**1. Fake News detection**

* Link: <https://www.kaggle.com/jruvika/fake-news-detection>
* Oszlopok:
  + URL
  + Head
  + Body
  + Label
    - 0: fake
    - 1: real
* Adatok: 4000
* Forrás: bbc, cnn, new york times, reuters, activistpost, beforeitsnews, dailybuzzlive, disclose, abcnews…
* Eloszlás: elvileg kiegyenlített, de a szerkezete nem engedi leellenőrizni… pofára kiegyenlített
* Megjegyzés:
  + Kiegyenlített adatbázis, jól használható, az adatok link alapján ellenőrizhetők
  + Címkézett
  + Kevés adat
  + 3-mal talán összevonható, és akkor megvan a 10k

**2. Political Social Media Posts**

* Link: <https://www.kaggle.com/crowdflower/political-social-media-posts/version/1>
* Oszlopok:
  + ID
  + Trusted judgement: 1-3
  + Dátum: last judgement
  + Audience:
    - national
    - constituency
  + Audience confidence: nagyrészt egy, néha 0.5 és 1 között (kb 20 eset)
  + Bias
    - Partisan
    - Neutral
  + Bias confidence: mint az audience-é
  + Message: 9 kategória: policy, attack, support, information, mobilisation, other…
  + Text: twitter üzenet
* Adatok: 7800
* Forrás: Twitter
* Eloszlás: nagy része 1, de a szöveg szerkezete itt is tragédia
* Megjegyzés:
  + Izgalmas téma, változatos adatok
  + Majdnem elég a minta elemszáma
  + A labelezés megbízhatósága és eloszlási egyenletessége kérdéses
  + „Csak” tweet-ek – rövid szövegek

**3. real\_or\_fake**

* Link: <https://www.kaggle.com/rchitic17/real-or-fake>
  + <https://github.com/GeorgeMcIntire/fake_real_news_dataset>
    - Másik oldal szerint 10ezer adat van benne (a formátum miatt ezt nehéz megállapítani)
    - Pros: Pretty neat, simpler, contains real news, data from both left and right wing sites
    - Cons: Less information for each entry, the labeling might not be 100% right(when is it ever 100%?), only two options ‘fake’ or ‘real’(no in between) - binary
* Oszlopok:
  + Title
  + Text
  + Label
    - FAKE
    - REAL
* Adatok: 6300
* Forrás: ismeretlen
* Eloszlás: kiegyensúlyozott – elvileg
* Megjegyzés:
  + Kiegyensúlyozott, közepes mennyiségű adat
  + Források nem ellenőrizhetők
  + Egyértelműen címkézett
  + 1-el talán összevonható, és akkor megvan a 10k

**4. Russian Troll Tweets**

* Link: <https://www.kaggle.com/vikasg/russian-troll-tweets#tweets.csv>
* Oszlopok:
  + User\_id, user\_key 🡪 felhasználókról segédadatbázis
  + Created – dátum
  + Retweet count, favorite count
  + Text
  + Tweet id
  + Hashtags
  + URLs
  + Mentions
* Adatok: 200.000
* Forrás: Twitter
* Eloszlás: egyoldalú – csak troll tweet-ek
* Megjegyzés:
  + Elvileg ezeket a felhasználókat törölték, mint orosz trollok – ez határozza meg a címkézés megbízhatóságát
  + Nagyon sok adat, bár rövidek (tweet-ek)
  + Nincs „nem-troll” tweet, de az szerezhető elvileg (api-val leszedni a másik kategória 200 ezer adatát)

**5. Fake News Challenge**

* Link: <https://github.com/FakeNewsChallenge/fnc-1>
* Oszlopok:
  + Cím
  + ID
  + Stance
    - Agree
    - Disagree
    - Discuss
    - Unrelated
* Adatok: 50.000 cím, 2500 cikk szöveg
* Eloszlás: egyenletlen – a többség unrelated (több, mint a fele)
* Megjegyzés:
  + Nagyon sok adat címekben
  + A body-val való összekötés elvileg sikerülhet, de a gyakorlat kérdéses
  + Nagy része unrelated címkével rendelkezik
  + Nincsenek meg a cikkek linkjei
  + Van egy gradient boostingot használó példa előttünk, baseline-nak
  + Az oldal kétlépcsős osztályozást javasol: unrelated? --> ha nem, akkor a maradék három osztály.
  + Eloszlás nagyon húz az unrelated felé, erre majd figyelni kell
  + Az egész lényege valahol ott van, hogy a cikk címe és a törzsszöveg tartalma hogy kapcsolódik szemantikailag (egyetért, nem ért egyet, nem foglal állást, eltérő téma). Hogy ez milyen módon kapcsolódik az álhírekhez, arról viszont szűkszavú a repo.

**6. Fake News Corpus**

* Link: <https://github.com/several27/FakeNewsCorpus>
* Oszlopok:
  + Id, Domain, URL
  + Content
  + Title
  + Authors
  + Keywords
  + Tags
  + Summary
  + Source
  + Type (11)
    - Fake news
    - Satire
    - Extreme bias
    - Conspiracy theory
    - State news (0 db)
    - Junk science
    - Hate news
    - Clickbait
    - Proceed with caution
    - Political
    - Credible
* Adatok: több millió
* Eloszlás: 6 kategória milliós méretnél, ebből kettő 2milliónál is
* Megjegyzés
  + Le kell tölteni, hogy tényleg minden működik-e
  + Hatalmas adatbázis, nagyszerűnek tűnik
  + Rengeteg lehetőség van benne

**7. Liar dataset**

* Link: <https://arxiv.org/pdf/1705.00648.pdf>
  + Letöltési link: <http://cs.ucsb.edu/~william/data/liar_dataset.zip>
* Oszlopok
  + Json – API-ról elvileg letölthető a body
    - Pl. <https://www.politifact.com//api/v/2/statement/8428/?format=json>
  + Label
    - False
    - Barely true
    - Half-true
    - Mostly-true
    - True
    - …
  + Title
  + Category – multihalmaz, több témakör is lehet
  + Néhány további adat
    - Column 1: the ID of the statement ([ID].json).
    - Column 2: the label.
    - Column 3: the statement.
    - Column 4: the subject(s).
    - Column 5: the speaker.
    - Column 6: the speaker's job title.
    - Column 7: the state info.
    - Column 8: the party affiliation.
    - Column 9-13: the total credit history count, including the current statement.
* Adatok: 10.000 training + 1200 validation + 1200 test
* Eloszlás: nagyjából ugyanannyi minden kategóriából
* Megjegyzés:
  + Ez tűnik az eddigi legjobbnak
  + Elég sok adat
  + Kiegyensúlyozott
  + Teljes szövegek is leszedhetők
  + Tudományos cikkben hivatkozott
  + Human labeled
  + Kicsit trükkösnek tűnik az API