# Fake and real news

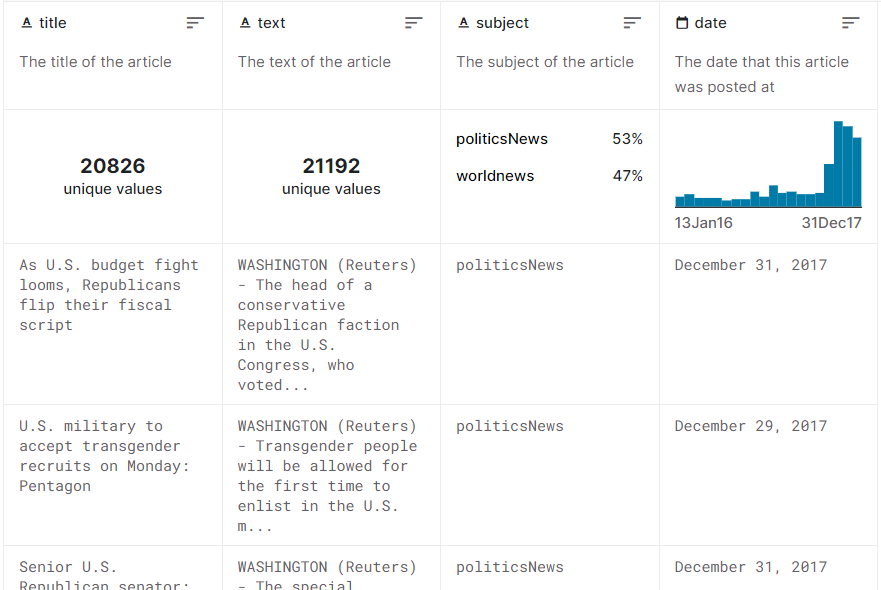
## Abstrakt

V tomto projekte sa zaoberám prácou s datasetom falošných a pravých správ a ich správnej kategorizácii pomocou supervised learning. Medzi iným, porovnávam vplyv veľkosti datasetu na presnosť predikcie, vplyv prítomností titulku a textu na presnosť predikcie a slová ktoré najviac naznačujú pravdivostnú hodnotu správ.

## Úvod

Cieľom tohto projektu je zistiť závislosť presnosti kategorizácie od veľkosti datasetu a jeho ďalšie podobné úpravy. Pri práci budem primárne využívať Python a jeho knižnice.

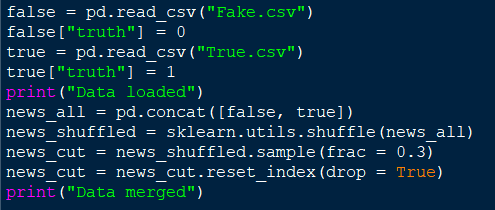
## Dataset

Dataset je získaný zo stránky [Kaggle](https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset/discussion/210551?resource=download&select=True.csv) a skladá sa z dvoch csv súborov, jeden obsahujúci pravdivé správy a druhý falošné. Dataset je tvorený 4 stĺpcami: titulok správy, text správy, téma správy a dátum vydania. Pri mojom projekte využívam len prvé dva stĺpce. Stĺpce téma správy a dátum vydania obsahujú informácie unikátne pre jednu z pravdivostných hodnôt a tým pádom robia klasifikáciu dát triviálnu.

## Metódy a popis kódu

### Načítanie datasetu

Na začiatku treba dáta načítať, na čo som využil knižnicu Pandas. Kedže dáta samé o sebe neobsahujú informáciu o pravdivosti, pridal som k nim ďalší stĺpec vyjadrujúci túto hodnotu, pre falošné správy som priradil 0 a pre pravé 1. Napriek tomu že tento stĺpec neskôr separovaný ako list Y pre samotné učenie. Jeho pripojením som uľahčil miešanie a vzorkovanie dát. Dáta však boli stále rozdelené do dvoch premenných., takže som ich musel zlepiť, premiešať pre ich rovnomerné rozloženie a preindexovať. Z výkonnostných dôvodov som ešte dodatočne zmenšil dataset na tretinu pôvodnej veľkosti. (Obr 1)



Obr

### Spracovanie datasetu

Takto upravený dataset, konkrétne prvé dva stĺpce, však bolo treba ďalej spracovať na formu spracovateľnú metódami knižnice sklearn. Stĺpce som spracovával oddelene na zachovanie informácie či ide o názov alebo text správy. Na spracovanie som sa inšpiroval jedným z praktických cvičený ktoré sa touto témou zaoberalo.

Ako prvé som odstránil takzvané *stop words*, teda slová, ktoré sa často používajú v jazyku, ale nemajú veľkú hodnotu pre analýzu textu. Zoznam týchto slov som získal z knižnice *nltk.* Všetok text som zároveň zmenil na malé písmená, čo zabráni duplicite slov len následkom ich polohy vo vete. Síce sa tým mierne stráca informačná hodnota dát, keďže písanie uppercasom je typické pre polopravdivé alebo falošné správy, účelom tohto projektu je kategorizácia na základe obsahu a nie formy.

Upravený text bol síce o poznanie kratší a mohol by sa použiť na učenie, avšak slová majúce viacero stále spôsobujú duplicitné skreslenie informácii. Bolo potrebné ich dať do ich základného tvaru, načo som využil metódu *PorterStemmer* z *nltk.stem.porter*. Z textu som následne ešte odstránil nechcené symboly ako bodky, čiarky a zátvorky .

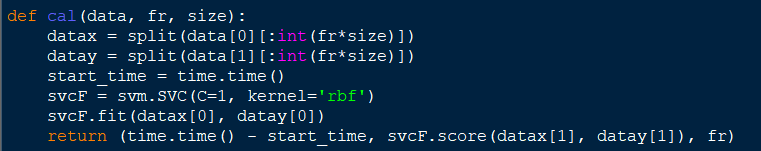
Z takto upraveného textu môžem vytvoriť moju slovnú zásobu, pričom titulok a text ju zdieľajú. Najprv som dáta tokenizoval, využil som na to zase knižnicu *nltk.* Zo zoznamu tokenov som vytvoril ich frekvenčnú tabuľku a vrchných 2500 som použil ako moju slovnú zásobu.

Po získaní slovnej zásoby som už len zmenil názov a text na polia, kde jednotlivé prvky vyjadrujú počty slov v texte. Tieto dva stĺpce som nakoniec zlúčil do jedného zoznamu, takže ich ako vstupné X mám zoznam zoznamov dĺžky 5000 prvkov (2 x dĺžka slovnej zásoby).  
Ako Y som jednoducho použil posledný stĺpec s pravdivostnými hodnotami.

### Trénovanie s rôznymi veľkosťami

V ďalšej časti som trénoval klasifikáciu s rôzne veľkými datasetmi. Dáta som rozdelil pomocou malej funkcie na trénovaciu a testovaciu časť a následne som natrénoval klasifikáciu pre 10 rôzne veľkých datasetov, pričom som zaznamenával čas a presnosť odhadu, ktoré som následne využil na vytvorenie 3 grafov. (Obr 2)

Okrem toho natrénujem dve separátne inštancie pre dataset, jednu bez textu a jednu bez názvu správy. Nakoniec využijem *Random Forest Clasifier*, pomocou ktorého môžem jednoducho zistiť ako majú jednotlivé slová vpliv na klasifikovanie správ.

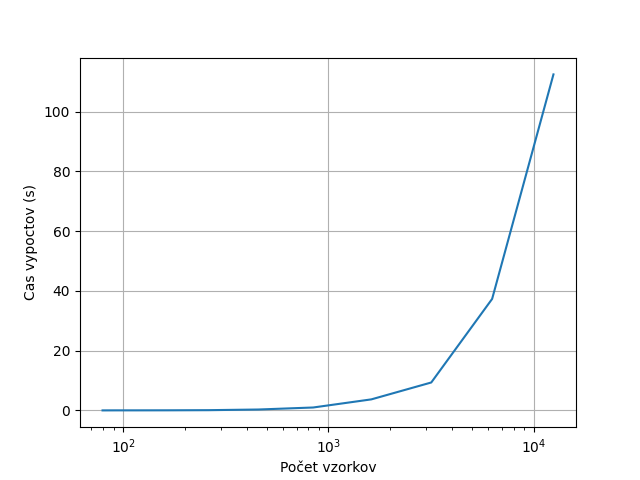


Obr

## Vyhodnotenie

### Vplyv veľkosti datasetu na klasifikáciu

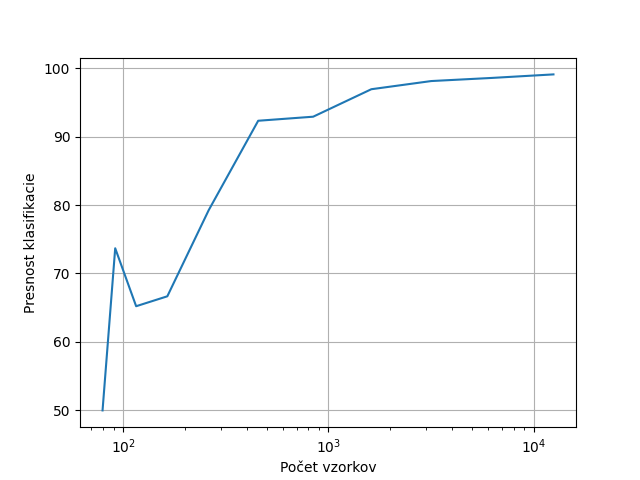
Ako prvé som vyhodnotil vplyv veľkosti datasetu na jeho klasifikáciu, konkrétne na rýchlosť výpočtov a presnosť následného modelu. Dáta som vyčítal primárne grafov.



Obr 3

Ako môžeme vidieť na obrázku 3, s veľkosťou datasetu rastie aj čas potrebný na správne fitnutie modelu. Tento rast je viditeľne exponenciálny a preto aj pomerne malý rozdiel v jeho veľkosti môže násobne urýchliť výpočty. Z obrázku môžeme vidieť že pri našom datasete dochádza k zlomu približne pri 3000-4000 vzorkách.

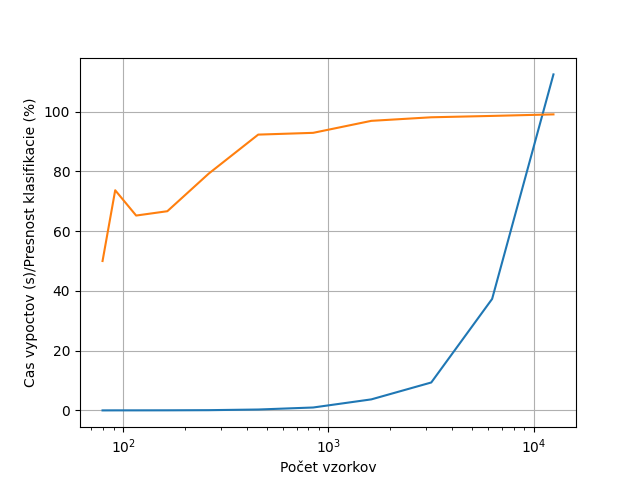
Na obrázku 4 môžeme vidieť závislosť presnosti klasifikácie od počtu vzoriek. Pri nižšom je klasifikácia nielen menej presná ale jej presnosť je ťahšie predvídateľná, väčší počet vzoriek nezaručí presnejšiu klasifikáciu. Model začína dosahovať presnejšie klasifikáciu pri 2000-3000 vzoriek, vyššie počty vzoriek síce zaručujú väčšiu presnosť ale majú klesajúce výnosy.



Obr

Pri výbere vhodnej veľkosti vzoriek teda musíme nájsť taký počet vzoriek, aby čas výpočtov nebol neprimerane veľký ale ani aby netrpela presnosť modelu. Keďže oba grafy zdieľajú jednu os, môžeme ich vykresliť do spoločného grafu (Obr 5).

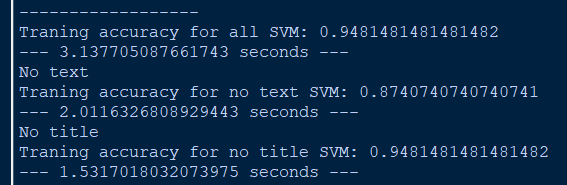
Z tohto grafu môžeme vyčítať že vhodný počet vzoriek je medzi 1000 až 3000. Tento výber je samozrejme ovplyvnený požadovanou presnosťou respektíve rýchlosťou klasifikovania dát, avšak pre moje účely som použil 2693 vzoriek, čo je približne 20% plnej trénovacej množiny. Pomocou týchto vzoriek som testoval ďalšie časti tohto projektu.



Obr

### Vplyv prítomnosti názvu/textu správy na presnosť predikcie

Ako ďalšie som skúmal potrebu prítomnosti názvu a textu správy pre korektnosť predikcie. Separácia dát názvu a textu bola pomerne jednoduchá, keďže prvých 2500 prvkov patrili k názvu a druhých 2500 k textu.



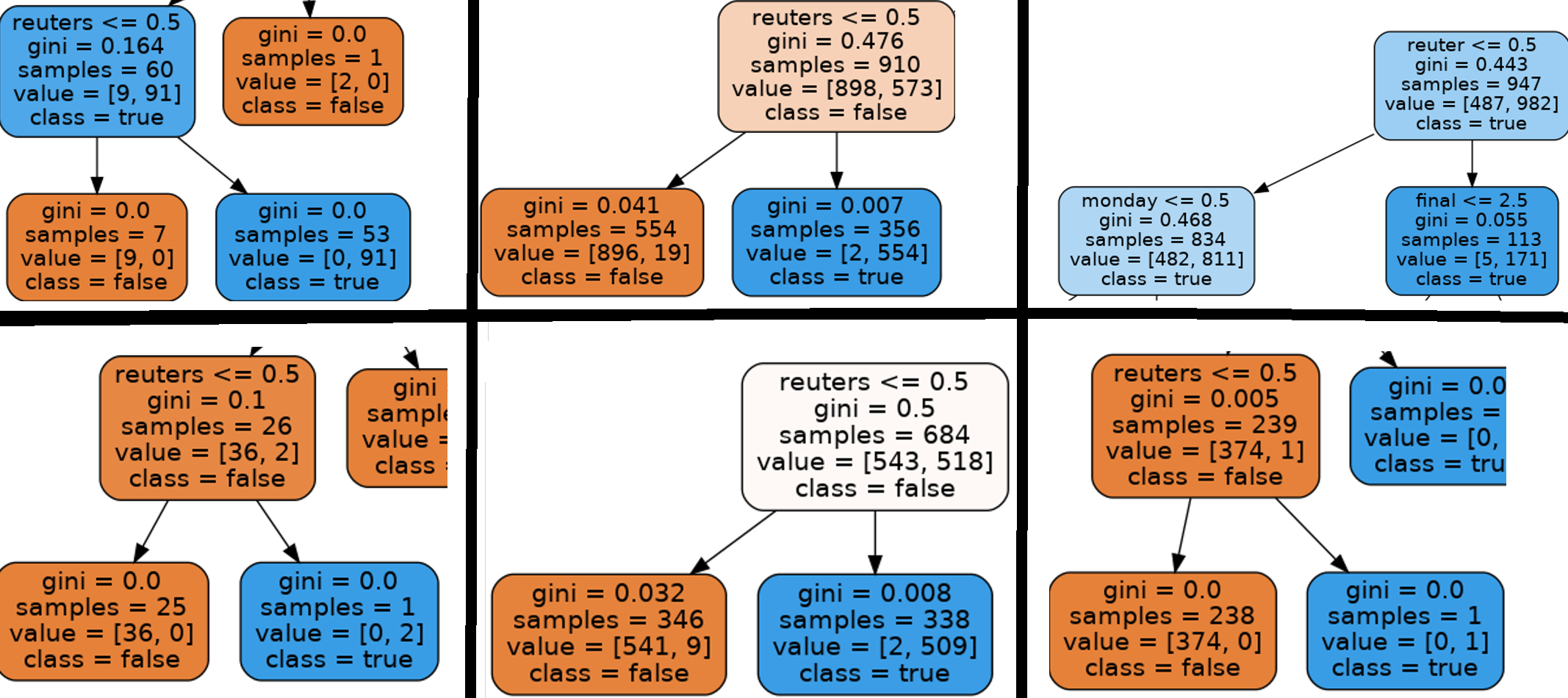
Obr 6

Ako je vidieť na obrázku 6, predikcia s názvom aj textom dosahuje vysoké hodnoty presnosti, avšak trpí v oblasti času. To bolo očakávané, keďže má k dispozícií najviac dát. Modeli trénované len pomocou titulku správy majú pomerné výrazné zníženie presnosti predikcie avšak o niečo rýchlejšie trénovanie. Používanie iba textu bez titulku naopak zachovalo presnosť plných dát pri našom rozsahu presnosti a tento výsledok bol dosiahnutý ešte v lepšom čase. Tento čas je ešte o niečo lepší než pri trénovaný len na názvoch napriek rovnako veľkému datasetu. To je pravdepodobne spôsobené lepšie separovateľnými dátami.

### Kľúčové slová

Pri rozhodovaný o pravdivosti datasetu na základe počtu slov, sú samozrejme niektoré slová viac významné ako iné. Pre odhalenie týchto slov som použil *Random Forest Clasifier* a následne som zo vzniknutých stromov vyčítal, ktoré slová majú aký vplyv. Obrázky celých stromov sú priložené v prílohe, keďže ich rozmer by ich spravil nečitateľnými.

Najviac významné je pravdepodobne slovo **Reuters** (spravodajská agentúra), v 5/6 stromov je jeho absencia príznakom falošných správ. (Obr. 7)



Obr

Medzi ďalšie významné slová patria