UniNE at PAN-CLEF 2019: Bots and Gender Identification

Catherine Ikae, Sukanya Nath, Jacques Savoy
University of Neuchatel, Switzerland

Task

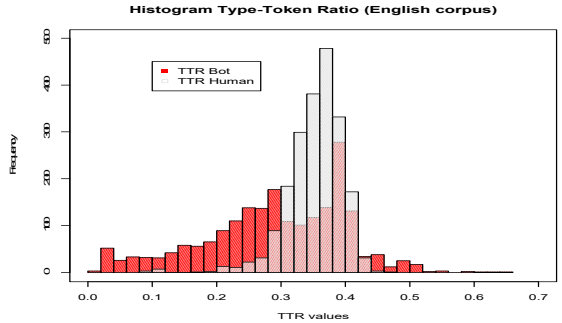
Given a Twitter feed, determine whether its author is a bot or a human. In case of human, identify her/his gender.

- Languages: English & Spanish
- Genre: Twitter feeds

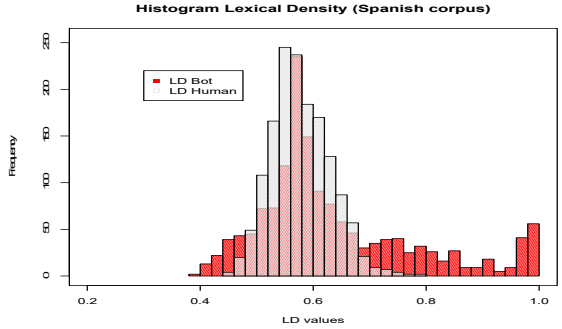
Overall statistics about the training data in both languages

	English		Spanish	
	Bots	Human M / F	Bots	Human M / F
Nb. doc.	2060	1,030 / 1,030	1,500	750 / 750
Nb tweets	205,919	102,842 / 102,930	149,968	75,000 / 75,000
Mean length	2,097	2,014 / 2,123	1,889	1,964 / 1,821
|Voc|	101,826	95,323 / 102,689 human: 162,384	119,965	95,590 / 89,141 human: 147,109

The Dataset : TTR



The Dataset : LD



Methods

The main steps of our automatic attribution system

- Unique vocabulary (VocUnC1, VocUnC2) belonging to the two categories (C1, C2) is determined.
- Determine the terms appearing frequently in one category but absent (or occurring rarely) in the second.
- The documents belonging to two categories are represented as vectors using a reduced set of features.
- The Zeta model is used to count the number of distinct terms appearing in C1 (VocUnC1) and in C2 (VocUnC2).
- If Zeta is unable to achieve a clear decision, the TTR value (Type-Token Ratio) is computed.
- If TTR fails to propose, the system calls the *k*-NN function.

```
preProcessing(trainDoc)
1   vocC1 = defineVoc(trainDoc)
2   vocC2 = defineVoc(trainDoc)
3   VocUnC1 = topVoc(vocC1, vocC2, top=200, min=3)
4   VocUnC2 = topVoc(vocC2, vocC1, top=200, min=3)
5   PtC1 = definePoints(trainDoc, C1)
6   PtC2 = definePoints(trainDoc, C2)
    return(VocUnC1, VocUnC2, PtC1, PtC2)

binaryClassifier(newD,VocUnC1,VocUnC2,PtC1,PtC2):
    decision = 0
1   dec = Zeta(newD, VocUnC1, VocUnC2, 0=3)
2   if (dec == 1) or (dec == 2):  return(dec)
3   aTTR = TTR(newD)
4   if (aTTR < 0.2): return(dec=1)
5   dec = k-NN(newD, PtC1, PtC2, k=13)
    return(dec)
```

Evaluation

The mean accuracy rates achieved considering *k*=13 or *k*=5 neighbors

- All words used in document surrogate: |V| = 162,384.
- The vocabulary size was reduced to consider only terms having a *df* value larger than 9: |V| = 14,728.
- "FS", the feature space was defined by our two-stage feature selection: |V| = 10,173.
- Finally, the feature space reduced to |V| = 100, 200, 300, 400, or 500 terms selected by the information gain.

English corpus

Evaluation of under different feature selection strategies

	English (dev set)			
	k = 13		k = 5	
	B/H	gender	B/H	gender
All voc	0.8807	0.7161	0.8863	0.7161
with df > 9	0.9032	0.7436	0.8976	0.7339
FS	0.9024	0.7557	0.8984	0.7420
100 IG	0.8927	0.7105	0.8960	0.7129
200 IG	0.8807	0.7081	0.8911	0.7145
300 IG	0.8831	0.7048	0.8815	0.6992
400 IG	0.8895	0.7177	0.8871	0.6992
500 IG	0.8960	0.7282	0.8911	0.7056

B/H: Bot or human identification

Official Evaluation of under different feature selection strategies

	TIRA test set 1 k=5		TIRA test set 1 k=13		TIRA test set 2 k=13	
Classifier	B/H	gender	B/H	gender	B/H	gender
FS+Zeta +TTR	0.8939	0.7689	0.8939	0.7992	0.9125	0.7371

Conclusion

- Bots can be viewed as repetitive, showing a low TTR value (usually lower than 0.25).
- Analyzing the emoji distribution, or the most frequent ones, we can infer that humans tend to employ them more frequently than bots.
- Emoji distribution: no significant difference between men and women.
- Our attribution approach is based on a cascade classifier: TTR + *k*-NN.
- A two-stage feature selection strategy was applied to reduce the feature space by one to three orders of magnitude.