

SU@PAN'2015: Experiments in Author Profiling

Yasen Kiproff

Sofia University, FMI

Momchil Hardalov

Sofia University, FMI

Preslav Nakov

Qatar Computing Research Institute, HBKU

Ivan Koychev

Sofia University, FMI

1. Introduction

We discuss the participation of the Sofia University team, kiprov15, in the 2015 edition of the Author Profiling task, part of the PAN 2015. We participated with a system for English and Spanish. We experimented with SVM classifiers using variety of features extracted from publicly available resources.

2. Data

The text of a set of tweets:

- 152 authors for English
~100 tweets each;
- 100 authors for Spanish
~100 tweets each;

3. Pipeline

1. Twitter tokenizer
2. RegEx sentence splitter
3. Language identifier
4. Language-specific POS tag
5. Gazetteer lookups
6. Rule-based feature extractors
7. Classifiers

4. Results

Language	GLOBAL	RMSE	Ranking
English	0.7211	0.1493	6
Spanish	0.7889	0.1495	2

Table 1. Summary of our official results for English and Spanish tweets.

Language	Age	Agreeable	Both	Consc.	Extrov.	Gender	Open	Stable
English	0.7254	0.1411	0.5915	0.1318	0.1416	0.8451	0.1198	0.2123
Spanish	0.7841	0.1249	0.7273	0.1386	0.1625	0.9091	0.1334	0.1884

Table 2. In detail look at our official results for English and Spanish tweets.

5. Features

Twitter-specific

- Total counts for each: **#hashtag**, **URL**, **retweet**, **@mention**;
- Number of tweets starting with a user mention;

Term-level

- **Word n-grams**: count of unigrams and bigrams;
- **Vocabulary size**: total number of different words used by a user;
- **POS tagging**:
 - **Tag frequency** for each tag;

We used GATE TwitIE for English and OpenNLP for Spanish

Orthographic

- **Elongated words** count;
- **Average sentence length**;
- **Letter case**: the number of lower-case, all-caps, and mix-case words;

Lexicons

- **NRC** Hashtag Emotion Lexicon: 16,862 terms with emotion;
- **Bad words** lexicon: a manually assembled dictionary;
- **World Well-Being Project** Personality Lexicon;

Lexicon based

- For each lexicon, **total matches count**;
- For the **NRC Lexicon**:
 - total score of all matching terms for each emotion type;
- For the **WWBP Lexicon**, for each of the big five traits:
 - Total positive terms score and count;
 - Total negative terms score and count;
 - Total terms score and count.

All of the above counts are calculated for all tweets of an author and are divided by their number.

6. Conclusion

- Standard approach with BoW and shallow features yields competitive results;
- Most useful features for all models were POS tag frequencies and 1-2 word n-grams;
- Dictionary-based features do not help much;
- 3+ n-grams hurt performance

7. Future Work

- Integration of additional resources: LIWC, clusters, embeddings, etc.
- Corpus segmentation by country, mother tongue, social groups