ABSTRACT

The unchecked spreading of fake news is an alarming phenomenon on every social media platform and information outlet. Since online content can have a decisive effect on users' political, social and business decisions and opinions, the identification and elimination of false information has become a critical prerogative for many online companies.

Natural language processing (NLP) encompasses linguistics, computer science, information engineering, and artificial intelligence. The main application of NLP is to train algorithms to analyze texts and extrapolated actionable insights based on text's word composition and frequency. Another way of examining a document is analyzing the document's intrinsic complexity. Text complexity is measured by several linguistic indexes such as Flesch Kincaid and Coleman Liau indexes.

Herein, I tested several machine learning models that leverage NLP techniques, linguistic indexes or both, to predict the news' truthiness. The best model I built was able to discriminate real from fake news with a balanced accuracy of 0.9779 and AUC score of 0.9992.

DATA OVERVIEW AND EDA

Data overview

The data utilized in this project consists of two separate csv files, one harboring all the real news (24417), one all the fake ones (23502 news) (https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset). The two files were compiled by two independent sources and thus could be inherently different. To minimize this suspected external bias, I applied a very aggressive data wrangling. The following features were available for each news:

- Date the news was published. The oldest news in the available dataset was published on March 31st, 2015, the last one on February 19th, 2018.
- Title of the news.
- Text of the news.
- Category of the news (Government News, Middle-east, News, US_News, left-news, politics, politicsNews or worldnews).

Data wrangling

Data wrangling of the dataset required numerous steps which are detailed in the Jupiter notebook available in the GitHub repository. Herein, I will mention some of most critical steps. I filtered out all news with text shorter than 8 characters. Both real and fake news presented cross site scripts which were completely removed. Next, Facebook, Twitter, Youtube, Bitly and Tmsnrt hyperlinks were converted into 11-character long tokens for further analysis. All the other hyperlinks were removed. Vulgarities were replaced with more politically correct words and verbal contractions were expanded out. At this point, text and title of the news were ready for sentences, words, syllables, long words, polysyllabic words and characters counts. Finally, all the linguistic indexes were calculated. Words in the clean text and title were then lemmatized. Standard English stop words were removed during this last step.

Before training and tuning the machine learning model, the number of features was trimmed down to reduce overall variables collinearity. 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. Table 1 summarizes the dataset features available 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 - World news	Title: - n. sentence - n. words - n. long words Text: - n. sentences - n. complex words	Title: - Flesh Kincaid Grade - SMOG - ARI - Gulpease Text: - Flesh Kincaid Grade - CLI - Gulpease	- Lemmatized title words - Lemmatized text words

Table 1 – Final Dataset Specification.

Explorative data analysis

The final dataset included 38614 observations evenly distributed between real (55%) and fake (45%) (Exhibit 1) news which confirms that the two classes are well balanced.

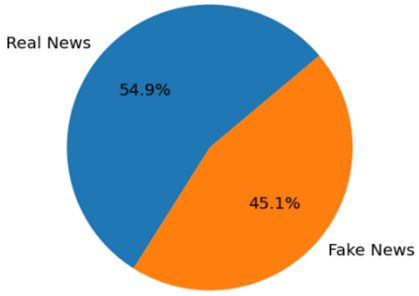

Exhibit 1. Dataset Classes distribution

Above I mentioned that every news was assigned to one or more of the following categories: Cat_Government News, Cat_Middle-east, Cat_News, Cat_US_News, Cat_left-news, Cat_politics, Cat_politicsNews, Cat_worldnews. The criteria utilized for this classification were not known. When data were grouped in real and fake news (Exhibit 2), real news belonged only to either politics-news or world-news categories. Meanwhile, all the fake news were assigned to the remaining six categories. This dichotomic categorization clearly reflects the different sources real and fake news were gathered from. This classification will therefore not be included in the training and testing of the ML classifier.

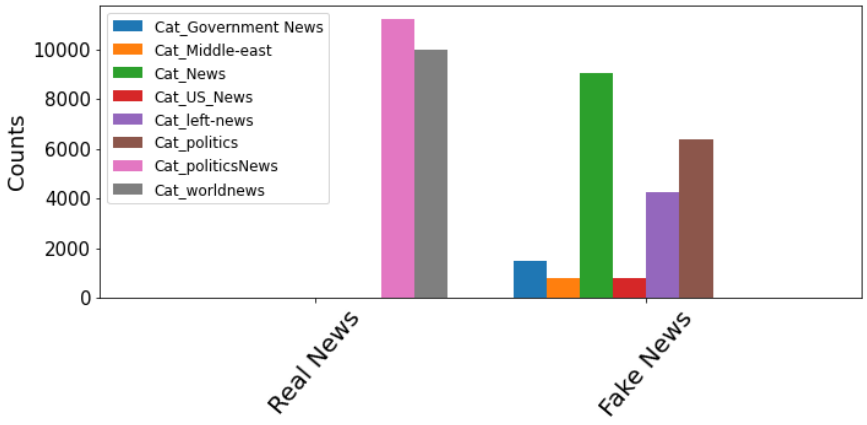

Exhibit 2. Real and fake new distribution among the eight categories.

During data wrangling I introduced artificial tokens to track and analyze specific hyperlinks embedded in the text. Their frequency is shown in Exhibit 3. Tmsnrt and Bitly hyperlinks are found only in the text of real news. Conversely, Facebook, Twitter and Youtube hyperlinks are presently exclusively in fake news. This dichotomy may be the result of the different sources fake and real news came from, further supporting the intuition that two datasets may have some intrinsic dissimilarities. However, since these hyperlinks are tokenized, they can be easily removed in the successive tuning steps of the ML model to assess their contribution to the overall prediction.

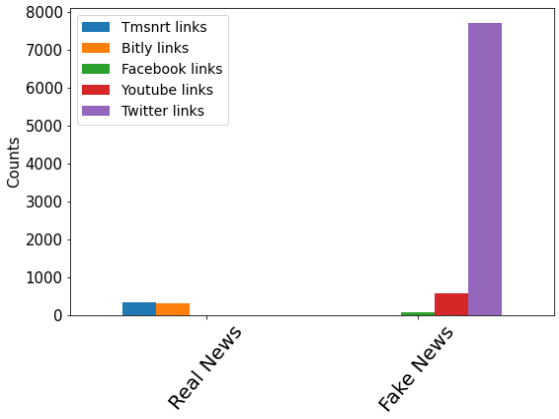

Exhibit 3. Hyperlinks frequency in real and fake news.

Linguistic indexes analysis

This FILE contains all the details of the exploratory data analysis. Below, I reported three of the most relevant findings.

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 4) (p~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.

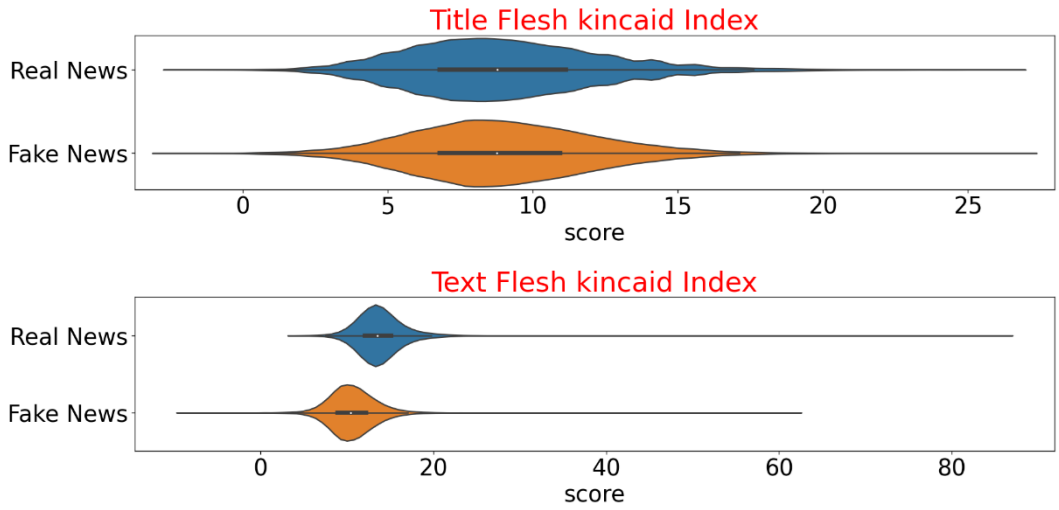

Exhibit 4. Flesh Kincaid index of title and text of real and fake news.

b) Gulpease index. The Gulpease score is a readability index normalized 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).

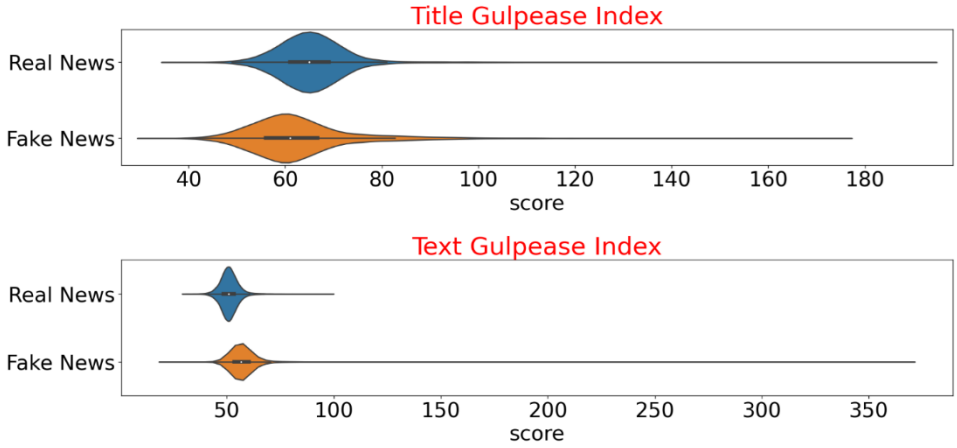

Exhibit 5. Gulpease score of title and text of real and fake news.

c) Most common words in the news.

Exhibit 6 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 news, since all the real news were indeed gathered from the Reuters website. To avoid the carryover of the news' source bias into the machine learning modeling, the word *reuters* was excluded from the vocabulary of the vectorizer (i.e it was treated as a stop word).

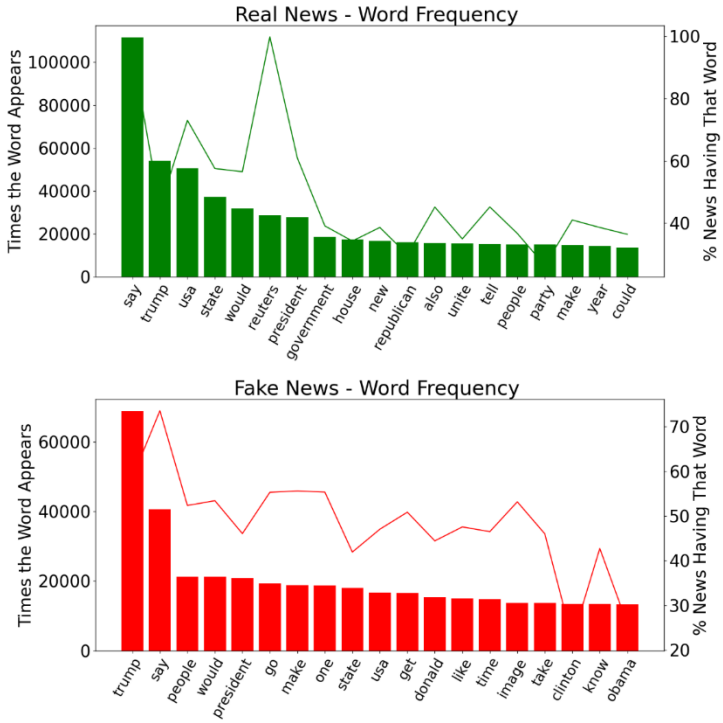

Exhibit 6 – Most frequent words in the news. The number of times the most common words are present in the news are reported on the left y-axis. The right y-axis shows the percentage of news having that word.

MACHINE LEARNING MODELING

Models employed. To predict the truthfulness of the news several models were implemented. Some models only exploited the linguistic characteristics of the text, some only the lemmatized words in the text, others both features.

GridSearch was deployed to find the best hyperparameters (see appendix for details). For details on the scripts please follow this link.

Model 1 - Multinomial Naïve Bayes with CountVectorized lemmatized text

First, I investigated the predictive power of the lemmatized words in the text of the news. The first parameter I adjusted was the lower cut-off for document frequency for the vectorizer (min_df). While lower values for min_df tend to improve prediction accuracy, it also is more computational expensive as the matrix representing the bag of the words is much larger. Exhibit 7 depicts how AUC score of the model changes with increasing values of the minimum document frequency.

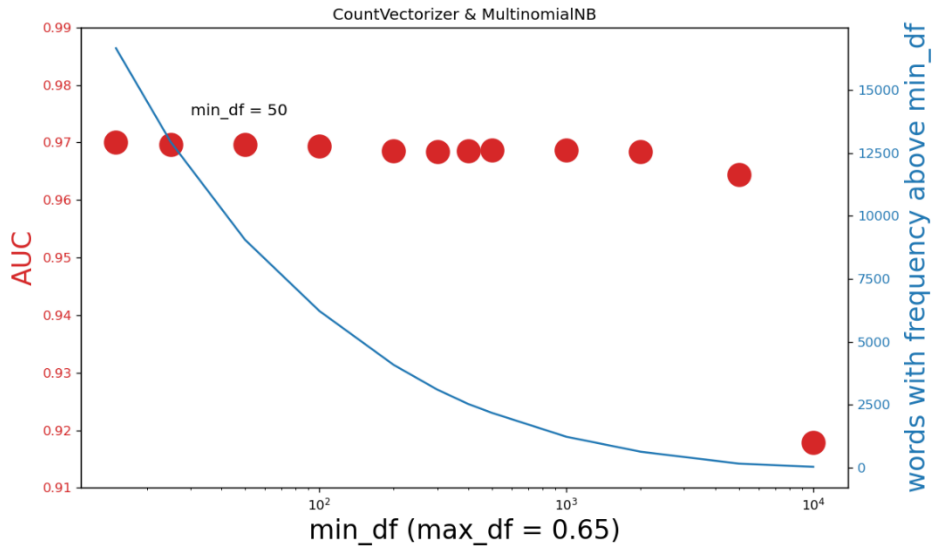

Exhibit 7. Model 2 performance with increasing document frequency.

Even though the best prediction is reached when each word appears in at least 15 documents, I selected the value of 50 for min_df for the reason hinted above: speed. The AUC score when min_df is 50 is just slightly lower than the AUC score obtained with min_df is 15 (0.9801 vs 0.9796), but the prediction is much faster.

Next, I extrapolated the 60 most predictive words for real and fake news (Exhibit 8):

- Fake news – key takeaways:
  - The presence of Twitter and YouTube links is strongly indicative of the fakeness of the news.
  - Photography related words (*getty*, *somodevilla, screenshot*, *flickr*, *getty, raedle*, *angerer*) are enriched in the fake news. Thus, if a news has an embedded picture in its text, is more likely to be fake.
  - The presence of vulgarities (*f**k, s**t*, *bigots*) in the text of the news is a hallmark of fake news.
- Real news – key takeaways:
  - Bitly/tmsnrt links are commonly embedded in the text of real news. Bitly links are customized short links that are commonly embedded into texts as replacement of long and distracting hyperlinks. Thus, if a text was cleaned up of intrusive long hyperlinks, probably belongs to a true news.
  - Name of non-mainstream politicians (*Odinga* - Kenyan politician, *Rajoy* - Spanish politician, *Hariri* - Lebanese politician, *Gulen* - Turkish Islamic scholar) and name of cities and countries rarely mentioned by the media (*Kirkuk* - city in Iraq, *Nairobi* -Kenya's capital, *Marawi* - Islamic city, *Harare* - capital of Zimbabwe, *Rakhine* - state in Myanmar) are highly common in real news. Therefore, if the news mentions the name of person or a city the reader heard only once or twice before, is probably real.

Several acronyms appear in both real and fake news. Their meaning is reported in the appendix.

Exhibit 8 – Word cloud for the words with highest predict power for real (left) and fake

To improve the performance of the model, different n-grams for the CountVectorizer were assessed. At the same time multiple values of the smoothing parameter alpha were evaluated. The model reached the highest ACU score 0.9954 with n-grams of 2, 3 and alpha = 0.01.

Model – 2. Multinomial Naïve Bayes with TFIDFVectorized lemmatized text.

The lemmatized words of the text of the news were fed into the TFIDFVectorizer and modeled with Multinomial Naïve Bayes. I found again that the AUC score is reached with the lowest min_df (Exhibit 9) but I decided to select the value of 50 for a faster prediction

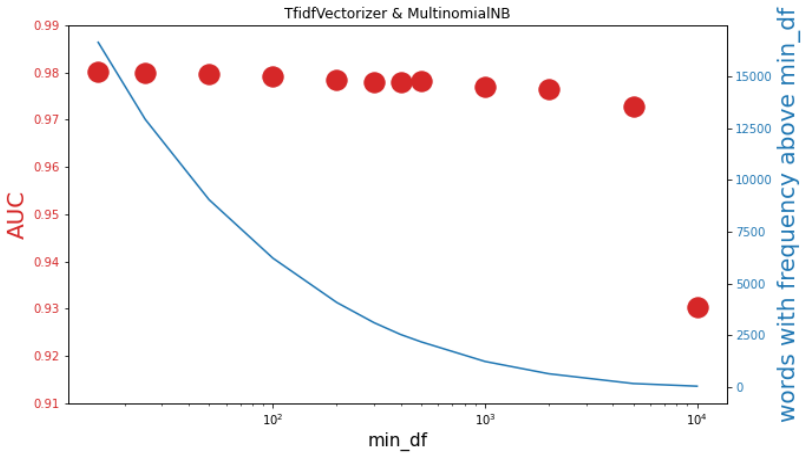

Exhibit 9. Model 2 performance with increasing document

Next, different n-grams and values of the smoothing parameter alpha were evaluated. Exhibit 10 graphs the performance of model 1 (MultinomialNB with CountVectorizer) and model 2 (MultinomialNB with TFIDFVectorizer) with different n-grams, side to side. The arrow indicates the most accurate model. The MultinomialNB classifier best performed with 2,3 n-grams TFIDFVectorizer.

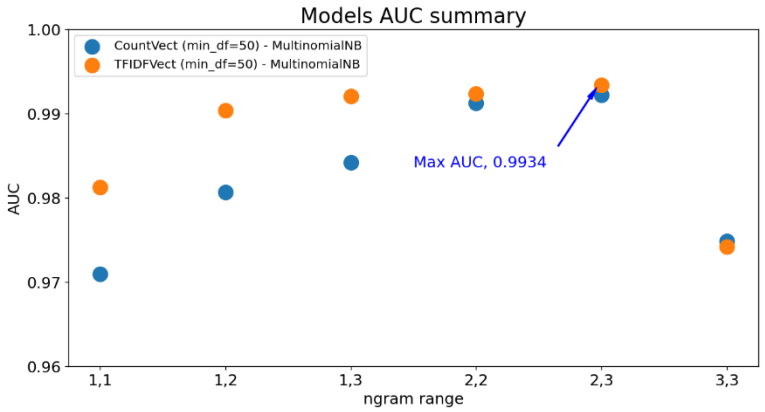

Exhibit 10. Model 1 & 2 balanced accuracy comparison with different n-grams.

Model -3. Logistic Regression with linguistic scores, text features

The main goal of this model is to highlight linguistic scores and text features with the highest predictive power on the truthiness of the news. Even though this model was the least accurate (balanced accuracy = 0.7649, AUC = 0.9431) among all the models tested, it helped to single out which linguistic features best separated real from fake news: number of words in the title, the Flesh Kincaid index of the text (Exhibit 11, top chart), CLI index of the text and the number of sentences in the title (Exhibit 11, bottom chart).

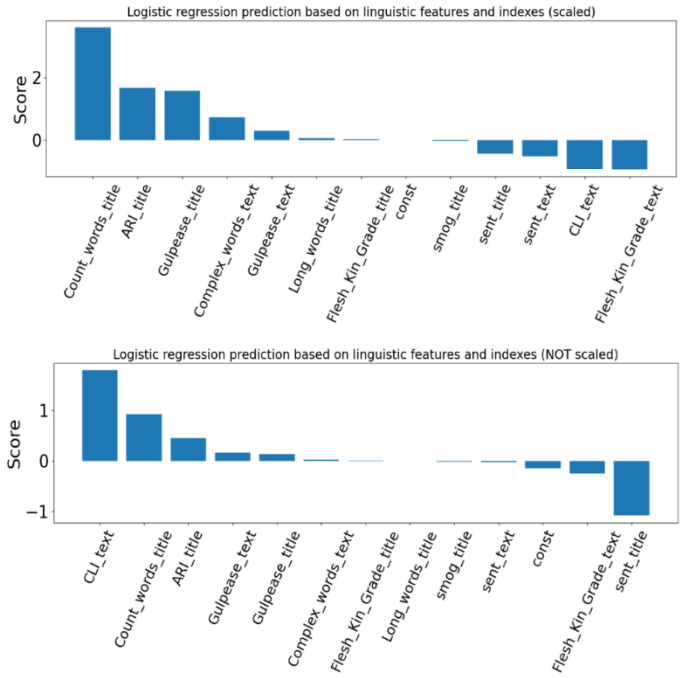

Exhibit 11 – Logistic regression coefficients for the linguistic indexes and text features.

Models 4 & 5. Logistic Regressor with lemmatized words, linguistic scores, basic text features

Next, I concatenated the vectorized lemmatized words with the linguistic features and trained a logistic regressor algorithm. Model 4 and 5 differs by which vectorizer was implemented. In model 4 I used the CountVectorizer, in model 5 the TFIDFVectorizer. For both vectorizers I set min_df to 50 and n_gram range to 2,3. Different value for the regularization hyperparameter C of the logistic regressor were tested. Among the two models, model 5 is the best performing classifier.

Models 6 & 7. Random Forest Classifier with lemmatized words, linguistic scores, basic text features

Finally, a random forest classifier was implemented. As input, I plugged in the vectorized lemmatized words concatenated with the linguistic features. For model 6 I used the CountVectorizer, in model 7 the TFIDFVectorizer. When the TFIDFVectorizer is utilized the random forest classifier performs slightly better, which perfectly aligns with the above observations.

Models comparison.

Table 2 summarizes the characteristics and performance of the 7 tested models. Model 5 has the highest AUC score and was utilized in few additional analysis.

	Classifier	Extraction Technique	Linguistic scores & text features	Balance Accuracy	AUC Score
Model 1	MultinomialNB (alpha =0.01)	CountVectorizer (min_df = 50, max_df =0.65, ngram_range = 2,3)	YES	0.9336	0.9922
Model 2	MultinomialNB (alpha =0.01)	TFIDFVectorizer (min_df = 50, max_df =0.65, ngram_range = 2,3)	NO	0.9262	0.9934
Model 3	Logistic Regressor (C=0.1)	None	NO	0.7649	0.9431
Model 4	Logistic Regressor (C=1)	CountVectorizer (min_df = 50, max_df =0.65, ngram_range = 2,3)	YES	0.9748	0.9987
Model 5	Logistic Regressor (C=200)	TFIDFVectorizer (min_df = 50, max_df =0.65, ngram_range = 2,3)	YES	0.9755	0.9992
Model 6	Random Forest (n_estimators = 100, min_samples_leaf = 1)	CountVectorizer (min_df = 50, max_df =0.65, ngram_range = 2,3)	YES	0.9632	0.9978
Model 7	Random Forest (n_estimators = 100, min_samples_leaf = 1)	TFIDFVectorizer (min_df = 50, max_df =0.65, ngram_range = 2,3)	YES	0.9637	0.9981

Table 2 – ML model summaries.

Balanced accuracy of model 5 is very high, but not 100 %. The confusion matrix shows that most of the misclassifications (80) are false negatives (65). Below I provide an excerpt of a fake news which was classified as real by the classifier:

> *"….i will still have that same big-shouldered chicago lust for power that drove me from greenwood avenue in hyde park to pennsylvania avenue in three short years. but if hillary replaces me in the oval, she and bill will take control of the democratic party it will become the clinton party once again and they will block me from having any future influence. i will end up like jimmy carter hammering away in appalachia for habitat for humanity. on the other hand, if a republican wins in 16, the clintons will be finished their foundation and their speaking fees will dry up and they will be a thing of the past. but i will still be the titular head of the party. i will be able to continue my push to transform america into a european-style socialist state. so, personally, i would be better off if the next president is a republican…"*

Assuming that no mistakes were made during compiling of the fake news dataset, what stands out is that the text above could easily belong to a real news. There is nothing in the text that would lead the reader to suspect fakeness. There are no vulgarities, no excess of punctuation, the grammar is correct, and the document even conveys a sound message. This is an interesting point as it shows that if a news can trick a human reader, can also do so with the classifier.

Next, I analyzed the false positives, e.i. real news classified as fake. Below I reported one of the 25 misclassified real news:

> *" washington (reuters) - actor and director george clooney, a supporter of hillary clinton's presidential bid, broke ranks over campaign financing on saturday to condemn the 'obscene' sums of money in usa politics and praised clinton's chief political rival in the process. clooney made the remarks in an interview with nbc news' 'meet the press' the day after he and his wife, amal, hosted a fundraiser on democratic party hopeful clinton's behalf friday night with a price tag of up to $353,400 per couple. 'we had some protesters last night when we pulled up in san francisco and they are right to protest. they are absolutely right. it is an obscene amount of money,' clooney said in excerpts*

released on saturday. the interview will air on sunday. bernie sanders, a usa senator from vermont and clinton's rival in the race for the democratic nomination to run for the white house in the nov. 8 election, has pounced on former secretary of state clinton over the big-ticket event and for accepting large sums of money for her campaign. 'the sanders campaign when they talk about it is absolutely right. it is ridiculous that we should have this kind of money in politics. i agree completely,' clooney said. in response to the dinner, the sanders campaign on friday evening sent out an email to supporters asking them to help reach their fundraising goals by chipping in $3. 53 apiece, instead. ”

The reader would immediately notice the presence of the word *reuters* which is a hallmark of every real news. One other thing that stands out is the shortness of news: it has indeed a small Flesh Kincaid score, 10.36 (the average text Flesh Kincaid score for real and fake news is 13.8 and 10.6, respectively). Exhibit 12 shows the distribution of text Flesh Kincaid scores for real and fake news. Every single vertical green line is a misclassified real news. From this simple analysis it appears the shortness and, thus, simplicity of the text may have contributed to the misclassification of a subset of real news. While this observation is not enough to establish a cause-effect relationship, it suggests that real news with shorter, simpler text are more prone to misclassification than real news with longer, complex text.

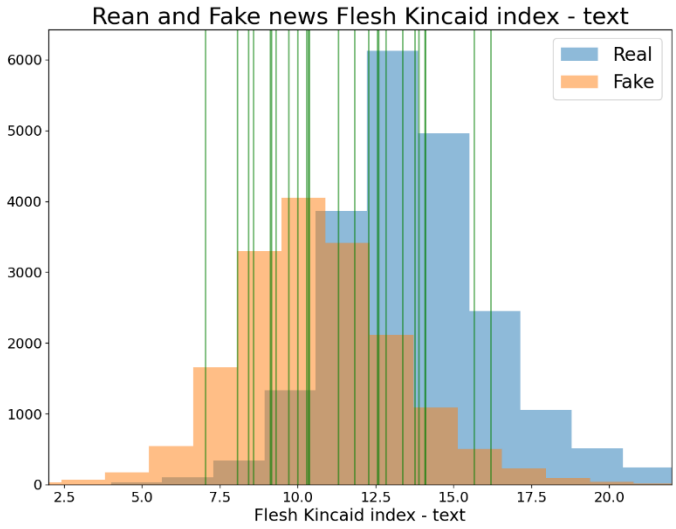

Exhibit 12. Flesh Kincaid Score for the misclassified real news

Next, I investigated how important the Facebook, Twitter, YouTube, Bitly and Tmsnrt hyperlinks are for the model prediction. As mentioned above the frequency distribution of these links among real and fake news is completely dichotomic. Removal of hyperlink tokens reduced the AUC score from 0.999211 to 0.999205 (Table 3), which corresponds to 0.0006 % decrease. Therefore, hyperlinks in the text significantly contribute to the overall predictive power of the model.

Tokens Excluded	Balanced accuracy	AUC
Yes	0.9741	0.999211
No	0.9738	0.999205

Table 3. Weight of the hyperlinks in the final prediction.

Finally, I identified the probability threshold that returns the highest balanced accuracy. This can be seen in Exhibit 13. Keeping all the other parameters of the model unchanged, balanced accuracy can be increased to 0.9779, by setting the probability threshold of the logistic regressor to 0.350.

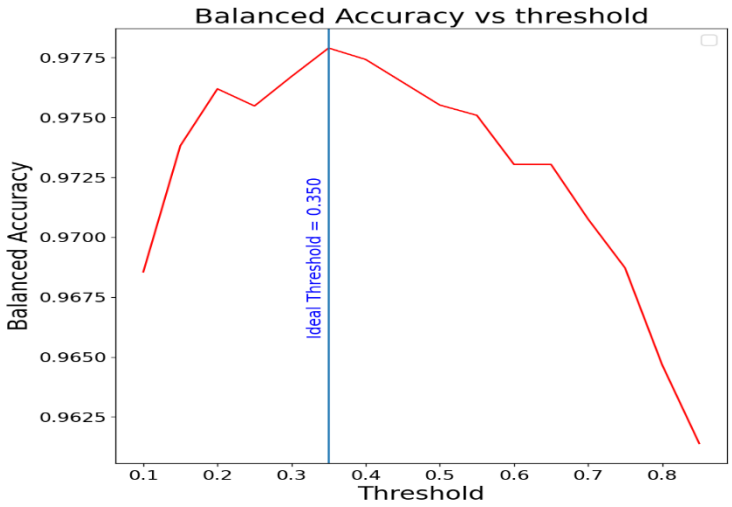

Exhibit 13. Balanced accuracy vs threshold plot

CONCLUSION AND FUTURE DIRECTION

The unchecked spreading of fake information has become a real problem for social media platforms and online information outlets, and it has prompted data science and machine learning communities to develop precise tools for identification and successive removal of untrue information. Herein, I created and tested several models for the accurate discrimination of real news from fake news. These models leveraged NLP techniques as well as linguistic complexity and indexes of title and text of the news.

One important limitation of this study is that real and fake news were collected from two different sources, suggesting that difference in these sources may have made prediction easier than it should be. I dove into the data to attempt to find any clear markers that differentiated these two sources, but even after these steps were taken my overall AUC score remained over .99 which seems unlikely to be true in reality. For future studies, ideally, we could either collect real and fake news articles (as labelled by a third party) from a similar distribution of sources, or we could pull articles from a few sources and use crowdsourcing to label them as real or fake.

There are additional steps that could be taken to improve the predictive power of the model which I couldn't take given the limitations of computational complexity and runtime. First, I could lower min_df to include rare words. Another feature that could be added is the amount of punctuation in the text of the document. Fake news' text tends to have an excess of exclamation marks, asterisks and other punctuation marks. The text could also be analyzed for spelling and grammar errors. Finally, a more sophisticated vectorizer such as word2vec could be implemented to maintain and analyze words' semantic in text and title of the news.

In conclusion I developed a high performing machine learning model for the discrimination of real from fake news. While my model may have some limitations, several courses of action exist to improve its robustness, reliability and predictive power.

APPENDIX

A) Linguistic indexes calculated:

1) Flesch Kincaid grade level

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

2) Flesch reading ease

$$206.835-1.015\frac{Total\ words}{Total\ Sentences}-84.6\frac{Total\ Sillables}{Total\ Words}$$

3) Smog index (Simple Measure of Gobbledygook) = measure of readability that estimates the years of education needed to understand a piece of writing

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

4) Gunning fog index (is a readability test for English writing)

$$0.4\left[\left(\frac{words}{sentences}\right)+100\left(\frac{complex\ words}{words}\right)\right]$$

Complex words = words with more than 3 syllables

5) Coleman Liau Index,

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

6) Automated readability index (ARI) readability index that gauges the understandability of a text

$$\text{ARI} = 4.71\frac{characters}{words}+0.5\frac{words}{sentences}-21.43$$

7) Lix (The Lasbarhetsindex Swedish Readability Formula).

$$LIX=\frac{words}{periods}+100\frac{long\ words}{words}$$

Long words = words with more than 6 characters.

8) Gulpease index. Readability index scaled on Italian language.

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

B) Most predictive words

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
```

C) GridSearch Parameters:

Multinomial Naïve Bayes	**CountVectorizer/ TfidfVectorizer**	**Logistic Regressor**	**Random Forest Classifier**

alpha = 0.01, 0.1, 1	min_df = 15, 25, 50, 100, 200, 300, 400, 500, 1000, 2000, 5000, 10000	C = 1, 10, 20, 100, 200, 260	n_estimators = 50, 100;
			max_depth = None, 2;
			min_samples_leaf = 1, 3.