Klasyfikacja fake newsów

Alicja Szpunar-Szałek

Dominika Kokoryk

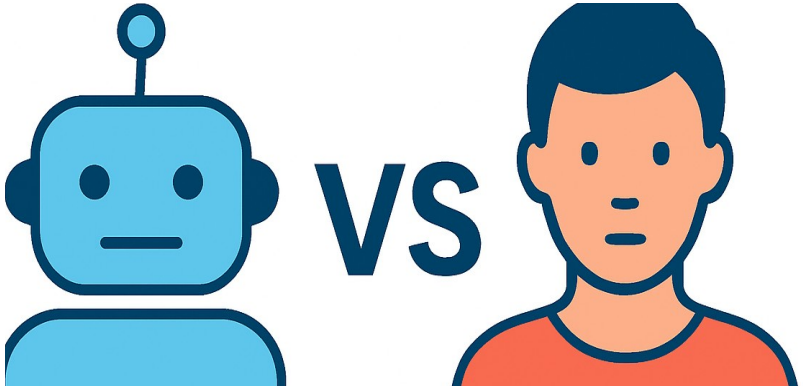
Adrian Komuda

Michał Alenowicz

Kampania społeczna

W dobie dynamicznego rozwoju technologii i powszechnego dostępu do internetu zjawisko rozprzestrzeniania się fałszywych informacji staje się coraz bardziej powszechne. Dodatkowo rozwój sztucznej inteligencji zwiększa ryzyko generowania fałszywych treści przez modele językowe, co utrudnia odróżnienie informacji prawdziwych od zmanipulowanych.

Dlatego postanowiliśmy zająć się stworzeniem modelu, który potrafi rozpoznawać fake newsy. Naszym celem jest **zwiększenie świadomości społecznej** i wsparcie użytkowników w krytycznym myśleniu oraz weryfikowaniu źródeł informacji.



Problem

- Umiejętność klasyfikowania krótkich tekstów w języku angielskim jako ***fake*** lub ***real*** news
- Klasyfikacja binarna tekstów jako fake = 0 lub fake = 1

Dane

Rzeczywiste teksty w języku angielskim opatrzone etykietami do uczenia nadzorowanego

Teksty z etykietami pozyskano z kilku datasetów:

- baza ClaimsKG
- ISOT
- WELFake Dataset
- LIAR Fake news dataset
- English Fake News Dataset

Charakter newsów:

- zakres czasowy: głównie 2010-2023
- tematyka:
  - polityka (ok. 50%)
  - nauka, odkrycia, zdrowie
  - rozrywka
  - ciekawostki, sensacje

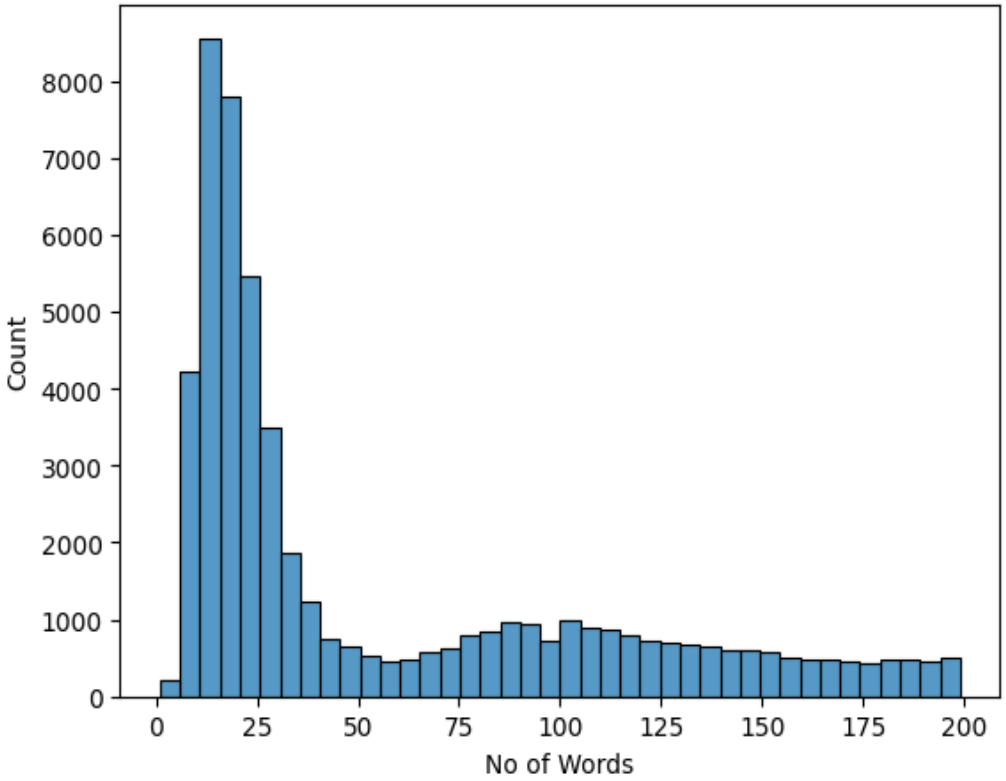
Czyszczenie danych

- Detekcja języka (niektóre bazy są wielojęzyczne),
- usunięcie duplikatów,
- usunięcie linków,
- zamiana wielu kategorii na dwie (dotyczy niektórych zbiorów, gdzie istniały etykiety typu ‘half-true’, ‘misleading’)

Długość tekstu / ilość danych

- Znaczny rozrzut długości!
- Długi ogon sięgający > 10.000 słów
- Wybrano zakres: 5-50 słów
- Trudność w klasyfikacji długich tekstów
- Efekt: ok. 34 tys. wierszy, po dodaniu ‘headlines’ z bazy ISOT uzyskano ok. 68 tys. wierszy

Histogram dla przedziału 0-200 słów



Wpływ długości tekstu na wyniki

	Accuracy	
	Krótkie <5; 50>	**Długie (50; 1200>**
	ok. 34K wierszy	*ok. 104K wierszy*
Logit	0,69	0,55
MultinomialNB	0,68	0,56
RF classifier	0,72	0,34
XGBclassifier	0,68	0,60
SVM	0,69	0,54

Wektoryzacja tf-idf vs. transformers

Dataset: 68K wierszy z tekstami <50 słów *Accuracy na defaultowych hiperparametrach modeli*			
	TF-IDF*	**Tf-idf (n-grams 1-2)**	**All-mpnet-base-v2**
	Wektor ok. 20 tys.	*Wektor ok. 430 tys.*	*Wektor 768*
Logit	0,78	0,80	0,80
00RF classifier	0,78	0,80	0,77
XGBclassifier	0,76	0,78	0,78
CatBoost	0,78	0,80	0,80
LGBM	0,76	0,79	0,78
SVM	0,79	0,81	0,84

* tokenizacja, stopwords, lemmatyzacja

Różne modele embeddingowe (sentence-transformers)

	‘all-MiniLM-L6-v2’	**‘all-MiniLM-L12-v2’**	**All-mpnet-base-v2**
	Wektor 384	*Wektor 384*	*Wektor 768*
Logit	0,76	0,77	0,80
RF classifier	0,74	0,73	0,77
XGBclassifier	0,76	0,76	0,78
CatBoost	0,78	0,78	0,80
LGBM	0,77	0,76	0,78
SVM	0,81	0,82	0,84
prosta NN	---	0,81	0,83

Poszukiwanie hiperparametrów

	Accuracy	Precision	f1-score
Logit	0,80	0,80	0,80
RF classifier	0,77	0,78	0,76
XGBclassifier	0,79	0,78	0,78
SVC	**0,84**	**0,84**	**0,84**
prosta NN	0,83	0,83	0,83

Najlepsze wyniki po GridSearchCV / Optuna
Metryki uzyskane na **danych testowych** (20%)

Wybór modelu

(tylko embeddingi)

Accuracy		
	Train (80%)	**Test (20%)**
Logit	0,80	0,80
RF classifier	0,80	0,77
XGBclassifier	0,81	0,79
SVC (‘poly’)	**0,91**	**0,84**
Prosta NN	0,90	0,83

Wybrany model - SVC

Parametry wybranego modelu:
- 'kernel': 'poly'
- 'C': 0,94
- 'coef0': 0,71
- 'degree': 2
- 'gamma': 'scale'

Threshold: 0,50

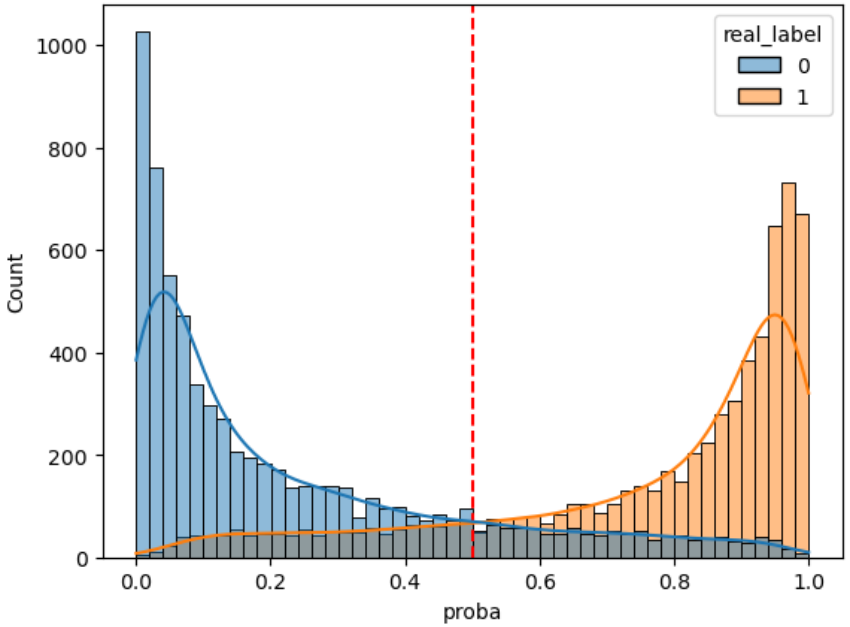
SVC on 50 words embeddings FINAL evaluation (TEST SET), threshold=0.5

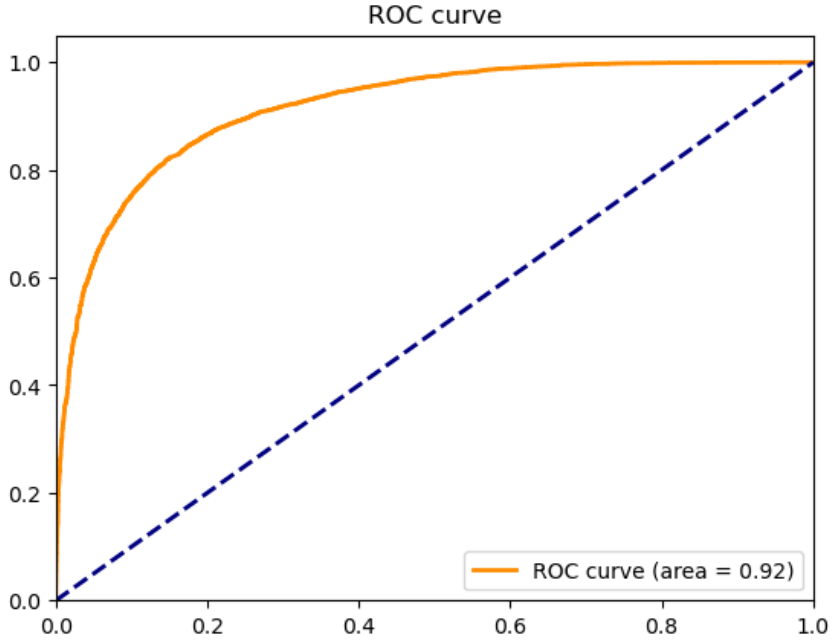
	precision	recall	f1-score	support
0	0.84	0.85	0.84	7010
1	0.84	0.82	0.83	6669
accuracy			0.84	13679
macro avg	0.84	0.84	0.84	13679
weighted avg	0.84	0.84	0.84	13679

F1 score: 0.831
Accuracy: 0.836
Precision: 0.837
Recall: 0.824

Wybrany model - SVC

Dataset testowy (20% danych)





Prompt użyty do uzyskania danych z chata GPT

I want you to prepare a validation dataset for my fake news classification neural network that I've trained. For starters, I want you to prepare a 200-row-long csv file with two columns: 'text' (strings, in English) and 'fake' (either 1 or 0; 1 for fakes, 0 for real news). Stick to the following requirements:

- texts should be 6-50 words long; I don't care about mean length, as long as you make the fakes and real news have similar mean length;
- the 'fake' **class should be balanced** (100 x fakes, 100 x real)
- **don't use padding or any "tacked on" gibberish** just to make texts longer - they should be grammatically and syntactically correct in English;
- text should pertain to: global politics, general health tips and COVID, entertainment, trivia, including potentially scandalous or sensational news (both fake and real);
- they should be either news headlines or simple (apparently factual) statements, possibly 2-3 sentence sub-headlines; use a mixture.
- **don't use any "giveaway words"** that would obviously label the news as fake (e.g. "allegedly", "claims", "unconfirmed" etc.) Both fake and real news should represent a journalistic tone (sometimes sensational or gossipy).
- Use real person names (politicians, actors, scientists, historical figures) and placenames, but don't tack them onto EVERY news - use sparingly.
- Don't try to "replicate" or "recycle" text templates by changing minor details and reusing the same basic sentence core (**use each unique template only once**).
- try to ensure syntactic variation by using various word ordering (within the limits of correct English syntax).
- IF POSSIBLE, **BASE BOTH REAL AND FAKE NEWS ON WEB RESEARCH OF THE MOST POPULAR SUBJECTS** IN THE FIELDS THAT I MENTIONED.

Kontrola człowieka nad danymi syntetycznymi do walidacji

- Połączenie 2 datasetów z Chata GPT (po ok. 170 próbek każdy)
- Ręczne usunięcie “near duplicates”
- Usunięcie twierdzeń dwuznacznych, zbyt ogólnych, półprawd
- Usunięcie twierdzeń wartościujących wiadomość lub źródło
- Nie modyfikowano treści - wiersz był albo usuwany, albo pozostawiany
- Na końcu losowe usunięcie kilku wierszy z etykietą, która dominowała
- Uzyskano zbalansowany dataset 150/150

Generalizacja

(tylko embeddingi)

Accuracy				
	Train (80%)	**Test (20%)**	**ChatGPT (extra)** Threshold = 0.5	**ChatGPT (extra)*** Threshold = ?
Logit	0,80	0,80	0.65	0,73
RF classifier	0,80	0,77	0,66	0,68
XGBclassifier	0,81	0,79	0,73	**0,81**
SVC ('poly')	**0,91**	**0,84**	**0,76**	0,76
Prosta NN	0,90	0,83	0,66	0,74

*maksymalne accuracy uzyskane po dostosowaniu threshold do innego charakteru danych

Wybrany model - XGBoost

Dataset ‘ChatGPT curated’ (300 próbek)

Parametry wybranego modelu:
- colsample_bytree: 0,61
- learning_rate: 0,03
- max_depth: 3
- min_child_weight: 14
- n_estimators: 705
- gamma: 0,01
- lambda: 9,98
- alpha: 0,55

Threshold: 0,36

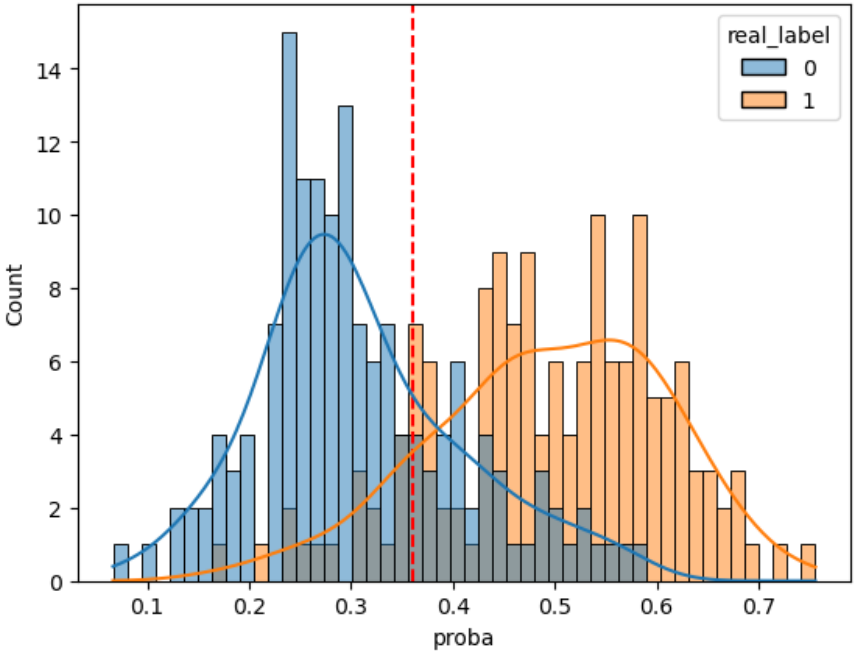
ChatGPT set evaluation

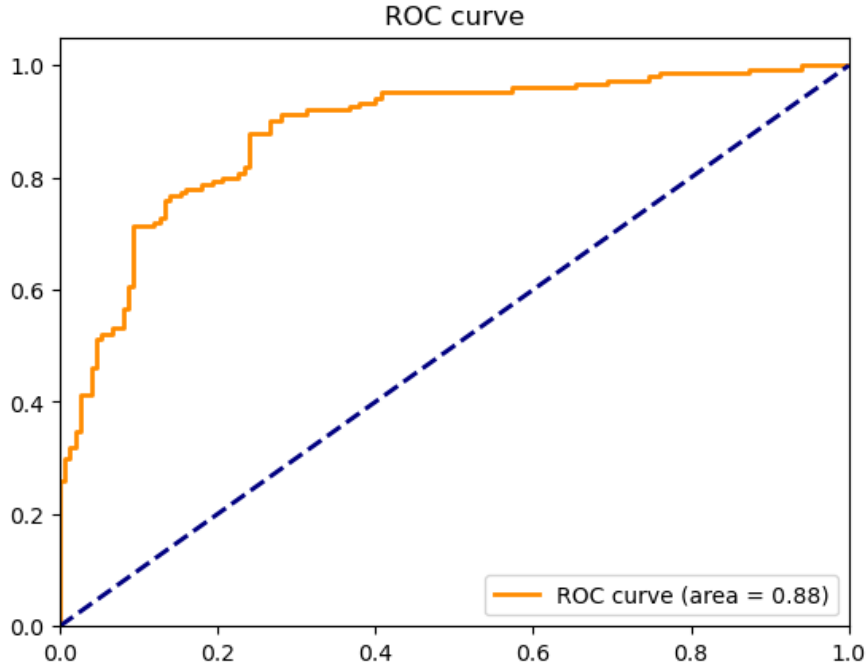
	precision	recall	f1-score	support
0	0.86	0.74	0.80	150
1	0.77	0.88	0.82	150
accuracy			0.81	300
macro avg	0.82	0.81	0.81	300
weighted avg	0.82	0.81	0.81	300

F1 score: 0.822
Accuracy: 0.810
Precision: 0.772
Recall: 0.880

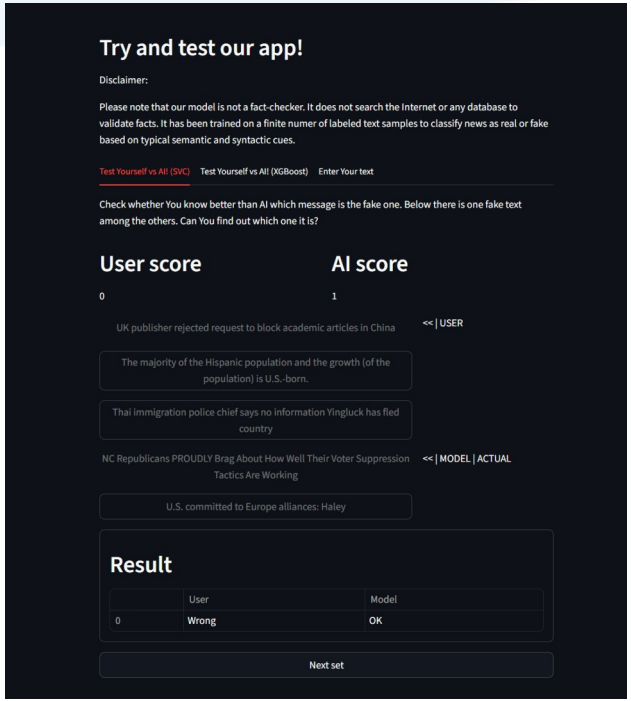
XGBoost

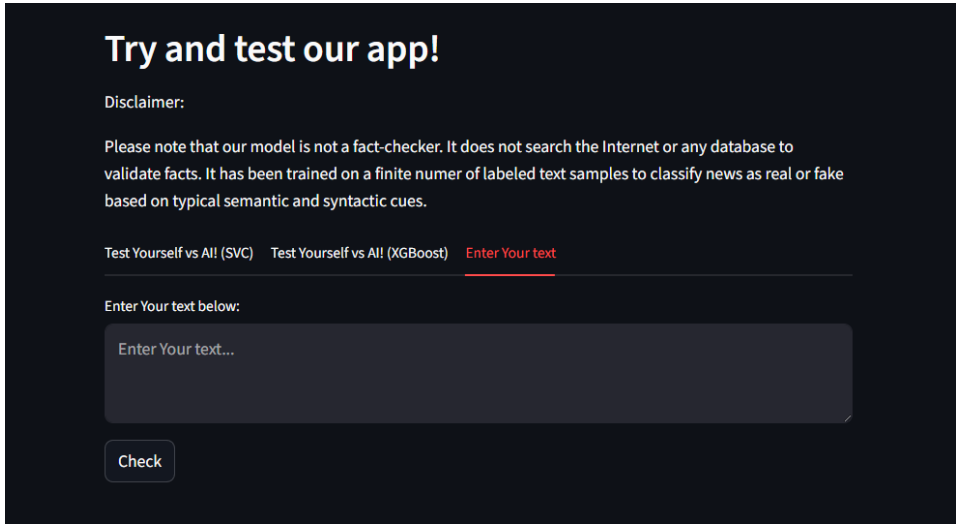
Dataset ‘ChatGPT curated’ (300 próbek)





Aplikacja


Try and test our app!
Disclaimer:
Please note that our model is not a fact-checker. It does not search the Internet or any database to validate facts. It has been trained on a finite numer of labeled text samples to classify news as real or fake based on typical semantic and syntactic cues.
Test Yourself vs AI! (SVC)
Test Yourself vs AI! (XGBoost)
Enter Your text
Check whether You know better than AI which message is the fake one. Below there is one fake text among the others. Can You find out which one it is?
User score
AI score
0
1
UK publisher rejected request to block academic articles in China
<< | USER
The majority of the Hispanic population and the growth (of the population) is U.S.-born.
Thai immigration police chief says no information Yingluck has fled country
NC Republicans PROUDLY Brag About How Well Their Voter Suppression Tactics Are Working
<< | MODEL | ACTUAL
U.S. committed to Europe alliances: Haley
Result
User
Model
0
Wrong
OK
Next set


Try and test our app!
Disclaimer:
Please note that our model is not a fact-checker. It does not search the Internet or any database to validate facts. It has been trained on a finite numer of labeled text samples to classify news as real or fake based on typical semantic and syntactic cues.
Test Yourself vs AI! (SVC)
Test Yourself vs AI! (XGBoost)
Enter Your text
Enter Your text below:
Enter Your text...
Check

Wnioski / F.R.I.N.

Wnioski:

- generalizacja osiągnięta dzięki połączeniu różnych datasetów;
- praca na dłuższych tekstach wymagałaby znacznego zwiększenia liczby próbek;
- po uzyskaniu embeddingów zadanie sprowadziło się do klasyfikacji binarnej - NN nie przebiła SVC.

Potencjalne ulepszenia:

- ensembling modeli;
- trening na większej ilości danych;
- zbadać dokładniej łączenie embeddingów z tf-idf.

Dziękujemy za uwagę