A Deep learning Model to predict gender, age and occupation of the celebrities based on tweets followers

Notebook for PAN at CLEF 2020

Roobaea Alroobaea[1], Ahmed H. Almulihi[1], Fahd S. Alharithi[2], Seifeddine Mechti[2], Moez Krichen[3], Lamia Hadrich Belguith[2]

[1] College of Computers and Information Technology, Taif University, Saudi Arabia
[2] The laboratory on Development and Control of Distributed Applications (ReDCAD)
[3] ANLP Group, MIRACL Laboratory, University of Sfax, 3018,Sfax Tunisia
r.robai@tu.edu.sa, a.almulihi@tu.edu.sa,
f.alshalawi@tu.edu.sa,mechtiseif@gmail.com, moez.krichen@redcad.org,l.belguith@gmail.com

Abstract. This paper present the methods are used for detection of celebrity profiles on Twitter when participating in PAN @ CLEF 2020. We have tried to predict the age, gender and occupation of celebrities based on their tweets followers. Our method is based on the use of deep learning techniques to discriminate between authors. The results obtained by the team "TUKSA20" are encouraging, indeed we obtained the 1st rank for the prediction of the gender, the second for the prediction of the occupation, and for the age the results obtained are less good.

1 Introduction

Author profiling is the study which consists in recognizing certain dimensions of the author's profile based on the stylistic features of their writings [Maharjan et al., 2014]. The dimensions targeted by profiling particularly concern the demographic and socio-cultural aspects of the author, such as his age, sex, personality, mother tongue, native region, and recently the occupation [Alroobaea, 2020; Wiegmann et al.,2019].
[Koppel, 2002; Argamon et al., 2009; Pennebaker, 2011].
Recently, the prediction of the author's occupation has been the focus of work on the author's profile detection. The question is whether the author is a politician, footballer, artist, singer, scientist, etc. predicting occupation has an important role in many domains such as websites bidding and forensic linguistics.
The Celebrity Profiling task 2020 is to develop a piece of software which predicts three demographics of a celebrity from the texts of their followers: occupation, age, and gender.
In this paper, we try to implement a system for detecting the age, gender and occupation of celebrities using deep learning techniques. The remainder of this paper is organized as follows; Section 2 presents a brief state of the art of author profile prediction methods. Section 3 introduces our method used to discriminate between authors. Section 4 presents the experiments as well as the evaluations. Finally, the summarization with conclusion will be mentioned.

2 Related Work

The study of [Wiegmann et al., 2019] distinguished two types of attributes: stylistic attributes (style-based features) and attributes based on content (content features).

Stylistic features

In determining the age or gender of a writer, it is important to consider functional words, such as prepositions, pronouns and determiners. These attributes have been shown to be effective in detecting an author's profile. [Cruz et al., 2013]. In other work, the authors use the frequency of punctuation marks, the frequency of capital letters and citations [Aleman et al., 2013]. Similarly, HTML attributes, such as image URLs or a web page link, have been used by [Sapkota et al., 2013]. In the work of [Patra et al., 2013] the authors are based on specific terms (foreign words) for the distinction between the authors. Similarly, other authors use a calculation of the frequency of emoticons as a discriminating attribute for the prediction of the authors [Irazu et al., 2013].

Content features

In addition to the style used, the occurrence of the themes used can differentiate several age categories. [Argamon et al., 2009].

The age of groups generally represents adolescents, young people and adults. Adolescents were grouped in the '10s' category presenting individuals between 13 and 17 years of age, young people in the '20s' category representing those between 23 and 27 years of age and finally adults in the '30s' category representing those aged between 33 and 47 years [Schler et al., 2006].

The work carried out by Koppel et al. have shown that at the gender level. Men prefer to categorize things using more determinants (le / la, ce / ce, un / une, etc.) and quantifiers (two, more, few, etc.). On the other hand, women frequently use personal pronouns in their writings (I, you, me, etc.) [Koppel et al., 2003].

Argamon et al. (2009) have done a study on the British national corpus using the elements of speech features. They obtained an 80% prediction for the determination of the genus [Argamon et al., 2009]. For their part, [Nguyen et al., 2013] have tried to predict age in the conversations of Dutch Twitter users. They used traits from the language model combined with parts of speech. 74% of the discussions were well classified. They obtained an average margin of error between 4.1 and 6.8 years for age prediction. Similarly, in [Zhang and Zhang, 2010], the authors worked on blog segments using word-type features, punctuations, average words per sentence, length of sentences, part of speech and the rate of prediction of the gender which amounts to 72%. Peersman et al. (2011) used a Netlog corpus by experimenting with the 2-gram, and 3-gram unigrams. The average obtained for the prediction of age and gender is 88% [Peersman et al., 2011].

In [Gaustad et al., 2007], the authors are interested in the automatic classification of mails. They obtained an accuracy rate of 81% for gender and 72% for age. We cite among others, the work of [Kose et al., 2007]; [Hariharan et al, 2011] which have shown promising results in detecting the author's gender in instant discussions. For their part, Maharjan et al. have developed a system based entirely on the Map Reduce technique for predicting age and gender. This technique is a parallel programming model allowing the processing of a very large mass of data (Big data). These real-time programs are executed on partitions (clusters) and are automatically parallelized [Maharjan et al., 2014]. This system which employs a large number of functionalities is capable of performing the prediction task in a fraction of time with good details. This approach demonstrates that the use of these systems can be a perfect solution for data quantity problems and / or a large number of attributes.

The objective of the study by Rangel et al. (2016) was to show how people use language to express their emotions and how it can help them identifying their age and gender. They used a graphical approach called "EmoGraph" which allows the text to be represented by a graph of parts of speech labeled by emotions and themes. The authors showed results comparable to those of the best systems presented at the PAN @ CLEF 2013 evaluation conference [Rangel et al., 2016].

3 Methodology

3.1 Preprocessing

Plain text should be filtered to remove noisy data, such as HTML tags and urls. This is because the presence of this noisy data could affect and reduce the accuracy of the entire analysis. [Derczynski et al., 2013] pointed out that "twitter text is difficult to part-of-speech tag: it is noisy, with linguistic errors and idiosyncratic style". Thus, preprocessing stage is important step. After that the cleaned data is then stored in a database.

3.2 text analysis

In this step, it is a matter of carrying out linguistic and statistical processing to find the useful information allowing to subsequently achieve the best classification of documents. To achieve our objective, we calculated the number of occurrences of all the words that occurred in the corpus and thus obtained a list of words sorted by frequency. Note that the text analysis step is repeated for each age group and for each gender. Then we tried to gather these terms into thematic classes by adopting a semi-automatic annotation which follows the following two steps:
-Automatic procedure: During this annotation, we based ourselves on the automatic search of thematic classes for the attributes based on the content and of syntactic classes for the attributes based on the style. Furthermore, the definition of thematic classes is based on the notion of a universe of discourse. Indeed, all the terms presenting a semantic relation are gathered around a centroid which bears the name of this class. In this annotation, we use the Wordnet-Similarity tool to extract the semantic distances between

terms. We use synonymy, hypernymy and hypomimy relationships to detect generalization and specification. At the end of this step all the terms having a semantic distance with the “synset” belong to the same class. It should be noted that the stylistic classes are obtained using the morpho-syntactic analyzer in order to be able to differentiate prepositions, verbs, pronouns, etc.

- Manual procedure: Since the automatic annotation does not allow assigning all the terms found in a class, we have had recourse to the manual annotation which is inspired by the thematic classification proposed by [Argamon et al., 2009]. It consists of manually grouping the most frequent terms in the same class. This operation is performed by three different linguists.

3.3 Features extraction

Recent methods differentiate two main categories of features that can be used to predict the profile of the author: Stylistic features and those based on content [Wiegmann, 2019] .In order to measure the relevance of a term in a given document in the corpus, we used the TF * IDF measure which is calculated as follows:
[Buckley et al., 1996]

$$1.1 \quad w_d = f_{wd} \times \log(|D|/f_{wd})$$

With:
d: documents,
W: the word,
D: one document ϵ D,
$f_{w,d}$: number of occurrence of w in d,
|D|: size of the corpora,
$f_{w,D}$: number of documents.

After forming the classes, the trait selection is used to elucidate which ones allow better discrimination. This step is essential since the initial calculation of the frequency of terms in the corpus does not provide information on the ability of the classes found to discriminate between the different age / gender categories. In addition, the selection of attributes is a very important factor which eliminates redundant attributes and which results in improved classification performance.

It should be noted that the selection of attributes can be done in two ways. The first is an "a priori" selection which allows you to filter the attributes before starting the classification step. In this case, it is a question of employing filtering algorithms making it possible to find, for example, the best K most discriminating attributes. As for the second, it is an "a posteriori" selection which exploits the classification process: each time it is a question of using a set of attributes for learning and of refining the results as and when. This iterative process stops when the classification result is optimal. Therefore, the attributes allowing to obtain the best results are retained as the most discriminating attributes.

3.4 Classification

Once the attributes have been defined, the classification of the texts is proceed. It is started with the construction of the two learning matrices relating to the gender and age dimension. However, we have opted to enrich our base of thematic classes. The idea is that each test corpus can contain terms semantically linked to existing terms in the learning corpus. Therefore, expanding these terms is a step towards improving relevance.

It should be noted that the enrichment of the base of thematic terms is defined as being the process of completing the original terms of the thematic classification proposed by associating them with semantically similar terms. It is one of the solutions envisaged to enrich the proposed classes. Similar to the text analysis stage, in this enrichment process, we have based ourselves on the exploitation of semantic links like synonymy, hyponymy or hypernymy for the enrichment process.

5 Experimentations and Results

The LSTM deep learning model is presented in figure 1:

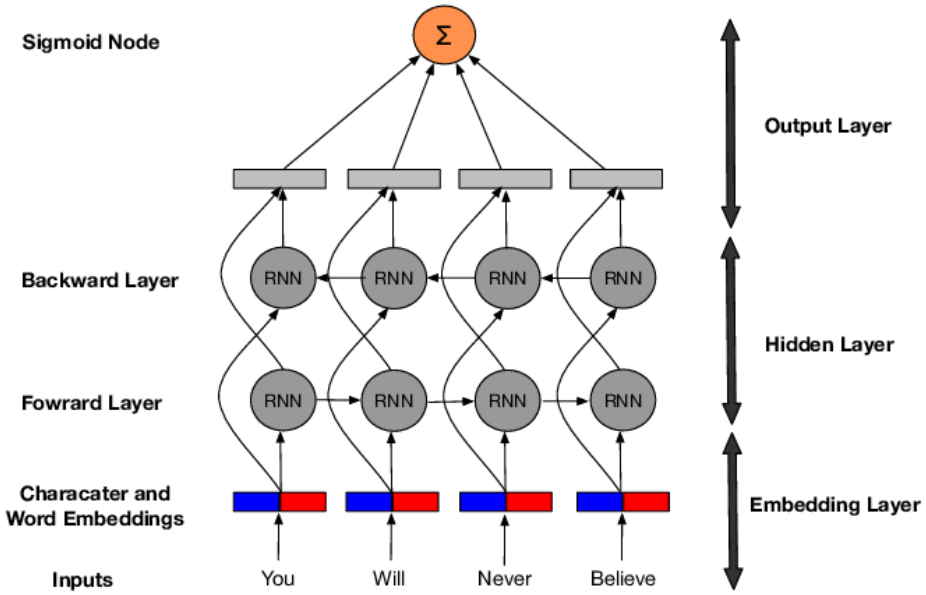

Fig 1: Architecture of the LSTM deep learning Model

In the evaluation process, Submissions are judged by a combined metric cRank, which is the harmonic mean of each label's metric.

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

All features are evaluated by their respective F_1. Precision and recall of birth-year are calculated leniently. If a prediction is within an m-window of the truth, it is counted as correct:

$$true\ birthyear - m \leq predicted\ birthyear \leq true\ birthyear + m$$

The window size m is based on the birth-year and increases linearly from about 3 years for 1999 to about 9 years for 1940.

After the evaluation process our team "TUKSA20" obtain third rank in term of CRANK and first rank for the prediction of the gender with 0.69. For the occupation we obtain second Rank VS Hodge20.

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
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 1. Official results of the PAN@CLEF celebrities profiling Task

6 Conclusion

This paper presents our method to predict gender, age and occupation of the celebrities based on tweets followers. For the prediction of gender and occupation, good results were obtain based on LSTM (Long Short term memory) architecture. Indeed, TUKSA20 ranked the first for the gender prediction. However, the age prediction results were less good because of the unbalanced data. As a perspective, other deep learning models need to be test such as CNN (Convolutional Neural Network) , GRu (Gated recurrent Units) and GAN (Adversarial Neural network).

References

Wiegmann M, Stein B, Potthast M. Celebrity Profiling. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), July 2019. ACL.

Wiegmann M, Stein B, Potthast M. Overview of the Celebrity Profiling Task at PAN 2019. CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org.

Maharjan S., Shrestha P. and Solorio T. *A* simple approach to author profiling in MapReduce. *In proceedings of the 4th Pan at conference and labs of the evaluation forum (CLEF)*, pp.1121-1128, England, 2014.

Koppel M., Argamon S. and Shimoni A. Automatically categorizing written texts by author, gender. *Literary and Linguistic Computing*, pp.401-412, 2003.

Argamon, S., Koppel M., Pennebaker J., and Schler J. Automatically Profiling the Author of an Anonymous Text. *Communications of the ACM*, 52(2), pp.119-123, 2009.

Cruz F., Haro. R., and *Ortega J. ITALICA at PAN 2013: An* Ensemble Learning Approach to Author Profiling. *In Proceedings of the 4th PAN at conference and labs of the evaluation forum (CLEF)* .Spain, 2013.

Sapkota U., Solorio T., Montes-y-Gómez M. and De-la-Rosa G. Author profiling for English and Spanish text. *In Proceedings of the 4th PAN at conference and labs of the evaluation forum (CLEF),* Spain, 2013.

Irazu D., Farias H., Guzman-Cabrera R., Reyes A. and Rocha M. Semantic-based Features for Author Profiling Identification. *In Proceedings of the 4th PAN at conference and labs of the evaluation forum (CLEF)*, Spain, 2013.

Schler J., Koppel M., Argamon S. and Pennebaker, J. Effects of age and gender on blogging. *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp. 191-197, 2006.

Peersman C., Daelemans W and Van Vaerenbergh L. *Predicting age and gender in online social networks*. In Proceedings of the 3rd international workshop on Search and mining user-generated contents, (SMUC), pp.37-44, USA, ACM, 2011.

Derczynski, L., Ritter, A., Clark, S. and Bontcheva, K., 2013, September. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013 (pp. 198-206).

Hariharan S. Gender prediction in chat based medium's using text mining. *In International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 1(1), pp. 18-22, 2011.

Alroobaea, R. An Empirical combination of Machine Learning models to Enhance author profiling performance. *In International Journal of Advanced Trends in Computer Science and Engineering*, 9(2), pp. 2130- 2137, 2020.

LI J., Zheng R. and Chen H. From fingerprint to writeprint. *Communication of the ACM*, 49(4), pp.76-82, 2006.

Potthast M and Gollub Tand Wiegmann M, Stein B ,Information Retrieval Evaluation in a Changing World,Nicola Ferro and Carol Peters, sep,Springer,TIRA Integrated Research Architecture, 2019

Wiegmann M and Stein B and Potthast M,Working Notes Papers of the CLEF 2020 Evaluation Labs, sep, CLEF and CEUR-WS.org, Overview of the Celebrity Profiling Task at PAN, 2020.