Real and Fake news: Machine Learning Models to Predict the Truthfulness of the News

EXPLORATIVE DATA ANALYSIS

1) Features overview. Correlation matrices were generated to identify collinear features. If two variables had a correlation of > 0.95, one was discarded. Next, variance inflation factor (VIF) was calculated for each remaining feature. Variables were removed until all VIF scores were below 10. Exhibit 1 shows that despite this, we still have some strong collinearity but because all VIF scores are below 10, we should still be able to reasonably trust our coefficients.

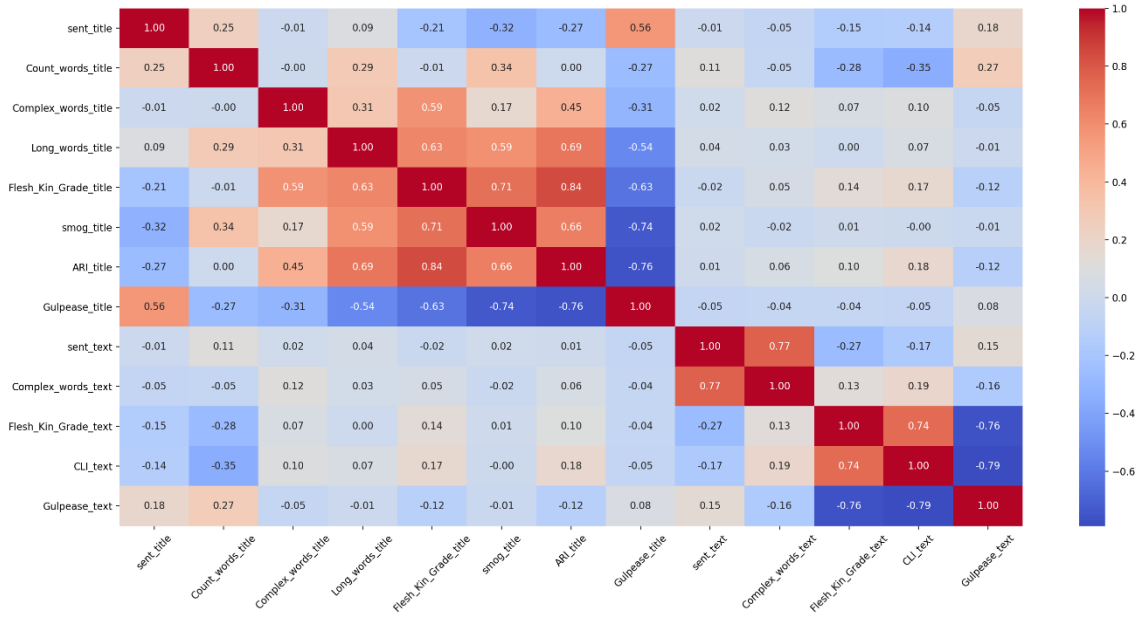

Exhibit 1. Correlation coefficients among basic text features and linguistic indexes

The table below summarizes the dataset features that will be examined during exploratory data analysis and utilized for training and testing of the machine learning classifiers.

Class	News Categories	Basic text features	Linguistic indexes	Processed text
Real (0) Fake (1)	- Government - Middle-east news - News - US news - Left news - Politics - Politics news	<u>Title:</u> - n. sentence - n. words - n. long words <u>Text:</u> - n. sentences - n. complex words	<u>Title:</u> - Flesh Kincaid Grade - SMOG - ARI - Gulpease <u>Text:</u> - Flesh Kincaid Grade	- Lemmatized title words - Lemmatized text words

	- World news		- CLI - Gulpease	

2) Classes distribution. The first analysis to be performed was examining whether the number of real and fake news was similar: the final dataset included 38614 observations evenly distributed between real (55%) and fake (45%) (Exhibit 2) news which confirms that the two classes are well balanced.

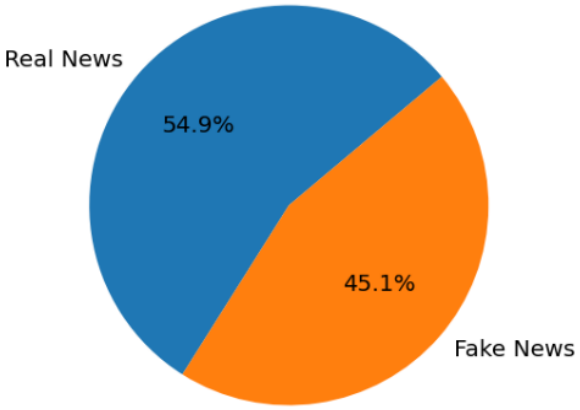

Exhibit 2. Dataset Classes distribution

3) Features analysis. Both basic text features and linguistic indexes were analyzed. Key findings are reported below.

a) Flesh Kincaid index. This readability index indicates how difficult a passage in English is to understand. It is calculated using the number of words, sentences and syllables (Appendix). The Flesh Kincaid index of real and fake news' titles is similar (Exhibit 3) ($p \sim 0.0813$). However, when the text is analyzed, real news have higher Flesh Kincaid score and, thus, more complex texts ($p<0.01$). The text of several fake news has very small, even negative, Flesh Kincaid index.

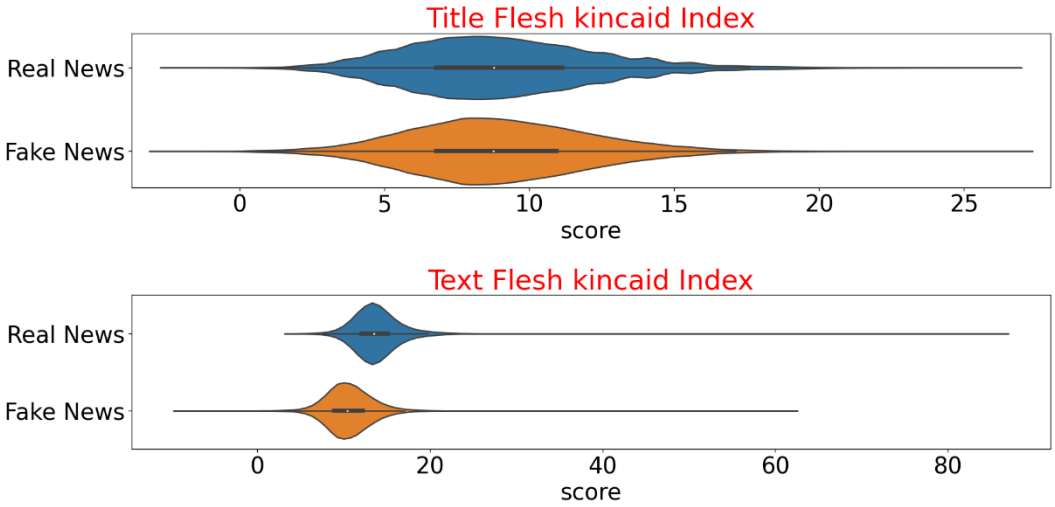

Exhibit 3. Flesh Kincaid index of title and text of real and fake news

Excerpt of the real news with the highest Flesh Kincaid index (~86.17): *"….. larry nichols co founder of devon energy corp james connaughton ceo of nautilus data technologies and a former environmental adviser to president george bush rick perry republican former texas governor navy admiral mike rogers director of the national security agency ronald burgess retired usa army lieutenant general and former defense intelligence agency chief robert cardillo director of the national geospatial intelligence agency pete hoekstra republican former usa representative from michigan rudy giuliani republican former mayor of new york debra wonyng a former usa attorney who was appointed by former president george bush ralph ferrara a securities attorney at law firm proskauer rose llp paul atkins a former sec commissioner who heads trump transition team for independent financial regulatory agencies daniel gallagher republican former sec commissioner john allison a former ceo of regional bank bb corp and former head of the cato institute a libertarian think tank paul atkins former sec commissioner thomas hoenig federal deposit insurance corp vice chairman and former head of the kansas city federal reserve bank dan dimicco former ceo of steel producer nucor corp robert lighthizer former deputy usa trade representative during the reagan administration mick mulvaney republican usa representative from south carolina david malpass former chief economist with investment bank bear stearns and a senior trump adviser rick perry former texas governor chuck conner a former acting secretary of the usa……"*

Excerpt of the fake news with the highest Flesh Kincaid index (~61.80): *"…orange county orange county riverside county sacramento county san bernardino county san diego county san francisco county san mateo county santa ana santa clara county santa cruz county sonoma countycoloradoarapahoe county aurora boulder county denver county garfield county grand county jefferson county larimer county mesa county pitkin county pueblo county routt county san miguel county weld countyconnecticuteast haven hartforddistrict of columbiawashingtonfloridaalachua county clay county hernando countygeorgiaclayton county dekalb countyiowabenton county cass county franklin county fremont county greene county ida county iowa city iowa city johnson county jefferson county marion county monona county montgomery county pottawattamie county sioux countyillinoischicago cook countykansasbutler county harvey county sedgwick county shawnee county story countylouisiananew orleansmassachusettsamherst boston cambridge lawrence northhampton somervillemarylandbaltimore montgomery county prince george countyminnesotahennepin countynebraskahall county sarpy countynew jerseymiddlesex county newark ocean county union countynew mexicobenalillo new mexico county jails san miguelnevadaclark county washoe countynew yorkfranklin county ithaca nassau county new york city omondaga county st lawrence county wayne countyoregonbaker county clackamas county clatsop county coos county crook county curry…"*

b) Coleman Liau Index (CLI). CLI calculation (Appendix) is similar to the Flesh Kincaid index's except that in CLI the number of syllables is substituted with the number of complex words (words with more than 6 characters). Real news have higher text CLI score than the fake news' text (Exhibit 4) ($p<0.01$).

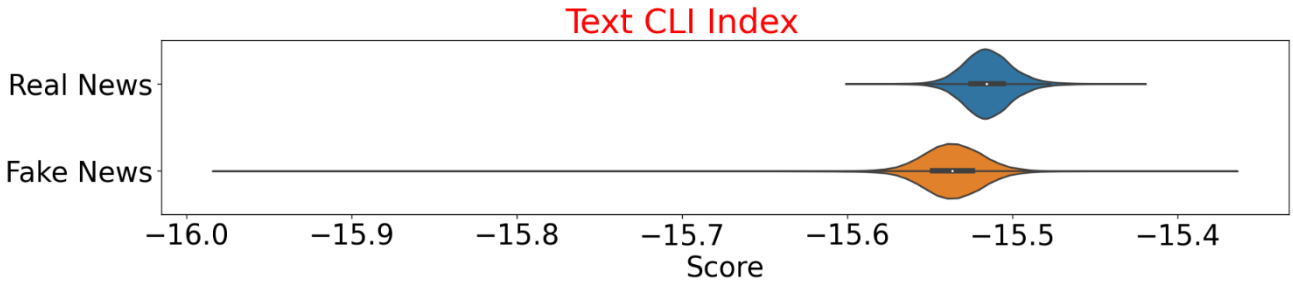

Exhibit 4. Text CLI score for real and fake news

c) Gulpease index. The Gulpease score is a readability index scaled on the Italian language. The title of real news has higher Gulpease score than title of fake news ($p<0.01$). This trend is inverted when the text is analyzed: Gulpease score is higher in the text of fake news (Exhibit 5).

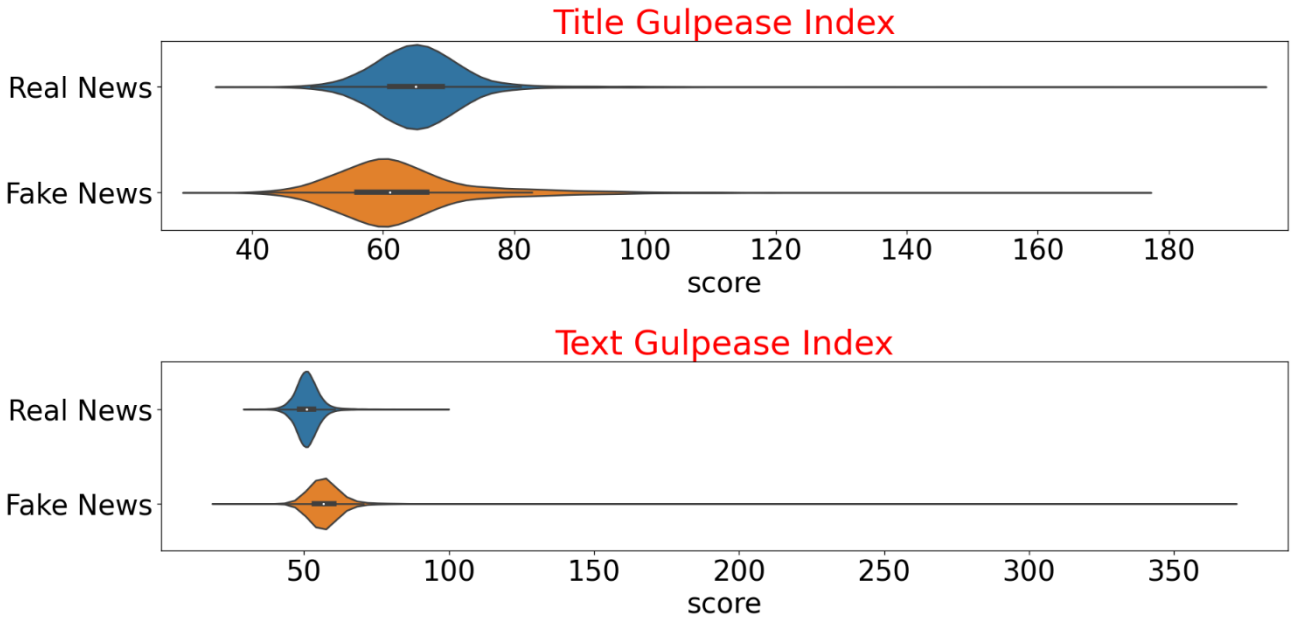

Exhibit 5. Gulpease score of title and text of real and fake news

***Real* news title with smallest Gulpease index (36.5):** *"trump considering representative barletta for transportation secretary: politico"*

***Fake* news title with smallest Gulpease index (32.68):** *"update: judge orders cancellation of redskins trademark registration – washington redskin's new stadium construction held hostage by petty obama administration"*

d) Scores comparison insight. In (b) it was shown that CLI can discriminate between real news from fake news' title, something that the Flesh Kincaid index could not achieve (a). This change is caused by the fact that every index selectively amplifies one or more linguistic properties within the same word and sentence and, thus, generates its own unique distribution (Appendix). Exhibit 6 shows an extreme example of how dramatically the score distribution can change. SMOG index includes the number of polysyllable and sentences, while Gulpease score utilized number of sentences, words and characters. While SMOG scores are spread according to a multimodal distribution, Gulpease score distribution is smooth and resembles a normal right-skewed distribution.

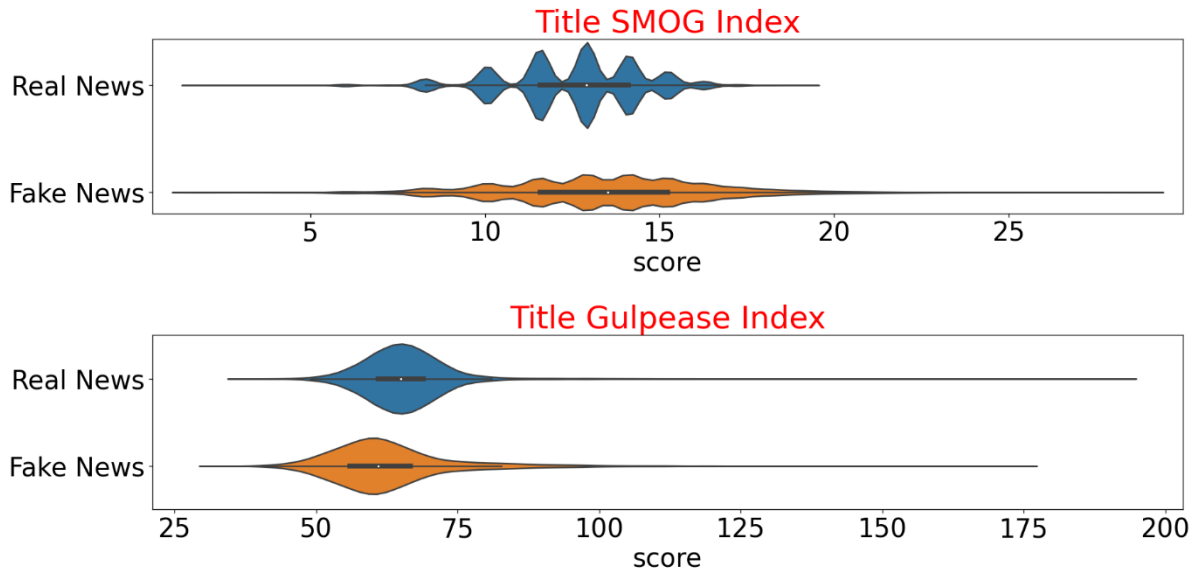

Exhibit 6 – Title SMOG and Gulpease scores comparison

d) **Relationship between linguistic indexes and truthiness of the news** – Scaled and unscaled basic text features and linguistic indexes (see section 1) were fed into logistic regression model to investigate their predictive power on the truthiness of the news. The most important predictive features are the number of words in the title, the Flesh Kincaid index of the text (Exhibit 7, top chart), the CLI index of the text and the number of sentences in the title (Exhibit 7, bottom chart).

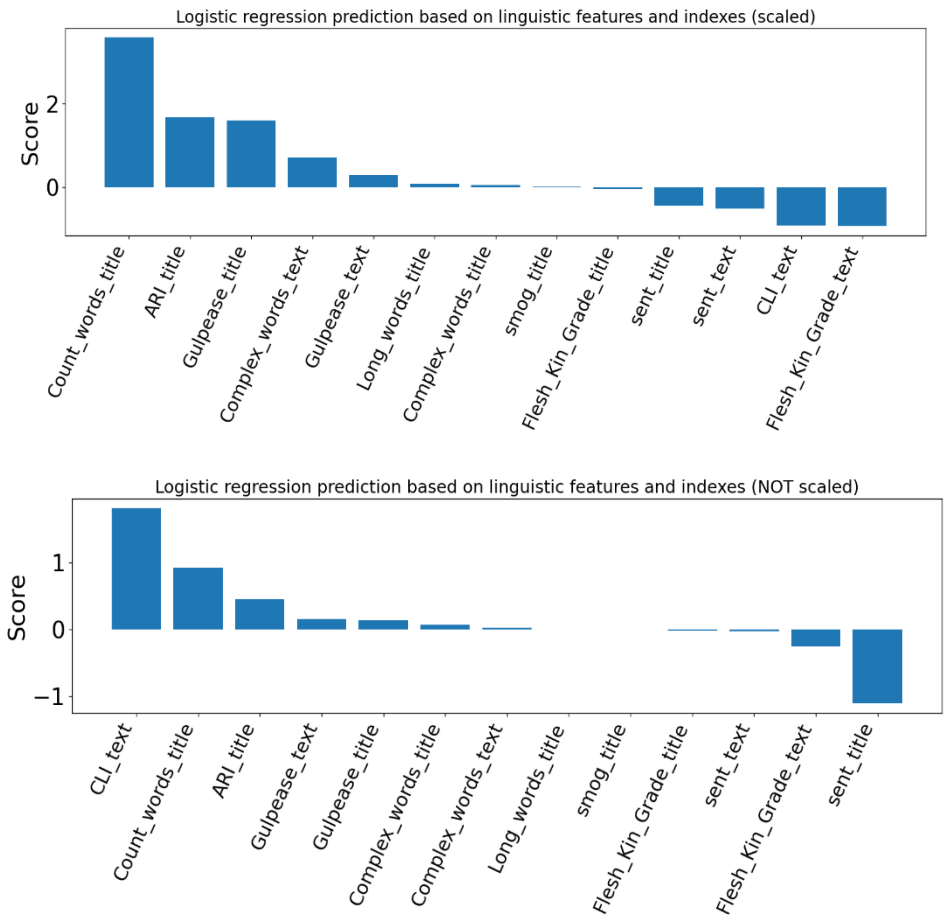

Exhibit 7 – Logistic regression coefficients for the linguistic indexes and text features

6) Lemmatized text analysis. The dictionary created from the lemmatized body of the news includes 96756 words. 80% and 90% of the words appearing less than 14 and 55 documents, respectively (Exhibit 8).

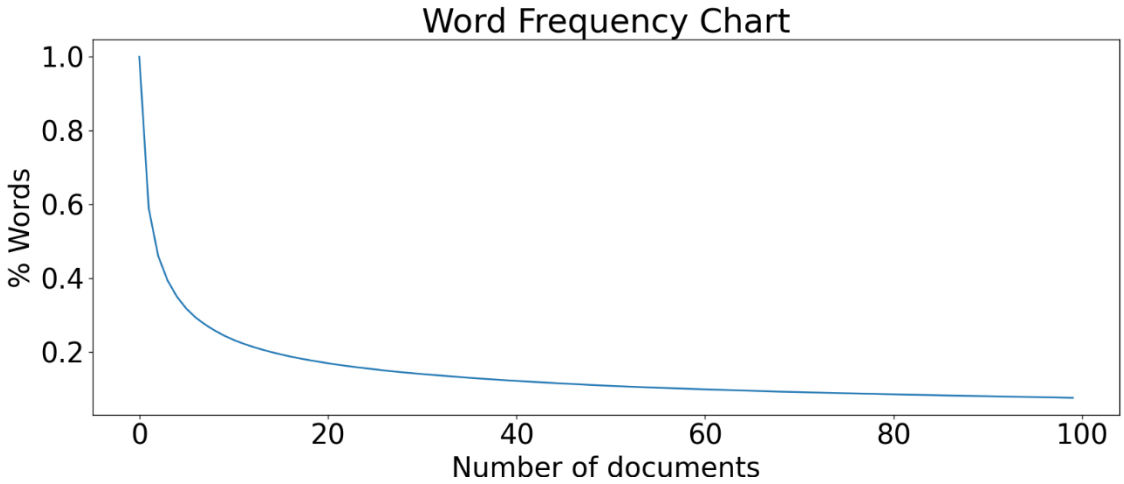

Exhibit 8. Graph showing the percentage of words (y-axis) appearing in at least n documents (x-axis)

Exhibit 9 shows the 20 most frequent words in real (top) and fake (bottom) news. The words *trump* and *say* are highly mentioned in both fake and real news. The word *reuters* is present in every real new. Tracking back to the source of the dataset containing the real news, led me to discover that all the real news were, indeed, collected from the Reuters website. This finding will be considered during testing of the optimization of the final predictive model.

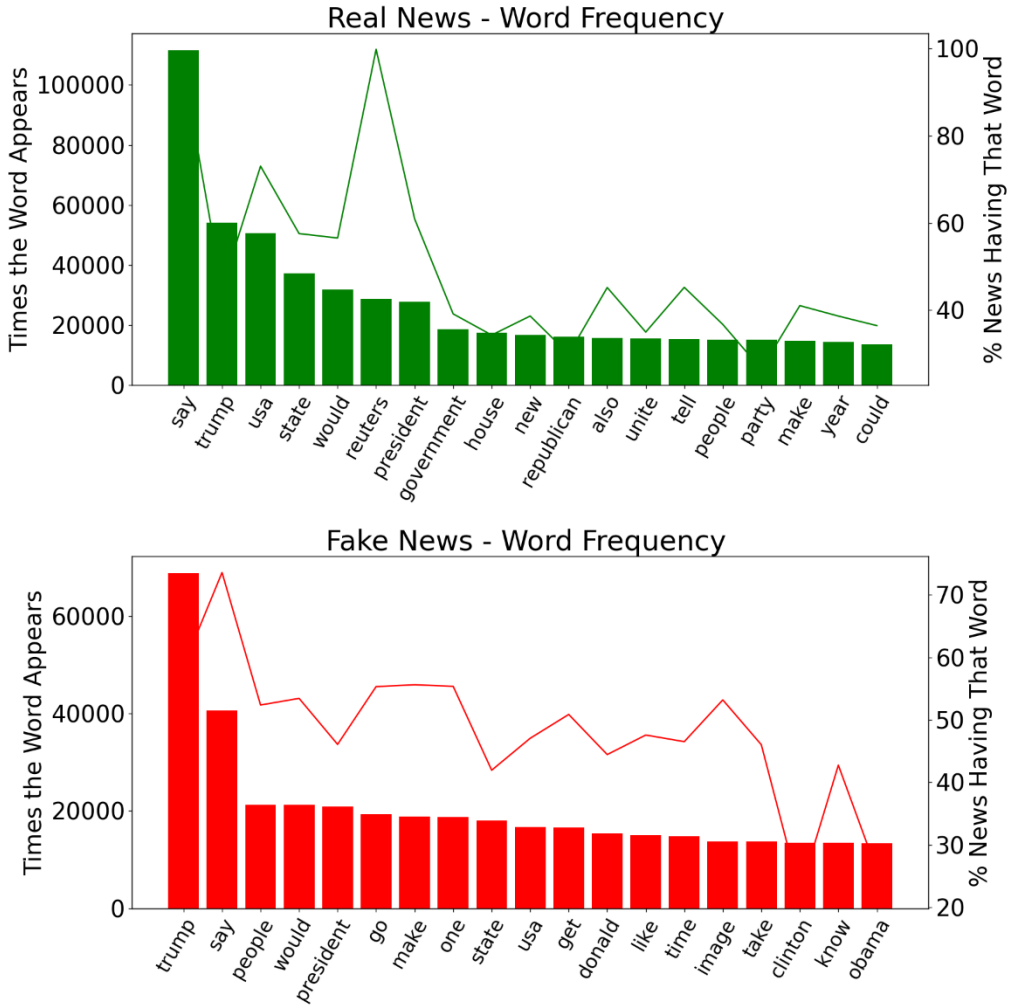

Exhibit 9 – Most frequent words in the news. The number of times the most common words are present in the news are reported on the left y-axis. The right y-axis shows the percentage of news having that word.

Next a predictive multinomial Naïve Bayes model was trained to determine which words have the most predictive power on truthiness of the news. The 60 most predictive words for real and fake news are depicted in Exhibit 10. Vulgarities, twitter/youtube links and flickr/getty related words (*getty*, *screenshot*, *flickr*, *getty, raedle*, *angerer*) are enriched in the fake news. On the other hand, real news are characterized by the presence of bitly/tmsnrt links and name

of non popular politicians (*Odinga* - Kenyan politician, *Rajoy* - Spanish politician, *Hariri* - Lebanese politician, *Gulen* - Turkish Islamic scholar) and name of cities and countries rarely mentioned by the mainstream media (*Kirkuk* - city in Iraq, *Nairobi* -Kenya's capital, *Marawi* - Islamic city, *Harare* - capital of Zimbabwe, *Rakhine* - state in Myanmar). Several acronyms appear in both real and fake news. Their meaning is reported in the appendix.

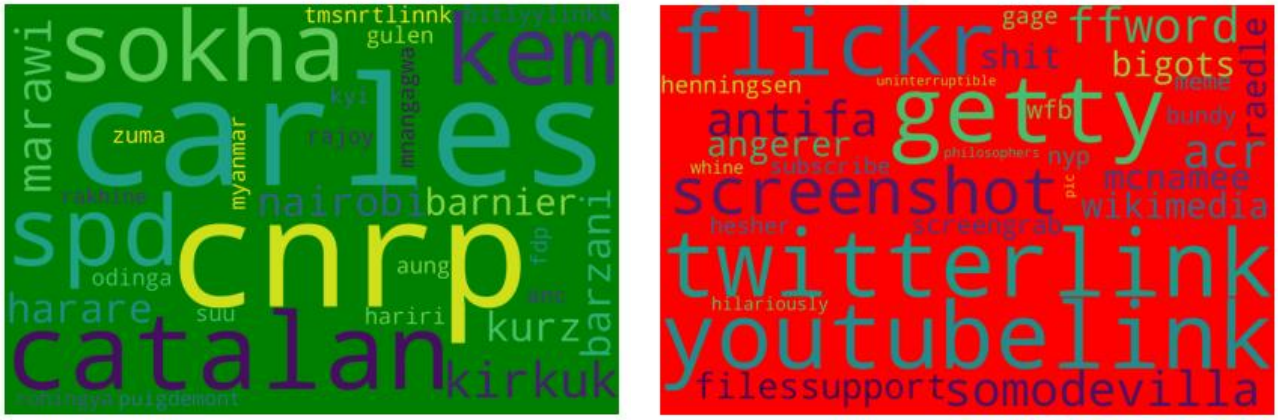

Exhibit 10 – Word cloud for the words with highest predict power for real (left) and fake (right) news.

APPENDIX

1) Flesch Kincaid index:

$$0.39\frac{Total\ words}{Total\ sentences}+11.8\frac{Total\ Syllables}{Total\ words}-15.59$$

2) Coleman Liau Index (CLI):

$$CLI=0.0588\frac{characters}{words}-0.296\frac{sentences}{words}-15.8$$

3) Smog index (Simple Measure of Gobbledygook):

$$1.0430\sqrt{number\ of\ polysyllables\frac{30}{number\ of\ sentences}}+3.1291$$

4) Gulpease index.

$$89+\frac{300*sentences-10*characters}{words}$$

5) Most predictive words **with hyperlinks' tokens**

```
prob -6.751 for the real-new-word:  Rohingya: Indo-Aryan ethnic group
prob -6.646 for the real-new-word:  Rakhine: state in Myanmar
prob -6.310 for the real-new-word:  Puigdemont: Catalan politician
prob -6.299 for the real-new-word:  zuma: Jacob Zuma South African politician
prob -6.151 for the real-new-word:  Myanmar: Southeast Asian nation
prob -6.111 for the real-new-word:  fdp: Free Democratic Party
prob -6.042 for the real-new-word:  kyi: San Suu Kyi, Nobel laurate
prob -6.032 for the real-new-word:  suu: San Suu Kyi, Nobel Peace
prob -5.828 for the real-new-word:  Mnangagwa: Zimbabwean revolutionary
prob -5.753 for the real-new-word:  anc: African National Congress
prob -5.720 for the real-new-word:  Odinga: Kenyan politician
prob -5.713 for the real-new-word:  rajoy: Spanish politician
prob -5.678 for the real-new-word:  hariri: Lebanese politician
prob -5.675 for the real-new-word:  gulen: Turkish Islamic scholar
prob -5.579 for the real-new-word:  tmsnrtlinnk: token link
prob -5.531 for the real-new-word:  aung: Burmese politician
prob -5.527 for the real-new-word:  bitlyylinkk: token link
```

```
prob -5.502 for the real-new-word:  barnier: French politician
prob -5.498 for the real-new-word:  Barzani: Kurdish politician
prob -5.424 for the real-new-word:  Harare: capital of Zimbabwe
prob -5.383 for the real-new-word:  kurz : chancellor of Austria
prob -5.299 for the real-new-word:  Nairobi: Kenya's capital
prob -5.283 for the real-new-word:  marawi: Islamic City of Marawi
prob -5.249 for the real-new-word:  Kirkuk: city in Iraq
prob -5.202 for the real-new-word:  kem: Kem Sokha, Cambodian politician
prob -5.191 for the real-new-word:  sokha: Kem Sokha, Cambodian politician
prob -5.071 for the real-new-word:  spd: german pollical party
prob -5.059 for the real-new-word:  catalan
prob -5.056 for the real-new-word:  cnrp: Cambodia National Rescue Party
prob -5.050 for the real-new-word:  carles: former covert cia
----------------------------------------
----------------------------------------
prob -0.011 for the fake-new-word:  pic
prob -0.011 for the fake-new-word:  philosophers
prob -0.011 for the fake-new-word:  uninterruptible
prob -0.011 for the fake-new-word:  whine
prob -0.011 for the fake-new-word:  hilariously
prob -0.010 for the fake-new-word:  hasher: acr hosts hesher
prob -0.009 for the fake-new-word:  subscribe
prob -0.008 for the fake-new-word:  nyp: new York post
prob -0.008 for the fake-new-word:  gage: gage skidmore
prob -0.008 for the fake-new-word:  meme
prob -0.008 for the fake-new-word:  bundy: American cattle rancher (and son)
prob -0.008 for the fake-new-word:  Henningsen: patrick Henningsen, writer
prob -0.008 for the fake-new-word:  wfb: via:wfb
prob -0.007 for the fake-new-word:  screengrab
prob -0.006 for the fake-new-word:  raedle: Joe Raedle getty images
prob -0.006 for the fake-new-word:  shit
prob -0.006 for the fake-new-word:  bigots
prob -0.006 for the fake-new-word:  wikimedia
prob -0.005 for the fake-new-word:  angerer: drew angerer getty images
prob -0.005 for the fake-new-word:  mcnamee: American businessman
prob -0.004 for the fake-new-word:  filessupport
prob -0.003 for the fake-new-word:  acr: alternate current radio
prob -0.003 for the fake-new-word:  ffword
prob -0.003 for the fake-new-word:  antifa
prob -0.003 for the fake-new-word:  somodevilla: photojournalists
prob -0.003 for the fake-new-word:  screenshot
prob -0.002 for the fake-new-word:  flickr
prob -0.002 for the fake-new-word:  youtubelink: token link
prob -0.001 for the fake-new-word:  getty
prob -0.000 for the fake-new-word:  twitterlink: token link
```

6) Most predictive words **when tokens for the hyperlinks are all removed**

```
prob -6.749 for the real-new-word:  rohingya
prob -6.644 for the real-new-word:  rakhine
prob -6.308 for the real-new-word:  puigdemont
prob -6.297 for the real-new-word:  zuma
prob -6.149 for the real-new-word:  myanmar
prob -6.108 for the real-new-word:  fdp
```

prob -6.040 for the real-new-word: kyi
prob -6.030 for the real-new-word: suu
prob -5.826 for the real-new-word: mnangagwa
prob -5.751 for the real-new-word: anc
prob -5.718 for the real-new-word: odinga
prob -5.711 for the real-new-word: rajoy
prob -5.676 for the real-new-word: hariri
prob -5.673 for the real-new-word: gulen
prob -5.529 for the real-new-word: aung
prob -5.500 for the real-new-word: barnier
prob -5.496 for the real-new-word: barzani
prob -5.422 for the real-new-word: harare
prob -5.381 for the real-new-word: kurz
prob -5.297 for the real-new-word: nairobi
prob -5.281 for the real-new-word: marawi
prob -5.247 for the real-new-word: kirkuk
prob -5.200 for the real-new-word: kem
prob -5.189 for the real-new-word: sokha
prob -5.069 for the real-new-word: spd
prob -5.057 for the real-new-word: catalan
prob -5.054 for the real-new-word: cnrp
prob -5.048 for the real-new-word: carles
prob -5.041 for the real-new-word: **provincial**
prob -5.014 for the real-new-word: **asean** Association Southeast Asian Nations
----
----
prob -0.012 for the fake-new-word: **masochists**
prob -0.012 for the fake-new-word: **hissy**
prob -0.011 for the fake-new-word: uninterruptible
prob -0.011 for the fake-new-word: pic
prob -0.011 for the fake-new-word: philosophers
prob -0.011 for the fake-new-word: whine
prob -0.011 for the fake-new-word: hilariously
prob -0.010 for the fake-new-word: hesher
prob -0.009 for the fake-new-word: subscribe
prob -0.008 for the fake-new-word: nyp
prob -0.008 for the fake-new-word: meme
prob -0.008 for the fake-new-word: gage
prob -0.008 for the fake-new-word: bundy
prob -0.008 for the fake-new-word: henningsen
prob -0.008 for the fake-new-word: wfb
prob -0.007 for the fake-new-word: screengrab
prob -0.006 for the fake-new-word: raedle
prob -0.006 for the fake-new-word: shit
prob -0.006 for the fake-new-word: bigots
prob -0.006 for the fake-new-word: wikimedia
prob -0.005 for the fake-new-word: angerer
prob -0.005 for the fake-new-word: mcnamee
prob -0.004 for the fake-new-word: filessupport
prob -0.003 for the fake-new-word: acr
prob -0.003 for the fake-new-word: ffword
prob -0.003 for the fake-new-word: antifa
prob -0.003 for the fake-new-word: somodevilla
prob -0.003 for the fake-new-word: screenshot
prob -0.002 for the fake-new-word: flickr
prob -0.001 for the fake-new-word: getty