Recepta na #popularność

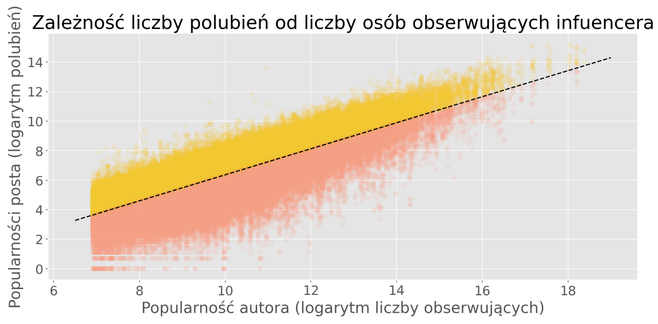
1. Opis problemu

Instagram jest potężnym medium społecznościowym, gdzie codziennie udostępniane są miliony różnorodnych postów. Każdy post składa się z centralnego elementu, jakim jest zdjęcie lub grafika, i dodatkowych informacji w postaci opisu lub hasztagów. Wiele osób marzy o zdobyciu sławy lub zarabianiu na życie za pośrednictwem Instagrama, jednak nie ma jednoznacznej recepty na sukces. Ocenienie, czy dany post "załapie", jest trudne i ciężko jest przewidzieć, co będzie miało największe znaczenie. Spójrzmy więc co na ten temat mówią czyste dane. Czy możemy przewidzieć popularność postu wyłącznie na podstawie jego cech?

2. Metodyka

Ze zbioru danych Instagram Influencer Dataset wybraliśmy reprezentatywną próbkę postów.
Następnie przeprowadziliśmy analizę eksploracyjną danych, by móc uzyskać odpowiedzi na postawioną hipotezę. Zaproponowaliśmy także trzy modele do predykcji popularności postów.

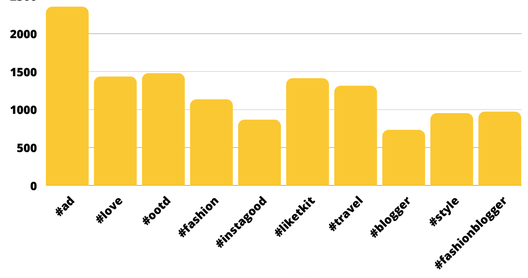
Zamiast modeli regresji, które słabo radzą sobie z wartościami o różnych rzędach wielkości, postanowiliśmy podzielić wszystkie posty na "popularne" i "niepopularne", a następnie wykonać klasyfikacje binarną.

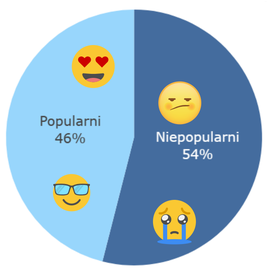


Ponieważ ilość polubień jest silnie skorelowana z liczbą osób obserwujących autora, za podział postów posłużyła nam regresja liniowa. Posty nad linią są rozpatrywane jako popularne, a pod linią znajdują się posty niepopularne.

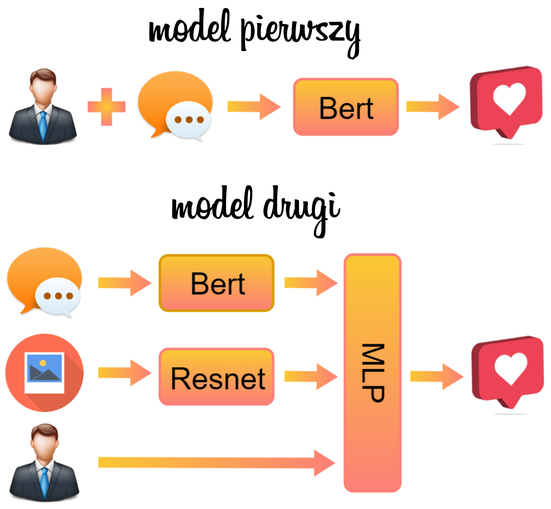
3. Dane

- **Instagram Influencer Dataset**
- **272000 postów**
- **zdjęcia, tekst, metadane postu i dane użytkownika**
- **>1000 followersów na użytkownika**
- **średnio 8 postów na osobę**





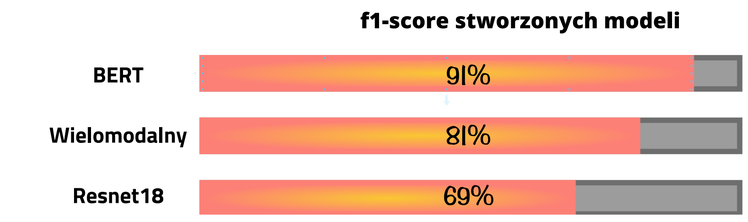
4. Modele

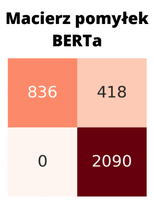


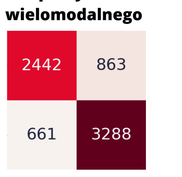
Zaproponowaliśmy dwa podejścia:

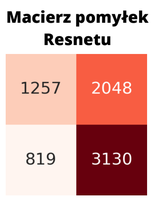
- **douczanie modelu językowego BERT w wersji wielojęzycznej. W tym celu dane tabelaryczne na temat autora i posta musiały zostać przekazane w postaci języka naturalnego.**
- **Model wielomodalny składający się z BERTa wielojęzycznego, Resnet18 oraz MLP. Resnet został wyuczony osobno i został wykorzystany jako ekstraktor cech. Natomiast BERT nie podlegał uczeniu.**

5. Wyniki





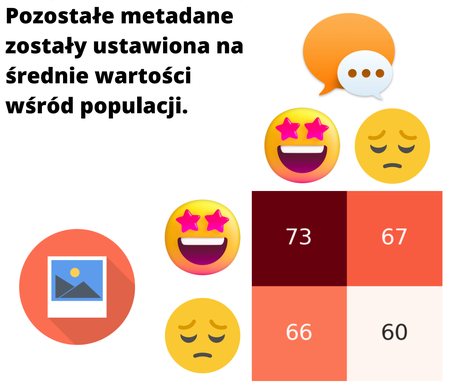




- **Najlepsze wyniki osiągnął BERT douczany do problemu.**
- **Model wielomodalny, mimo korzystania z pełnych danych osiągnął nieco gorsze rezulataty niż BERT.**
- **Resnet18 uzyskał aż 69% mimo działania wyłącznie na obrazach, co pokazuje, że zdjęcia mają znaczenie.**

6. Analizy

W celu sprawdzenia istotności poszczególnych mód, dokonaliśmy krzyżowej zamiany zdjęć oraz tekstów zaczerpniętych z postów o skrajnie niskim i wysokim współczynniku popularności. Pozostałe metadane zostały ustawiona na średnie wartości wśród populacji.



Z przeprowadzonego badania wynikło, że zarówno tekst jak i zdjęcia mają bardzo duży wpływ na popularność posta.

Autorzy:
Wojciech Kozłowski 246992@student.pwr.edu.pl
Sylwia Zemsta 269869@student.pwr.edu.pl
Opiekun:
Tomasz Kajdanowicz

Politechnika Wrocławska

Katedra Sztucznej Inteligencji