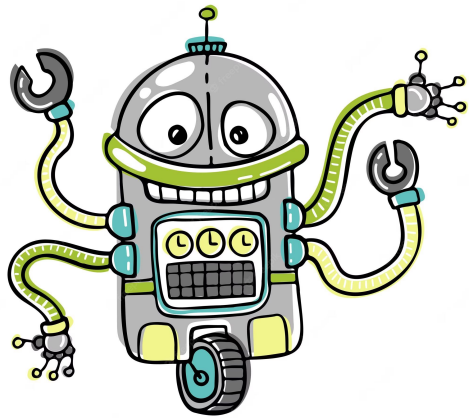
Fake News Detection

Wstęp do uczenia maszynowego

Alicja Charuza, Mateusz Gałęziewski



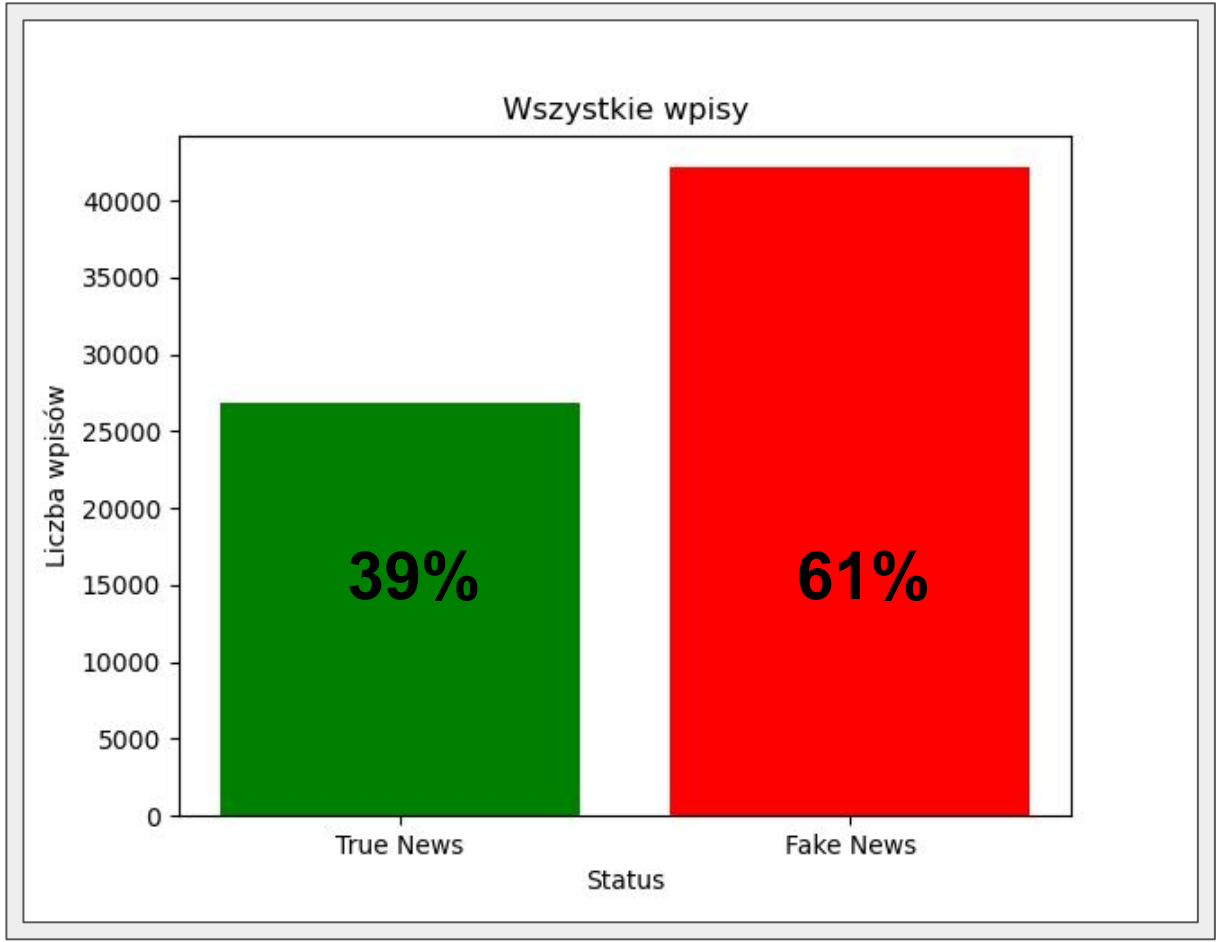
Źródło danych: https://www.kaggle.com/datasets/mohammadaflahkhan/fake-news-dataset-combined-different-sources

O zbiorze danych:

- Dane zawierają **tytuł**, **tekst właściwy** oraz **etykietę fake/true**

- Ilość wszystkich zarejestrowanych wpisów: **69045**

- Wpisy pochodzą z różnych platform



Krok 1: detekcja języków

W jakich językach występują wpisy w zbiorze?

Co możemy zauważyć?

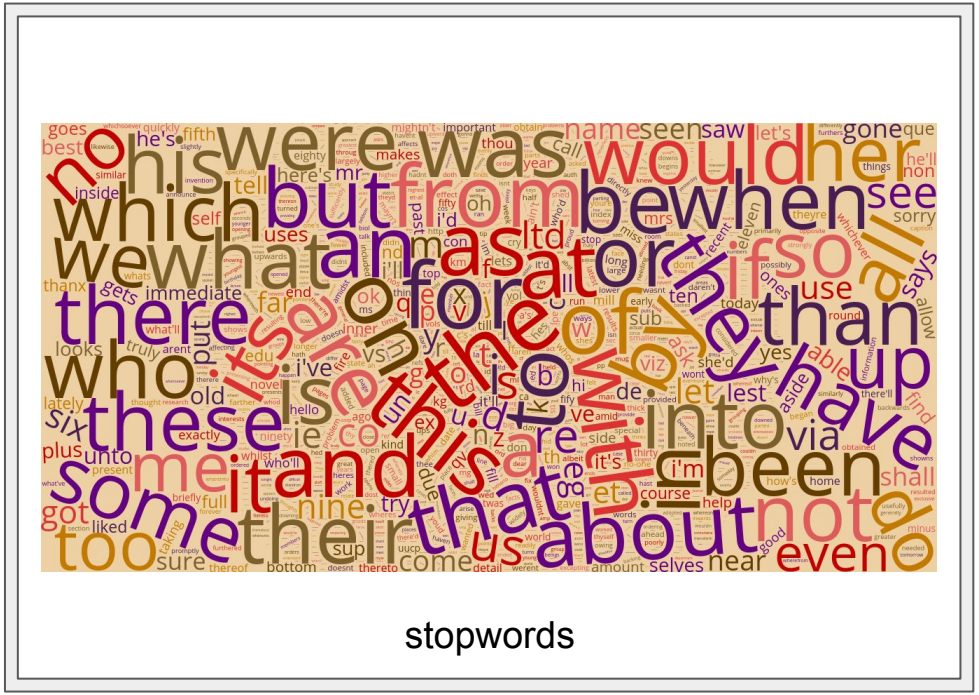
- Łączna ilość wpisów po angielsku: **67 419**
- Prawie wszystkie wpisy w innym języku są fałszywe!

language	Ground Label	
af	fake	3
ar	fake	22
ca	fake	3
cy	fake	3
da	fake	2
de	fake	133
el	fake	4
en	fake	40538
	true	26881
es	fake	175
et	fake	1
fi	fake	5
fr	fake	53
hr	fake	3
hu	fake	1
id	fake	2
it	fake	11
lt	fake	1
nl	fake	4
no	fake	4
pl	fake	4
pt	fake	15
ro	fake	5
ru	fake	203
sl	fake	1
so	fake	8
sv	fake	1
sw	fake	14
tl	fake	2
tr	fake	10
unknown	fake	925
	true	5
vi	fake	2
zh-cn	fake	1

Przygotowanie danych tekstowych w j. angielskim

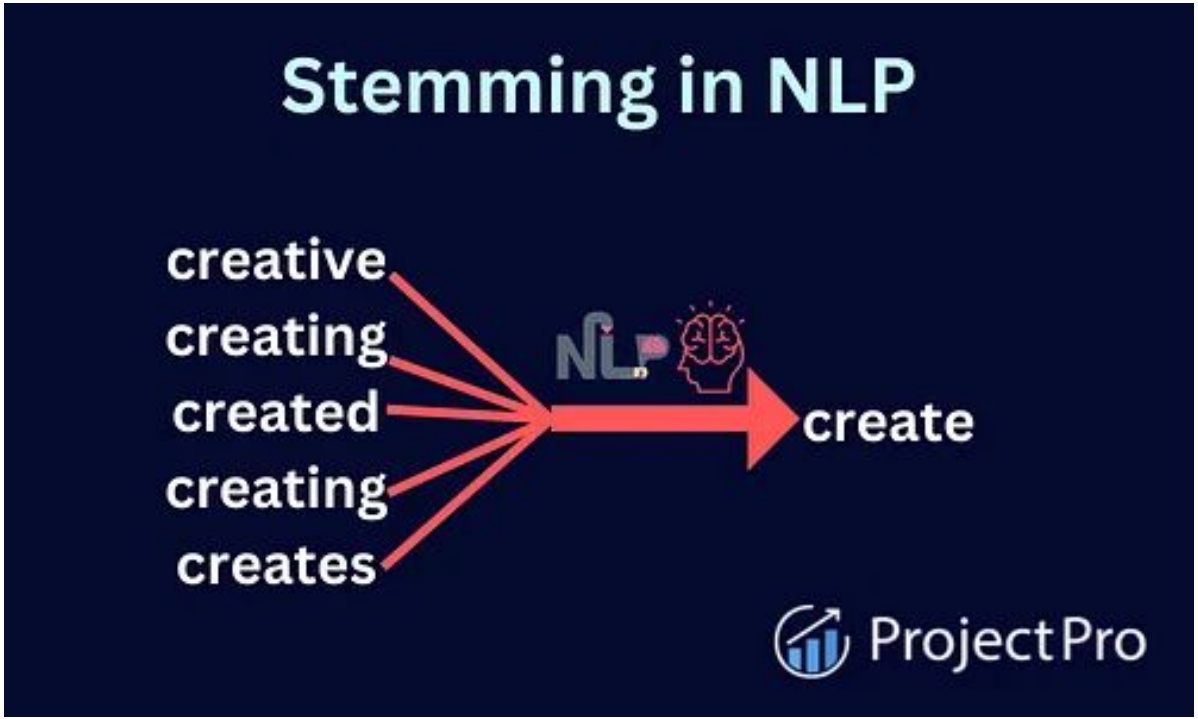
usunięcie :

- interpunkcji
- cyfr
- pojedynczych znaków tj. ‘(^|).$
- stopwordów w języku angielskim



stopwords

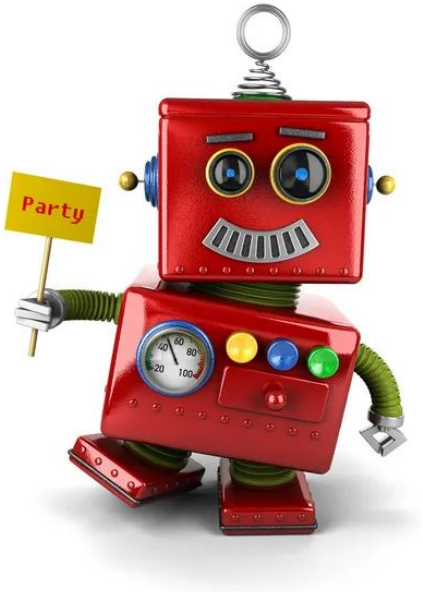
Stemizacja słów



Źródło:https://www.projectpro.io/article/stemming-in-nlp/780

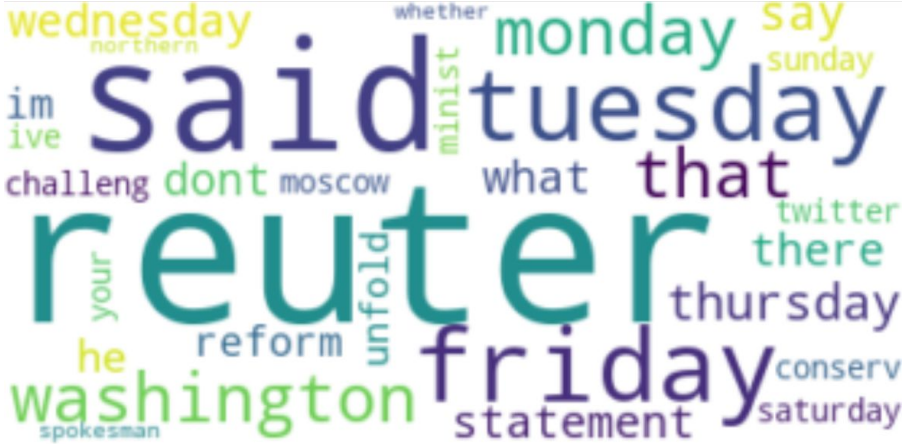
Modele wdrożeniowe

1. Regresja logistyczna
2. Drzewo decyzyjne
3. Las losowy



1. Regresja logistyczna

Słowa najbardziej wpływające na klasyfikację wpisu jako **prawdziwy news**



Słowa najbardziej wpływające na klasyfikację wpisu jako **fałszywy news**

2. Drzewo decyzyjne

Słowa najbardziej wpływające na klasyfikację jako **prawdziwy** lub **fałszywy** news

3. Las losowy

Słowa najbardziej wpływające na klasyfikację jako **prawdziwy** lub **fałszywy** news

Czas na pytania

Dziękujemy za uwagę :)