Celebrity Profiling based on follower's tweets

Hachem Sfar
Sana Ben Hassouna
Ahmed-Adnane Qassemi

University of passau
Chair of Data Science

5981P TEXT MINING PROJECT
Summer Semester-2020

Prof. Dr. Michael Granitzer
Dr. Jelena Mitrovic
Supervised by:
Dr. Fellicious Christofer

Outline

1. Introduction
    - Preamble
    - Problematic
    - Goal and Task
2. Data preprocessing
3. Models
4. Evaluation
    - Gender prediction results
    - Occupation prediction results
    - Birth year prediction results
    - Overall results and comparison with other teams in the PAN competition
5. Conclusions and Future Work
6. References

1. Introduction
    - Preamble
    - Problematic
    - Goal and Task
2. Data preprocessing
3. Models
4. Evaluation
    - Gender prediction results
    - Occupation prediction results
    - Birth year prediction results
    - Overall results and comparison with other teams in the PAN competition
5. Conclusions and Future Work
6. References

Preamble

Author profiling technology predicts personal or demographic traits of an author based on the expression of these traits in an authors text. This analysis requires the availability of a constructed corpora and a large quantity of ground truth such that Celebrities.

Who is a Celebrity ?

Celebrities are prolific and highly influential users on social media acting as hubs for the like-minded. In social media context, a person may attain a celebrity status, if he or she possesses a verified account.

Problematic

- At PAN, celebrity profiling from there **own text** has been studied in 2019, covering different and many demographics including gender, age and occupation.
- However, **the strong homophily** within a celebrities community opens the way to study and explore the interplay of author characteristics and community expression in author profiling and predicting traits of the celebrity of users **without own texts using only their follower network**.

Task

What is our goal ?

Given the Twitter feeds of the followers, we will predict three demographics of a celebrity: the age, occupation and gender.

1. Introduction
   - Preamble
   - Problematic
   - Goal and Task
2. Data preprocessing
3. Models
4. Evaluation
   - Gender prediction results
   - Occupation prediction results
   - Birth year prediction results
   - Overall results and comparison with other teams in the PAN competition
5. Conclusions and Future Work
6. References

Data preprocessing

- Create balanced Dataset
- Removing stopwords, and all symbols except for letters, numbers, @, and ,all retweets of a follower, and all accented characters because we only want to analyze the English language.
- Replacing all hyperlinks with a $< url >$ token, all user tagging with a $< user >$ token, all the tabs, repeated punctuation with a single character, and new lines and consecutive white spaces with single white space.
- Expands contractions (e.g can't= cannot) and tweet slang (eg. omg = oh my god)
- Transfer emojis into texts by using emoji library
- Convert string labels to numerical values (eg creator :0, sports :1)
- Discretize the variable using a window size m based on the birth year. The value increases linearly from about 3 years for 1999 to about 9 years for 1940.

1. Introduction
   - Preamble
   - Problematic
   - Goal and Task
2. Data preprocessing
3. Models
4. Evaluation
   - Gender prediction results
   - Occupation prediction results
   - Birth year prediction results
   - Overall results and comparison with other teams in the PAN competition
5. Conclusions and Future Work
6. References

Models

- Machine Learning Algorithms :
    - Convert a collection of raw documents to a matrix of TF-IDF features.
    - Build many model using Logistic regression, Linear SVC, MultinomialNB, Random forest.
- Neural network:
    - Produce word embedding by using word2vec
    - Use various model:
      - Bidirectional LSTM
      - Convolutional 1D $->$ Max pooling $->$ LSTM
      - Convolutional 1D $->$ max pooling layer
      - Fully connected neural network
- Transfer learning using BERT

1. Introduction
   - Preamble
   - Problematic
   - Goal and Task
2. Data preprocessing
3. Models
4. Evaluation
   - Gender prediction results
   - Occupation prediction results
   - Birth year prediction results
   - Overall results and comparison with other teams in the PAN competition
5. Conclusions and Future Work
6. References

Evaluation

- To avoid overlap between training and test subsets and then avoid overfitting a 5-fold cross validation protocol was applied for the evaluation.
- The evaluation process will be based on a combined metric cRank, which is the harmonic mean of each label's metric.

$$cRank = \frac{3}{\frac{1}{F_{1,occupation}} + \frac{1}{F_{1,gender}} + \frac{1}{F_{1,age}}}$$

The experimental results form tests on all mentioned models:

Gender prediction results

	Training set F1	Test set F1
BERT	**0.74089**	**0.68447**
Neural Network	0.555,0.5126	0.51,0.49
Linear SVC	0.56284	0.49174
Logistic regression	0.563118	0.517457
MultinomialNB	0.562294	0.50896
Random forest classifier	0.519747	0.428324

Table 1: Results from experiments on gender prediction using different models

Occupation prediction results

	Training set F1	Test set F1
BERT	**0.57179**	**0.5557**
Neural Network	0.3834,0.318	0.34,0.307
Linear SVC	0.405073	0.3839
Logistic regression	0.407459	0.3726
MultinomialNB	0.400455	0.3755
Random forest classifier	0.30959	0.2947

Table 2: Results from experiments on occupation prediction using different models

Birth year prediction results

	Training set F1	Test set F1
BERT	**0.3842**	**0.33574**
Neural Network	0.2557,0.2129	0.196, 0.19
Linear SVC	0.270562	0.2167
Logistic regression	0.273701	0.212307
MultinomialNB	0.270436	0.241433
Random forest classifier	0.222645	0.20192

Table 3: Results from experiments on birthyear prediction using different models

Overall results and comparison with other teams in the PAN competition

Team	Test-Dataset			
	CRANK	AGE	GENDER	OCCUPATION
baseline-ngram celebrity-tweets	0.631	0.500	0.753	0.700
hodge20	0.577	0.432	0.681	0.707
koloski20	0.521	0.407	0.616	0.597
our-team	**0.480**	**0.335**	**0.684**	**0.555**
tuksa20	0.477	0.315	0.696	0.598
baseline-ngram follower-tweets	0.469	0.362	0.584	0.521
random	0.333	0.333	0.500	0.250

Table 4: Results of other teams

PAN team results source: ***https://pan.webis.de/clef20/pan20-web/celebrity-profiling.html***

1. Introduction
    - Preamble
    - Problematic
    - Goal and Task
2. Data preprocessing
3. Models
4. Evaluation
    - Gender prediction results
    - Occupation prediction results
    - Birth year prediction results
    - Overall results and comparison with other teams in the PAN competition
5. Conclusions and Future Work
6. References

Conclusions and Future Work

Challenges

We faced several challenges that are worth highlighting:

- The birth year prediction task was the most critical task in which the number of labels was reduced, creating five groups for better accuracy.
- Dataset of followers' tweets had an evident imbalance.
- Dataset volume: the need to process more than 15 GO of tweets. To deal with that, we worked on the server of the university to extract a subset chosen randomly of tweets per follower profiles from our corpus.

Conclusions and Future Work

Future Work

- Analyze the models with all the volume of dataset using a cluster of servers.
- Use another technique in the preprocessing step consisting in leaving and analysing only hashtags instead of analyzing all the tweets.
- Use XLNet which is a new transfer learning model that appear to overcome some of the BERT limits.

1. Introduction
    - Preamble
    - Problematic
    - Goal and Task
2. Data preprocessing
3. Models
4. Evaluation
    - Gender prediction results
    - Occupation prediction results
    - Birth year prediction results
    - Overall results and comparison with other teams in the PAN competition
5. Conclusions and Future Work
6. References

References

[1] Francisco Manuel Rangel Pardo, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd author profiling task at pan 2015. In CLEF 2015 Evaluation Labs and Workshop Working Notes Papers, pages 1–8 (2015).
[2] Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, andWalter Daelemans. Overview of the 2nd author profiling task at pan 2014. In CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014, pages 1–30 (2014).
[3] Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at pan 2013. In CLEF Conference on Multilingual and Multimodal Information Access Evaluation, pages 352–365. CELCT (2013).

Thank you for your attention!