Celebrity Profiling

Sana Ben Hassouna
benhas02@ads.uni-passau.de

Hachem Sfar
sfar01@ads.uni-passau.de

Ahmed-Adnane Qassemi
qassem01@ads.uni-passau.de

ABSTRACT

This paper is written in the context of the Text mining project in the University of Passau for the summer semester 2020 and aims to describe our approach to solve the problem of Celebrity Profiling. The task is defined in the following way: Given the Twitter feeds of the followers, determine the age,occupation, and gender of a celebrity. There are two classes for gender, four for occupation and the age range is from 1940 to 1999.

KEYWORDS

celebrity profiling, twitter, social groups, follower network

1 INTRODUCTION

Author profiling has become a topic of great interest nowadays due to its strong impact and to it wide range of applications including politics, social science and Internet security. This aims to analyses texts written by the target person, build a profile and predicting as much information as possible. This analysis requires the availability of a constructed corpora and a large quantity of ground truth. In this regard, Celebrity are one source of ground truth, available in large quantities with a near-perfect label reliability.
In recent years, multiple papers are dealing with Celebrity profiling in different contexts.
At PAN, author profiling has been studied since 2013, covering different and many demographics including personality [1], age and gender [2–4], language variety [4], genres including blogs, reviews, and social media posts [4].
However, previous research are based on Author profiling where own texts are used in the analyses process. This approach has its own limitations because it's based on one source, the celebrity texts. In this context, a long standing question in author profiling is, how much the assessed expression used by algorithms depends on characteristics of an individual author and how much on the expression of social groups and communities.
As celebrities are among the most prolific users and rallying followers they represent an ideal condition to study the interplay of author characteristics and community expression in author profiling. For instance, our proposal goes further demonstrating the correlation between follower expression and demographics of the celebrity and predicting traits of the celebrity without own texts using only their follower network. Studied traits consist of demographics including the occupation, age, and gender.

2 THE WEBIS CELEBRITY CORPUS

This section defines our Webis Celebrity Corpus, identifying the source and detailing the content and specification of your corpus. A corpus prepossessing follows.

2.1 Who is a Celebrity?

In twitter context, a person may attain a celebrity status, if he or she possesses a verified Twitter account. Twitter identify "that an account of public interest is authentic" (Twitter, 2018) by putting a blue checkmark badge. A celebrity is usually followed by a huge number of persons named followers who supports and admires him and likes his ideas.

2.2 Corpus source

The corpus is provided by the PAN CLEF 2020, which is a series of scientific events and shared tasks on digital text forensics and stylometry.

2.3 Exploratory Corpus Analysis

The datasets contain three files containing JSON objects of size 15 Go . The first file contains the inputs, which are the English tweets of followers. The second file represent the expected output containing the occupation, gender and the age that should be predicted. In additional file was provided containing the Twitter timelines of the original celebrities which can be used for additional study.

We have done an exploratory data analysis to summarize the datasets main characteristics and to check weather there is missing values and highlight the correlation between the twitter feeds of the followers and each demographics of the celebrity to predict.

First, we looked at the number of Celebrities in our datasets by counting the number of unique ids and we found that the number of Celebrities is 10185. Then, we counted the total number of tweets we found: 25732538 and each tweet contains 14.39 words. We found that the dataset contains: 7275 Male and 2910 Female Celebrity. And the age average of these persons is 42.83. Then, we created a word cloud:
Word cloud represent the frequency or the importance of each word.
First, we created a cloud of all the tweet in the dataset .
Then, we created more word clouds for each occupation and gender, to see the vocabulary words used by each category:
With 4 different occupation classes to choose from. Celebrities are usually famous for the activities they perform in their given occupation, and it seems that a celebrity and his followers would often be writing about such activities and use words that are clear indicators of the given field.

Label	Value	#Tweets
Gender	Male	7275
	Female	2910
Occupation	Creator	1724
	Politics	480
	Performer	4944
	Sports	3037

Table 1: Characteristics of dataset

3 LITERATURE

At PAN, author profiling has been studied since 2013, covering different demographics including age and gender, personality, language variety, genres including blogs, reviews, and social media posts, and cross-domain prediction.
The subject of research in PAN 2019 was to predict the four demographics gender, year of birth, degree of fame, and occupation of a celebrity from

their history of tweets on Twitter. Eight participants submitted software to this task. An overview of the Celebrity Profiling Task at PAN 2019 and a comparative study was conducted [5]. The comparative study shows that the approach which achieved the best performance is the Radivchev et al. approach [6].

4 PREPROCESSING AND SUBMITTED MODEL

After taking a look at the literature, we have decided to involve the following techniques:

4.1 Preprocessing and Baselines

Since it is required to load a 15 GB data set into memory, we requested access to the university server. Due to the limited laptop resources we extracted a subset chosen randomly of tweets per follower profiles from our corpus for our experiments. We chose to split the dataset into a training and test set in a ratio of 80:20.

We will start with the data preprocessing phase applied to the twitter feeds of the followers. The applied preprocessing involved the following operations:

- Convert the imbalanced dataset into balanced dataset for our three prediction task: occupation, gender and age prediction.
- Lowercase the text
- Removing all symbols except for letters, numbers, @, and
- Replacing all hyperlinks with a <url> token
- Replacing all user tagging with a <user> token
- Removing all the tabs, new lines and multiple consecutive white spaces with single white space.
- Removing accented characters because we only want to analyze the English language.
- Expands contractions (e.g can't = cannot) to help with text standardization
- Removing Stopwords
- Expands tweet slang (OMG = Oh My God)
- Replaces repeated punctuation with a single character (Wow!!!!! = Wow!)
- Removing HTML tags
- Removing all retweets of a follower
- Transfer emojis into texts by using emoji library (substituted emojis with text)

- Label Encoding, we converted each value in a the labels into the machine-readable form: numerical form as shown in Table 2.

Category	Label	Numerical value
Occupation	creator	0
	performer	1
	politics	2
	sports	3
Gender	female	0
	male	1

Table 2: Label Encoding for Occupation and gender

For the task of identifying a celebrity's age, the birthyear trait has the highest number of possible classes, covering the years from 1940 to 1999. It seemed unwise to train a fine granular classifier to distinguish between all these classes. The task also acknowledges this difficulty by accepting answers as correct as long as they are inside a certain interval around the actually correct value. Following the interval task formula in the competition evaluation script, we will determine 5 intervals as shows the Table 4.1 to cover the entire range, and then we will reclassify each training example to the "middle".

$$truebirthyear - m \leq predictedbirthyear \leq truebirthyear + m$$

Interval range	Class	Mean year
1990-1999	0	1994
1981-1989	1	1985
1970-1980	2	1975
1956-1969	3	1962
1940-1955	4	1947

Table 3: Birthday interval ranges used

4.2 System description

In this section, we describe the predictive model used in our submission. The model used for our task was designed to identify three type of classes: gender, age and occupation. Submitted models used the described above preprocessing. In the following, we will discuss our attempted techniques and models to predict each the celebrity features.

4.2.1 Common Machine Learning Algorithms. In this section, we present the list of common machine learning algorithms we used in the prediction of the celebrity characteristics from the follower' tweets. First, we vectorized the follower' tweets using a TF-IDF vectorizer, taking into account the top 10,000 features from word bigrams.

We used different machine learning algorithms, which are well studied and have extensively been used in several text classification tasks. In particular, we used: Random Forest, Linear SVC, MultinomialNB and Logistic regression were used as models for each different task.

4.2.2 Neural network. We tried then to replace machine learning models by using different neural network architecture.
The first model was Bidirectional LSTM.
The second model starts with Convolutional 1D , followed by Max pooling then LSTM. The third model starts also with convolutional 1D followed by max pooling layer.
The final model represent a fully connected neural network.

4.2.3 Transfer learning using BERT. BERT is one of the new machine learning methods, released in late 2018, where a model designed for a task is reused as the starting point for a model on a second task. Since re-training takes less time and requires less data than training a model from scratch we decided use this methods to reach our goals. The models can both extract high quality language features from text, or fine-tuned on a specific task (classification, question answering, etc.).

4.3 Libraries used

- **Scikit Learn**
  A free software machine learning library and intended for the Python programming language.
  For machine learning, various algorithms such as clustering including support vector machines, regression or classification algorithms are available.
  Scikit-learn is used with SciPy and NumPy scientific python libraries. The library is characterized by its robust and well-documented functions and it provides Simple and efficient tools for data mining and data analysis.
- **Nltk**
  Natural Language ToolKit platform used for processing human language using text processing libraries for classification, tokenization,

stemming, tagging, parsing and other natural language processing techniques. In our project we used it for tokenization and Part Of Speech tagging.

- **Pandas**
  Pandas is a famous library for python data analysis, for python programming language it provide a high performance, flexibility, powerful and easy to use open source data analysis and data structures. Pandas is to handle dataframes.
  Pandas allow importing data of file in different formats like csv, excel .... It allows to make many different operations such as groupby, join, merge, melt, concatenation as well as data cleaning features like filling missing values in dataframes.
- **NumPy**
  It's a python library which used to make high level mathematical functions.
- **Matplotlib**
  By using python scripts, Matplotlib python library is used for creating plots and 2D graphs. Matplotlib contain a module called pyplot which is useful and makes things easy for plotting by providing feature to control linestyles,font properties,formatting axes etc.
  It supports a many different graphs like charts, histogram etc.

5 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 are described and summed up in Table 4, Table 7 and Table 5 below.

	Training set F1	Test set F1
BERT	**0.74089**	**0.68447**
Neural Network	0.555,0.3333,0.5126,0.5046	0.51,0.29,0.49,0.48
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 4: Results from experiments on gender prediction using different models

We can see that BERT achieved the best results on the gender prediction task.

	Training set F1	Test set F1
BERT	**0.57179**	**0.5557**
Neural Network	0.3834,0.101,0.3207,0.318	0.34,0.8,0.29,0.307
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 5: Results from experiments on occupation prediction using different models

BERT achieved also the best results on the occupation prediction task.

	Training set F1	Test set F1
BERT	**0.3842**	**0.33574**
Neural Network	0.2557,0.0665,0.068,0.2129	0.196,0.0571,0.054,0.19
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 6: Results from experiments on birthyear prediction using different models

Again BERT overtop other model in the prediction of birthyear and achieved the best results on this task.

Considering best results founded in each task we computed the evaluation metric cRank defined above and compared our results with other teams results founds for the same task in the computation PAN Celebrity Profiling 2020. Our model outperform random, baseline-ngram-follower-tweets and tuksa20 but we achieved lower results then koloski20, hodge20 and baseline-ngram-celebrity-tweets.

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	**0.480833864**	**0.33574**	**0.68447**	**0.5557**
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 7: Results of other teams

6 CONCLUSIONS AND FUTURE WORK

The task of celebrities profiling using followers' tweets generated several challenges that are worth highlighting. We have to predict three characteristics the most critical was the birth year prediction task in which the number of labels was reduced, creating five groups for better accuracy. In addition to that the training dataset of followers' tweets had an evident imbalance. In the other hand, the dataset volume was another important challenge it was necessary 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. As future work, we plan to analyze the models with all the volume of dataset using a cluster of servers and use another technique in the preprocessing step consisting in leaving and analysing only hashtags instead of analyzing all the tweets since hashtags are usually the information carrier .
In addition, we plan to use XLNet which is a new transfer learning model that appear to overcome some of the BERT limits.

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, and Walter 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).

[4] Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al.*, pages 750–784 (2016).

[5] Matti Wiegmann, Benno Stein, Martin Potthast, L Cappellato, N Ferro, D Losada, and H Müller. Overview of the celebrity profiling task at pan 2019. In *CLEF* (2019).

[6] Victor Radivchev, Alex Nikolov, and Alexandrina Lambova. *Celebrity profiling using tf-idf, logistic regression, and svm.*