Cross-genre gender prediction

Eva Vanmassenhove
Amit Moryossef
Alberto Poncelas
Andy Way
Dimitar Shterionov

The good, the bad and the ugly data…

- +/- 1.800.000.000 websites
- Information
  - Relevant
  - Irrelevant
  - Fake/misleading

Author profiling

- The analysis of content in order to predict author's demographics such as gender, age, personality, native language, or political orientation

- Marketing – improve targeted advertising
- (Digital) forensics – generate additional evidence in criminal investigations
- Security – identify potential threats

- PAN, EVALITA, CLIN2019


Social Media
BLOG

Cross-genre gender prediction… say whaaat?

- Males and females use different language in terms of style and syntax.
  - Different word choices and grammar rules
  - Females use more adverbs and adjectives while writing compared to males
  - Females tend to write more about wedding styles and male tends to write more about technology and politics

Content Based Features

Style Based Features

Topic Based Features

Cross-genre gender prediction… say whaaat?

- Males and females use different language in terms of style and syntax.
  - Different word choices and grammar rules
  - Females use more adverbs and adjectives while writing compared to males
  - Females tend to write more about wedding styles and male tends to write more about technology and politics

Content Based Features

{cf1, cf2, cf3, …}

Style Based Features

{sf1, sf2, sf3, …}

Topic Based Features

{tf1, tf2, tf3, …}

Cross-genre gender prediction… say whaaat?

- Males and females use different language in terms of style and syntax.
  - Different word choices and grammar rules
  - Females use more adverbs and adjectives while writing compared to males
  - Females tend to write more about wedding styles and male tends to write more about technology and politics
- Language diversity:
  - Between languages
    (e.g., English, Dutch, vs., French, German)
  - In genres/domains
    (News, Twitter, YouTube)

Content Based Features

`{cf1, cf2, cf3, …}`

Style Based Features

`{sf1, sf2, sf3, …}`

Topic Based Features

`{tf1, tf2, tf3, …}`

Cross-genre gender prediction… say whaaat?

- The road so far:
  - On Twitter data for English accuracies 80%-85%
  - PAN-RUS Gender prediction across different domains accuracies 65%-93%
  - EVALITA cross-domain with accuracies 51%-64%
  - Russian, Italian => gender agreement with the first person is common
- CLIN2019: Gender prediction for Dutch
  - No agreement with the first person
  - News, Twitter, YouTube

An armada of models to address the challenge

- Neural Networks
  - SpaCy Text Categorizer models
  - Convolutional Neural Network
  - Long Short-Term Memory
  - Long Short-Term Memory with Attention
  - Region-based Convolutional Neural Networks
  - Recurrent Neural Network
  - Self Attention
- Traditional approaches:
  - Statistical Language Models
  - Support Vector Machines
  - K-Nearest Neighbour
  - Logistic Regression
  - Random Forest
  - Naive Bayes

An armada of models to address the challenge

- Neural Networks
  - SpaCy Text Categorizer models
  - Convolutional Neural Network
  - Long Short-Term Memory
  - Long Short-Term Memory with Attention
  - Region-based Convolutional Neural Networks
  - Recurrent Neural Network
  - Self Attention
- Traditional approaches:
  - Statistical Language Models
  - Support Vector Machines
  - K-Nearest Neighbour
  - Logistic Regression
  - Random Forest
  - Naive Bayes

+ Ensembling:
  - Combine multiple systems
  - 5 top best systems

+ Features:
  - Word n-grams
  - Character n-grams
  - Part-of-speech tags
  - Article counts
  - Clusters
  - Words used more by men
  - Words used more by women
  - Diminutives

How did the armada perform?

- Best scores without additional training data

Test set	in-genre	cross-genre
Twitter	64.75%	57.89%
Youtube	62.47%	56.98%
News	66.60%	53.50%
AVG	64.61%	56.12%

- Best scores with additional training data

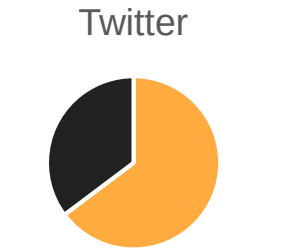
Test set	in-genre	cross-genre
Twitter	65.01%	55.89%
Youtube	63.49%	57.10%
News	66.30%	55.80%
AVG	64.94%	56.26%

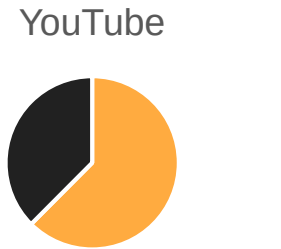
Accuracy: how many predictions are actually correct

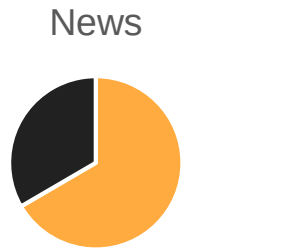
- Above 50% - more correct than incorrect
- What does 53.50% mean?

How did the armada perform?

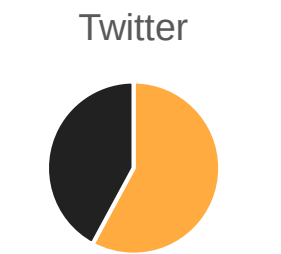
- In-domain

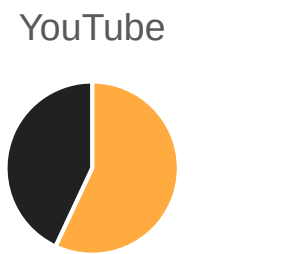


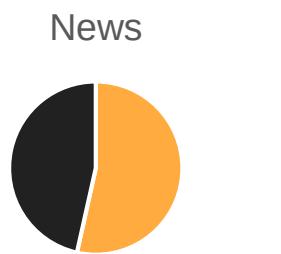




- Cross-genre







What did we win?

- Best performance in 5 out of 6 subtasks
- Best systems are Neural
- Best systems are not that good => more work is needed?

- Understanding about language and gender
  - Generally, male and female speakers/writers tend to use different language
  - Generally, these differences are identifiable
  - … but not transferable from one domain to another

Can we, humans, do better?

Ik weet ook gewoon niet meer wie en wat ik ben en ik wil alles gewoon even niet meer
I just don't know anymore who or what I am and I simply don't want any of it anymore.

Een prachtig domein. Een geweldig onthaal. Een heel prettig weerzien. — bij Anno1000 Resort-Country House https://t.co/Z4YXZu00pO
A beautiful venue. An amazing reception. A very nice reunion. – at Anno1000 Resort-Country house https://t.co/Z4YXZu00pO