Profiling microblog authors using concreteness and sentiment

Know-Center at PAN 2016 author profiling

Oliver Pimas, Andi Rexha, Mark Kröll, and Roman Kern

Know-Center GmbH Graz, Austria {opimas, arexha, mkroell, rkern}@know-center.at

Abstract The PAN 2016 author profiling task is a supervised classification problem on cross-genre documents (tweets, blog and social media posts). Our system makes use of concreteness, sentiment and syntactic information present in the documents. We train a random forest model to identify gender and age of a document's author. We report the evaluation results received by the shared task.

1 Introduction

The paper at hand presents a description of our approach to author profiling task at PAN 2016. The author profiling task includes two separate classification problems: gender classification and and age group classification. The latter is a multi-class (18-24, 25-34, 35-49, 50-64, 65-xx) classification problem. The classification problem can be described as follows: An author profile in the context of the task is defined as an author's gender and age group. Given a set of documents with author profiles known, learn to identify the author's profile of documents of unknown authorship.

The PAN 2016 author profiling task is cross-genre, meaning that the training documents will be on one genre and the evaluation will be on another genre. While this resembles real-world problems more closely, it also makes the task more challenging. The training corpus is a collection of tweets in English, Spanish and Dutch. However, our approach only focuses on documents in English language.

This notebook paper is outlined as follows: in section 2 we describe our classification approach. In section 3 we present the results. Finally, we present the conclusion in section 4.

2 Approach

As mentioned in section 1, we consider the problem a supervised classification problem. We pre-process each document, extracting features and thus vectorising the input. Once all the features a extracted, we train a random forest model. The random forest implementation we use is provided by the class *RandomForest* from the machine learning framework WEKA [8]. We did not tune any parameters, but used WEKA's default settings. In the following we describe our main feature types used in the classification task.

2.1 Concreteness

A number of features are based on the concreteness of words within tweets. The base of this features is a dataset assembled by Brysbaert et al. with the help of Amazon Mechanical Turk [2]. The dataset comprises over 37 thousand words, which are known by at least 85% of the raters. Thus the contained words can be considered to be known to a large share of the English speaking population. Concreteness is defined in this context, whether a word refers to a perceptible entity. This concept is driven by the intuition that concrete words are easier to remember and to process than words that refer to abstract concepts.

Concreteness has been studied in a variety of scenarios, with an emphasis on topics like age-of-acquisition. There is some work on the link between the tendency to use words with varying degree of concreteness and the age and gender of the person [3]. More research is needed to arrive at an answer to which extent gender or age are related to the use of concrete words. In our work we may give an answer to this question, based on the results and a deeper analysis of the results.

Our set of concreteness features consists of nine individual numeric features, based on three different scores being computed on a per word basis:

- 1. Mean concreteness: The score reflects the concreteness of the words within a tweet. Concreteness thereby ranges from 5 to 1.
- Standard deviation concreteness: This score encodes how strong the individual annotators agreed on the concreteness score. For words were all raters agreed, the score will be low.
- 3. Percent known: This score represent the percentage of all raters, who indicated that they know the word. This score ranges from 0.85 to 1.

In order to arrive at features at tweet level, all word based scores are aggregated. Therefore the minimum, the maximum and the arithmetic mean are computed for each of the three types of scores.

2.2 WordNet Domains

The motivation for this feature is to encode the main topics of a tweet in a concise way. It is based on the publicly available WordNet Domains corpus ¹ [1,9]. This specialized dataset is an augmented version of the WordNet ² corpus and provides an assignment of words to so called domains. There are about 200 different domains, which are organised in a hierarchy.

We developed an algorithm that creates a set of domains for a given short snippet of text. If available, the part-of-speech of the words can be utilised to narrow down the appropriate synset for each word. All domains of all words are combined while keeping track of a weight. The weight reflects how ambiguous the domain mappings are, thus words with many domains will yield lower weights.

Finally, the hierarchy of the domains is exploited, where each sub-domain distributed a share of its current weight to its parent. The ranked list of domains is finally

¹ http://wndomains.fbk.eu/

² https://wordnet.princeton.edu/

pruned. All domains with a lower weight than half of the weight of the top ranked domain will be removed. On average a short snippet of text will yield a set of 1 to 5 domains.

In order to convert the set of domains into features we created a binary feature for each domain. If a tweet is associated with a certain domain, the corresponding feature will be set to true.

2.3 Sentiment

Sentiment in text in general, and more particular in tweets, might help to discriminate different age groups as well as different genders. Based on this hypothesis, we generate features that capture the polarity (whether positive or negative sentiment) of the tokens in the tweet. For this task we use the well known sentiment library called Senti-WordNet [5]. SentiWordNet specifies different polarities of words, depending on their context and provides a linear score between -1 and 1. Words with a negative polarity do have a negative score, and the ones with the positive score do have a positive polarity.

In order to capture the polarity of the tweet and learn from its feature, we extract the score for each token. If the tokens aren't defined in SentiWordNet we ignore them. We get the score of the most used context of the token. As a final step we model the polarity as four numeric features: we collect the tweet's

- 1. maximum polarity,
- 2. minimum polarity,
- 3. average polarity, and the
- 4. standard deviation of polarity of all terms with polarity mapping.

These features represent the polarity distribution of a tweet seen as a bag of words.

2.4 Hashtags

Twitter provides some specific features like hashtags, retweets and replies. Especially hashtags are easy to use, but lack a direct equivalent in other blogs or message services. We expect users familiar with twitter to make use of these service specific features more often. We encode the usage of hashtags as three features:

- 1. Existence: whether one or more hashtags were used.
- 2. Count: the number of hashtags used.
- 3. Ratio: the ratio between non-hashtag terms and hashtags.

2.5 Token Length

The motivation behind the token length feature is somewhat similar to the hash tag usage. Users familiar with micro blogging or texting are used to the 140 character limit. As a consequence, we expect more frequent usage of abbreviations and acronyms. We encode the mean token length and the median token length.

2.6 Instance Selection

In our approach we combine a number of different features into a single feature space. Therefore it is highly likely that the feature space itself will not be linearly separable. Depending on the actual classification algorithm this might be problem. Algorithms that have a low bias and a high variance will tend to cope easier with such scenarios. For example, a 1-NN algorithm does not impose the requirement of a linearly separable feature space. Apart from the implications of a high variance, there might be another culprit of such system.

In machine learning, a single object from which a model can be learned, or to which a model can be applied, is called an instance. In our case, an instance is a vector representation of a single document (i.e. a single tweet, blog or social media post).

It has been discovered that in many real-life datasets some instances behave different to others. More precisely, certain instances have the tendency to be over-represented in the neighbourhood of the remaining instances. These instance effectively behave like hubs, hence the term hubness has been introduce to describe this phenomenon [12,6]. Furthermore, it has been shown that down-regulating the influence of these hubs will improve performance.

In order to deal with such hubs we introduce an optional step in the feature engineering pipeline. This additional step is conduced after the feature space has been created and before the actual classification. Instead of identifying individual hubs, we try to detect regions, where multiple hubs are expected to be found. For this we utilize a density based clustering algorithm, more precisely DBSCAN [4]. This clustering algorithm has a number of advantageous properties, for example its excellent runtime complexity and the fact that the number of clusters does not have to be specified beforehand. Additionally, the algorithm separates regions of high density from regions with a lower density.

We make use of this property by filtering out all instances from high density regions. This is motivated by the intuition that instance, that are similar to each other will be less helpful for the learning algorithm than instances, that capture certain characteristics not present in the instances from the high density areas. The parameter ϵ of the DBSCAN algorithm can be effectively used to control the amount of instances being filtered out in this step.

3 Results

We report the results as shown by the PAN 2016 evaluations done on TIRA [10][7]. After training our model with the training set provided in TIRA, we ran classification on both English training sets.

As we ran into memory problems using the 4gb provided by the virtual machines on TIRA, we had to deactivate features in order to be able to successfully train a model. As the task at hand is cross-genre, we decided to deactivate the WordNet domains feature group. We expect the use of topics to be of minor help when dealing with tweets across different genres.

Table 1 shows the evaluation results obtained from TIRA.

Table 1. The evaluation results obtained from TIRA.

Dataset	Gender	Age Class	Both
pan16-author-profiling-test-dataset2-english-2016-05-07	0.5769	0.3205	0.1410
pan16-author-profiling-test-dataset1-english-2016-03-08	0.0201	0.0086	0.0057

While the results (see table 1) on 'pan16-author-profiling-test-dataset2-english-2016-05-07' are where we expected them to be, the results on the results on 'pan16-author-profiling-test-dataset1-english-2016-03-08' are extremely low. We cannot comment on this yet, as we have no further details on how the test sets look like.

An overview [11] of the shared tasks will be made available, including the author profiling results.

4 Conclusion

In this paper we presented our software developed for the PAN 2016 author profiling task. By extracting features like concreteness and sentiment, we trained a RandomForest to identify the gender and age class of an unknown tweet author. This is an initial approach towards authorship profiling. While our system achieved results in the region we expected on one of the test sets, it greatly underperformed on the other. Lacking the details on the test sets, we are not yet able to analyse the reasons for this.

4.1 Future Work

In the future we will experiment with different combinations of the features. We also had a lot of problems with memory usage, which led us to remove some feature groups from the final evaluations. We plan to improve on this, thus being able to validate new features and use the full extraction pipeline available.

5 Acknowledgements

The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

- Bentivogli, L., Forner, P., Magnini, B., Pianta, E.: Revising the wordnet domains hierarchy: semantics, coverage and balancing. In: Proceedings of the Workshop on Multilingual Linguistic Ressources. pp. 101–108. Association for Computational Linguistics (2004)
- Brysbaert, M., Warriner, A.B., Kuperman, V.: Concreteness ratings for 40 thousand generally known english word lemmas. Behavior research methods 46(3), 904–911 (2014)

- Calais, L.L., Lima-Gregio, A.M., Arantes, P., Gil, D., Borges, A.C.L.d.C.: A concreteness judgment of words. Jornal da Sociedade Brasileira de Fonoaudiologia 24(3), 262–268 (2012)
- Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. vol. 96, pp. 226–231 (1996)
- Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: 5th Language Resources and Evaluation Conference (LREC 2006). pp. 417–422 (2006)
- Flexer, A., Schnitzer, D.: Can shared nearest neighbors reduce hubness in high-dimensional spaces? In: Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on. pp. 460–467. IEEE (2013)
- Gollub, T., Stein, B., Burrows, S., Hoppe, D.: TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments. In: Tjoa, A., Liddle, S., Schewe, K.D., Zhou, X. (eds.) 9th International Workshop on Text-based Information Retrieval (TIR 12) at DEXA. pp. 151–155. IEEE, Los Alamitos, California (Sep 2012)
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software : An Update. SIGKDD Explorations 11(1), 10–18 (2009)
- 9. Magnini, B., Cavaglia, G.: Integrating subject field codes into wordnet. In: LREC (2000)
- Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
- Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: 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, CLEF and CEUR-WS.org (Sep 2016)
- Tomasev, N., Radovanovic, M., Mladenic, D., Ivanovic, M.: The role of hubness in clustering high-dimensional data. Knowledge and Data Engineering, IEEE Transactions on 26(3), 739–751 (2014)