Tulane University
SCHOOL OF
SCIENCE AND ENGINEERING

Identifying 'Disinformation' on Twitter

@matt_fein
B.S. Computer Science and Political Science

@noah_hendlish
B.S. Computer Science and Finance

@dr._zizhan_zheng
Research Advisor, Department of Computer Science

#Introduction

- **Millions of people rely on social media as a primary news source.**
  - Social media platforms are easily manipulated to spread false information, to large amounts of naïve users.
- **Disinformation is false information intended to mislead, especially propaganda issued by a government organization to a rival power or the media.**
  - This was most famously illustrated during the 2016 US Presidential Election
- **Disinformation accounts tweet with distinct patterns**
  - Using established Machine Learning techniques we have created an accurate and precise algorithm that detects illegitimate tweets
- **To adequately combat malicious accounts, information must be accessible and straightforward for average users**
  - We developed a plugin for Google's Chrome Browser that allows Twitter users to run our algorithm on tweets that are suspicious to that user.

#TheProblem

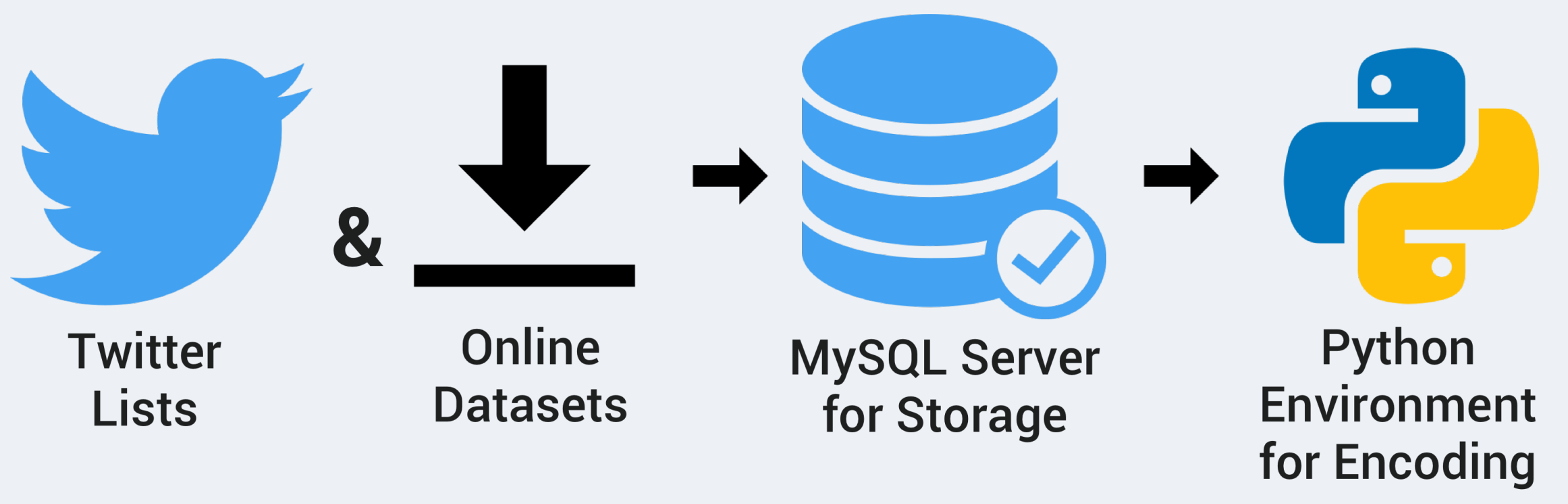
The Internet Research Agency (IRA) is a Russian government agency responsible for much of the political disinformation spread on American media platforms. After the success of Russia's initiative, countries such as Iran and Saudi Arabia have also begun spreading disinformation.

A popular misconception about these programs is that they have a conservative bias. This, however, is not true. In reality, these accounts aim to confuse and diminish people's belief in democratic institutions across the world.

Although this issue is being widely discussed, social media companies currently are not doing enough to curb the influence of illegitimate accounts. Since these accounts do not affect these companies' profits, it is the responsibility of others to ensure that this problem is dealt with. We hope that our tool will be one piece in beginning to end the influence of 'troll' accounts on the internet.

#DataCollection

- **When training a machine learning algorithm, labeled training data is necessary.**
  - This means we needed accurate malicious data and accurate non-malicious data
- **In October 2018 Twitter released a dataset of over 10 million confirmed 'troll' tweets**
- **It was more difficult to find a representative 'clean' dataset**
  - A general query of tweets would presumably contain 'troll' tweets
  - Twitter Lists became our solution
    - Using Twitter's API, we pulled tweets from lists made by reputable sources (Ex: ESPN NBA Journalists)
- **Working with millions of tweets quickly became unwieldly**
  - All tweets were stored on a MySQL database to ease problems with large datasets
- **Tweets were then exported to our Python environment for feature encoding**



#FeatureEncoding

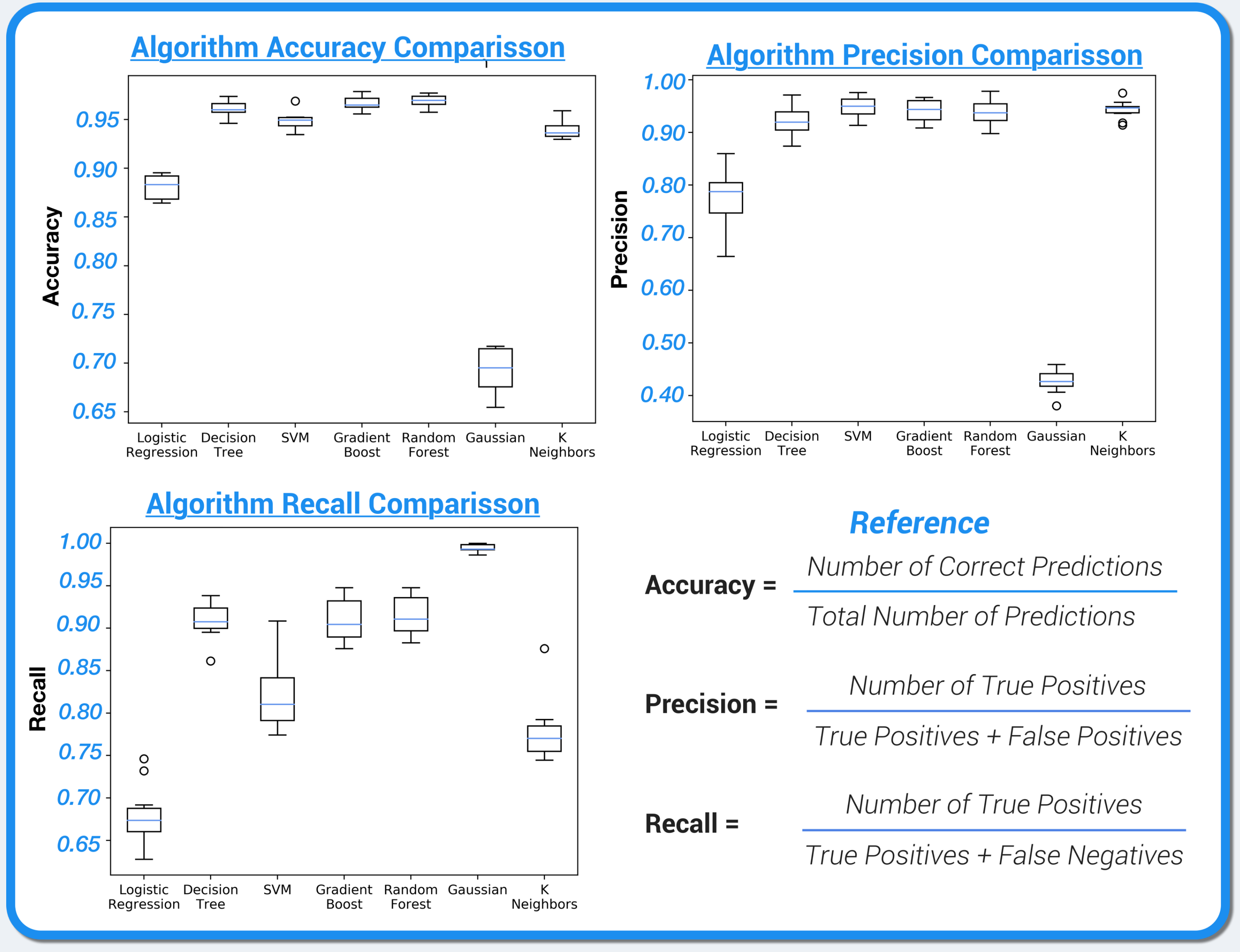
- **Machine Learning Algorithms cannot take in strings of data and compute accurate results**
  - Each tweet's features, or characteristic, had to be numerically encoded
- **Our model is trained for the following features:**

Account Language	# of User Mentions in Bio	# of Emojis in Tweet
Tweet Language	# of Emojis in Bio	# of Hashtags in Tweet
User Location	# of Links in Tweet	Is the User Verified?
# of Hashtags in User Bio	# of User Mentions in Tweet	Tweet Topic
Tweet Client (e.g. Twitter for iPhone)	# of Words in Tweet	# of Links in Bio

Fig 1: *Included Features*

#MachineLearning

- **Before inputting the data into the classification algorithms, language processing determines the tweet topic through Topic Classification**
  - This is done through a dictionary trained on general Tweets and the Latent Dirichlet Allocation (LDA) model
- **After this, the following four classification algorithms take the encoded data as input:**
  - Decision Tree
  - K-Neighbors
  - Support Vector Machine (SVM)
  - Random Forest
- **All four algorithms output a decision**
- **Mean output is calculated by a weighted voting system**
  - The weights are based on the accuracy and precision of the individual algorithm during testing
- **The voting system outputs a result indicating likelihood that tweet is illegitimate**



Reference

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

$$\text{Precision} = \frac{\text{Number of True Positives}}{\text{True Positives + False Positives}}$$

$$\text{Recall} = \frac{\text{Number of True Positives}}{\text{True Positives + False Negatives}}$$

Fig 2: Accuracy, Precision, and Recall Comparisson

***Fig. 3**: Feature Importance in Model*

Feature	Significance to Classification
Tweet Client	27%
Is the User Verified?	20%
# of Words in Tweet	15%
Tweet and Account Language	12%
Account Location	9%

#Conclusion

- **The spread of disinformation is not going away**
  - Although our project's scope was limited to Twitter, the problem extends to other platforms like Facebook and Reddit
- **Awareness is key**
  - Tools like ours can only help so much
  - Awareness of disinformation accounts ultimately will reduce their efficacy
- **On Twitter be aware of following characteristics for disinformation accounts:**
  - Suspicious Location Info:
    - Ex: Many accounts will use 'Louisiana, USA' instead of 'New Orleans, LA'
  - Tweet Client
    - Be on the lookout for 3rd Party tweet clients that you have never heard of
    - Clients like 'Twitter for Web' are more likely to be legitimate than clients like 'generationπ' or ' iziaslav'.
  - Hashtag Use
    - Many of these accounts use multiple hashtags in their Twitter Bios. This is uncommon for legitimate users
    - These accounts will also use multiple unnecessary hashtags in their tweets

References

1. https://www.justice.gov/opa/pr/grand-jury-indicts-thirteen-russian-individuals-and-three-russian-companies-scheme-interfere
2. https://about.twitter.com/en_us/values/elections-integrity.html#data
3. https://medium.com/@katestarbird/a-first-glimpse-through-the-data-window-onto-the-internet-research-agencys-twitter-operations-d4f0eea3f566

We would like to thank Dr. Zizhan Zheng, Dr. Brian Summa , and Dr. Ramgopal Mettu for their extraordinary support.