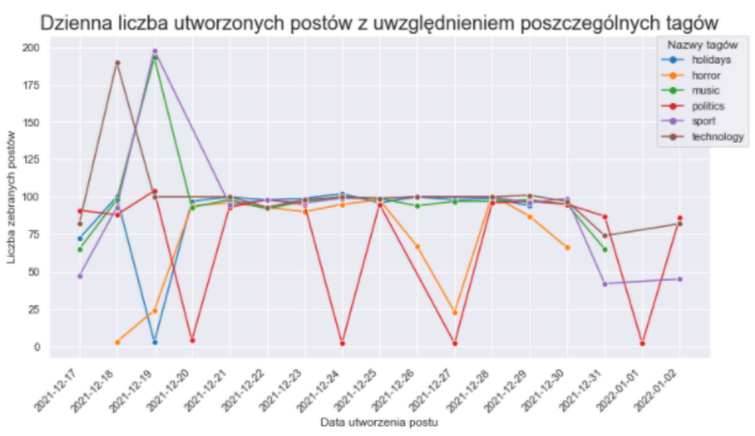
Rys. 2:(a) Nieznany sentyment (b) Nieznany sentyment (c) Nieznany sentyment

- Żeby zaradzić problemowi małych ilości danych jak i trudności adnotacji skorzystamy z danych które posiada platforma Instagram.
- Platformy sieci społecznościowych zawierają w dzisiejszych czasach biliony danych o użytkownikach, które można wykorzystać w celach naukowych bądź badawczych.
- Instagram to platforma która zawiera ogrom danych z różnych dziedzin. Przeciętna struktura danych instagrama to wpisy, które zawierają zdjęcia, opis i komentarze.

Pozyskanie danych

Proces pozyskiwania danych z platformy Instagram składał się z następujących kroków:

- Selekcja tagów poddanych analizie
- Przygotowanie trzech użytkowników na platformie Instagram
- Przygotowanie skryptu pobierającego około 100 postów co godzinę dla jednego z tagów
- Automatyczna adnotacja sentymentu pobranych postów na podstawie opisu, komentarzy bądź tagu przy użyciu dostępnych modeli do przetwarzania języka naturalnego
- Przeprowadzenie procesu zbierania danych przez okres 23 dni



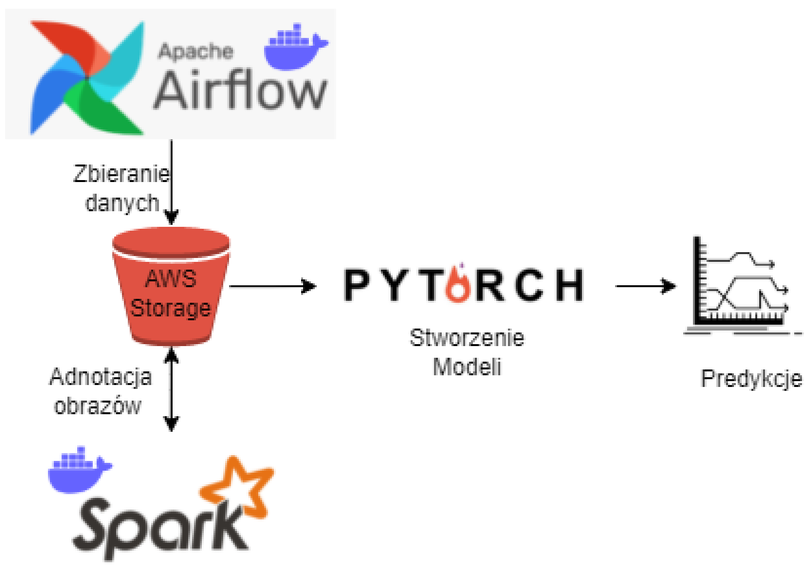
Rys. 3:Dokładny przebieg zbierania danych dla wyznaczonych tagów przez okres 23 dni. Dane zbierane między 17 grudnia 2021, a 2 stycznia 2022.

Informacje kontaktowe

- Strona katedry: `http://kio.pwr.edu.pl/`
- filip.drewnowski@gmail.com
- vladimirzaigrajew@gmail.com

Środowisko pracy

W tej sekcji zaprezentowano diagram jak wyglądało środowisko pracy. Z platformy Instagram dane zostały zebrane przy użyciu narzędzia Airflow i umieszczone w AWS Storage, następnie przy użyciu narzędzia Spark wpisy zostały adnotowane i podzielone w odpowiednie zbiory danych. W ostatnim etapie zbiory danych zostały pobrane na lokalne maszyny. w ostatnim kroku nastąpiło stworzenie i wyuczenie modeli oraz wykorzystanie ich w zadaniu predykcji sentymentu dla popularnego zbioru danych.

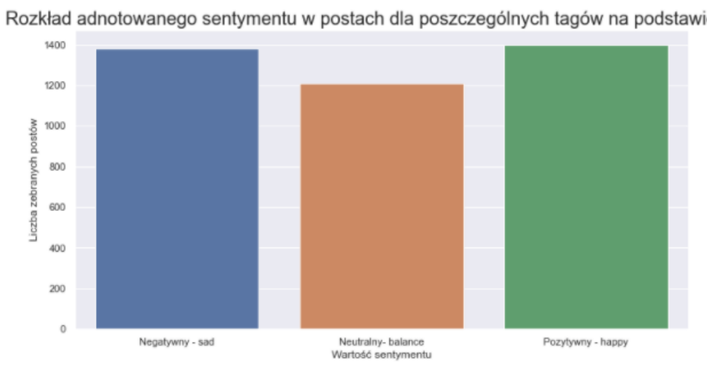


Rys. 4:Poszczególne komponenty rozwiązania i oddziaływanie między sobą

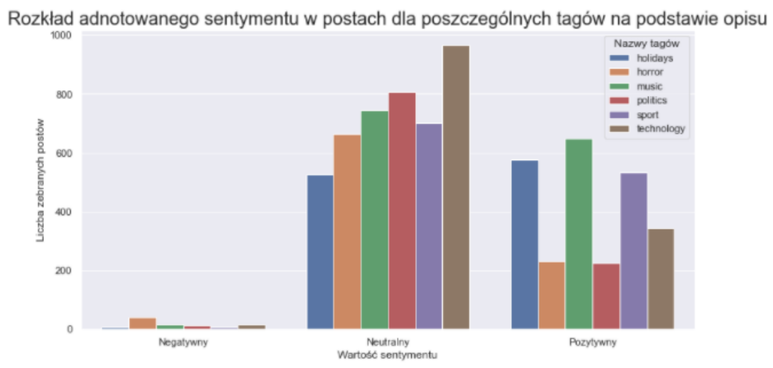
Adnotacja sentymentu dla obrazów

Poniżej przedstawiono jak rozkładała się wartość sentymentu w pobranych danych. Adnotacja obrazów była prowadzona trzema metodami:

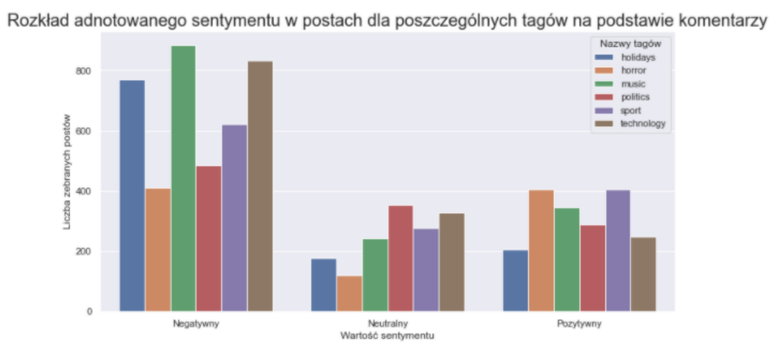
1. Adnotowanie na podstawie tagu - na podstawie wydźwięku danego tagu automatycznie przypisywana jest wartość sentymentu dla obrazu.
2. Adnotowanie na podstawie opisu - metoda, która wykorzystuje wyuczony model wykrywania wartości sentymentu z tekstu i określa go na podstawie opisu danego wpisu.
3. Adnotowanie na podstawie komentarzy - metoda, która tak jak w poprzednim rozwiązaniu wykorzystuje wyuczony model wykrywania sentymentu z tekstu, aby obliczyć średnią wartość sentymentu na podstawie komentarzy zamieszczonych pod wpisem.



Rys. 5:Rozkład uzyskanych danych dla zbioru adnotowanego na podstawie tagu



Rys. 6:Rozkład uzyskanych danych dla zbioru adnotowanego na podstawie postu



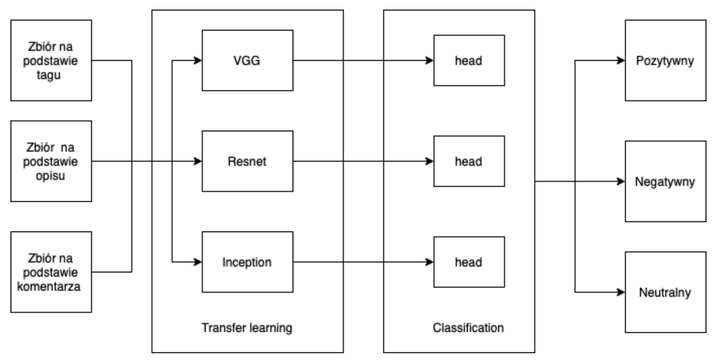
Rys. 7:Rozkład uzyskanych danych dla zbioru adnotowanego na podstawie komentarzy

Wykorzystanie zebranych danych

W tej części skorzystaliśmy z trzech powszechnych modeli do klasyfikacji obrazów:

- Resnet
- VGG
- Inception

Do przetestowania jak dobrze znane modele radzą sobie z zadaniem klasyfikacji sentymentu na uzyskanych zbiorach. Dodatkowo przetrenowane modele były testowane na publicznym zbiorze danych **The Sentiment Dataset** w celu sprawdzenia jakości zebranych zbiorów. Autorzy testowego zbioru danych wskazują, że najlepsze wyniki jakie udało im się osiągnąć dla ważonej miary F1 to 89.07.



Rys. 8:Proces trenowania modeli

Testowanie modeli

W tabeli poniżej przedstawiono wartości miary osiągnięte przez wyuczone modele dla zbioru testowego. Miary, które zostały zaprezentowane to zbiorcza ważona miara F1 oraz miara F1 dla poszczególnej klasy pozytywnego, neutralnego i negatywnego sentymentu (w nawiasie).

Model	Sentyment przypisany na podstawie tagu	Sentyment pozyskany z opisu	Uśredniony sentyment z komentarzy
VGG	0.01 (0.15 0.0 0.0)	0.23 (0.00 0.57 0.00)	0.35 (0.00 0.00 0.68)
resnet	0.16 (0.16 0.20 0.10)	0.23 (0.00 0.57 0.00)	0.35 (0.00 0.00 0.68)
inception	0.09 (0.16 0.06 0.11)	0.23 (0.00 0.57 0.00)	0.35 (0.00 0.00 0.68)

Wnioski

Problem tworzenia zbiorów jak i ich adnotowanie jest ciężkim i pracochłonnym procesem. Występuje dużo technik, które pozwalają zmniejszyć nakład pracy jednak nadal poziom złożoności procesu jest bardzo wysoki. Wykorzystanie mediów społecznościowych pozwala na zdobycie dużej ilości danych jak i wiedzy o tych danych, dlatego w celu pozyskania i adnotacji danych została wykorzystana wiedza z Instagram. Zbiory uzyskane z tej platformy były jednak mało reprezentatywne co bezpośrednio miało wpływ na końcowy etap, gdzie skorzystano z modeli do klasyfikacji sentymentu na podstawie obrazu. Pierwszy zbiór stworzony na podstawie tagu najgorzej reprezentuje dane co jest wynikiem naiwnego założenia, że zdjęcia z postów posiadają taką samą wartość sentymentu co tagi do nich przypisane. Mimo tego, że możliwe było uzyskanie zbalansowanego zbioru to zbiór na podstawie takiej adnotacji nie nadawał się do zadania klasyfikacji. Kolejny zbiór oparty na opisach postów od początku wskazywał na tendencje, że opisy są pisane w pozytywnym, bądź neutralnym kontekście. Dzieje się tak, bo twórcy starają się opisywać swoje zdjęcia w nienegatwyny sposób. Przez co rozkład klas w zbiorze jest niezbalansowany, a to prowadzi do niskiej jakościowo reprezentacji zbioru. Ostatnie podejście to zbiór adnotowany na podstawie komentarzy, który przynosił najwięcej informacji, ponieważ wykorzystana została tutaj uśredniona opinia użytkowników dla danego wpisu. Jednak uzyskane wyniki dla zbioru opartego o komentarze nadal nie sugerują, że tak wyciągnięta informacja odpowiednio opisuje zdjęcie, które jest mu przypisane.

Podsumowanie

Media społecznościowe to miejsce, w którym można pozyskać masowo wiele danych i wiedzy na temat świata. Jednak treści prezentowane w mediach cyfrowych mogą posiadać inny wydźwięk niż zdjęcia, które zostały zrobione bez konkretnego kontekstu. Jak pokazały badania informacje, które te media prezentują nie zawsze są dobrymi źródłami danych do niektórych problemów naukowych.

Powiązane prace

- The Sentiment dataset `https://datasets.simula.no/image-sentiment/`
- Model wykrywający sentyment z tekstu VaderSentiment `https://pypi.org/project/vaderSentiment/`
- Pretrenowane modele CNN z biblioteki PyTorch `https://pytorch.org/vision/stable/models.html`