1. Introduction

The Author Profiling (AP) task consists in knowing as much as possible about an unknown author, just by analyzing a given text [2], for example: age and gender.

The PAN13 AP task consists in profiling age and gender in social media data. The AP task can be approached as a classification problem.

Differences with other classification tasks are:

- The used textual features.
- The representation.

The standard Bag of Terms (BOT)

Some shortcomings of BOT like representations are:

- High dimensionality.
- High sparseness of the representation.
- They do not preserve any kind of relationship among terms.

Our proposal

We propose the use of very simple but highly effective meta-attributes. These textual features highlight the relationships that terms and documents hold with profiles.

These attributes are inspired in some ideas from CSA [3] to represent documents in text classification.

2. Document Representation

Document Profile Representation (DPR) is built in two steps:

1. Terms representation in a space of profiles.
2. Documents representation in a space of profiles.

1) Terms representation

For each term \( t_j \) in the vocabulary, we build a term vector \( \vec{t}_j = (wtp_{j1}, \ldots, wtp_{kj}) \), where \( wtp_{kj} \) is a value representing the relationship of the term \( t_j \) with the profile \( p_i \). For computing \( wtp_{kj} \), first:

\[
 wtp_{kj} = \log_2 \left( 1 + \frac{tf_{kj}}{len(d_k)} \right)
\]

Some term vectors have stronger peaks.

Highly descriptive term vectors for specific profiles.

There are other similar term vectors for specific profiles for example:

- "j" for detecting young people (e.g. profiles 10s, and 20s).
- "game" for the prediction of males.

2) Documents representation

Add term vectors of each document. Documents will be represented as:

\[
 \overrightarrow{d}_k = \sum_{t_j \in D_k} \frac{tf_{kj}}{\text{len}(d_k)} \times \vec{t}_j
\]

where \( D_k \) is the set of terms of document \( d_k \).

Examples of highly descriptive term vectors.

Differentiation of terms for each profile, gender, and age.

3. Evaluation

Corpus description using our features.

Evaluation

We use the 50K most frequent terms from each information source.

We used a LIBLINEAR classifier [1], and a 10-fold-CV in the training set for preliminary evaluation of our approach.

Final results

Second Order Attributes (SOA) and BOT computed over the 50,000 most frequent terms on the datasets.

4. Conclusions

1. The best method at PAN13 to predict age profiles in blogs (for both corpus).
2. Our results overcome the conventional BOT and holds the first position for both languages (overall accuracy).
3. More than 454 times faster than the method in one position below, 166 times faster than the method in first position.
4. This is the first time that AP is addressed using such dense attributes vectors that represent relationships with profiles.

5. References

