Fake News Detection

Big Data for Official Statistics
February 2021

Oluwatoyin Yetunde SANNI
Matricola: 1871835

What is fake news?

Fake news—news articles that are intentionally and verifiably false designed to manipulate people's perceptions of reality—has been used to influence politics and promote advertising. But it has also become a method to stir up and intensify social conflict.

Why should we care if a news is fake?

It has been used to influence politics and promote advertising. But it has also become a method to stir up and intensify social conflict.

It is now known that false information played a major role in the last American presidential election.

Fake news—news articles that are intentionally and verifiably false designed to manipulate people's perceptions of reality—has been used to influence politics and promote advertising.

Fake news—news articles that are intentionally and verifiably false designed to manipulate people's perceptions of reality—has been used to influence politics and promote advertising.

Impacts of Fake news spreads across different sectors around the world

Democratic impacts

Will I vote differently if I know the Pope endorses a politician's candidacy?

Fear

Will I have insomnia? Will my quality of life decrease?

Health Impacts

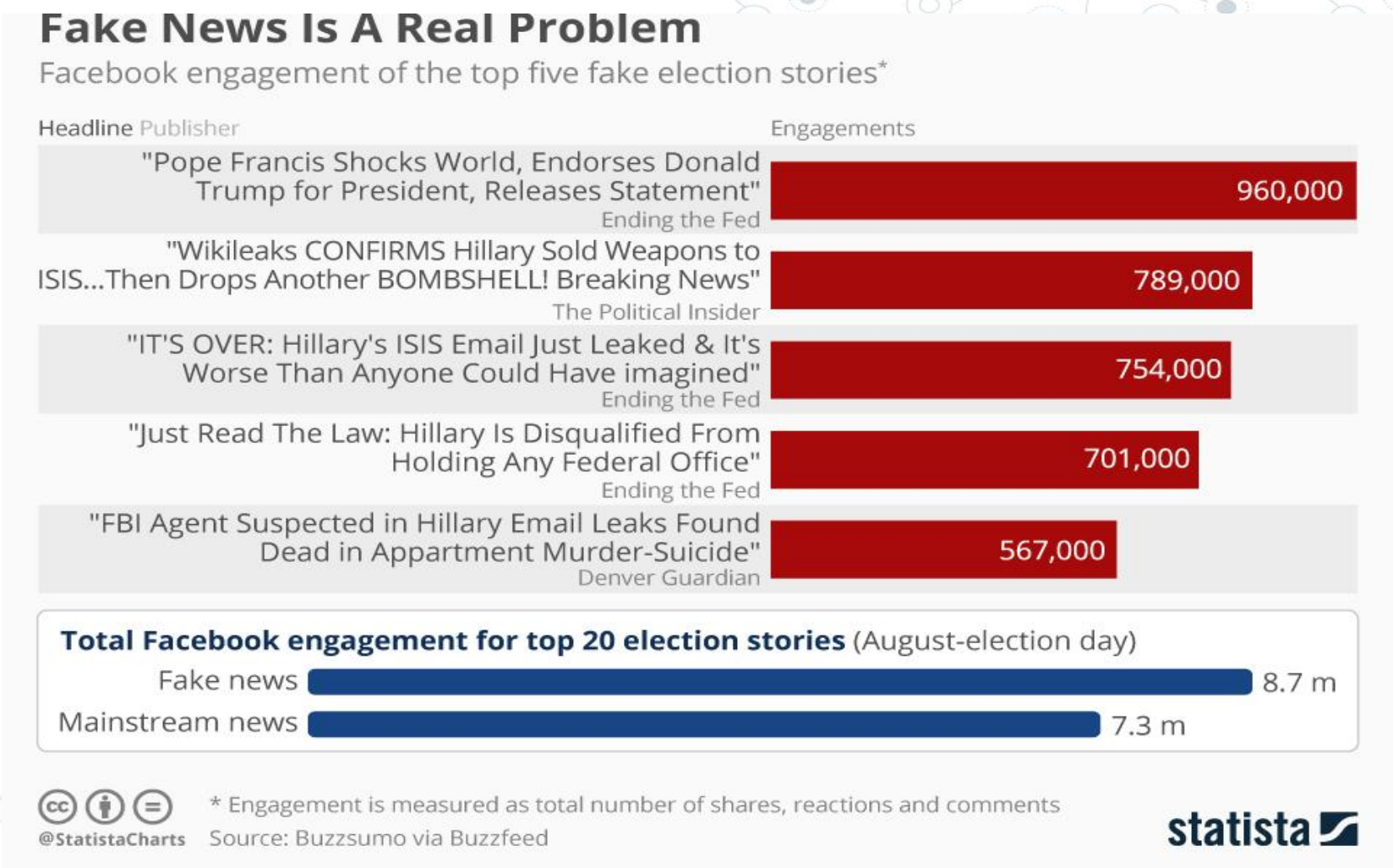
Will I take the Covid19 Vaccine if I read a news article that it killed or has negative lasting effects?

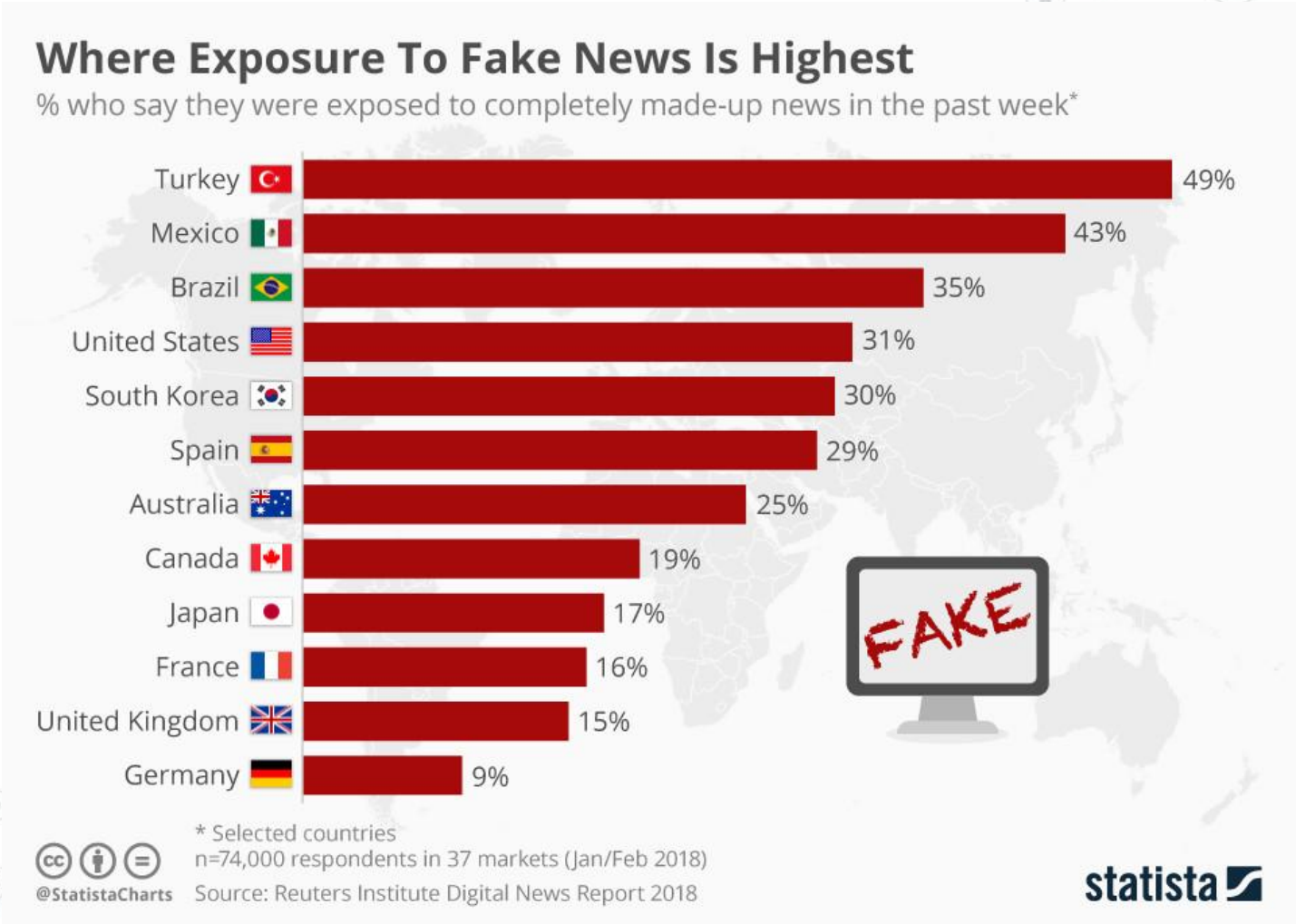
Financial impacts

If I like that blogger and her post pushes me to buy a certain product?

TOP
5 FAKE
NEWS

During
trump
election





TOP
5 FAKE
NEWS

During
trump
election

The major goal of what I have done in this project is to classify news content as Fake/Real

Hence, I did the following in this project:

1. Obtain the datasets.
2. Preprocess the data and prepare it in a format that machine learning models can work with.
3. Did some Data exploratory analysis to understand my Data.
4. Built different machine learning and deep learning models to learn the prepared data to compare results and achieve the best possible classification results.

Dataset and Preprocessing

“As data scientists, our job is to extract signal from noise.” —Daniel Tunkelang

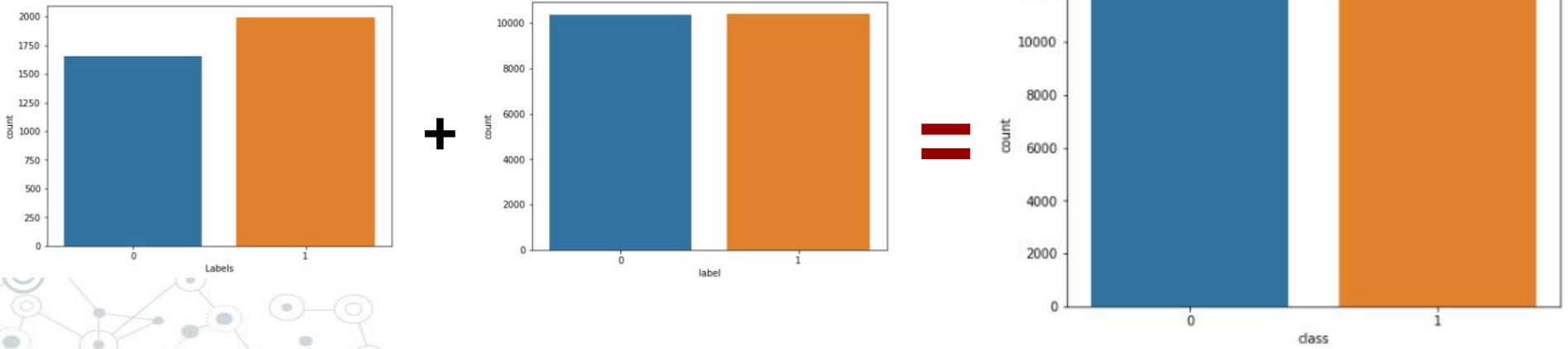
Dataset

- I combined two datasets from MachineHack and Kaggle. They both consists of texts and the corresponding binary classification.

- Since this is a news article, the datasets isn't heavily noisy. However, it is still extremely difficult to learn the specific words and sequences that matter to detect a false news content. This also makes the vocabulary bloated , and the tweet vectors sparse.

Dataset

Interestingly, both have a normal distribution with an even distribution of the target. This gives us a balanced dataset which was splitted into 70% for training and 30% for cross validation.



Data Cleaning and Preprocessing

- Since our task takes consideration of context and semantic meanings, I parsed the dataset to remove words that will provide no intuition. Words like punctuations, stop words, were cleaned off as well.
- Words with accents were represented in Unicode from the dataset, so I removed them.
- I also converted all the words to lowercase, also to reduce the vocabulary size.

Data Cleaning and Preprocessing

Our models can only understand numbers, hence representing our texts in numeric is a crucial step. This is called feature extraction or vectorization. For this project, I explored Count Vectorizer and TF-IDF(Term Frequency Inverse Document Frequency).

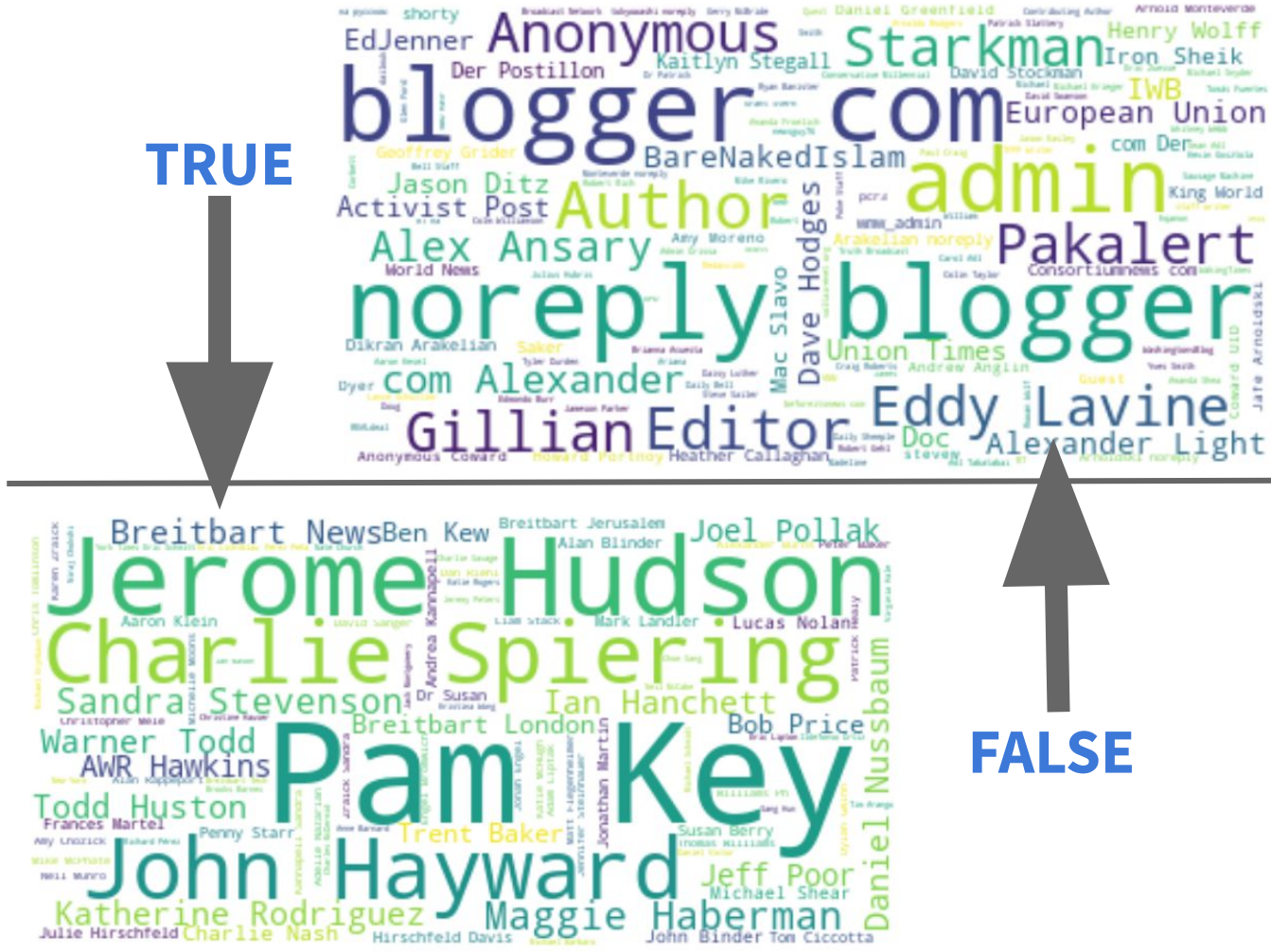
Count Vectorizer

	blue	bright	sky	sun
Doc1	1	0	1	0
Doc2	0	1	0	1

TD-IDF Vectorizer

	blue	bright	sky	sun
Doc1	0.707107	0.000000	0.707107	0.000000
Doc2	0.000000	0.707107	0.000000	0.707107

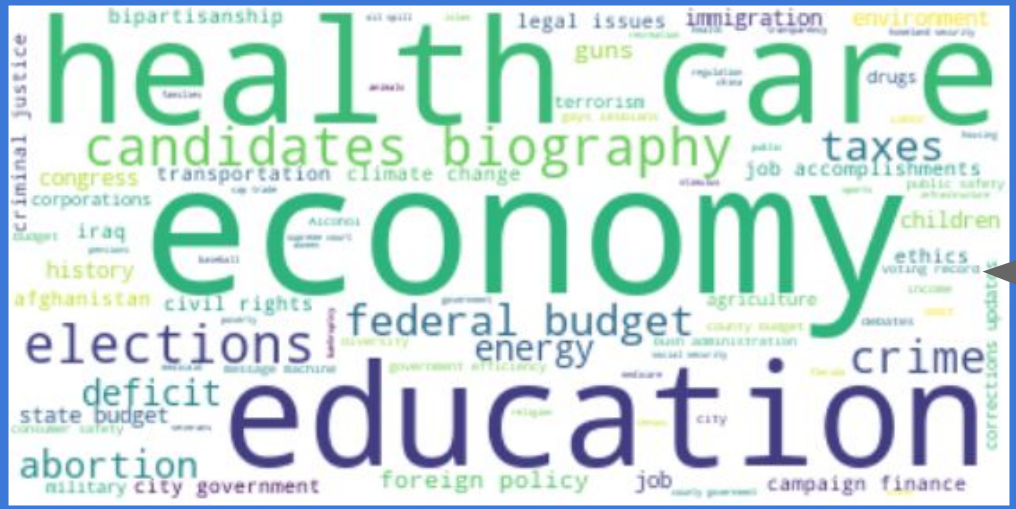
Word Count of the most frequent Authors for False and True News content



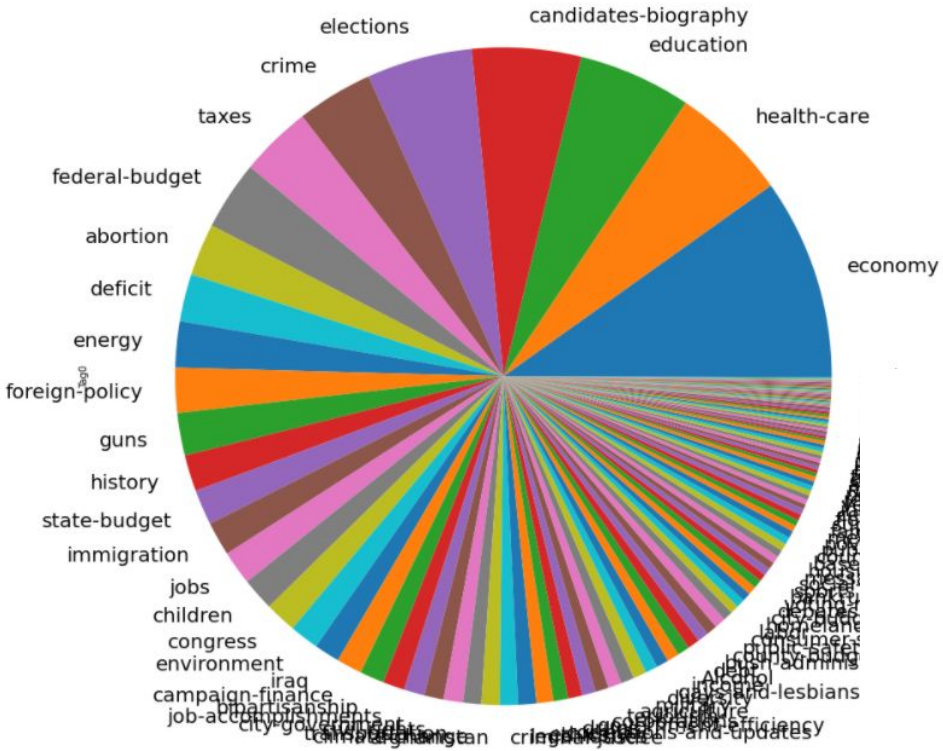
Word Count of the most frequent tags for False and True News content

FALSE

TRUE

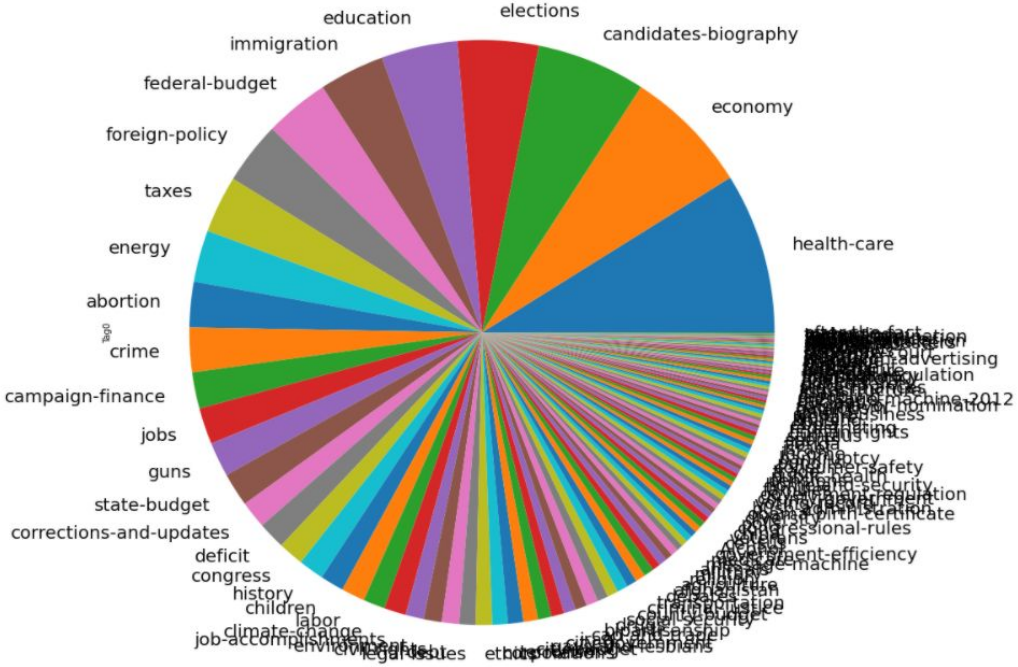


The Pie Chart gives us more insights to our previous Word Count



TRUE

FALSE

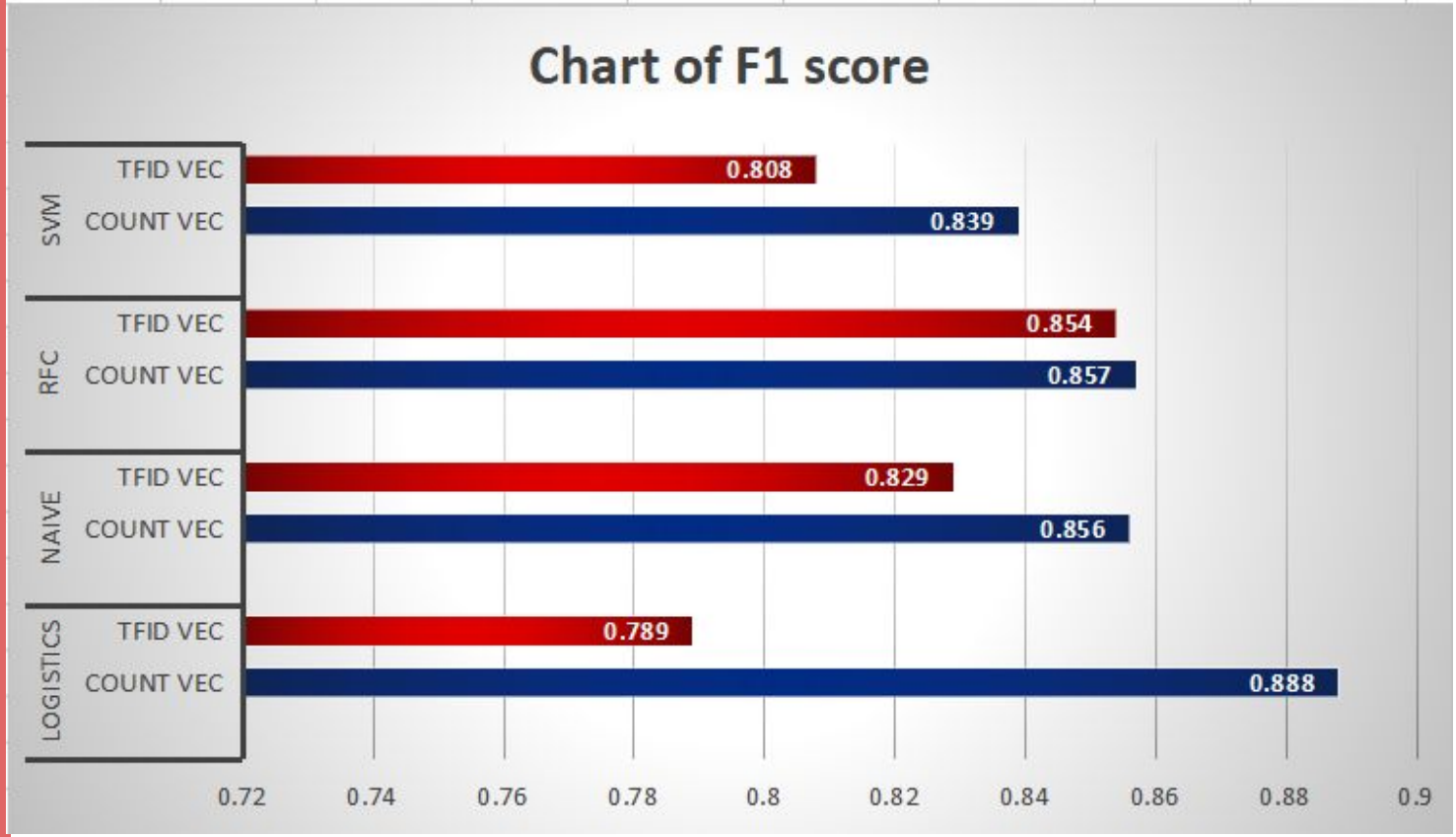


Experiments and Results

“All knowledge - past, present, and future - can be derived from data by a single, universal learning algorithm.” — Pedro Domingos

Experiment One

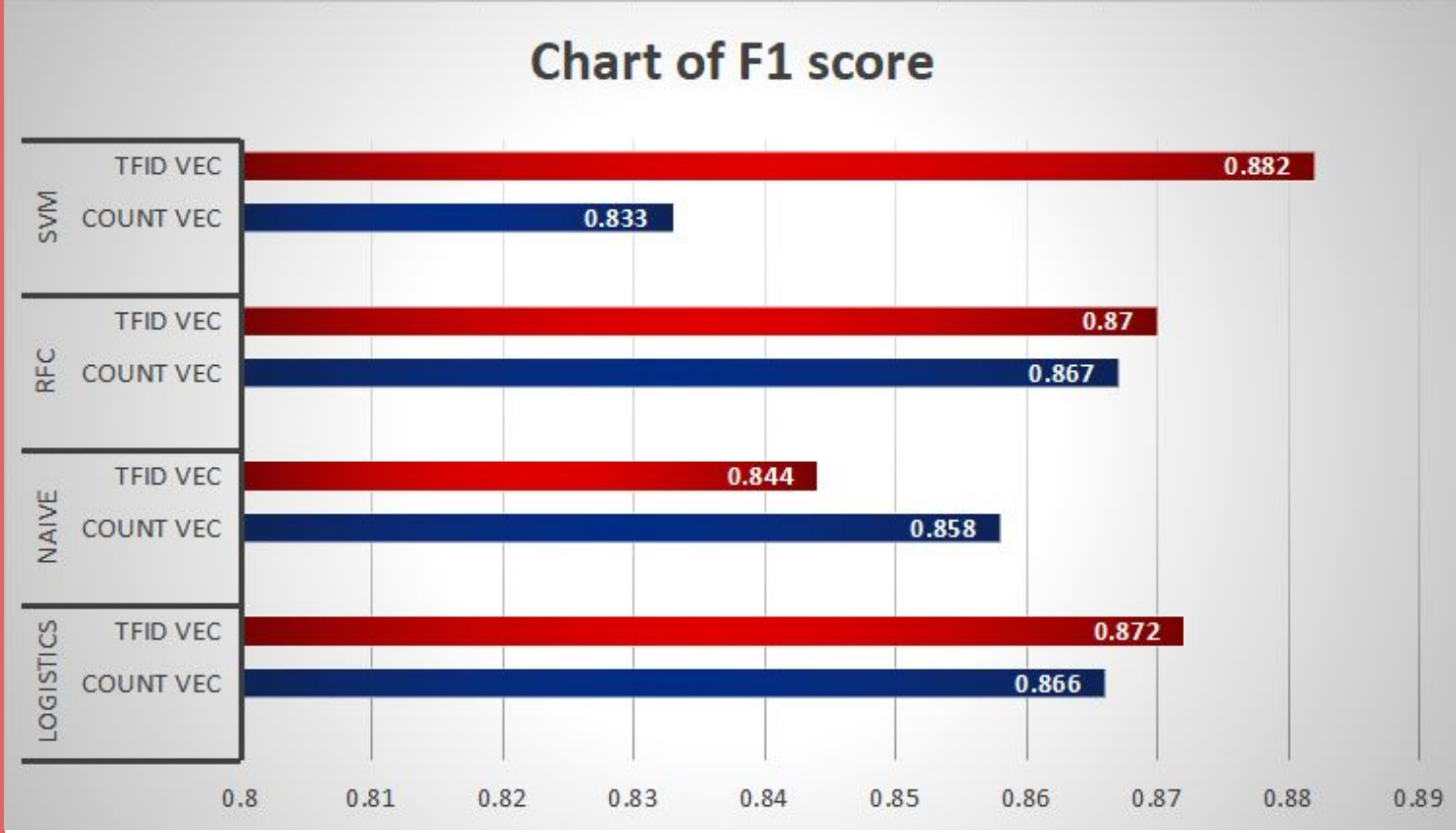
Selecting the best Statistical ML model among 4 classifiers.



The chart above shows the F1 scores on test data **without applying the data cleaning process**. Logistics Regression with count vector features features better generalization of the data.

Experiment Two

Selecting the best Statistical ML models among 4 classifiers.

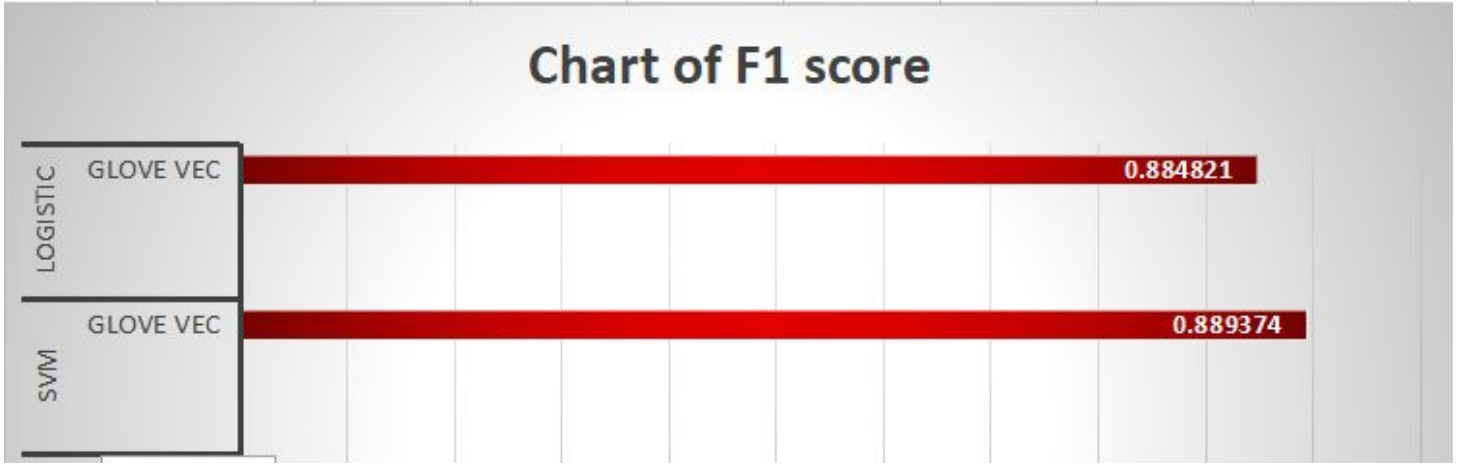


The chart above shows the F1 scores on test data from this experiment. **SVM with TF IDF** weighted features shows better generalization of the data.

Experiment Three

Selecting the best Statistical ML model among 4 classifiers.

I also combined the use of Glove Embeddings with SVM and Logistics Regression, this experiment shows a slight improvement in the dataset.



The chart above showing the F1 scores on test data **reveals SVM** performed better.

Selecting the best Statistical ML model

Model	SVM	Logistics
Cleaned Data	No	Yes
Embedding used	TF IDF	Count Vectorizer
F1 Score	**0.882**	**0.888**

I conducted a parameter grid search on the SVM model was able to improve the F1 score on dataset 2 to 0.735 with the TF IDF vectorizer parameters max_df =0.5 and ngram_range =(1, 2)).

SVM: 0.904	**LOGISTICS: 0.882**
max_df = 0.5, **ngram_range =(1, 2))** **C value = 1.**	**max_df = 0.5,** **ngram_range =(1, 2))** **C value = 0.1**

Experiment Four

Can Neural Networks do better using BiLSTMs?

Bi LSTMs have performed really well for various NLP tasks because of their ability to learn word sequences, so I explored two Bi-LSTM model configurations, and a transfer learning on a pre trained model.

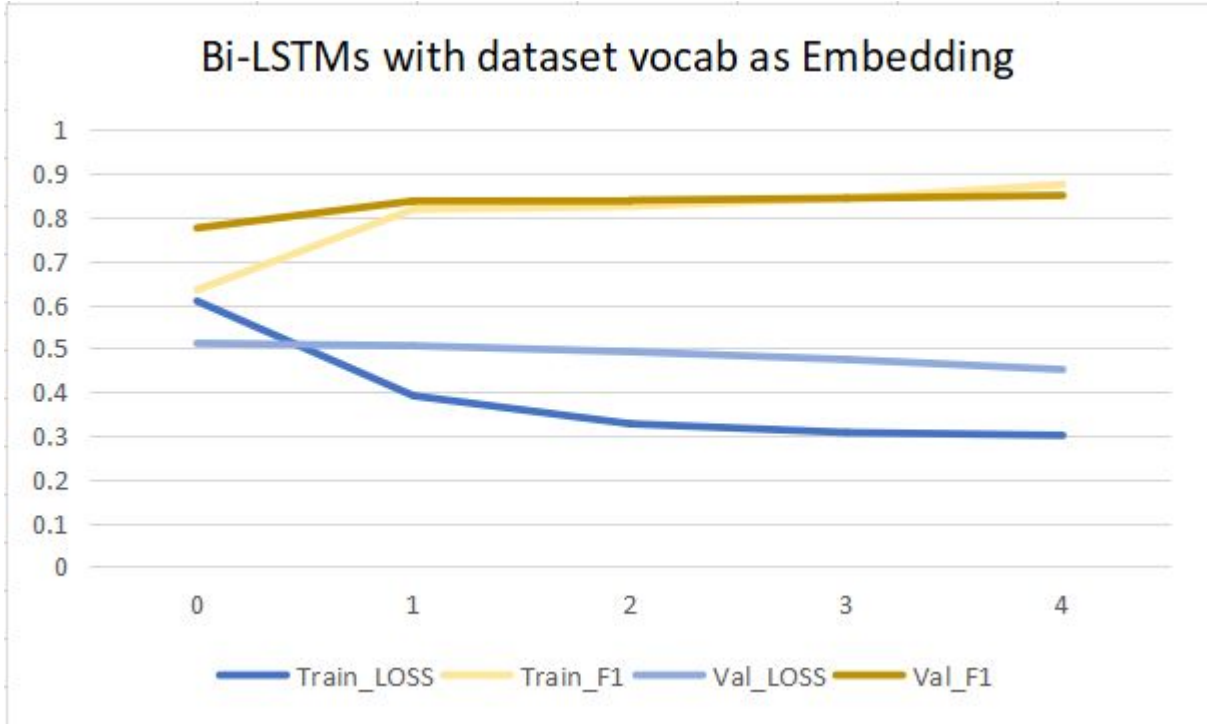
1. Bi LSTMs with our dataset vocabulary as the **embedding layer**.
2. Bi LSTMs with **GloVe pre trained embedding** word Vectors.

.

Experiment Four

Can Neural Networks do better using BiLSTMs?

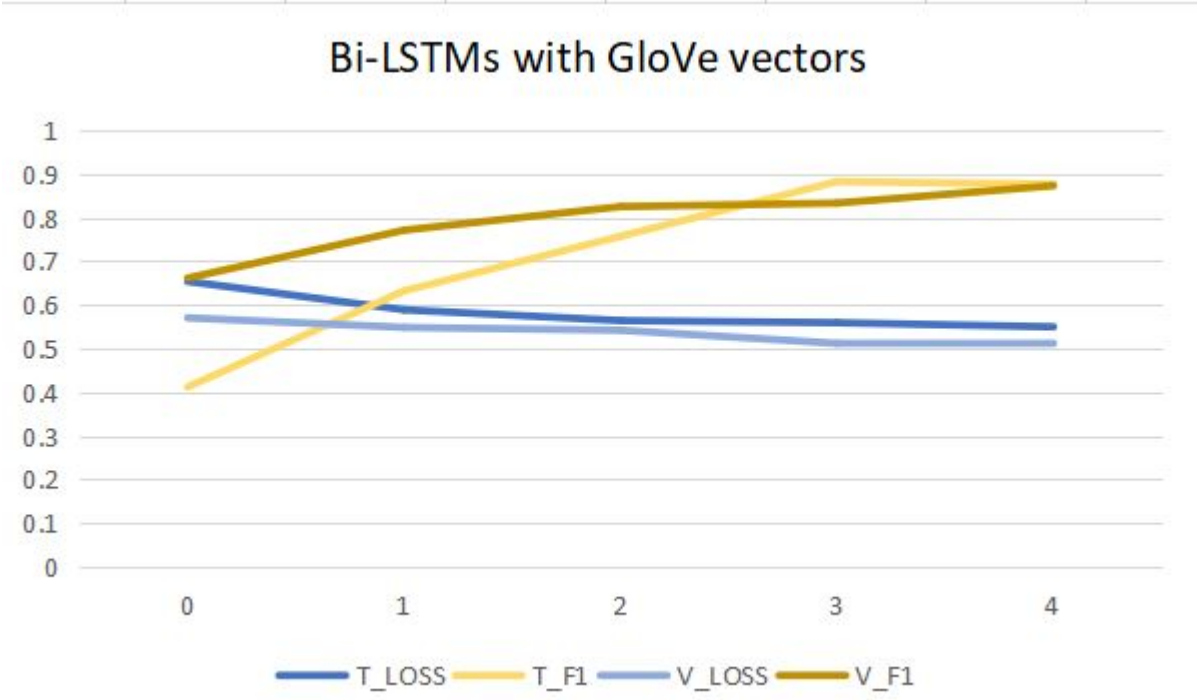
Bi-LSTMs with dataset vocab as Embedding	Bi-LSTMs with GloVe vectors
0.936	0.947



Experiment Four

Can Neural Networks do better using BiLSTMs?

Bi-LSTMs with dataset vocab as Embedding	Bi-LSTMs with GloVe vectors
0.936	0.947



CONCLUSION FROM EXPERIMENTS

1. Statistical ML models performs well on sparse data representations while deep learning approaches on sparse data representations because they have no information of context semantics and sequence.
2. RNNs perform greatly with sequential data especially when the data size is large. Hence, combining our dataset was a good call.
3. Word embeddings are good for representing semantics and meaning, this is a fact why the deep learning models do well with this representation.

THANKS FOR LISTENING

References

1. https://www.webwise.ie/teachers/what-is-fake-news/
2. https://www.cits.ucsb.edu/fake-news/danger-social
3. https://monkeylearn.com/text-classification/#:~:text=Some%20of%20the%20most%20popular,SVM)%2C%20and%20deep%20learning.
4. https://medium.com/the-innovation/sentiment-analysis-using-word2vec-and-glove-embeddings-5ad7d50ddb0d
5. http://nadbordrozd.github.io/blog/2016/05/20/text-classification-with-word2vec/