# Towards the Automatic Extraction of Plant Traits from Textual Descriptions

*Onatkut Dagtekin, University of Manchester*

*William Ulate, Missouri Botanical Garden*

*Riza Batista-Navarro, University of Manchester*

# Motivation

- Aid in answering ecological questions
  - Plant to plant interactions
  - Finding suitable species for restoration
- World Flora Online wants to catalogue plant traits
- Human error
- Domain knowledge needed
- Characteristics of descriptions

# Characteristics of descriptions

- Different lengths
  - Morphological: *…Leaves usually densely covered with small scales below,…*
  - Habitat: *Grassland.*
  - Distribution: *Tanqua Karoo to Prince Albert.*

# Characteristics of descriptions

- Specific locations
  - ***Richtersveld,*** *northern* ***Namaqualand*** *to* ***Bitterfontein.***
- High number of adjectives & domain vocabulary
  - *…Leaves pinnately* ***3-foliolate,*** *leaflets narrowly* ***lanceolate,*** *shortly stalked,* ***tomentose*** *below…*

# Proposed Solution: Machine Learning

- Automate the segmentation/classification process
  - Remove human error
  - No domain knowledge needed
- Segment into pieces and label each as:
  - Morpohological
  - Habitat
  - Distribution

# Methodology

- Naïve Bayes (NB)
- Conditional Random Fields (CRF)
- Neural Networks (LSTM)
- Principal Component Analysis for feature extraction

# Experiments

- NB
  - Type of NB = *Gaussian, Bernoulli*
- CRF
  - # of iterations = *[100,…, 6000]*
  - c value = *[0.1,…,6]*
- LSTM
  - # of hidden layers = *[1, 2]*
  - # of nodes per layer = *[5], [20, 15]*

# Evaluation

- 10-fold Cross-validation used for CRF & NB
- Training/test split for LSTM

# Metrics

- F-score
- ROC analysis

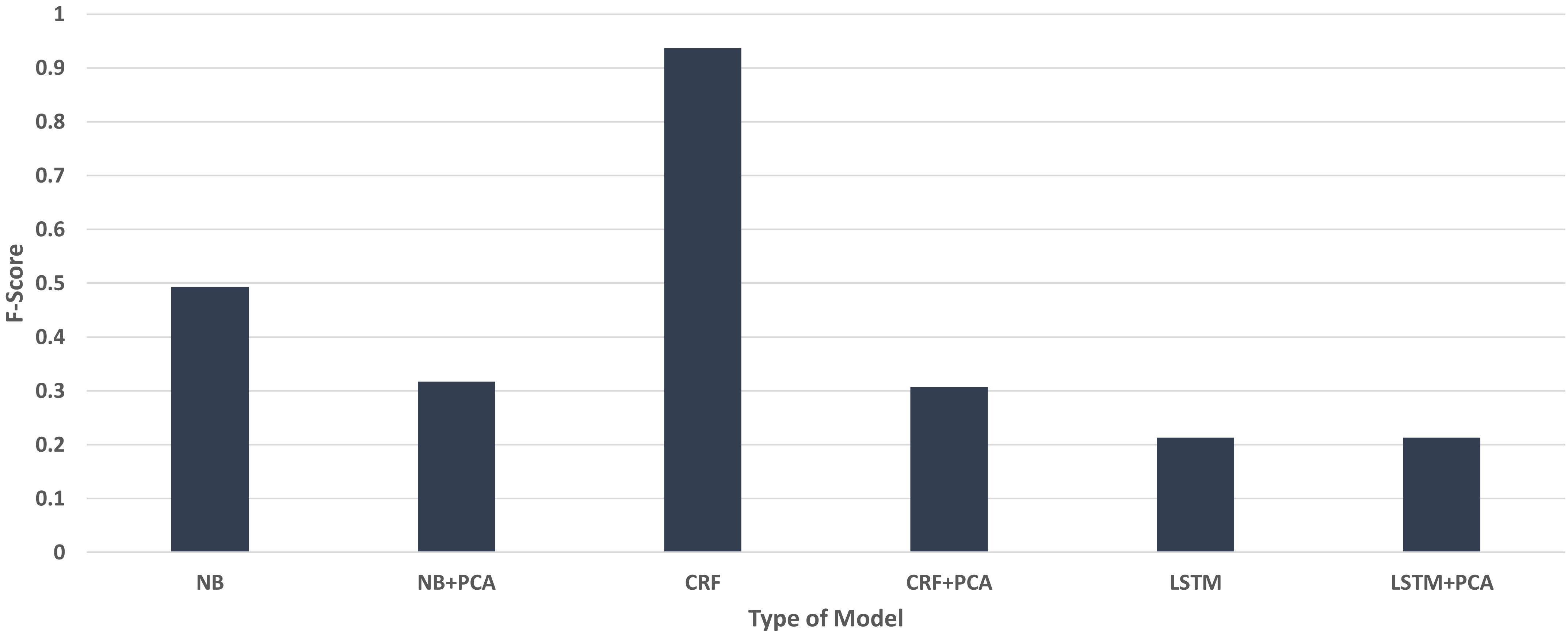
# Features

- Named Entities: LOC & GPE + List
  - *Cedarberg, Humansdorp…*
- Part of Speech Tags
- Ontology Matches: Habitat
  - Only with PoS tags: NN, NNP, NNPS, NNS
  - *Swamp, forest, grassland…*

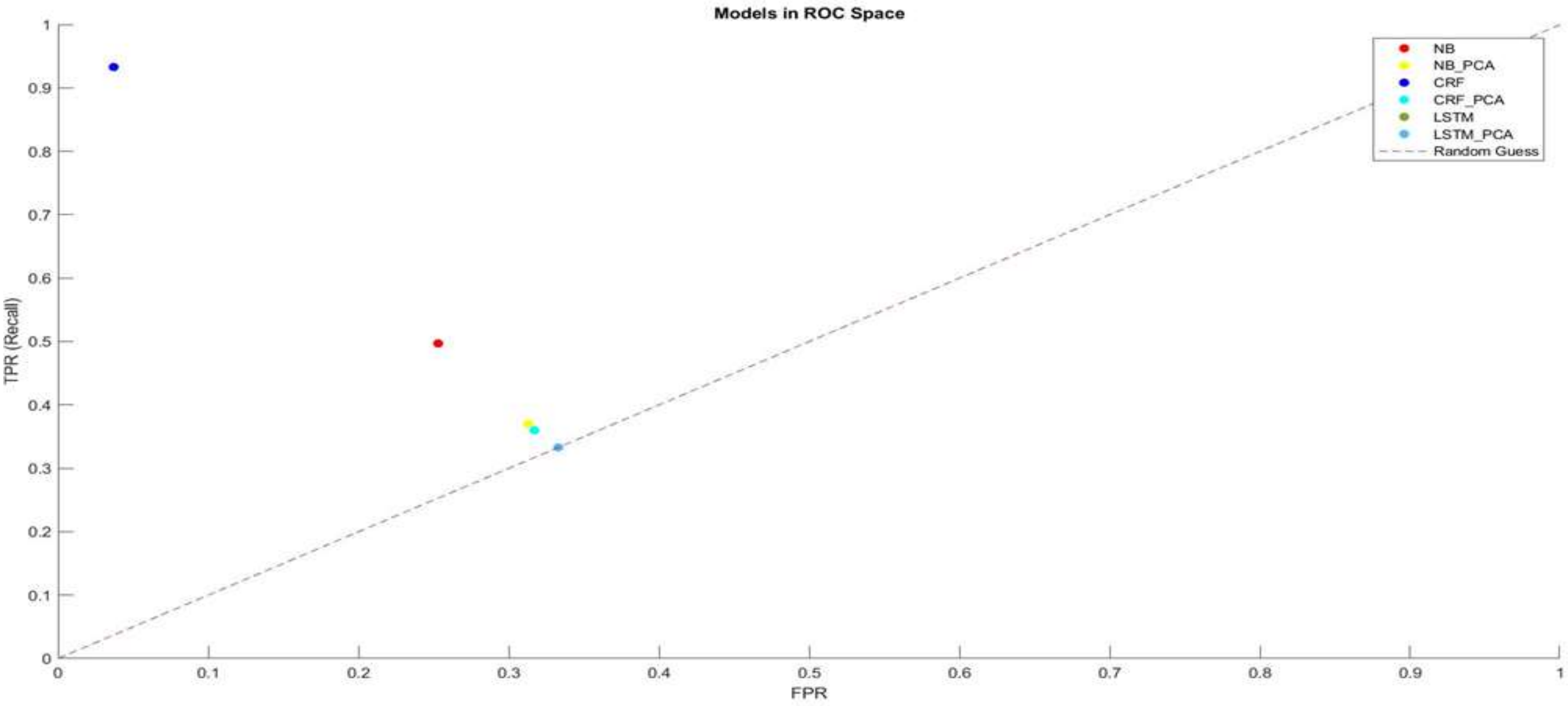
# Features

- Regular expressions: numbers, number ranges, punctuation
  - *(0-9)*-(0-9)*, string.punctuation,...*
- Lists: directions, measurements, continents & oceans
  - *[NE,SW,S,W,...], [mm,µm,...],...*

# Results

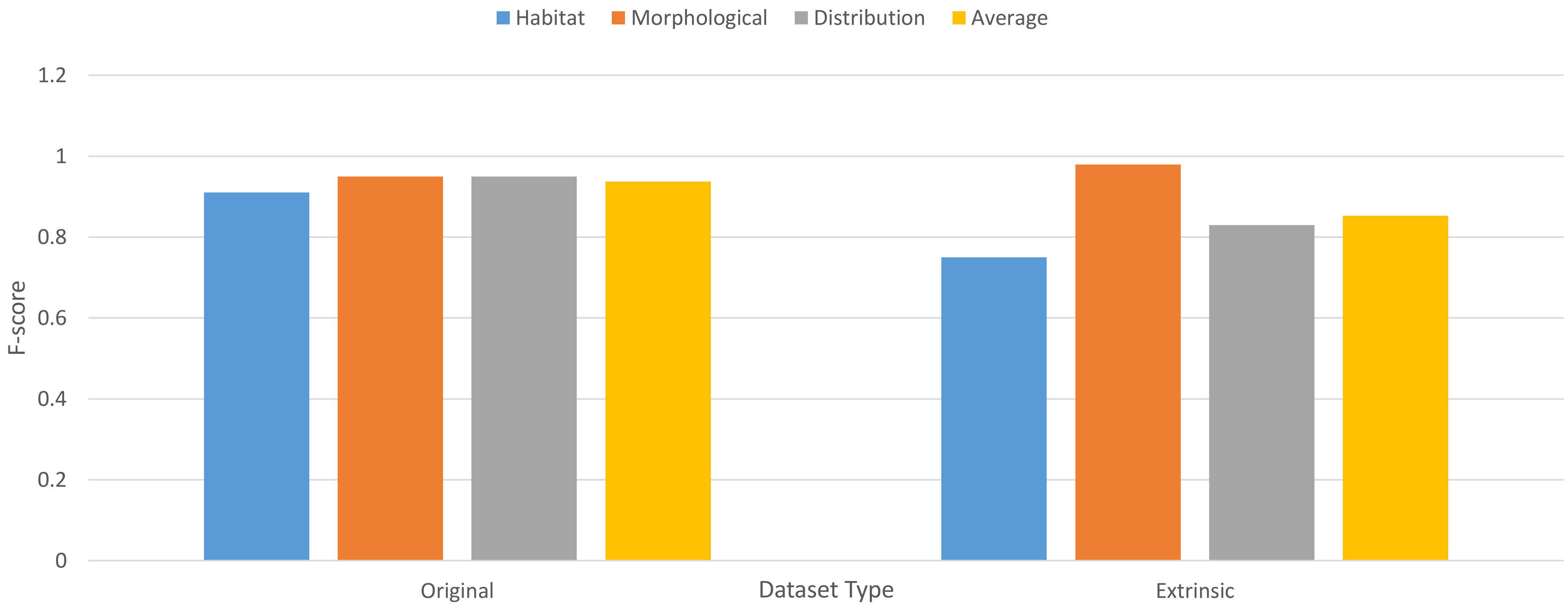

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
F-Score
NB
NB+PCA
CRF
CRF+PCA
LSTM
LSTM+PCA
Type of Model

# ROC Analysis



# Extrinsic Testing

- Best algorithm: CRF
  - Chosen by: mean & variance of f-score



# Behaviour of Models

| Token | NB | CRF | LSTM | True Label |
|---|---|---|---|---|
| ( | Morphological | Habitat | Distribution | Habitat |
| Rocky | Distribution | Habitat | Distribution | Habitat |
| ) | Morphological | Habitat | Distribution | Habitat |
| Grassland | Habitat | Habitat | Distribution | Habitat |
| Or | Distribution | Habitat | Distribution | Habitat |
| Open | Distribution | Habitat | Distribution | Habitat |
| woodland | Habitat | Habitat | Distribution | Habitat |
| . | Habitat | Habitat | Distribution | Habitat |

Table 1: Sample of tokens and the labels for each of the models

# Behaviour of Models

- Gaussian NB:

$$P(x|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\sigma_y^2}\right)$$

- Bernoulli NB:

$$P(x_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i)$$

# Behaviour of Models

- CRF: linear chain
- LSTM:
  - Last 1% of samples:
    - 126 Morphological
    - 190 Habitat
    - 1980 Distribution

# PCA Results

- Harms classifiers
- Non-linear
- Variance not important

# Limitations

- Dataset limited to Southern Africa
  - Cascading errors
- Implementation of algortihms
  - word n-grams, character n-grams

# Conclusion

- Best model: CRF
- PCA is not beneficial

# Future Work

- Multilabel Format
- Segmentation Format
- Bidirectional LSTM, Semi-CRF
- Feature extraction/selection