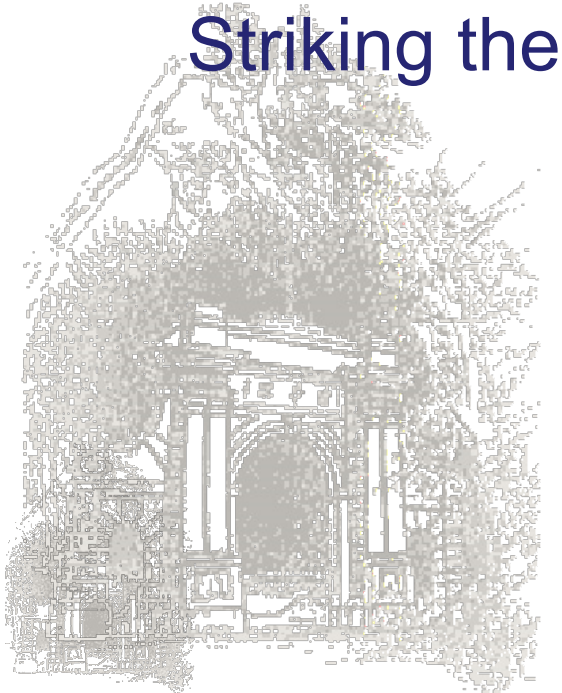
A New Approach to Bot Detection:

Striking the Balance Between Precision and Recall



Fred Morstatter et al.

Presented by Jun Yang

2017.3.29

What is a bot?

- Social media accounts that are controlled by software.

- Self-declared bots.
- Spambots.
- Socialbot.

What is a bot?

- Innocuous.
- Post up-to-date weather, news, historical events, etc.

- Nocuous.
- Infiltration.
- Influence trending.
- Repost or follow specific user.

How many?

- Over half of the accounts on Twitter are not human.
- 5-9% bots produce 24% tweets on Twitter.
- 28% of accounts created in 2008 and half of the accounts created in 2014 have been suspended by Twitter.

Influence

- Harvest private users' data.
- Sway discussion.
- Influence trending hashtags and user statistics.
- **Lose user experience and trust.**
- **Social media researches.**

Bots detection

- Classification tasks.

Bots detection

- Classification tasks.
  - Content
    - Different from normal users.
    - URLs.
    - Sentiment.
    - Length.
    - Similarity.
    - Original tweet.

Bots detection

- Classification tasks.
  - User profile
    - Automatically generated accounts with detectable patterns.
    - E-mail addresses.
    - Creation times.
    - Life time.
    - Screen name and verified name.
    - Human typing.

Bots detection

- Classification tasks.
  - Activities
    - Request frequency.
    - IP addresses.
    - Multiple login location.

Bots detection

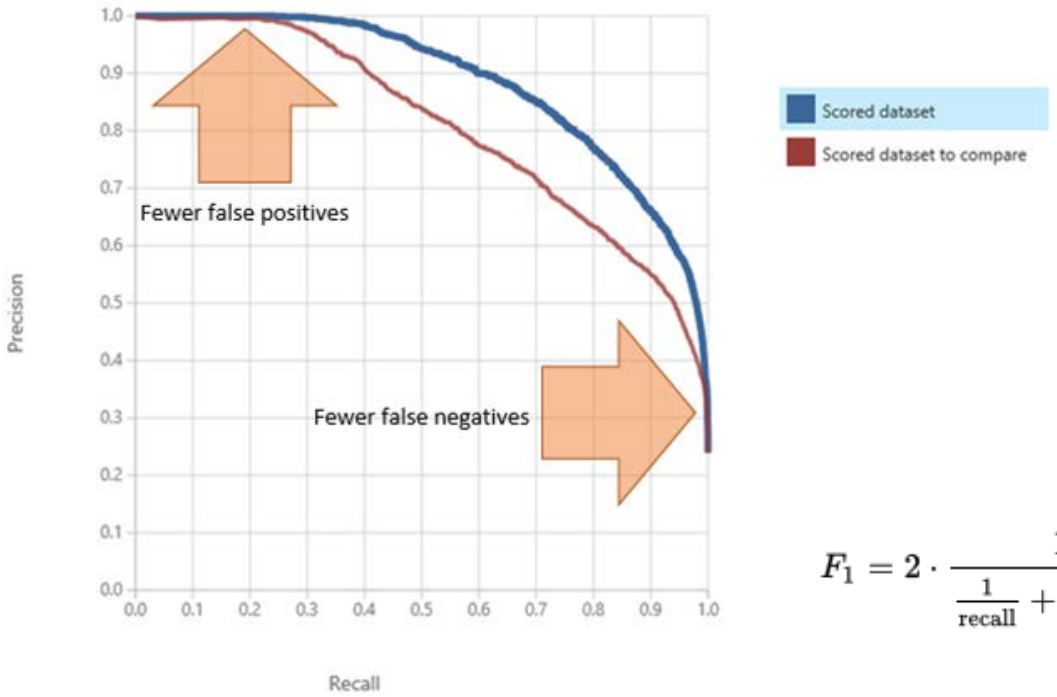
- Classification tasks.
  - Network structure and connection
    - Mass following and unfollowing behaviors.
    - Statistical and structural features.

Ground Truth Acquirements

- Manual annotation
- Suspended users list.
- Honeypots.

Precision vs Recall

- An undetected bot vs an angry user?



$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Contribution

- Two labeled datasets by different techniques.
- Textual features by LDA.
- Modified approach for higher Recall and F1.

Dataset

- **Lybya**
- Querying keywords of *Arab Spring*.
- Collect accounts from 2011.2 to 2013.2
- Check whether suspended or removed in 2015.2
- 7.5% accounts as bots.

Dataset

- **Arabic Honeypot**
- Random tweet or retweet Arabic phrases.
- Measures to avoid suspension.
- Collect human users that tweet same phrases.
- Balanced dataset.

Dataset

- **Arabic Honeypot**
- Random tweet or retweet Arabic phrases.
- Measures to avoid suspension.
- Collect human users that tweet same phrases.
- Balanced dataset.

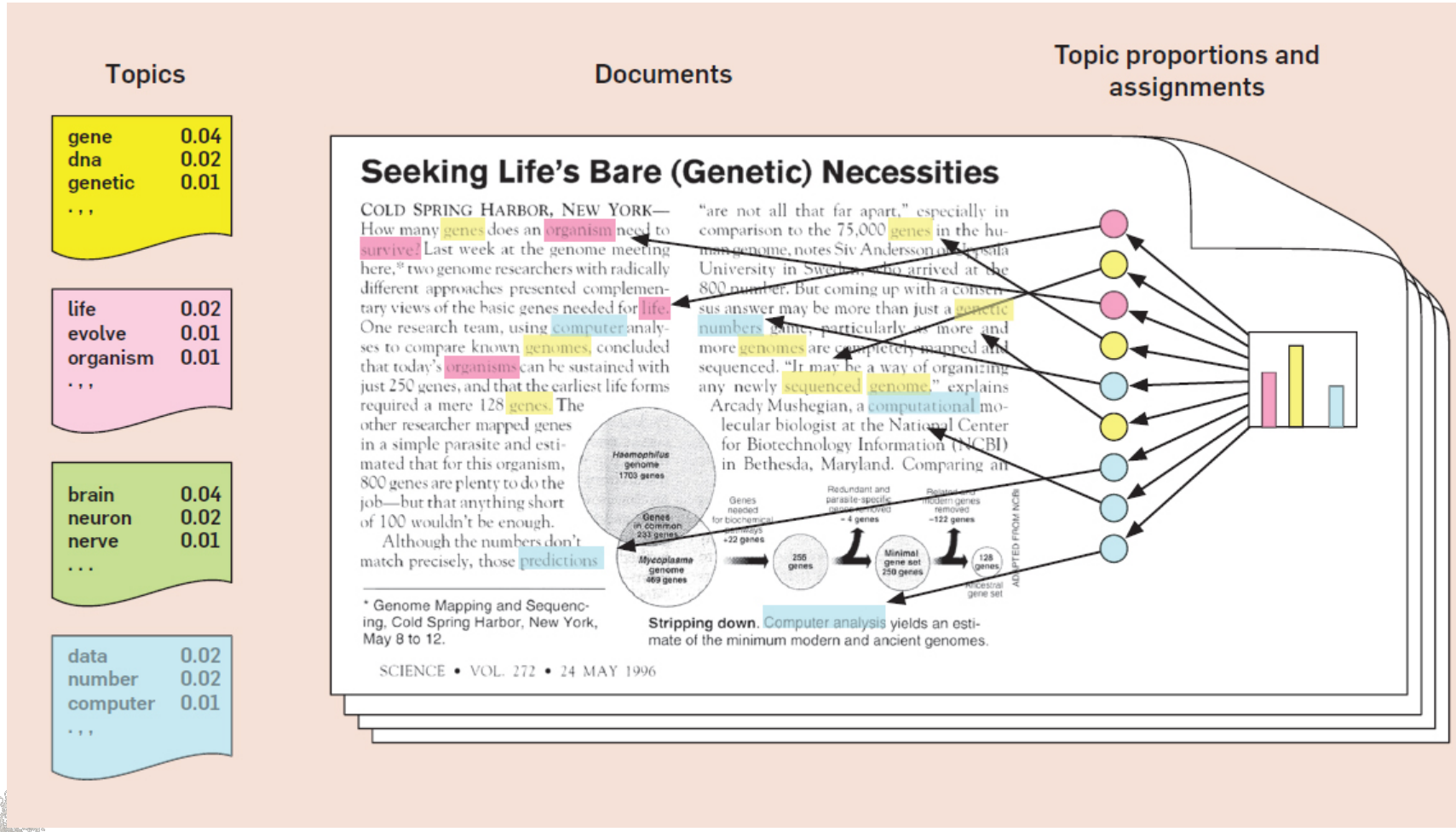
TABLE I: Statistics of the data used in this study.

Property	Libya	Arabic Honeypot
Tweets	1,150,192	504,679
Retweets	576,167	220,500
Unique Users	94,535	6,285
Labeling Approach	Suspended Accts	Honeypots
Bot Ratio	7.5%	49.0%

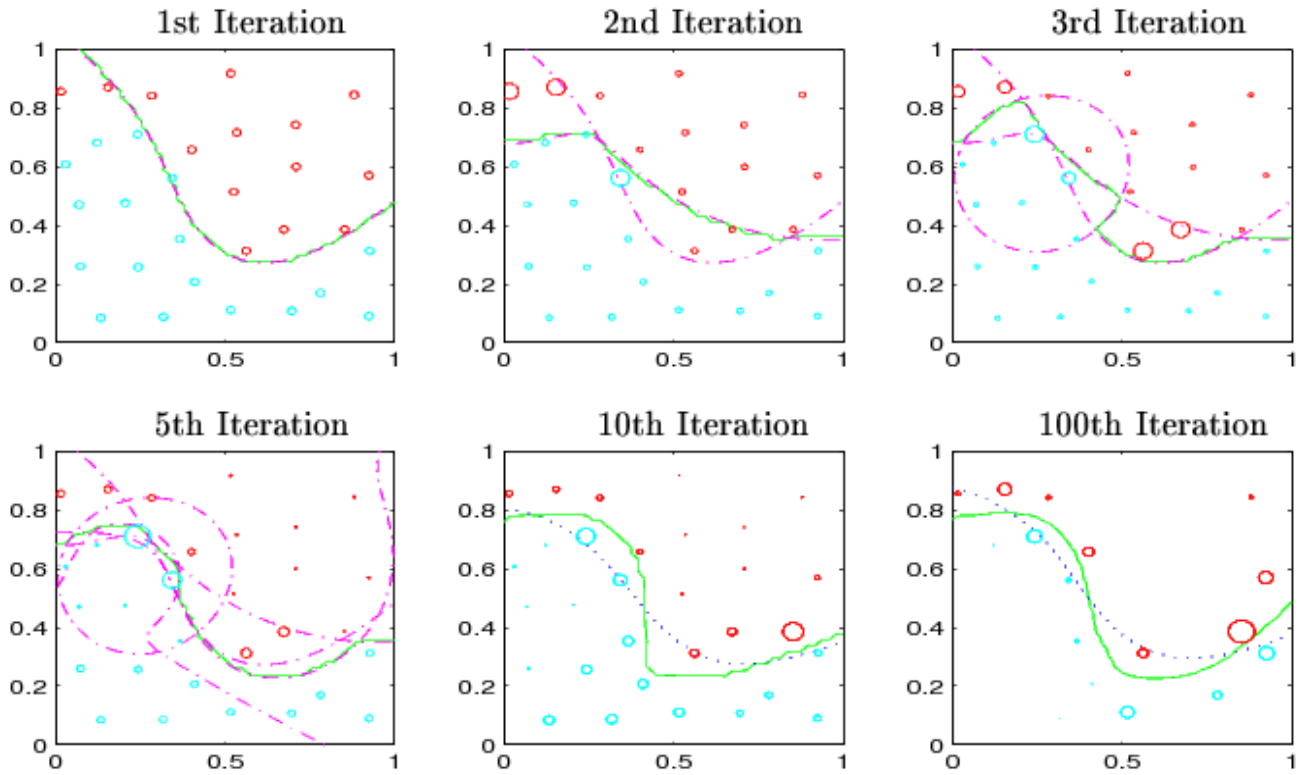
Baselines

- **Heuristics**
- Retweet fraction.
- Average tweet length.
- URLs fraction.
- Average time interval.

LDA


Topics
gene 0.04
dna 0.02
genetic 0.01
life 0.02
evolve 0.01
organism 0.01
brain 0.04
neuron 0.02
nerve 0.01
data 0.02
number 0.02
computer 0.01
Documents
Topic proportions and assignments
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.
Although the numbers don't match precisely, those predictions
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways ~22 genes
Redundant and parasite-specific genes removed ~4 genes
Related modern genes removed ~122 genes
255 genes
Minimal gene set 250 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
SCIENCE • VOL. 272 • 24 MAY 1996

AdaBoost



- Different weak classifiers will focus on different bots.

BoostOR

- A modified AdaBoost algorithm to improve Recall.

BoostOR

- A modified AdaBoost algorithm to improve Recall.

TABLE IV: The Precision, Recall and F_1 measure of different models on Libya dataset.

Method	Precision	Recall	F_1
$Heuristic_{URL}$	6.74%	65.12%	12.21%
$Heuristic_{Retweet\%}$	7.73%	53.63%	13.51%
$Heuristic_{Length}$	7.74%	53.63%	13.51%
$Heuristic_{Time}$	7.48%	99.89%	13.91%
SVM	29.24%	8.78%	13.53%
$AdaBoost$	75.25%	7.48%	13.61%
$BoostOR$	75.41%	8.14%	**14.69%**

TABLE V: The Precision, Recall and F_1 measure of different models on Honeypot dataset.

Method	Precision	Recall	F_1
$Heuristic_{URL}$	49.69%	96.39%	65.58%
$Heuristic_{Retweet\%}$	50.05%	99.33%	66.56%
$Heuristic_{Length}$	50.00%	99.82%	66.63%
$Heuristic_{Time}$	49.99%	99.96%	66.65%
SVM	62.41%	62.52%	62.47%
$AdaBoost$	79.76%	72.41%	75.91%
$BoostOR$	71.42%	82.48%	**76.55%**

Discussion of F1-score

- Balanced test set.

	Precision	Recall	F1
C1	90.00%	70.00%	78.75%
C2	70.00%	90.00%	78.75%

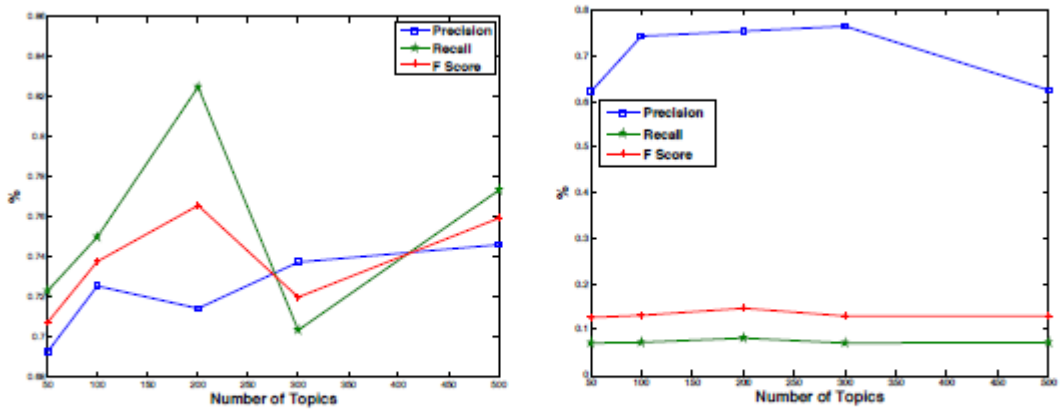
- Positive:Negative = 3:1

	Precision	Recall	F1
C1	96.43%	70.00%	81.12%
C2	87.50%	90.00%	**88.73%**

- Positive:Negative = 1:3

	Precision	Recall	F1
C1	75.00%	70.00%	**72.41%**
C2	43.75%	90.00%	58.88%

Number of topics



(a) Honeypot dataset (b) Libya dataset

Fig. 2: Precision, recall, and F_1 score of BoostOR with varying number of topics.

Thank you!

Questions?