Detection of Fake News With Machine Learning

Gianluca Turcatel, PhD

Capstone Project, September 2020

The Problem

- *Definition: Fabricated stories with no verifiable facts, sources or quotes*
- *Not a new problem (Condorcet 1795)*
- Goal: manipulate and mislead the reader
- Fake news spread faster then real news
- Widespread problem:

  - 2018, Statistica Survey: 52% of respondents said that "*online website reports fake news regularly"* (2018, Statistica)

  - https://www.tweetgen.com/

Who Might Care?

Social Media Platforms

Media Outlets

The Reader

The Government

The Data

- Data Source :

  - Real News: Reuters (24417)

  - Fake News: Kaggle dataset(23502)

- Data Structure:

  - Date: March 31st , 2015, to February 19th, 2018

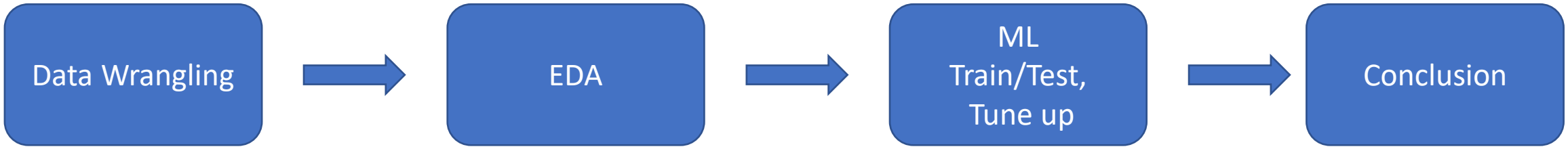
  - Title

  - Text

  - Category (Government News, Middle-east, News, US_News, left-news, politics, politicsNews or worldnews)

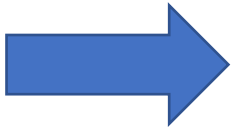
https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

How the problem was tacked



Data Wrangling - Overview

STEP 1:
- Merge datasets
- Drop duplicates, incomplete, missing values



STEP 2:
- Removal of cross site scripts
- Tokenize most frequent hyperlinks
- Tokenize vulgarities
- Expand out verbal contraction
- Linguistic indexes calculation
- Lemmatize text

Data Wrangling - Removal of cross site scripts

……The children s story was so inspiring that it was picked up by the popular Facebook page Have a Gay Day. // < ![CDATA[// < ![CDATA[// < ![CDATA[(function(d, s, id) { var js, fjs = d.getElementsByTagName(s)[0]; if (d.getElementById(id)) return; js = d.createElement(s); js.id = id; js.src = "//connect.facebook.net/en_US/sdk.js#xfbml=1&version=v2.3"; fjs.parentNode.insertBefore(js, fjs);}(document, 'script', 'facebook-jssdk')); //]]>This.Posted by Have A Gay Day on Tuesday, ……

…. Congress just lacks a spine to do it! (function(d, s, id) { var js, fjs = d.getElementsByTagName(s)[0]; if (d.getElementById(id)) return; js = d.createElement(s); js.id = id; js.src = "//connect.facebook.net/en_US/sdk.js#xfbml=1&version=v2.3"; fjs.parentNode.insertBefore(js, fjs);}(document, 'script', 'facebook-jssdk'));Senator Elizabeth Warren goes….

Data Wrangling: Hyperlinks Tokenization

Common hyperlinks:
- Facebook
- Bitly
- Youtube
- Twitter
- Tmsnrt

Replaced with **11-character long** tokens

….we were an enemy of the people. Here s the video via YouTube:https://www.youtube.com/watch?v=KyipRvdPyLE&feature=youtu.beDonald Trump is the true ….

….we were an enemy of the people. Here s the video via YouTube: **youtubelink**. Donald Trump is the true ….

Data Wrangling – Linguistic Indexes Calculation

STEP 1: Calculate number of:

Sentences
Words
Characters,
Syllables,
Polysyllabic words
Long word*
Complex Words**
Monosyllabic words

STEP 2: Compute the linguistic scores:

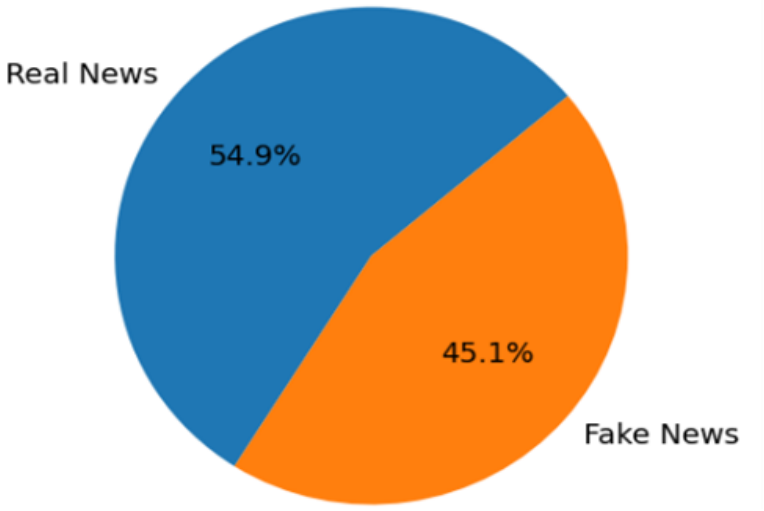
Flesh Kincaid
Flesh Read Easy
Simple Measure of Gobbledygook (SMOG)
Gunning Fox
Lasbarhetsindex (LIX)
Coleman-Liau (CLI)
Automated Readability Index (ARI)
Gulpease

Example of linguistic index:

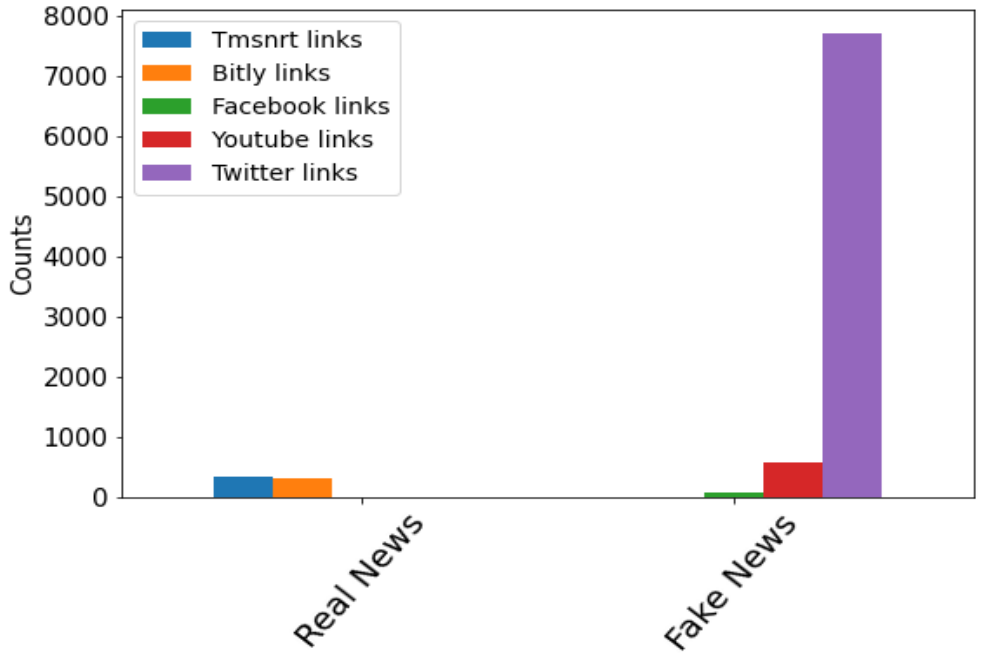
$$CLI = 0.0588\frac{characters}{words} - 0.296\frac{sentences}{words} - 15.8$$

*Words with more than 6 characters;
** Words with more 3 syllables

EDA – News and Hyperlinks Distribution



The two classes are well balanced

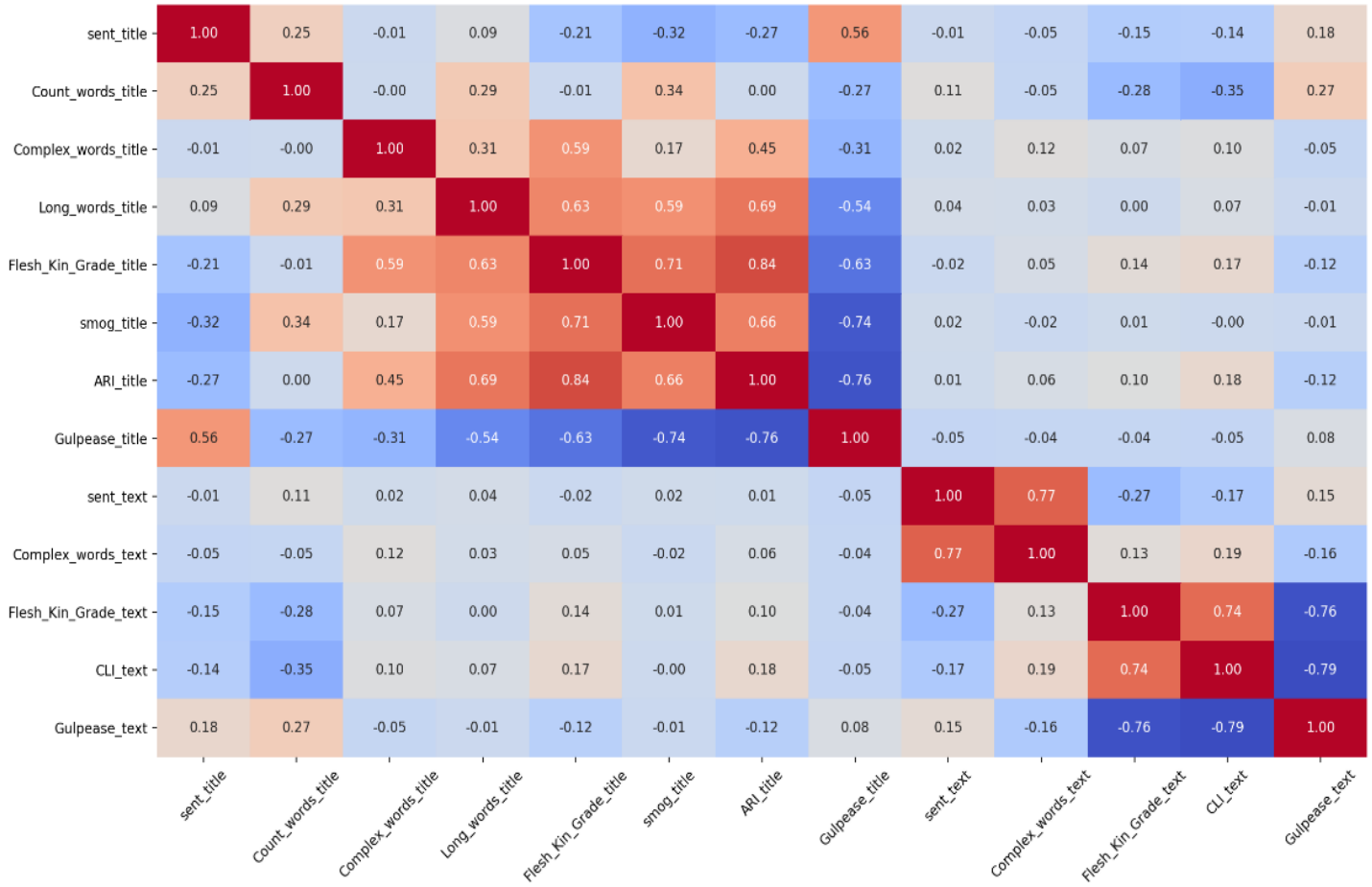


Dichotomic distribution of the hyperlinks

EDA – Linguistic Features Trimming

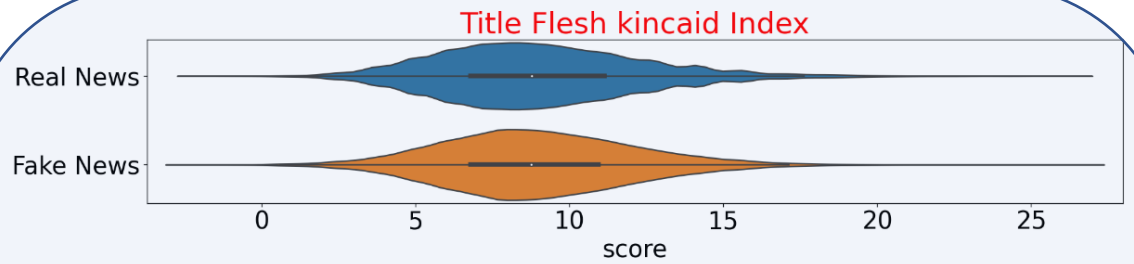
32 text, linguistic Features:
- 16 for the title
- 16 for the text

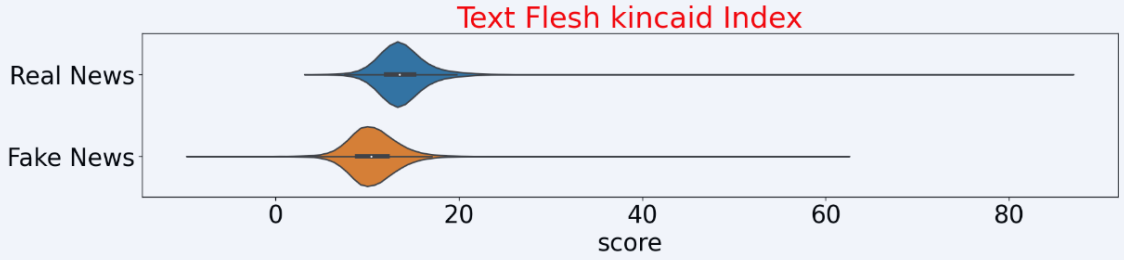
13 text, linguistic Features:
- 8 for the title
- 5 for the text



EDA – Indexes' Insight

Flesh Kincaid Index

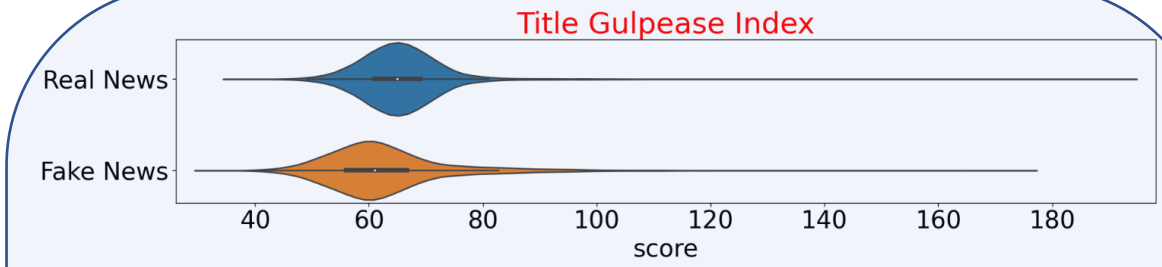


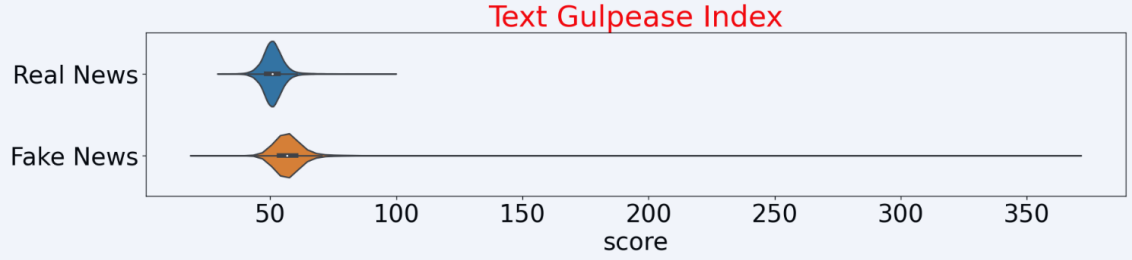


$$0.39\frac{Total\ words}{Total\ sentences} + 11.8\frac{Total\ Syllables}{Total\ words} - 15.59$$

Text of real news is more complex/sophisticated

Gulpease Index



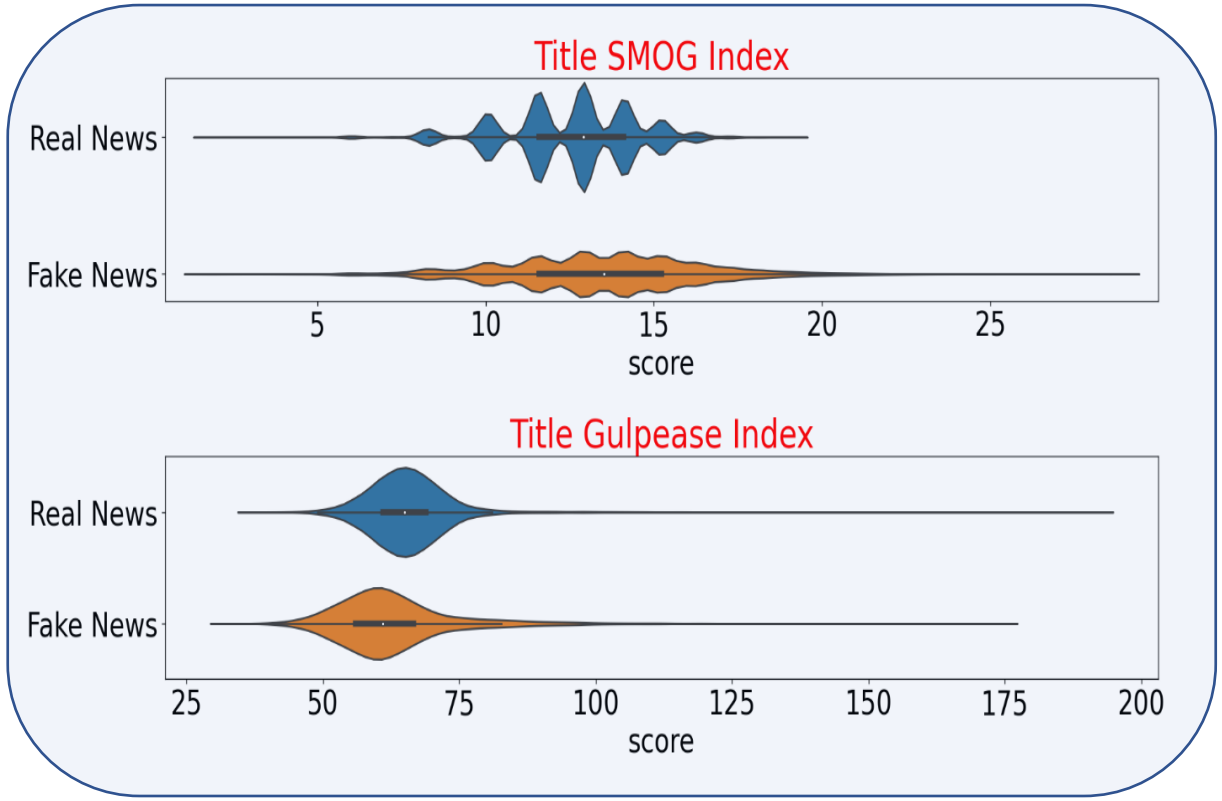


$$89 + \frac{300 * sentences - 10 * characters}{words}$$

Title of real news is shorter/simpler
Text of real news is longer/more complex

EDA – Indexes comparison

SMOG index vs Gulpease Index



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

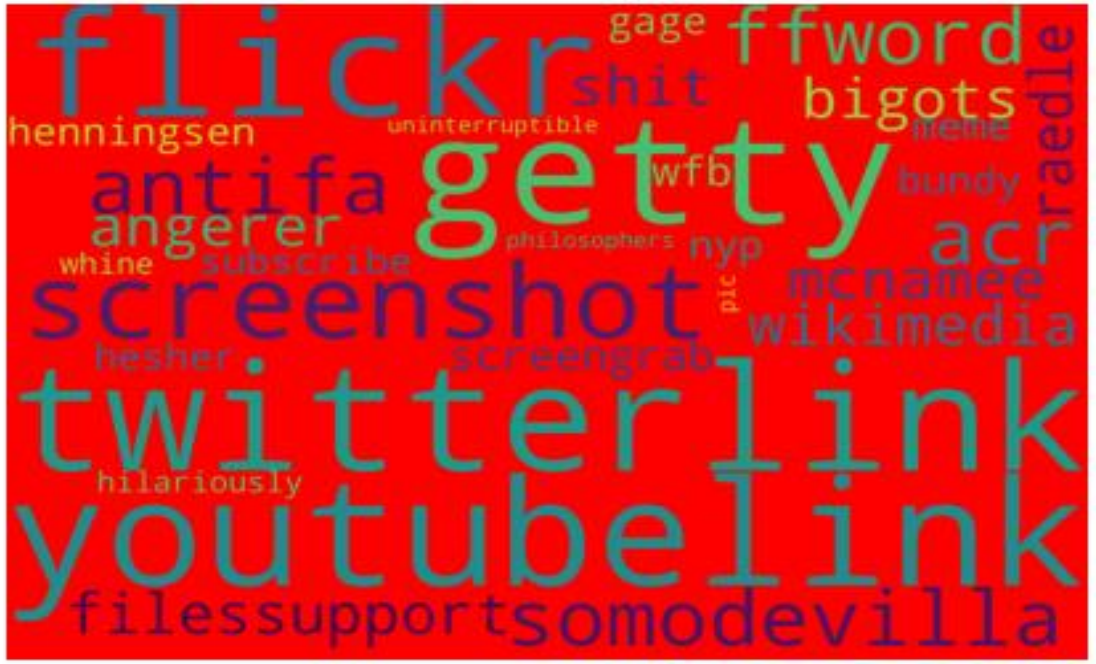
$$89 + \frac{300 * sentences - 10 * characters}{words}$$

Each index amplifies one or more specific linguistic properties in the word and sentence and, thus, generates its **own unique distribution**

EDA - Predictive words

Real News

Fake News



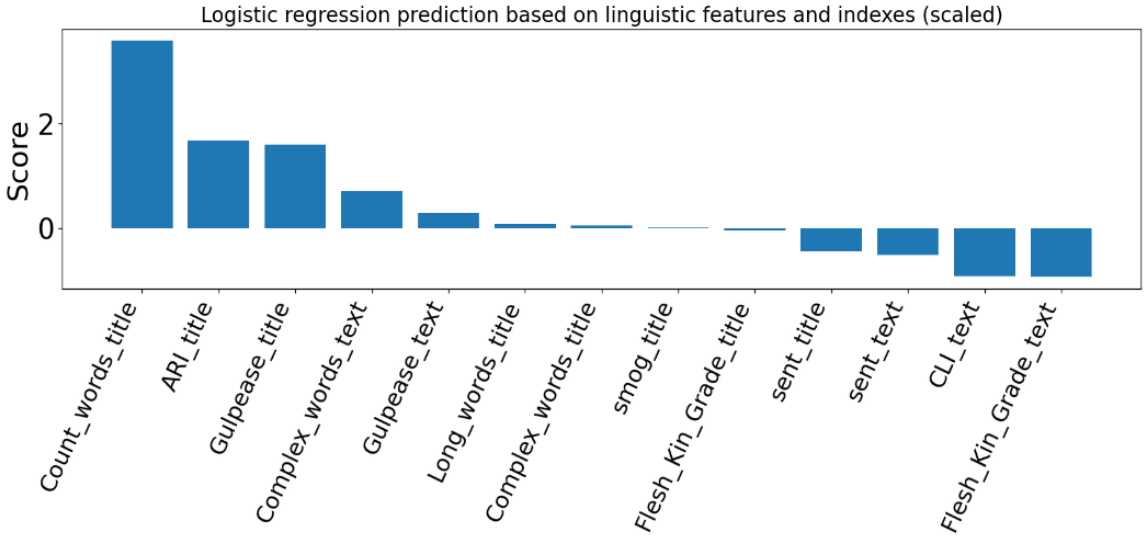
Real news – **key takeaways:**

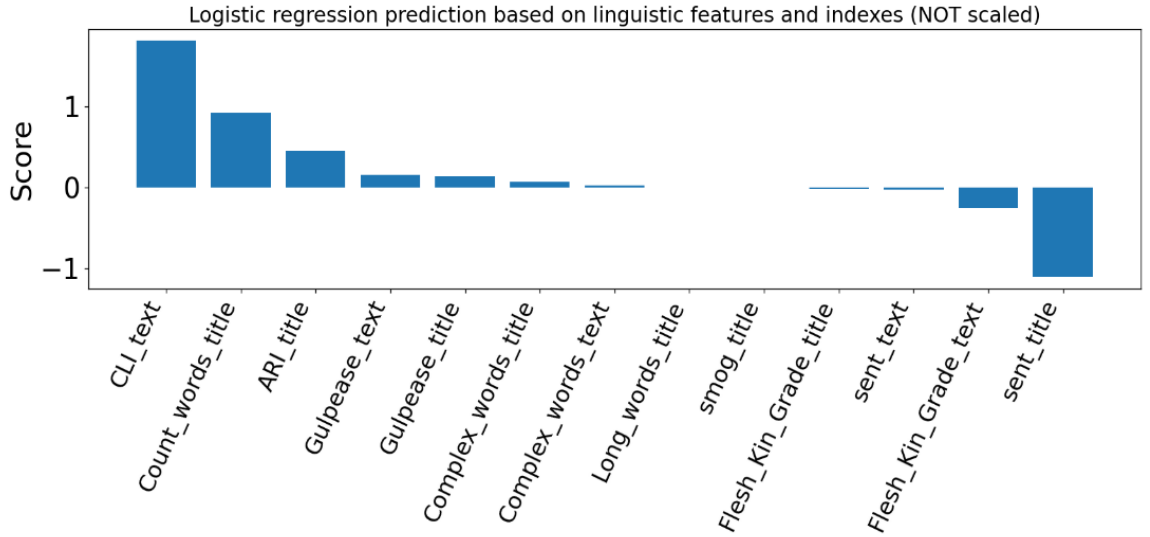
- **Bitly/tmsnrt hyperlinks.**
- Name of **non-mainstream politicians** (Odinga - Kenyan politician, Rajoy - Spanish politician, Hariri - Lebanese politician, Gulen - Turkish Islamic scholar) and **name of cities and countries rarely mentioned by the media** (Kirkuk - city in Iraq, Nairobi -Kenya's capital, Marawi - Islamic city, Harare - capital of Zimbabwe, Rakhine - state in Myanmar).

Fake news – **key takeaways:**

- **Twitter and YouTube hyperlinks.**
- **Photography related words** (*getty*, *somodevilla, screenshot*, *flickr*, *getty, raedle*, *angerer*).
- **Vulgarities** (*f**k, s**t*, *bigots*).

EDA - Most predictive linguist indexes, text features





Words in the title:

➡ The longer the title, the higher the prob the news is fake

CLI index in the text:

➡ The higher the complexity of the text, the lower the prob the news is fake

Text Flesh Kincaid index :

➡ The higher the Flesh Kincaid, the higher the prob the news is real

Number of sentences in title:

➡ The longer the title, the higher the prob the news is fake

Machine Learning Modeling - Overview

Models 1 & 2: Multinomial NB
- **Input: Lemmatized text**
- **Optimized for alpha, n-grams, min_df**

1 – CountVectorizer, 2 – TfidfVectorizer

Models 3: Logistic Regression
- **Input: linguistic features**
- **Optimized for C**

Models 4 & 5: Logistic Regression
- **Input: Lemmatized text, and linguistic features**
- **Optimized for C**

4 – CountVectorizer, 5 – TfidfVectorizer

Models 6 & 7: Random Forest
- **Input: Lemmatized text, and linguistic features**
- **Optimized for number of estimators, min sample leaf**

6 – CountVectorizer, 7 – TfidfVectorizer

Machine Learning Modeling - Results

	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

Model 5 – False Negatives

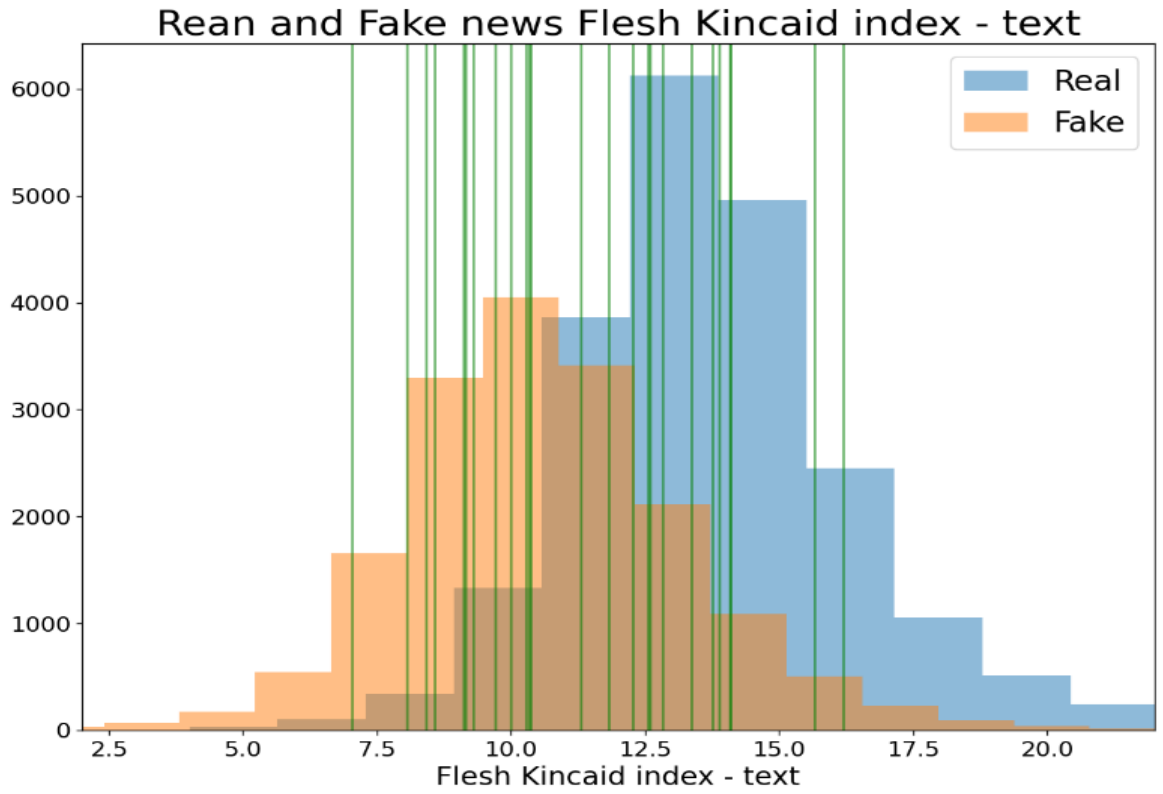
Confusion Matrix

```
[4186   25]
[  65 3440]
```

Excerpt of fake news classified as real

"….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…"

Model 5 – False Positives



Real news Flesh Kincaid index mean: **13.8**

Fake news Flesh Kincaid index mean : **10.6**

False positive Flesh Kincaid index mean: **11.39**

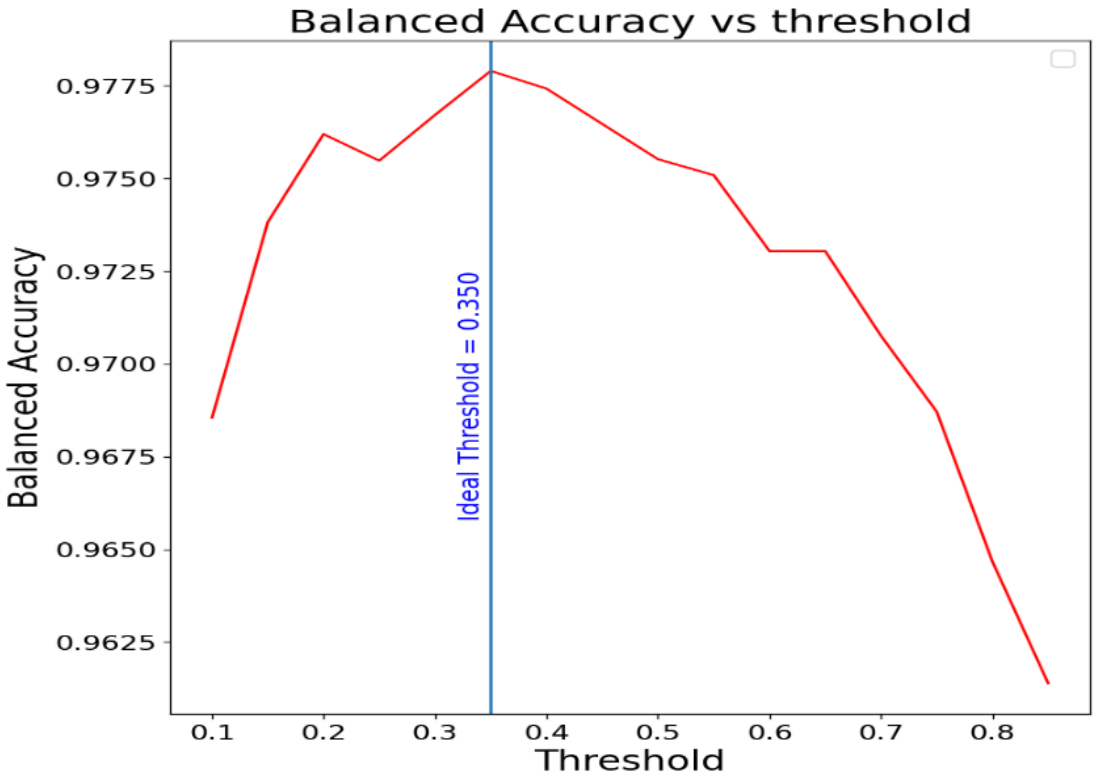
Real news with shorter, simpler texts are prone to misclassification

Model 5 – Hyperlinks weight in prediction

Tokens Excluded	Balanced accuracy	AUC
Yes	0.9741	0.999211
No	0.9738	0.999205

Drop of <u>0.0006%</u>

Model 5 – Probability Threshold Tuning



Threshold=0.5: **Balanced Accuracy 0.9755**

Threshold=0.35: **Balanced Accuracy 0.9779**

Conclusion

- All models are very accurate, **actually too much accurate**: real and fake news were collected from two different sources.

- In future real and fake news articles (as labelled by a third party) could be collected 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.

- **Possible improvements of the classifier:**
  - Reduce min_df hyperparameter of the Vectorizer
  - Detection of punctuation patterns
  - Maintaining words semantic
  - Detection of spelling and grammar error