# Gender Identification through Multi-modal Tweet Analysis using MicroTC and Bag of Visual Words

## INGEOTEC participation in User Profiling Task@PAN18

http://github.com/INGEOTEC
http://www.ingeotec.mx

**Eric S. Tellez, Sabino Miranda-Jiménez, Daniela Moctezuma, Mario Graff, Vladimir Salgado, and José Ortiz-Bejar**
**CONACyT-INFOTEC-CentroGEO, México**

## Introduction

This poster reports our participation in the multi-modal Author Profiling task of PAN'18. We use our $\mu$TC tool to tackle the text sub-task, and a variant of Bag of Visual Words to deal with the user's visual content. Finally, our multi-modal approach use a convex combination of both textual and visual information.

## Modeling Users

### Text based Author Profiling

**In general, we model each user as an array of her/his tweets**. We use MicroTC to perform the text modelling of each user.

- MicroTC ($\mu$TC) is our generic framework for text classification task, i.e., it works regardless of both domain and language aspects.
- The main idea behind $\mu$TC is to select a competitive configuration from a vast universe of possible ones. Each configuration is composed of:
  - **Text transformations**
    - Hashtags, numbers, urls, user mentions, and emoticons (with three value options: remove, group, none).
    - Remove: diacritic, character duplication, punctuation, and case normalization (with two value options: activate or not-activate).
  - **Tokenizers**
    - $n$-grams of words ($n = 1, 2, 3$)
    - $q$-grams of characters ($q = 1, 3, 5, 7, 9$)
    - Skip-grams: $(2, 1)$, $(2, 2)$, and $(3, 1)$
  - **Weighting schemes**
    - Raw frequency
    - TFIDF
    - Entropy
- Finally, a Support Vector Machine is used as classifier.

This approach is a kind of **black box model**, but we are currently dealing with how to extract **valuable information from the generated models**.

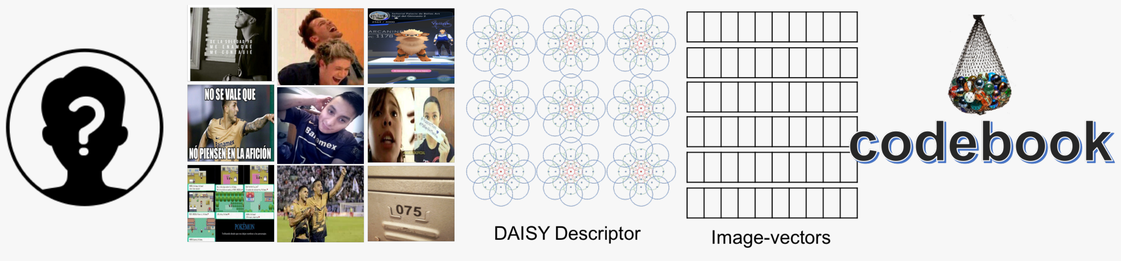
You can download $\mu$TC from our GitHub page: http://github.com/INGEOTEC/microTC.

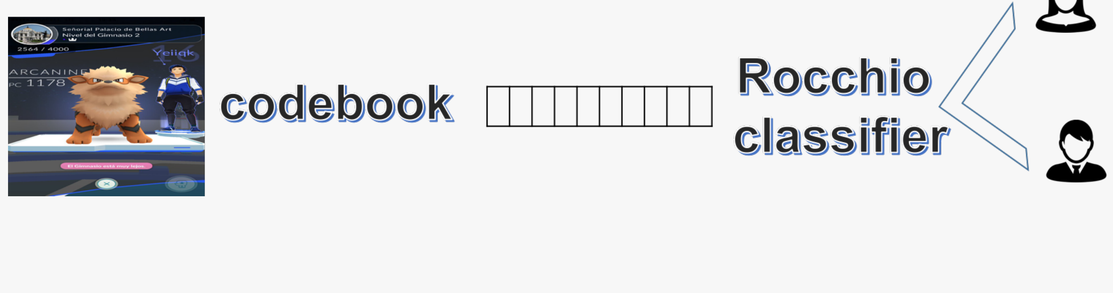
### Image based Author Profiling

**In the image problem, we model each user as an array of her/his images to convert them to text.**

- The image to text transformation has three main steps:
  1. We use DAISY [2] to compute an array of feature descriptors, for each image.
  2. An efficient clustering algorithm is used to create a codebook.
  3. The codebook is used to create a text representation of each image.

Finally, we perform text classification over the generated text using an algorithm inspired by Rocchio.



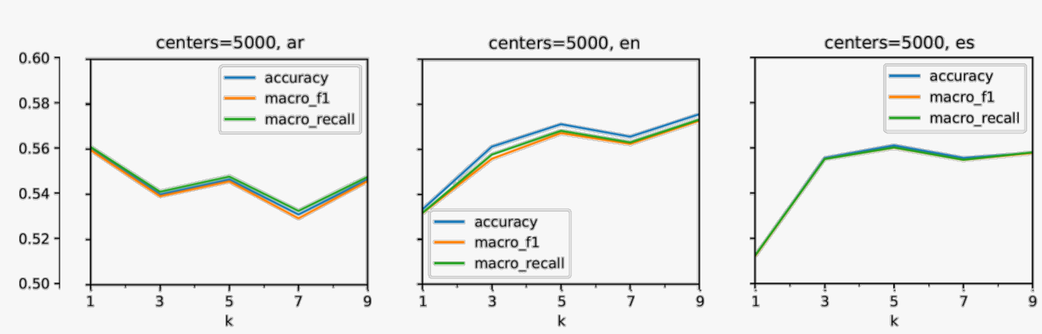

## Results

### User profiling results using MicroTC with text

| dataset | setup | accuracy | macro-F1 | macro-Recall |
|---|---|---|---|---|
| Arabic | Arabic@PAN'17 | **0.8378** | **0.8377** | **0.8385** |
| English | Arabic@PAN'17 | **0.8267** | **0.8266** | **0.8284** |
| Spanish | Arabic@PAN'17 | **0.7933** | **0.7933** | **0.7943** |

### User profiling results using BoVW with images

The image-based profiling uses our Bag of Visual Words with $5000$ centers and $k = 7$ (nearest centroids), with this configuration, our approach produced an accuracy of $0.5691$, $0.5468$, $0.5900$ for Spanish, English, and Arabic languages, respectively.




### Results of the Text and Image Combination

| dataset | $\alpha$ | accuracy | macro-F1 | macro-Recall |
|---|---|---|---|---|
| Arabic | 0.99 | 0.8400 | 0.8399 | 0.8408 |
| English | 0.95 | 0.8278 | 0.8278 | 0.8293 |
| Spanish | 0.925 | 0.8033 | 0.8033 | 0.8042 |

## Conclusions

- We used our MicroTC ($\mu$TC) framework [1] to deal with text content, and a variant of BoVW to deal with image content.
- Regarding text, a gross analysis shown that q-grams are among the highest weighted features; however, they are also among the lowest weighted tokens; that means, is not easy to understand why classifier choose to label an user as female or male.
- Regarding images, we observed that women tend to share *selfies* and images with text-content, while men share cartoons, humorous images, and landscape photos.

## References

[1] Eric S. Tellez, Daniela Moctezuma, Sabino Miranda-JimÃľnez, and Mario Graff. An automated text categorization framework based on hyperparameter optimization. *Knowledge-Based Systems*, 149:110 – 123, 2018.

[2] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, May 2010.