Assignment 3: Covid-19 Fact Checking

Team Coders For Hire:
Ihsaan Malek, 40024975 & Olivier Racette 40017231

Datasets

- Very dirty/noisy
  - Emojis
  - Twitter Hyperlinks -> Unique
  - Different Languages
  - Grammatical
    - Slang
    - Plurals vs singular, ex: case vs cases
  - Numbers
    - How do you determine if a number is relevant or correct
  - Stopwords
- These all affect the probability

Regular vocab vs Filtered Vocab

- Sample Length:
  - Regular:
    - Yes: 3316
    - No: 1640
  - Filtered:
    - Yes: 1113
    - No: 732
    - Takes out unique elements

- Leads to a better model as it removes any non unique word. Boost the conditional weight of all words with frequency >= 2

Summary of Model Performances

MODEL	*Accuracy*	F1-SCORE		*Recall*		*Precision*	
		YES	**NO**	**YES**	**NO**	**YES**	**NO**
NB - Regular	0.67	0.77	0.44	0.91	0.32	0.67	0.70
NB - Filtered	0.75	0.82	0.59	0.94	0.46	0.72	0.83
LSTM (averages)	0.73	0.81	0.55	0.90	0.45	0.73	0.77

Extra: Sanitizer

Cleans strings of:

- emoji
- &amp
- url
- punctuation

Sample:

“day 6 of quarantine: my dad gave a talk to the cats about covid-19 😭😭😭😭😭😭😭😭 https://t.co/mLdFXzcE2P”

“day 6 of quarantine my dad gave a talk to the cats about covid19”

(Only really works for english; accents and non latin/roman characters are lost)

Performance

MODEL	*Accuracy*	F1-SCORE		*Recall*		*Precision*	
		YES	**NO**	**YES**	**NO**	**YES**	**NO**
Sanitized - Unfiltered	0.745	0.811	0.611	0.909	0.5	0.732	0.786
Unsanitized - Filtered	0.75	0.82	0.59	0.94	0.46	0.72	0.83

"More than 2" Filter acts like a less complex sanitizer

Contributions

Description of contributions and responsibilities

Main.py

Olivier, Ihsaan

model_eval.py

Olivier, Ihsaan

PPT Slides

Ihsaan, Olivier