N-GrAM: New Groningen Author-profiling Model

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim
Overview

Meet the Team
Task and Data
Our approach
Data insights
Conclusion
MEET THE TEAM
During and after writing

Malvina Nissim  
(Head honcho)

Hessel Haagsma  
Masha Medvedeva  
(PAN Veterans)

Gareth Dwyer  
Josine Rawee  
Angelo Basile  
(PAN Newbies)
TASK AND DATA
Task and data

Twitter data:
- ~100 tweets/ author
- 600 authors / variety

<table>
<thead>
<tr>
<th>Language</th>
<th>Varieties</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>4</td>
<td>2400</td>
</tr>
<tr>
<td>English</td>
<td>6</td>
<td>3600</td>
</tr>
<tr>
<td>Portuguese</td>
<td>2</td>
<td>1200</td>
</tr>
<tr>
<td>Spanish</td>
<td>7</td>
<td>4200</td>
</tr>
</tbody>
</table>
Task and data

Twitter data:
- ~100 tweets/ author
- 600 authors / variety

<table>
<thead>
<tr>
<th>Language</th>
<th>Varieties</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>4</td>
<td>2400</td>
</tr>
<tr>
<td>English</td>
<td>6</td>
<td>3600</td>
</tr>
<tr>
<td>Portuguese</td>
<td>2</td>
<td>1200</td>
</tr>
<tr>
<td>Spanish</td>
<td>7</td>
<td>4200</td>
</tr>
</tbody>
</table>

Gender and Language Variety profiling
- Is the author Male or Female?
- What language variety are they using?
OUR APPROACH
N-grams + SVM

Start with basic system
- Word and Character n-grams
- TF-IDF
- Linear Support Vector Machine
Gotta use it all
Gotta use it all

More data

- Previous PAN data
- Twitter14k dataset
Gotta use it all

**More data**
- Previous PAN data
- Twitter14k dataset

**More features**
- Tokenizers
- POS tags
- Twitter Handles + Place Names
- Emojis
Gotta use it all

**More data**
- Previous PAN data
- Twitter14k dataset

**More features**
- Tokenizers
- POS tags
- Twitter Handles + Place Names
- Emojis

**More classifiers**
- Fast Text, Decision Trees, Neural Networks
More data is better data!
Gotta use it all

More data
- Previous PAN data
- Twitter14k dataset

More features
- Tokenizers
- POS tags
- Twitter Handles + Place Names
- Emojis

More classifiers
- Fast Text, Decision Trees, Neural Networks
More Data

Adding data from previous pan years
- Train on 2016, test on 2017
- Vice versa
- :(  

More Data

Adding data from previous pan years
  - Train on 2016, test on 2017
  - Vice versa
  - :(  

Add Twitter 14k dataset
  - Typically ‘male’ and ‘female’ words
  - :(  
Adding features will help!
Gotta use it all

More data
- Previous PAN data
- Twitter14k dataset

More features
- Tokenizers
- POS tags
- Twitter Handles + Place Names
- Emojis

More classifiers
- Fast Text, Decision Trees, Neural Networks
More Features

Tokenizers
- TweetTokenizer (NLTK)
- Happy Fun Tokenizer (emoticons)
- :(  

More Features

Tokenizers
- TweetTokenizer (NLTK)
- Happy Fun Tokenizer (emoticons)
- :( 

POS Tags
- :( 

More Features

Tokenizers
- TweetTokenizer (NLTK)
- Happy Fun Tokenizer (emoticons)
- :( 

POS Tags
- :( 

Twitter Handles + Place names (Variety)
- Collect corpus of associations with common towns/ handles
- :( 

More Features (2)

Emoji
- SwiftKey report
- :( 
More Features (2)

Emoji
- SwiftKey report
- :( 

GronUP
- Punctuation, word length, capitals, vocabulary, etc, etc
- :( 

---

SwiftKey report
- Poop
- Gun
- Cheers
- Money

SwiftKey
Gotta use it all

More data
- Previous PAN data
- Twitter14k dataset

More features
- Tokenizers
- POS tags
- Twitter Handles + Place Names
- Emojis

More classifiers
- Neural Networks (!!!!!)
More Classifiers

FastText
- It’s fast!
- :( 
More Classifiers

FastText
- It’s fast!
- :( 

scikit-learn MLP
- Not so fast
- :( 
More Classifiers

FastText
- It’s fast!
- :(  

scikit-learn MLP
- Not so fast
- :(  

Keras
- Had fun with generative models
- :(  

<table>
<thead>
<tr>
<th>Data</th>
<th>Features</th>
<th>Classifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MORE MORE MORE MORE

Data

Features

Classifiers
MORE MORE MORE

Data

Features

Classifiers
Grid search for results

64 cores, 1TB RAM, 1 day
Grid search for results

64 cores, 1TB RAM, 1 day

Tune parameters per language / task?
- Not necessary this time
Grid search for results

64 cores, 1TB RAM, 1 day

Tune parameters per language / task?
- Not necessary this time

Scikit-learn defaults are well chosen
- $\text{min\_df}=2$, $\text{sublinear\_tf}=\text{True}$
Start

Tweets → TF-IDF
  character n-grams (3-5 grams)
  word n-grams (1-2 grams) → Linear SVM
# Results

<table>
<thead>
<tr>
<th>Task</th>
<th>System</th>
<th>Arabic</th>
<th>English</th>
<th>Portuguese</th>
<th>Spanish</th>
<th>Average</th>
<th>+ 2nd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variety</td>
<td>N-GrAM</td>
<td>0.8313</td>
<td>0.8988</td>
<td>0.9813</td>
<td>0.9621</td>
<td>0.9184</td>
<td>0.0013</td>
</tr>
<tr>
<td></td>
<td>LDR</td>
<td>0.8250</td>
<td>0.8996</td>
<td>0.9875</td>
<td>0.9625</td>
<td>0.9187</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>N-GrAM</td>
<td>0.8006</td>
<td>0.8233</td>
<td>0.8450</td>
<td>0.8321</td>
<td>0.8253</td>
<td>0.0029</td>
</tr>
<tr>
<td></td>
<td>LDR</td>
<td>0.7044</td>
<td>0.7220</td>
<td>0.7863</td>
<td>0.7171</td>
<td>0.7325</td>
<td></td>
</tr>
<tr>
<td>Joint</td>
<td>N-GrAM</td>
<td>0.6831</td>
<td>0.7429</td>
<td>0.8288</td>
<td>0.8036</td>
<td>0.7646</td>
<td>0.0101</td>
</tr>
<tr>
<td></td>
<td>LDR</td>
<td>0.5888</td>
<td>0.6357</td>
<td>0.7763</td>
<td>0.6943</td>
<td>0.6738</td>
<td></td>
</tr>
</tbody>
</table>
Stereotypes ahead!

WARNING

YOUR ASSUMPTIONS ABOUT ME MAY BE DISTORTED BY EXPOSURE TO OUTDATED GENDER STEREOTYPES
English gender visualisation

Made with https://github.com/JasonKessler/scattertext
English variety visualisation

Colour/color? lift/elevator? Toilet/Loo/WC/Dunny?
English variety visualisation

Colour/color? lift/elevator? Toilet/Loo/WC/Dunny?
English variety visualisation

Colour/color? lift/elevator? Toilet/Loo/WC/Dunny?

“Australia”
Colour/color? lift/elevator? Toilet/Loo/WC/Dunny?

“Australia”, “Dublin”
English variety visualisation

Colour/color? lift/elevator? Toilet/Loo/WC/Dunny?

“Australia”, “Dublin”, “NZ”
English variety visualisation

Colour/color? lift/elevator? Toilet/Loo/WC/Dunny?

“Australia”, “Dublin”, “NZ”, “Edinburgh”
English variety visualisation

Colour/color? lift/elevator? Toilet/Loo/WC/Dunny?

CONCLUSION
Conclusion

N-grams + SVM is (still?) a powerful combo
Conclusion

N-grams + SVM is (still?) a powerful combo

Adding data and features doesn’t always help (and can harm)
Conclusion

N-grams + SVM is (still?) a powerful combo

Adding data and features doesn’t always help (and can harm)

Neural Networks are tricky
Conclusion

N-grams + SVM is (still?) a powerful combo

Adding data and features doesn’t always help (and can harm)

Neural Networks are tricky

Assumptions are wrong
FIN
Questions?
Suggestions?
Answers?
Money?

*With apologies to James Connan*