Novel Approaches to Analyzing and Distinguishing Fake and Real News to Mitigate the Problem of Disinformation

Alina Vereshchaka, Seth Cosimini, and Wen Dong

University at Buffalo
Department of Computer Science and Engineering

University at Buffalo
Department of English

University at Buffalo
The State University of New York

Overview

Identifying fake news has become an important challenge. Increasing usage of social media has led an increase in the number of people who can be influenced, thus the spread of fake news can potentially impact important events. Fake news has become a major societal issue and a technical challenge for social media companies to identify.

Our goal

Distinguish between the real and fake news.

What is disinformation?

False information deliberately and often covertly spread (as by the planting of rumors) in order to influence public opinion or obscure the truth.

Our approaches

We addressed the problem of fake news identification using three approaches to make it manageable and more accurate:

1. Sociocultural and textual approach. It allows us to identify the rhetorical and textual characteristics that distinguish "real" or "fake" information.
2. Data science approach. It helps to dig into the data analytic by building the words and phrases frequencies,
3. Deep learning approach. We built a binary classifiers that extract features from fake and real news using deep learning models, such as Long Short Term Memory (LSTM), Recurrent Neural Network (RNN) and Gated Recurrent Unit (GRU).

Dataset

We used the dataset extracted using the FakeNewsNet* tool. The final dataset contains both fake and real news in the political domain.

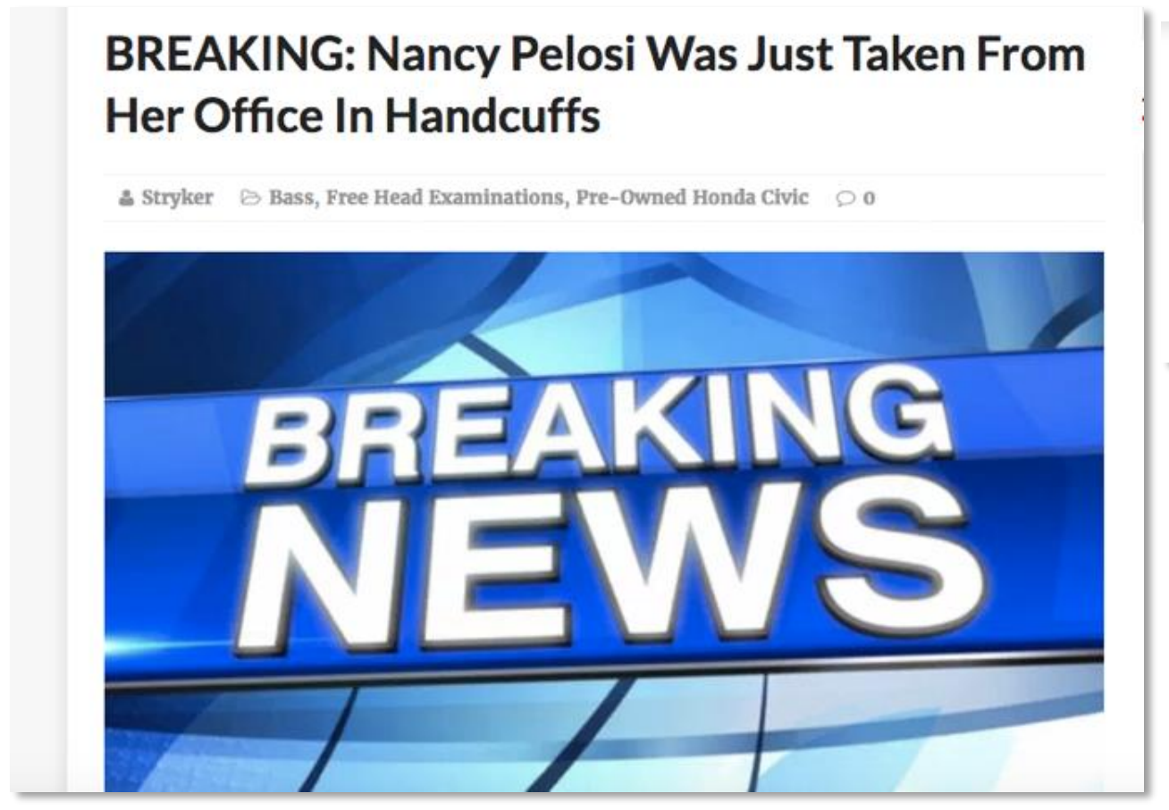
Total Number	Fake	Real
News articles	432	624
News articles with text	420	528
News articles with images	336	447

*https://github.com/KaiDMML/FakeNewsNet

Sociocultural textual analysis

HEADLINES

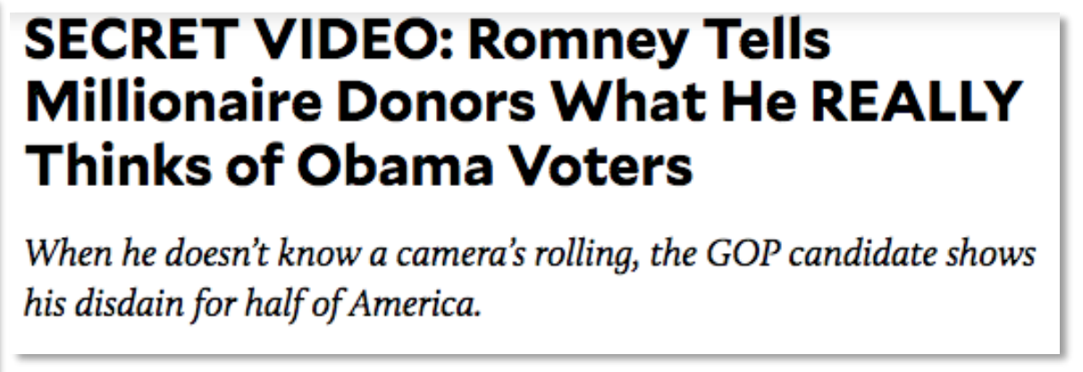
In both real and fake news articles, headlines will often begin with a word like "breaking" in all capital letters to catch a reader's attention and communicate an urgency to the information in the article.



BREAKING: Nancy Pelosi Was Just Taken From Her Office In Handcuffs

Stryker   Bass, Free Head Examinations, Pre-Owned Honda Civic   0

Fake news



SECRET VIDEO: Romney Tells Millionaire Donors What He REALLY Thinks of Obama Voters

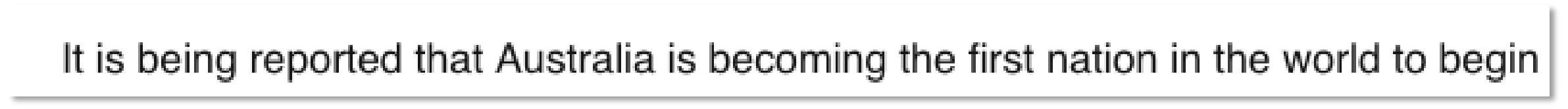
When he doesn't know a camera's rolling, the GOP candidate shows his disdain for half of America.

Real news

Example of the use of all capital letters in both fake and real news story headlines to communicate urgency

SOURCE

Fake news will often report information without including a source, simply using phrases such as "told reporters" or "is being reported" to signify credibility rather than offering actual sourcing.

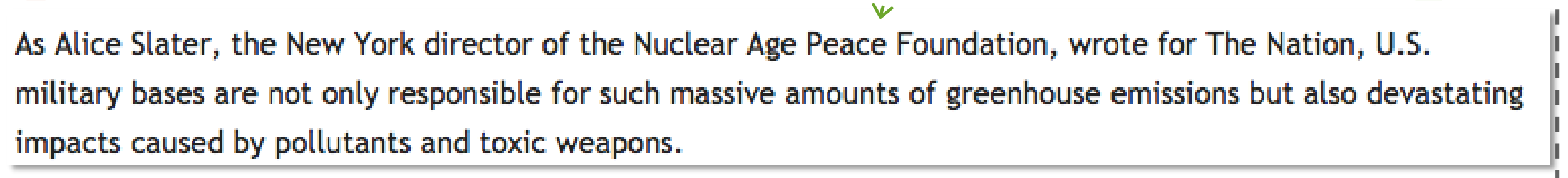


It is being reported that Australia is becoming the first nation in the world to begin

Fake news

"It is being reported" rather than citing a reputable source for this information

Directly referencing source, even if no hyperlink is provided

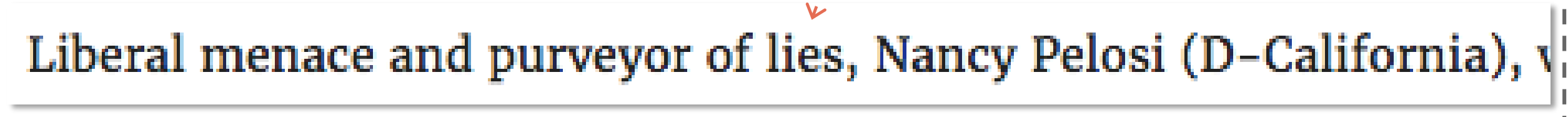


As Alice Slater, the New York director of the Nuclear Age Peace Foundation, wrote for The Nation, U.S. military bases are not only responsible for such massive amounts of greenhouse emissions but also devastating impacts caused by pollutants and toxic weapons.

Real news

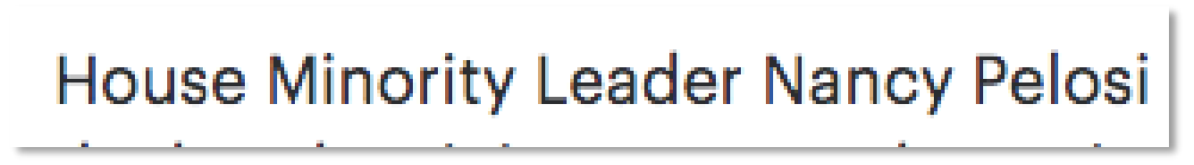
USE OF ADJECTIVES

Adjectival phrases providing non-factual, politically-motivated description



Liberal menace and purveyor of lies, Nancy Pelosi (D-California), v

Fake news



House Minority Leader Nancy Pelosi

Example of adjectival phrase used to provide contextual information on real news page

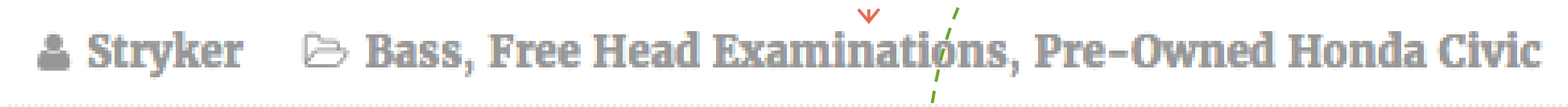
Real news

WRITTEN INFORMATION

- Missing author biography
- Author biography that provides no information about their journalistic affiliation
- Irrelevant or content-less photos
- Missing publication date
- Erroneous metadata

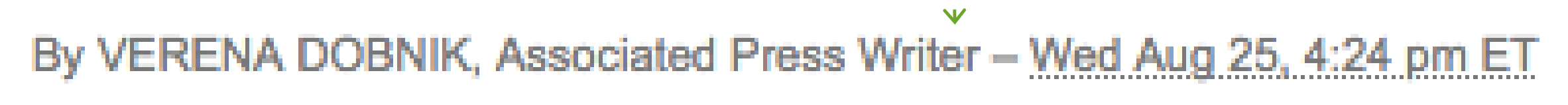
Example of erroneous metadata

Example of full author name, affiliation, and publication date of article



Stryker   Bass, Free Head Examinations, Pre-Owned Honda Civic

Fake news

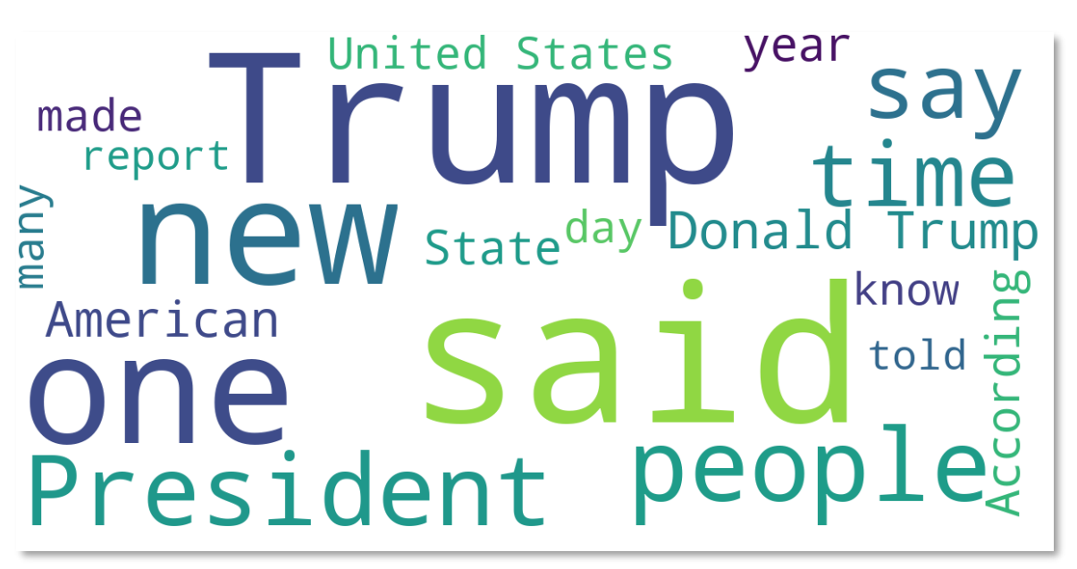


By VERENA DOBNIK, Associated Press Writer – Wed Aug 25, 4:24 pm ET

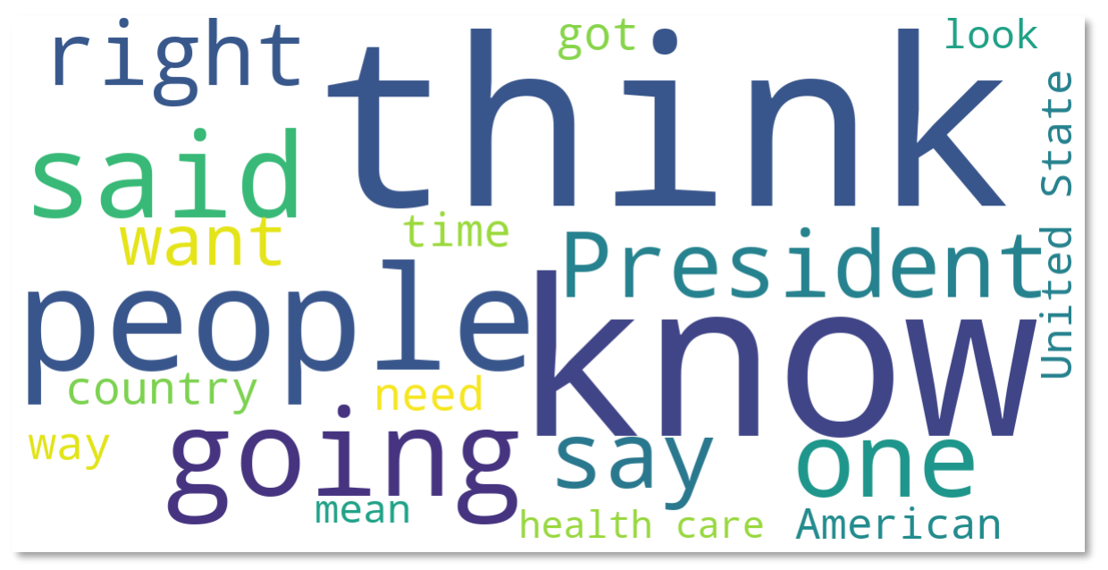
Real news

Data Analysis

WORDCLOUDS



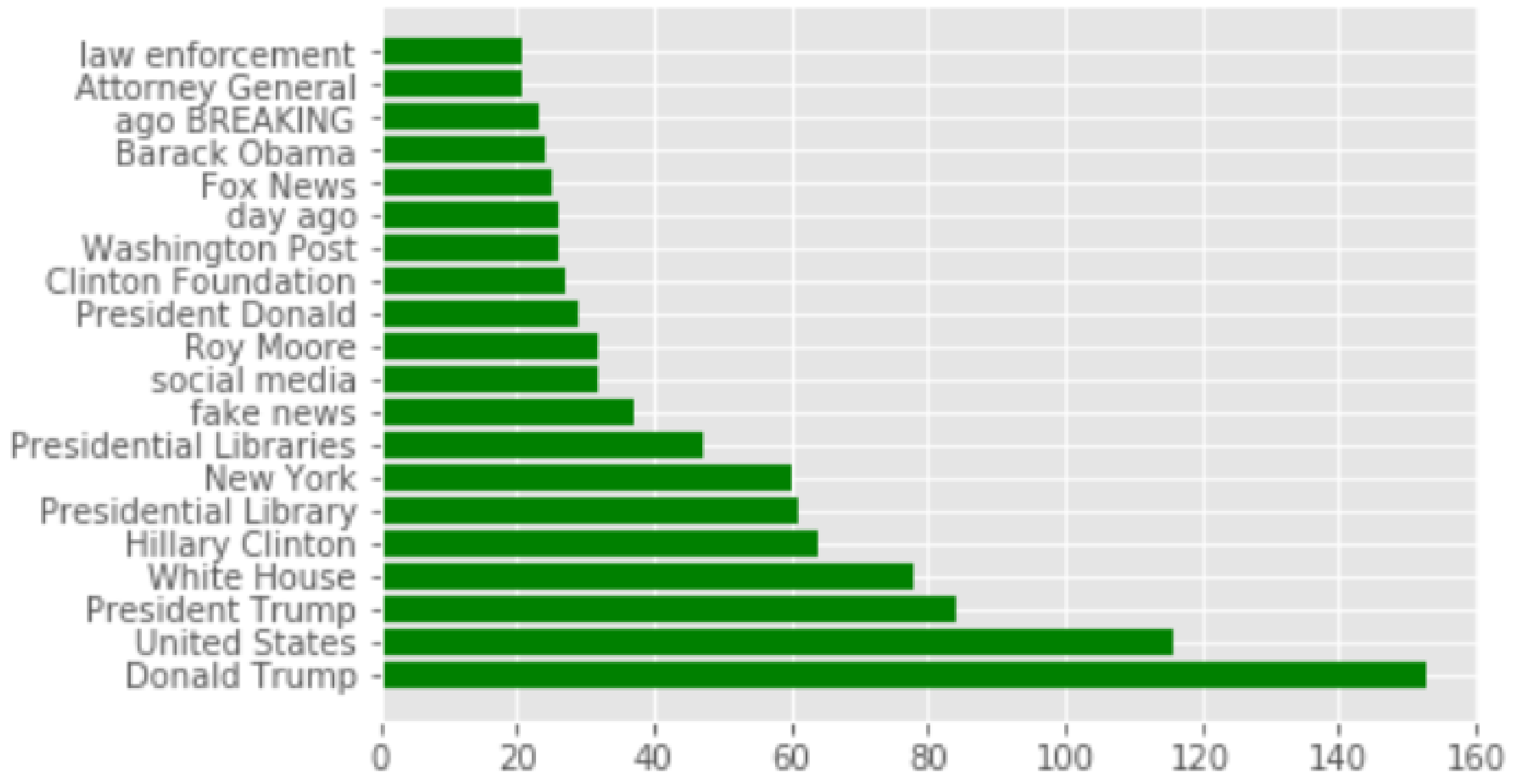
Fake news



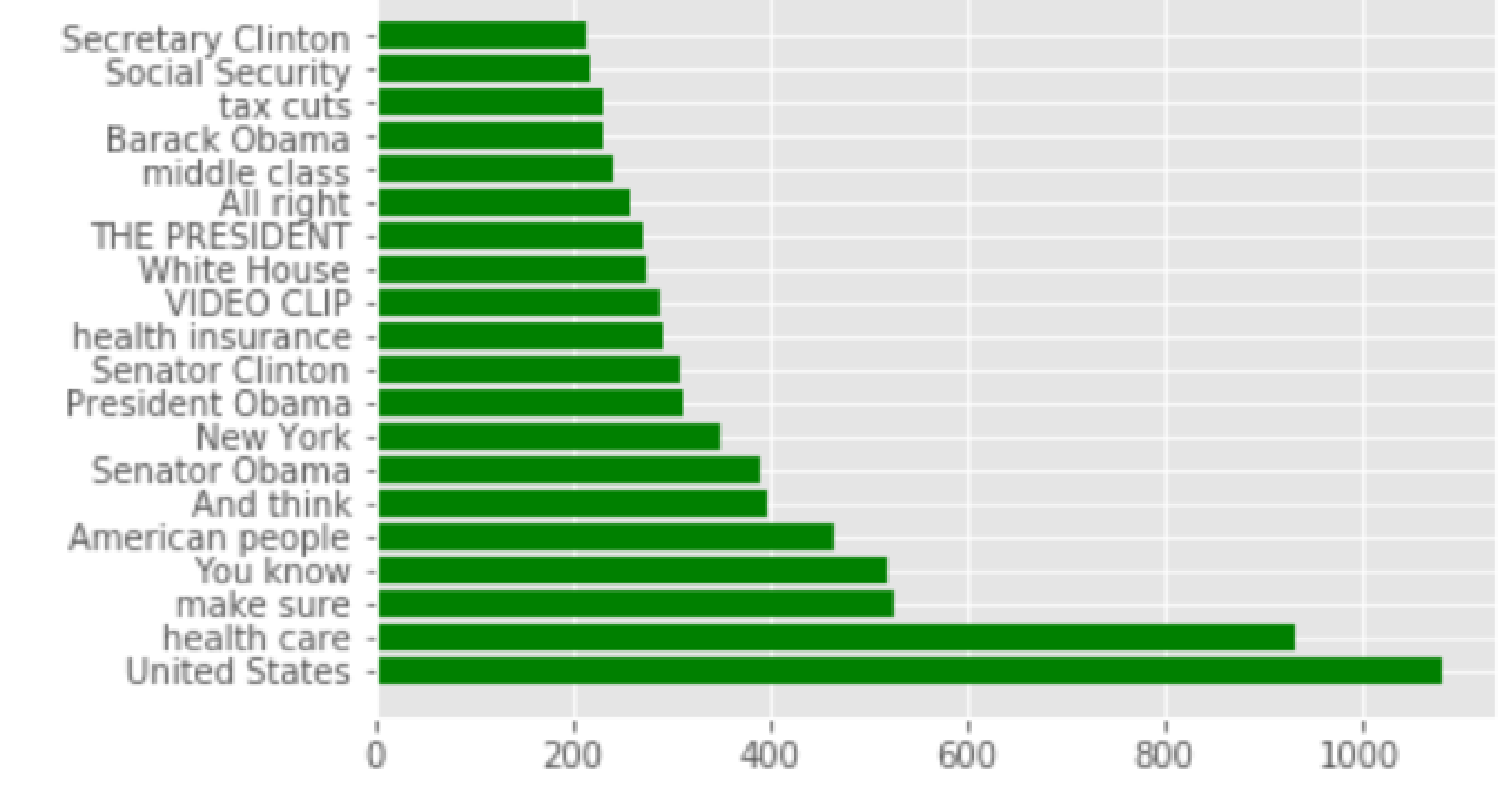
Real news

In fake news the word "Trump" word is more prevalent, while in real news "president" is used to address the key news maker. This distinction between the office of the president and the individual begins to show that fake news has a stronger investment in affective and ideological approaches than real news.

BIGRAMS



Top 20 word phrases for Fake news



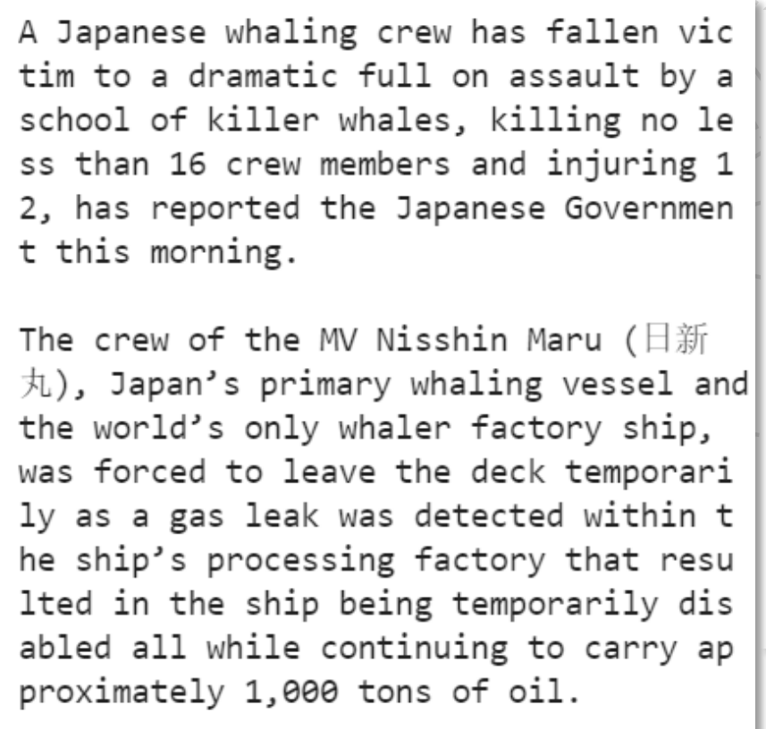
Top 20 word phrases for Real news

Similar to the wordcloud, the bigrams show in real news, public figures are often referred to by official titles: Secretary Clinton, Senator Obama, Senator Clinton. In fake news, however, first names are given instead, or in addition, such as "President Donald."

Deep Learning

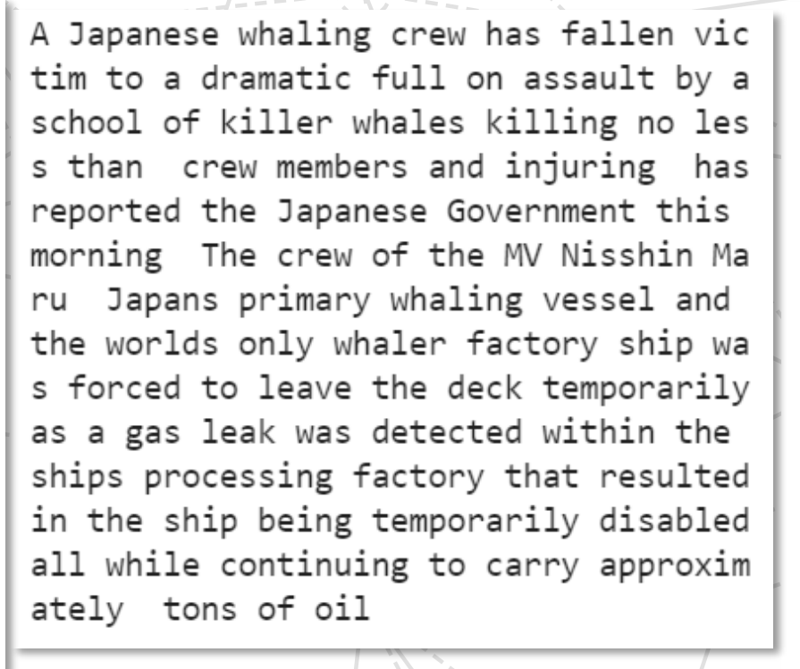
WORD EMBEDDING

We encoded the dataset using Byte Pair Encoding (BPE). BPE is a simple data compression technique that iteratively replaces the most frequent pair of bytes in a sequence with a single, unused byte.



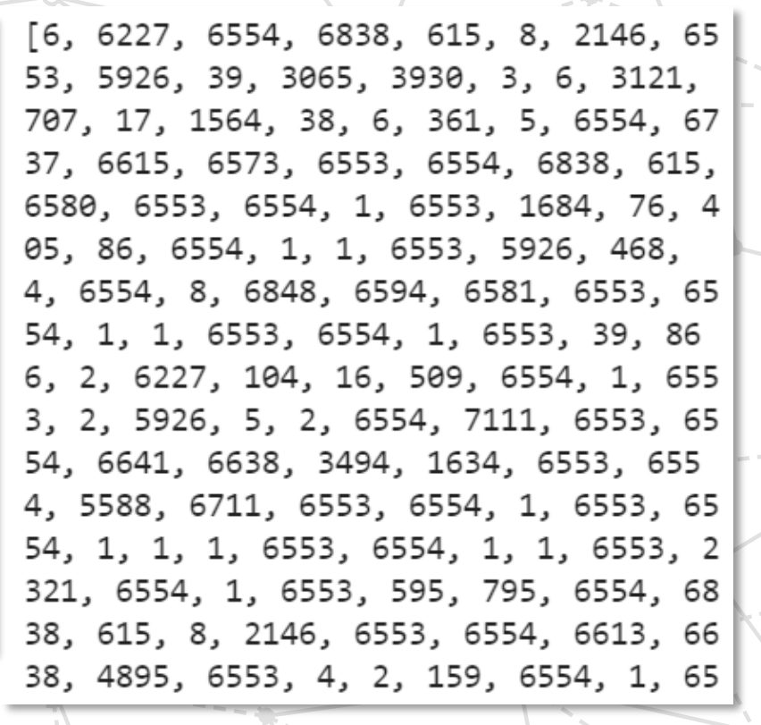
```
A Japanese whaling crew has fallen vic
tim to a dramatic full on assault by a
school of killer whales, killing no le
ss than 16 crew members and injuring 1
2, has reported the Japanese Governmen
t this morning.

The crew of the MV Nisshin Maru (日新
丸), Japan's primary whaling vessel and
the world's only whaler factory ship,
was forced to leave the deck temporari
ly as a gas leak was detected within t
he ship's processing factory that resu
lted in the ship being temporarily dis
abled all while continuing to carry ap
proximately 1,000 tons of oil.
```

Original text



```
A Japanese whaling crew has fallen vic
tim to a dramatic full on assault by a
school of killer whales killing no les
s than  crew members and injuring  has
reported the Japanese Government this
morning  The crew of the MV Nisshin Ma
ru  Japans primary whaling vessel and
the worlds only whaler factory ship wa
s forced to leave the deck temporarily
as a gas leak was detected within the
ships processing factory that resulted
in the ship being temporarily disabled
all while continuing to carry approxim
ately  tons of oil
```

Cleaned text



```
[6, 6227, 6554, 6838, 615, 8, 2146, 65
53, 5926, 39, 3065, 3930, 3, 6, 3121,
707, 17, 1564, 38, 6, 361, 5, 6554, 67
37, 6615, 6573, 6553, 6554, 6838, 615,
6580, 6553, 6554, 1, 6553, 1684, 76, 4
05, 86, 6554, 1, 1, 6553, 5926, 468,
4, 6554, 8, 6848, 6594, 6581, 6553, 65
54, 1, 1, 6553, 6554, 1, 6553, 39, 86
6, 2, 6227, 104, 16, 509, 6554, 1, 655
3, 2, 5926, 5, 2, 6554, 7111, 6553, 65
54, 6641, 6638, 3494, 1634, 6553, 655
4, 5588, 6711, 6553, 6554, 1, 6553, 65
54, 1, 1, 1, 6553, 6554, 1, 1, 6553, 2
321, 6554, 1, 6553, 595, 795, 6554, 68
38, 615, 8, 2146, 6553, 6554, 6613, 66
38, 4895, 6553, 4, 2, 159, 6554, 1, 65
```

Encoded text

DEEP LEARNING MODELS

We have applied three deep learning models to do feature extraction and perform binary classification.

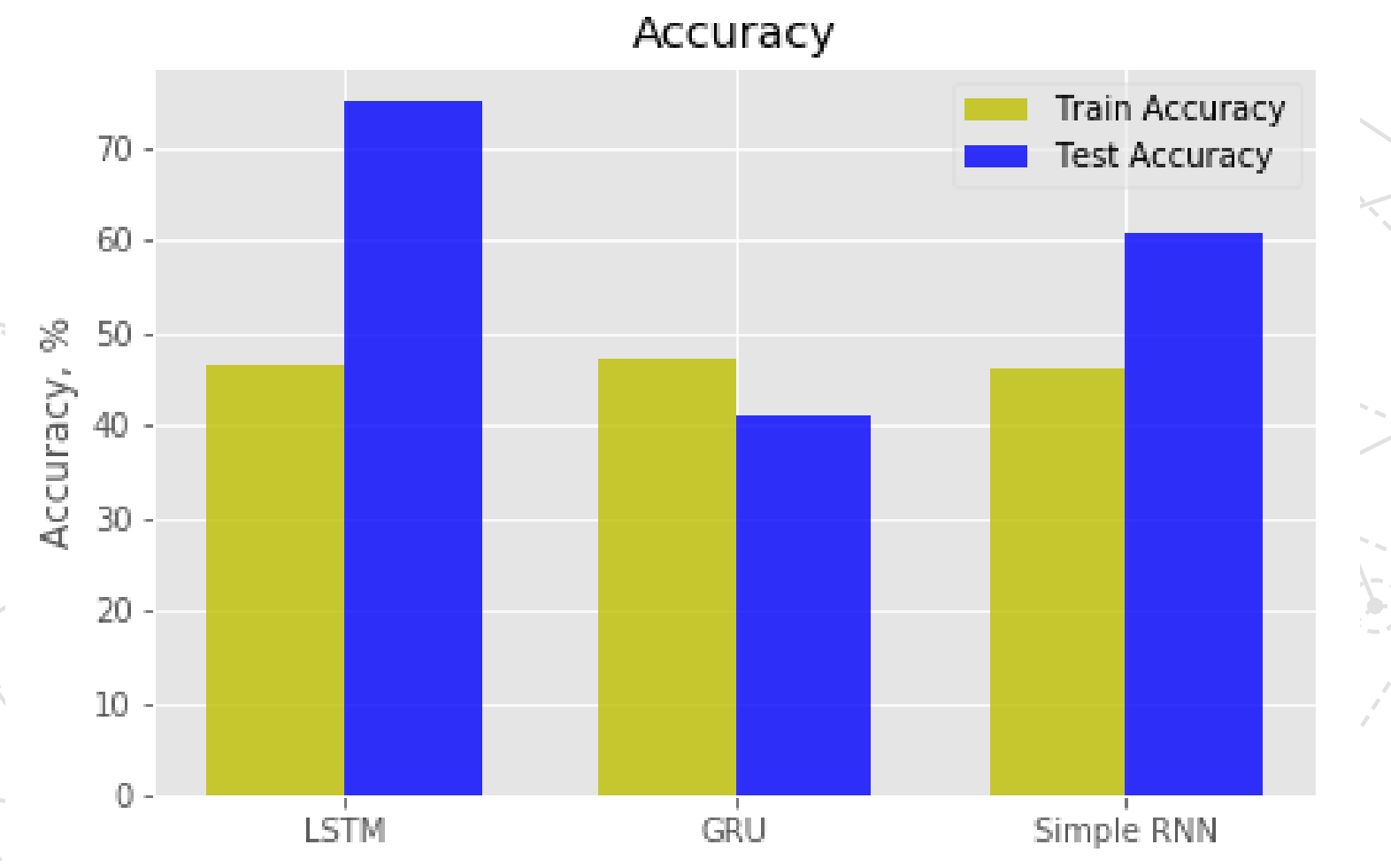
Hyperparameters for training the models

Parameter	LSTM	GRU	RNN
Layers	1	2	1
Activation	-	SeLu	SeLu
Activation output	Sigmoid	Sigmoid	Sigmoid
Optimizer	Adam	Adam	SGD
Epochs	3	1	5
Nodes 1-layer	15	15	15
Nodes 2-layer	-	1	5

RESULTS

Test Accuracy

LSTM: 75%
GRU: 41.07 %
RNN: 60.71%



Conclusion

We can notice one of characteristics of disinformation is its ideological context. This is the first time this kind of sociocultural textual analysis has been conducted using this dataset. Deep learning models showed reasonable results, but it might not be generalized to other types of datasets.

References

Alina Vereshchaka, Seth Cosimini, and Wen Dong. Novel Approaches to Analyzing and Distinguishing Fake and Real News to Mitigate the Problem of Disinformation, 2019

Contact us: {avereshc, sethcosi, wendong}@buffalo.edu