Topic models and n–gram language models for author profiling

Adam Poulston, Mark Stevenson and Kalina Bontcheva
Department of Computer Science, University of Sheffield

Advanced Computing Research Centre

1. Task

- A set of Twitter users and their posts were provided.
- Set was divided into four languages: Italian, English, Dutch, Spanish.
- Each user's gender, age and personality were given.
- Task was to predict age, gender and personality of unseen users; given a single set of these known users.

2. Data and processing

- Users were equally balanced by author gender. No guarantee of equal balance for other attributes, i.e. age had a marked imbalance.
- As authors may have different numbers of tweets. Up- (and down-) weighting of users was tested to avoid over-fitting to particular authors.
- Investigating effects of Hyperlinks was addressed in two ways:
  - Collecting domain of hyperlinks,
  - replacing hyperlinks with special token.
- The effect of shares and retweets were not considered in this approach.
- A single document representing each user was formed by aggregating their respective tweets.
- Each document was tokenized with a Twitter-aware tokenizer.

3. Feature extraction

- N-gram language model:
  - Word n–grams with n in the range 1–3 were extracted.
  - Weighted using TF-IDF (term frequency–inverse document frequency).
  - Character level n-grams were not considered.
- Topic model:
  - Topic models identify hidden themes in a document.
  - LDA (Latent Dirichlet Allocation) was employed. This is a generative model in which documents are treated as a finite mixture of topics, such that each word in a document must be generated by one of its topics.
  - A topic model trained on input data was used to label every topic as present or not present within each document.

4. Architecture

- Two feature sets—n-grams and topics—were combined to train Support Vector Machines with a linear kernel from package scikit-learn.
- Resulting model was then presented with previous unseen documents; performing judgements on the author attributes it was trained with.

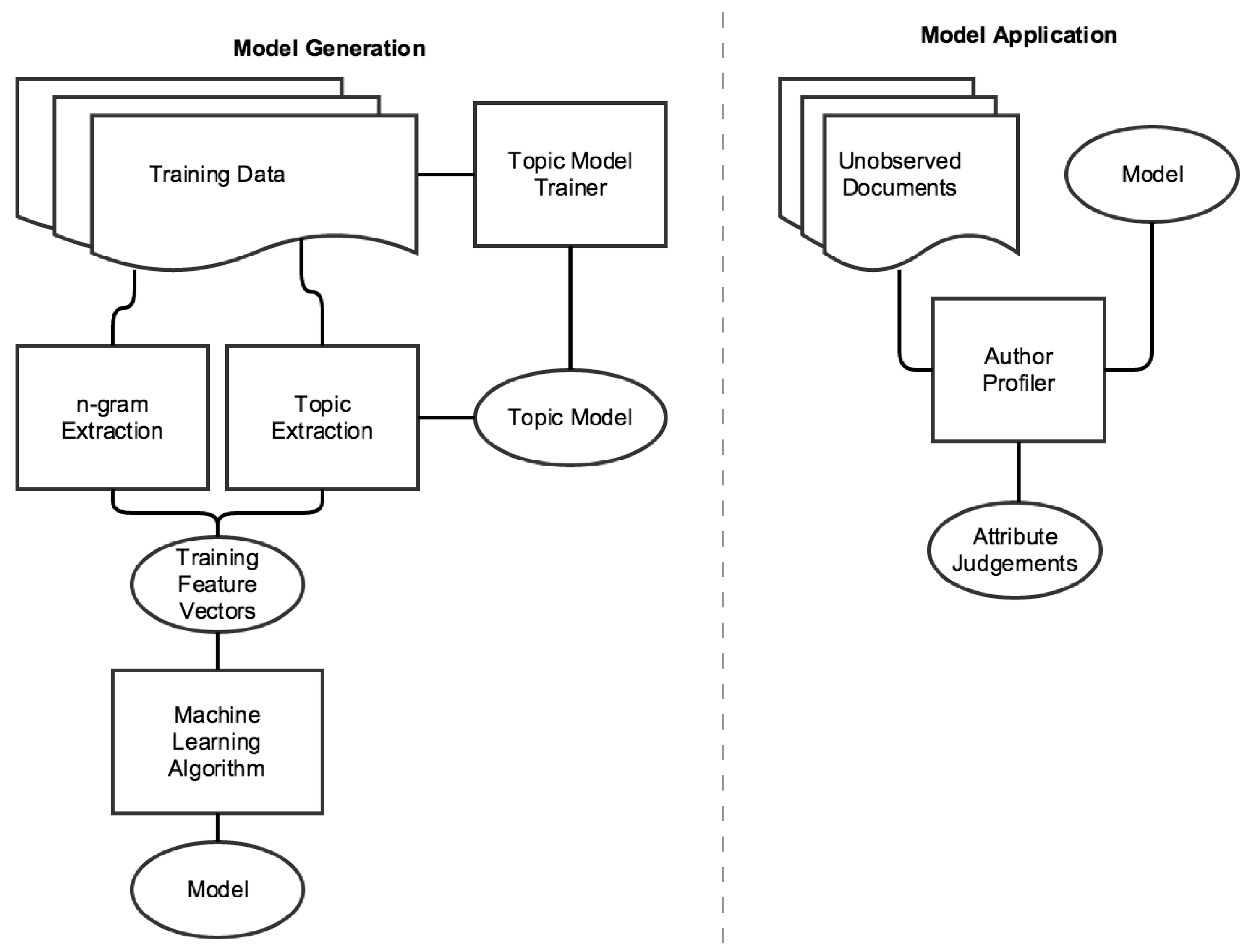

Figure 1: Architecture of presented system.

5. Results

		English	Spanish	Italian	Dutch
	Global Ranking	0.6743	0.6918	0.8061	0.6796
	Average RMSE	0.1725	0.1619	0.1378	0.1409
Accuracy	**Gender**	0.6901	0.8409	0.7500	0.5000
	Age	0.7394	0.5909	N/A	N/A
	Joint	0.5211	0.5455	N/A	N/A
RMSE	**E**	0.1381	0.1669	0.1279	0.1752
	N	0.2223	0.2285	0.1923	0.1511
	A	0.1918	0.1398	0.1257	0.1444
	C	0.1749	0.1412	0.1187	0.1344
	O	0.1352	0.1329	0.1243	0.0993

Table 1: Results of final software submission including global rankings and individual attribute performance.

- Age, gender and their combination were scored using the accuracy metric for each of the four languages.
- Personality aspects (E=Extraversion, N=Neuroticism, A=Agreeableness, C=Conscientiousness and O=Openness) were scored with RMSE (root mean squared error). An average RMSE for each language is also provided.
- Global ranking is a combination of the joint (age, gender) accuracy and the average personality RMSE.
- These results show that n-grams and topic models are useful in developing multiple language compatible author profiling systems, as consistent results are achieved over the four languages.
- Manipulating hyperlinks was found to have no affect on system performance.
- Up- (and down-) weighting of users to avoid author over-fitting also had no affect on performance.

6. Further Work

- Attempt to generate more robust topic models by training on a large external corpus.
- Assess effect of additional stylometric features such as readability.
- Investigate network and behavioural features for author profiling on social media.

References

- D.M. Blei, A.Y. Ng, M.I. Jordan (2003) Latent Dirichlet Allocation *J. Mach. Learn. Res.* (3) pp. 993–1022
- C. Manning, R. Prabhakar, H. Schutze (2008) Introduction to Information Retrieval, *Cambridge University Press*
- F. Pedregosa, *et al.* (2011) Scikit-learn: Machine Learning in Python *J. Mach. Learn. Res.* (12) pp. 2825–30
- F. Rangel, P. Rosso, M. Potthast, B. Stein, W. Daelemans (2015) Overview of the 3rd author profiling task at PAN 2015 *in* L. Cappellato, N. Ferro, J. Gareth, E. San Juan (Eds.) CLEF 2015 Labs and Workshops, Notebook Papers
- R. Řehůřek and P. Sojka (2010) Software Framework for Topic Modelling with Large Corpora *in* Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks *ELRA* pp. 45–50

Acknowledgments

- This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-14-1-0333.

Contact: Adam Poulston
Email: arspoulston1@sheffield.ac.uk
Web: www.acrc.com

The University Of Sheffield.