

Klasyfikacja wydźwięku wypowiedzi za pomocą uczenia maszynowego

Paweł Narolski

Praca pod kierunkiem dr. inż. Piotra Sygi

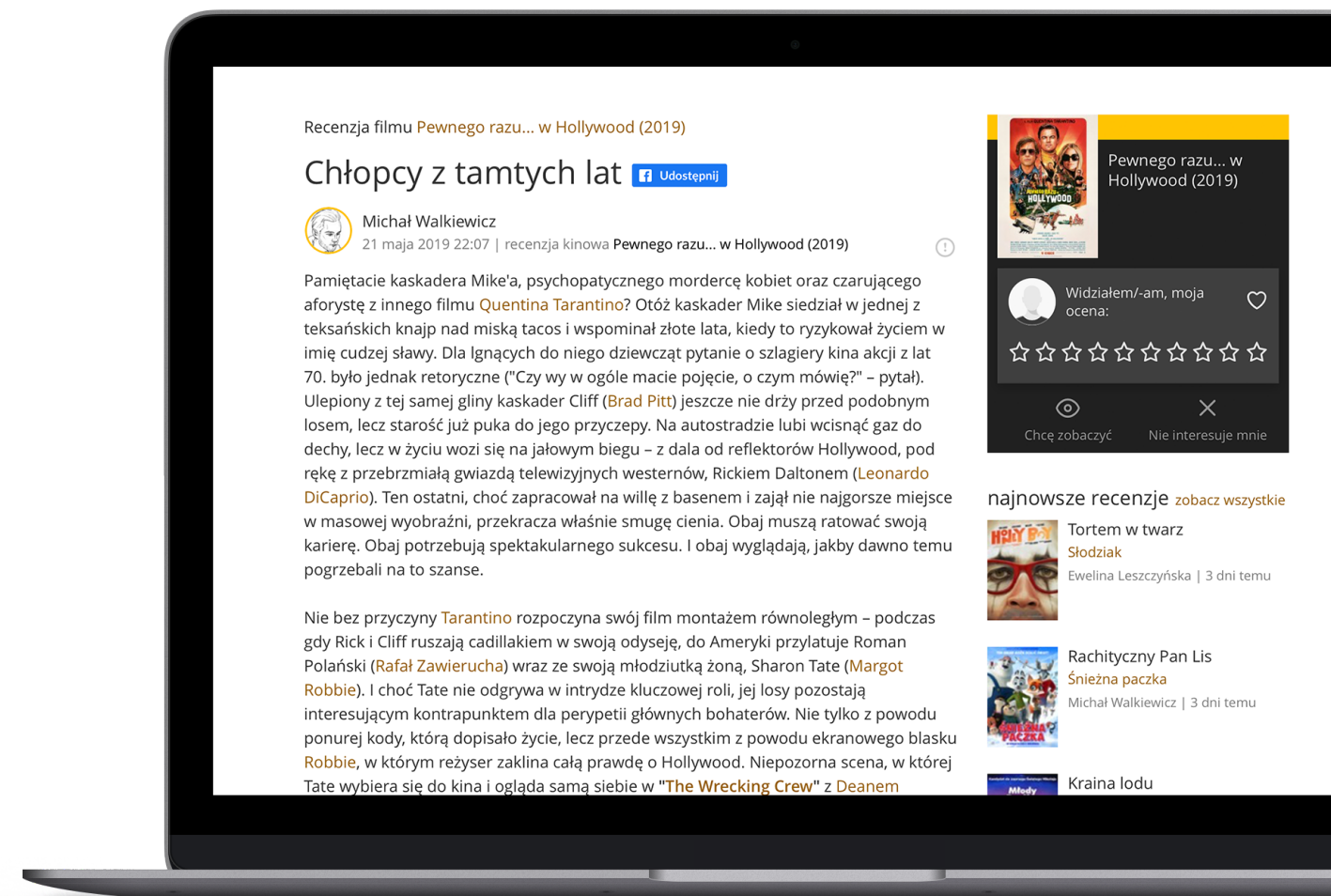
**91% osób w wieku 18-34 lat
ufa recenzjom internetowym tak samo,
jak rekomendacjom od bliskich**

Aby osiągnąć sukces jako firma, musimy rozumieć,
jak ludzie wypowiadają się o naszych produktach.

Cel pracy

Klasyfikacja wydźwięku wypowiedzi dokumentów tekstowych w języku polskim


- Opracowanie rozwiązania pozwalającego na klasyfikację wydźwięku wypowiedzi **długich dokumentów tekstowych w języku fleksyjnym** przy użyciu głębokich, rekurencyjnych sieci neuronowych
- Wykorzystano nowatorskie metody **uczenia transferowego oraz uczenia nienadzorowanego** skuteczne w zastosowaniach z ograniczoną ilością dostępnych danych treningowych
- Rozwiązanie zastosowano w **d dziedzinie recenzji filmowych**



Fragment recenzji „Chłopcy z tamtych lat”, Michał Walkiewicz, Filmweb, 2019

Cel pracy

Klasyfikacja wydźwięku wypowiedzi dokumentów tekstowych w języku polskim



Four empty rectangular boxes stacked vertically, likely for a list or notes.

Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates, Kudo, 2018
Universal Language Model Fine-tuning for Text Classification, Howard et al., 2018

Cel pracy

Klasyfikacja wydźwięku wypowiedzi dokumentów tekstowych w języku polskim

Skuteczna metoda trenowania modelu

Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates, Kudo, 2018
Universal Language Model Fine-tuning for Text Classification, Howard et al., 2018

Opracowanie klasyfikatora z wykorzystaniem metody uczenia nienadzorowanego i transferowego ULMFiT

- Użycie ogólnie dostępnych, nieoetykietowanych zbiorów danych i oetykietowanych danych w celu poprawy dokładności docelowego, dziedzinowego klasyfikatora
- Opracowanie **dziedzinowego klasyfikatora** w oparciu o dziedzinowy model językowy, dostrajany na bazie ogólnego modelu językowego
- Wykorzystanie skutecznych strategii **regularyzacji** rekurencyjnych sieci neuronowych **AWD-LSTM**
- Implementacja w języku **Python**

100x

Mniej oetykietowanych przykładów potrzebnych do uzyskania dokładności klasyfikacji porównywalnej z *Virtual*, *oh-LSTM*, *CoVe* przy użyciu ULMFiT.

Cel pracy

Klasyfikacja wydźwięku wypowiedzi dokumentów tekstowych w języku polskim

Skuteczna metoda trenowania modelu

Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates, Kudo, 2018
Universal Language Model Fine-tuning for Text Classification, Howard et al., 2018

Cel pracy

Klasyfikacja wydźwięku wypowiedzi dokumentów tekstowych w języku polskim

Skuteczna metoda trenowania modelu

Przygotowanie zbiorów danych

Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates, Kudo, 2018
Universal Language Model Fine-tuning for Text Classification, Howard et al., 2018

Rozwiązanie

Autorskie zbiory danych



plwiki

Ogólny zbiór danych
oparty na zawartości polskiej Wikipedii,
zawierający **110038 nieoetykietowanych**
artykułów i 119754538 słów



Filmweb+

Dziedzinowy zbiór danych zawierający
7655 nieoetykietowanych i 19235
oetykietowanych recenzji filmowych
o średniej długości **514 słów/dokument**

*Wikipedia jest zarejestrowanym znakiem towarowym Wikimedia Foundation
Filmweb jest znakiem towarowym Filmweb Sp. z o. o. Sp.k.*

Cel pracy

Klasyfikacja wydźwięku wypowiedzi dokumentów tekstowych w języku polskim

Skuteczna metoda trenowania modelu

Przygotowanie zbiorów danych

Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates, Kudo, 2018
Universal Language Model Fine-tuning for Text Classification, Howard et al., 2018

Cel pracy

Klasyfikacja wydźwięku wypowiedzi dokumentów tekstowych w języku polskim

Skuteczna metoda trenowania modelu

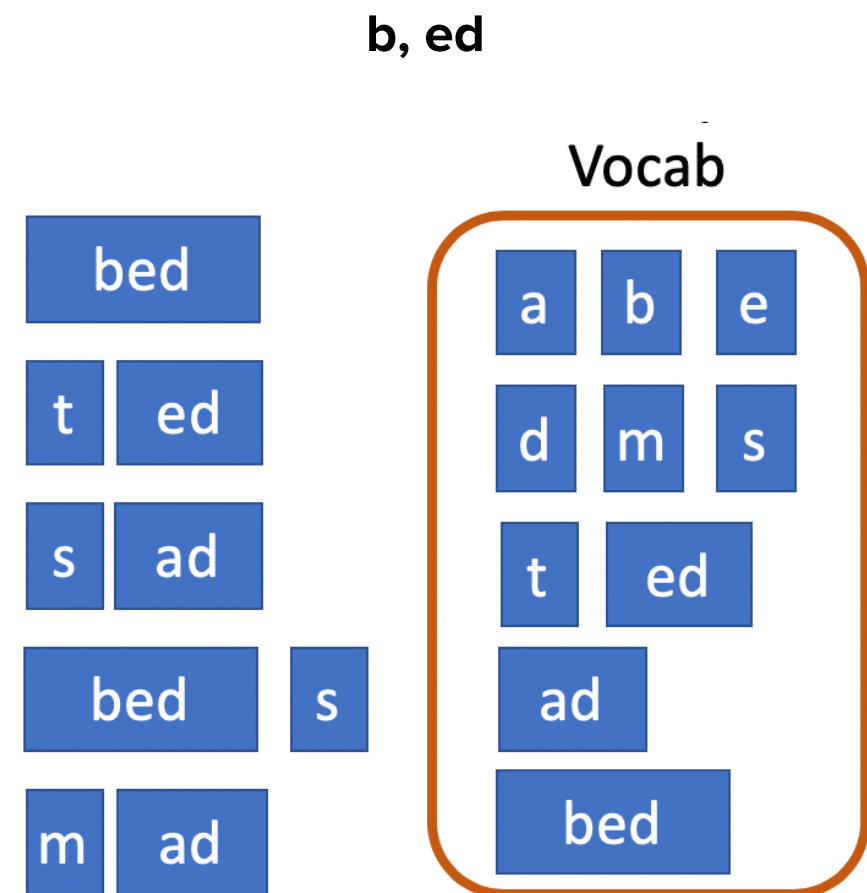
Przygotowanie zbiorów danych

Skuteczna metoda tokenizacji danych

Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates, Kudo, 2018
Universal Language Model Fine-tuning for Text Classification, Howard et al., 2018

Tokenizacja oparta o algorytm segmentacji unigramów sentencepiece

- Korpus dzielony jest na tokeny składające się z pojedynczych znaków
- Dopóki słownik tokenów nie jest zapełniony, **znajdowane są najczęściej występujące bigramy** (pary tokenów), **które zostają scalone**, tworząc nowy token w słowniku
- Rozmiar słownika tokenów określono doświadczalnie na **32000 tokenów**
- Nie stosujemy **lematyzacji** ani **stemmingu**



Cel pracy

Klasyfikacja wydźwięku wypowiedzi dokumentów tekstowych w języku polskim

Skuteczna metoda trenowania modelu

Przygotowanie zbiorów danych

Skuteczna metoda tokenizacji danych

Fragment recenzji „Chłopcy z tamtych lat”, Michał Walkiewicz, Filmweb, 2019

Cel pracy

Klasyfikacja wydźwięku wypowiedzi dokumentów tekstowych w języku polskim

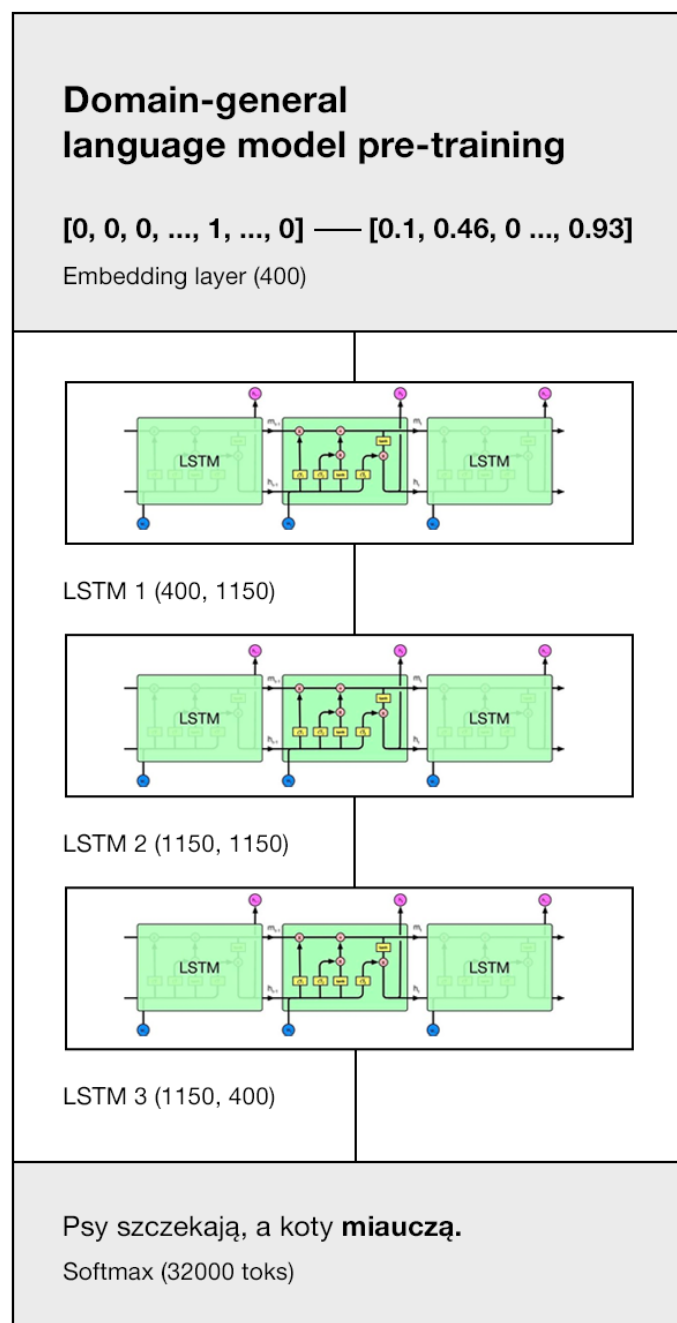
Skuteczna metoda trenowania modelu

Przygotowanie zbiorów danych

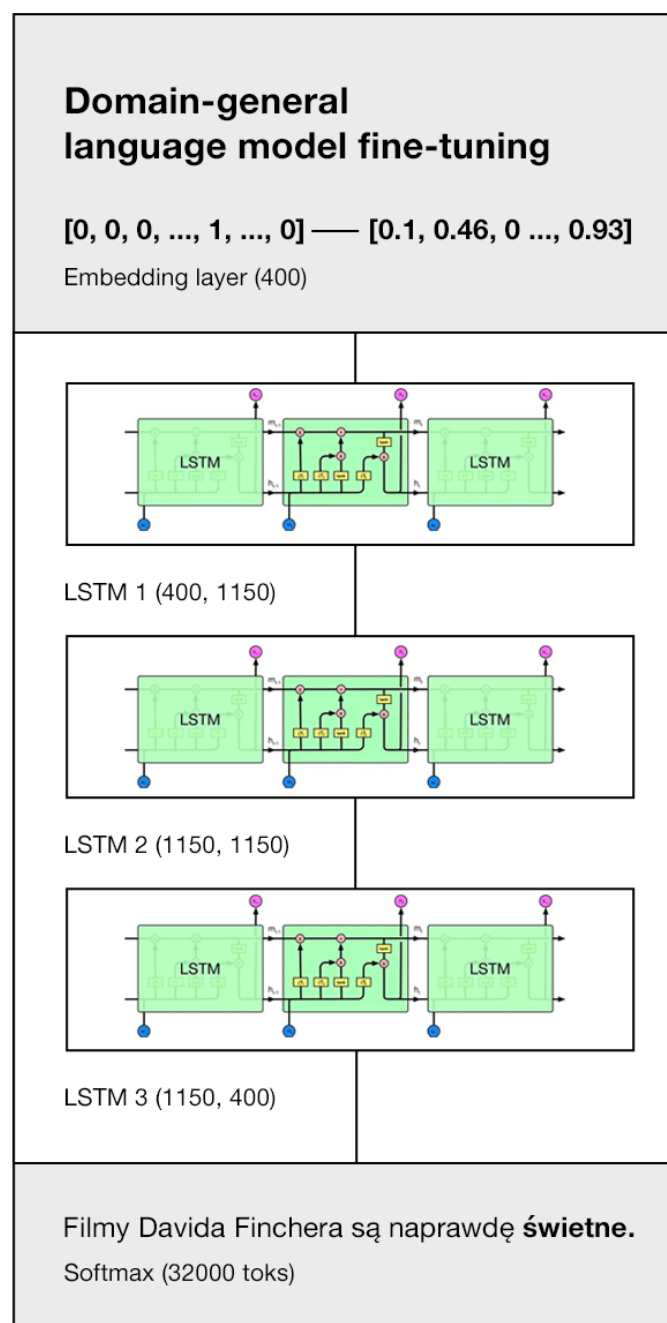
Skuteczna metoda tokenizacji danych

Opracowanie docelowego klasyfikatora

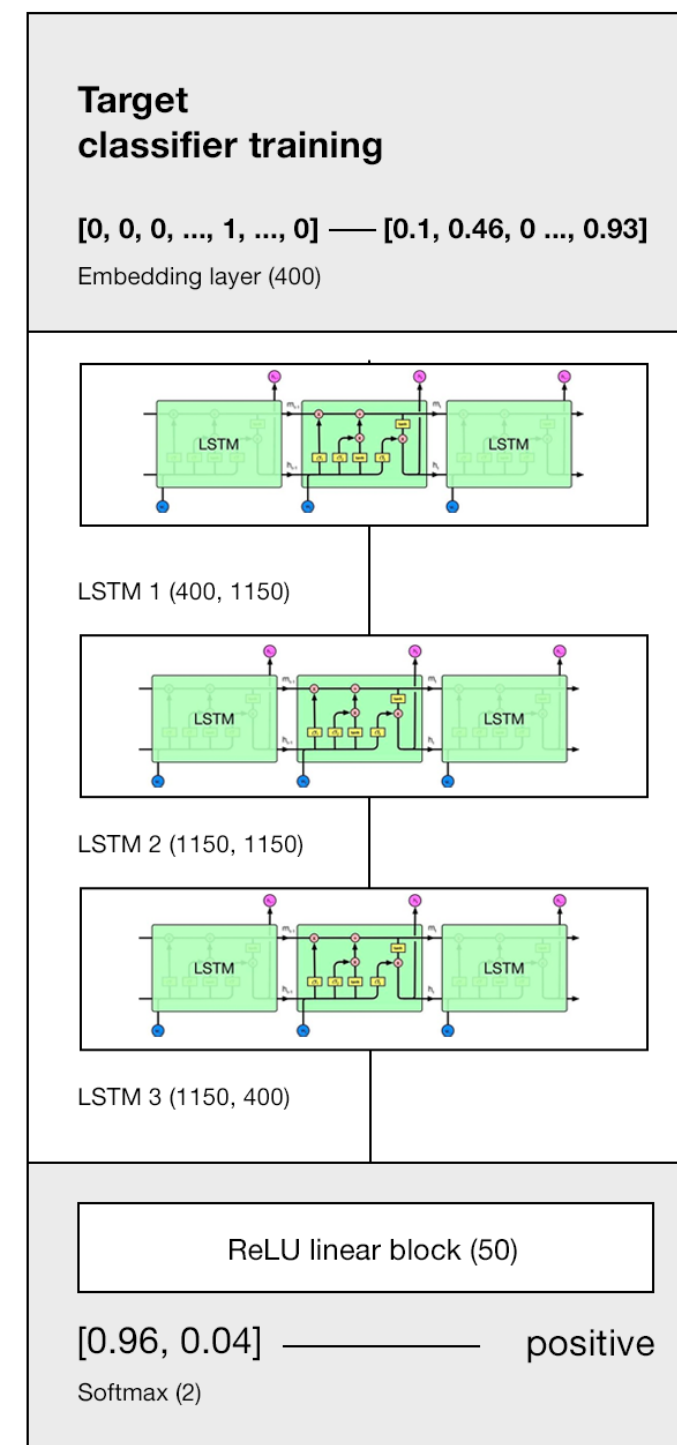
Fragment recenzji „Chłopcy z tamtych lat”, Michał Walkiewicz, Filmweb, 2019



Trenowanie przez
7 godzin i 32 minuty



Dostrajanie przez
1 godzinę i 43 minuty



Dostrajanie przez
5 minut i 13 sekund

Najlepszy wynik dokładności klasyfikacji wydźwięku dla długich dokumentów tekstowych w języku polskim

- **Dokładność klasyfikacji wynosi 94,19%** (mierzona na zbiorze walidacyjnym)
- Większa dokładność uzyskiwana tylko dla klasyfikacji krótkich dokumentów (np. tweetów)
- Wynik osiągnięto używając zaledwie **3085** oetykietowanych przykładów recenzji należących do **pozytywnej klasy** i **3085** przykładów należących do **negatywnej klasy** pod względem wydźwięku wypowiedzi

Autor, praca, zbiór, rok	Dokładność
n-waves, <i>ULMFiT, poleval2019</i> (2019)	90,10%
Chlasta, <i>Sentiment Analysis Model For Twitter, własny</i> (2015)	77.32%
Wawer, <i>Predicting Sentiment of Short Texts</i> , ELMO+RF, skład-TW (2019)	97.9%

Przykładowa klasyfikacja recenzji filmowej w języku polskim

Zarówno w kadrze, jak i poza nim, „Bogowie” Łukasza Palkowskiego to opowieść o przewyciężaniu pewnego impasu. [...]

Nagle okazuje się, że da się zrobić u nas bezpretensjonalne, lekkie kino gatunkowe, które trzyma w napięciu i „się ogląda”; które śmieszy, kiedy ma śmieszyć, i wzrusza, kiedy ma wzruszać. Serce rośnie. [...]

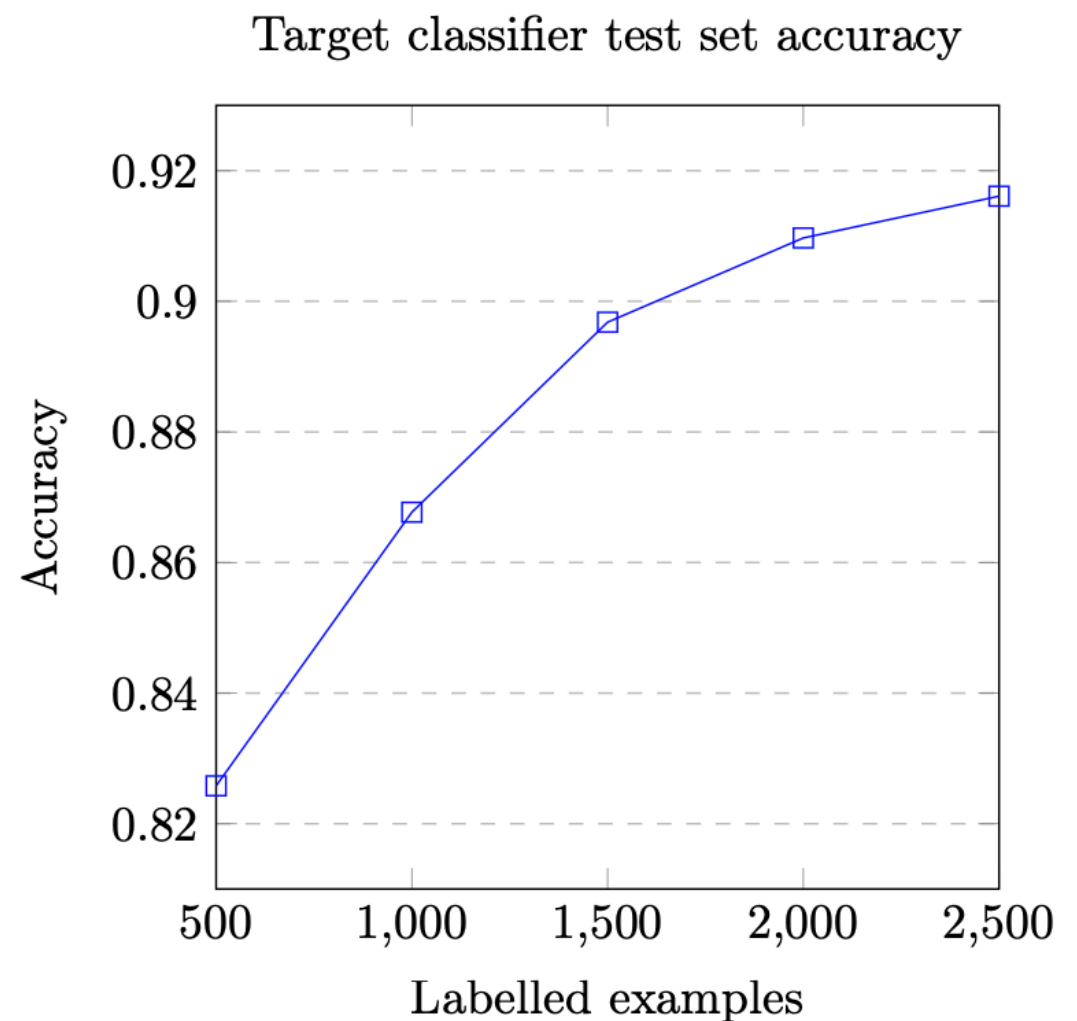
„Bogowie” to film nakręcony według gatunkowego podręcznika. Tu zawiązanie, akcji, tam zwrot, tu zgrabne scenariuszowe analogie, tam nośne one-linery. Żeby osiągnąć zamierzoną lekkość, potrzeba jednak – zaiste – chirurgicznej precyzji realizacyjnej. I Palkowski daje radę, pierwszorzędnie reanimując szablon oraz tłocząc w niego zapas świeżej krwi.

→ **pozytywna**
pewność 99,54%

Jakub Popielecki
ocenia ten film na: ★ 8/10 *bardzo dobry*

Przeprowadzone eksperymenty

- **Wpływ rozmiaru słownika tokenów** na dokładność klasyfikacji i modeli językowych
- **Wpływ ilości oetykietowanych danych** na dokładność klasyfikacji
- **Wpływ hiperparametrów treningowych (ilość epochów, iteracji)** na dokładność klasyfikacji
- **Czas trenowania modeli** w zależności od rozmiaru słownika tokenów



Wnioski z pracy i dalsze możliwe kierunki rozwoju

- **Możliwe jest osiągnięcie dużej dokładności klasyfikacji tekstów w języku polskim** przy użyciu ograniczonej liczby oetykietowanych przykładów treningowych i wielu przykładów nieoetykietowanych
- Opracowany system można wykorzystać jako podstawę **komercyjnego rozwiązania** do klasyfikacji wydźwięku recenzji filmowych
- System można także zastosować dla **różnych dziedzin, różnych problemów klasyfikacji i różnych języków dokumentów**



Dziękuję Państwu za uwagę.